THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

New molecular simulation methods for quantitative modelling of protein-ligand interactions.

Stefano Bosisio

THE UNIVERSITY
of EDINBURGH

Thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy
to the
University of Edinburgh — 2018

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Stefano Bosisio
27th July 2018

Lay Summary

When we get sick, we take a medicine, usually in form of pills, to get relief. A pill just contains a simple drug which has shown a good effect to beat a specific disease. However, to understand which drug should be used and how should it be, it is not an easy task. In particular, to get to the final "pill stage" it could take up to 25 years of intensive and expensive research. This work is trying to improve the "drug research" to get more accurate and fast results, leading to a quicker "pill" creation and a less expensive research.

In order to help the research, this thesis proposes a variety of protocols to compute in a more accurate and more robust way the "free energy" of a drug, namely how a drug could be efficacious against a specific target-disease. The protocols are presented in a logical way, from the easier to the harder "free energy" problems.

Abstract

The main theme of this work is the design and development of new molecular simulation protocols, to achieve more accurate and reliable estimates of free energy changes for processes relevant to the structure-based drug design. The works starts with an insight into the reproducibility problem for alchemical free energy calculations. Even if simulations are run with similar input files, the use of different simulation engines could give different free energy results. As part of a collaborative effort, the implementation details of AMBER, GROMACS, SOMD and CHARMM simulation codes were studied and free energy protocols for each software were validated to converge towards a reproducibility limit of about 0.20 kcal·mol^{-1} for hydration free energies of small organic molecules.

Following, new simulation methods for the estimation of lipophilicity coefficients (log P and log D) for drug like molecules were developed and validated. log P values were computed for a dataset of 5 molecules with increasing fluorination level. Predictions were in line with the experimental measures and the simulations also allowed new insights into the water-solute interactions that drive the partitioning process. Then, as part of the SAMPL5 challenge, log D values for 53 drug-like molecules were computed. In this context two different simulation models were derived in order to take into account the presence of protonated species. The results were encouraging but also highlighted limits in alchemical free energy modelling.

As an additional task of the SAMPL5 contest, three different protocols were validated for predicting absolute binding affinities for 22 host-guest systems. The first model yielded a free energy of binding based on free energy changes in solvated and complex phase; the second added the long range dispersion correction to the previous model; the third one used a standard state correction term. All three protocols were among the top-ranked submission in SAMPL5, with a correlation coefficient R^2 of about 0.7 against experimental data.

Finally, the origins and magnitude of the finite size artefacts in alchemical free energy calculations were investigated. Finite size artefacts are especially

predominant in calculations that involve changes in the net-charge of a solute. A new correction scheme was devised for the Barker Watts Reaction Field approach and compared with the literature. Hydration free energy calculations on simple ionic species were carried out to validate the consistency of the scheme and the approach was further extended to host-guest binding affinities predictions.

Acknowledgements

These four years have had plenty of amazing moments, countless laughters and wired scientific discussions with hundreds of people passed by room 234 and I would like to express my gratitude to all of them.

First of all, I am grateful to my supervisor Dr Julien Michel. Thanks very much for relying on me during this fantastic experience (despite my knowledge of chemistry). Julien always supported and guided me through my research with infinite patience, feeding my curiosity and enthusiasm with new projects and challenges. *Merci beaucoup!* Thanks also to my second supervisor, Prof. Philip J. Camp, who helped me to tackle Vivas with his hard questions and his support.

Many thanks to all the members of the Michel group. Thanks to Charis and Pattama ($\epsilon\nu\chi\alpha\rho\iota\sigma\tau\omega$ $\phi\iota\lambda\epsilon$, *kop-khun-krup*) for your friendship and all the food provided at every BBQ! $\epsilon\nu\chi\alpha\rho\iota\sigma\tau\omega$ to a bunch of Greek people, who brought a pure Mediterranean soul in the lab: Haralambos, Michalis, Georgia and Sofia - without you the lab would have been too quiet and gray. *Muchas gracias* Joan, Jordi, Salome, Cesar, *tapadh leat* Lisa, *danke* Toni, *namaste* Arun, *merci* Marie and Alessio for all the sacred and profane discussions we had, ranging from Pokemon to high scientific levels. Thanks to the old computational "batch" Gaetano, George, Doig, Joshua and Julien.

Many thanks to my mum and dad. *Grazie* for understanding my choices and for your support. *Grazie mille* also to my "extended" family, Graziella, Roberto, Ilario and Elena for the countless funny situations in Edinburgh and for the kilos of salami brought here! Thanks to Luigi and Ilaria, who always supported and unconditionally believed in me.

Finally, I would like to thank the lovely "humanistic" side of my life, Federica. Without you all of this would have been impossible. Thanks for supporting and encouraging me in every decision. Thanks for our laughs and your love. *Grazie di cuore, realizziamo i nostri sogni ora.*

Contents

1 Introduction **10**
- 1.1 The drug discovery process . . . 10
- 1.2 Statistical mechanics concepts: ensemble and free energy . . . 16
- 1.3 Classical potentials . . . 19
- 1.4 Barker-Watts Reaction Field . . . 23
- 1.5 Molecular Dynamics . . . 25
- 1.6 Alchemical free energy calculations . . . 29
  - 1.6.1 Outline of the thesis . . . 36

2 Reproducibility of Free Energy Calculations Across Different Molecular Simulation Software **39**
- 2.1 Introduction . . . 39
- 2.2 Methods . . . 43
  - 2.2.1 Alchemical Free Energy Implementations . . . 43
  - 2.2.2 RAFE Setup . . . 45
  - 2.2.3 Free Energy Estimations . . . 52
- 2.3 Results . . . 53
  - 2.3.1 AMBER . . . 58
  - 2.3.2 CHARMM . . . 62
  - 2.3.3 GROMACS . . . 64
  - 2.3.4 SOMD . . . 67
- 2.4 Discussion and Conclusions . . . 70

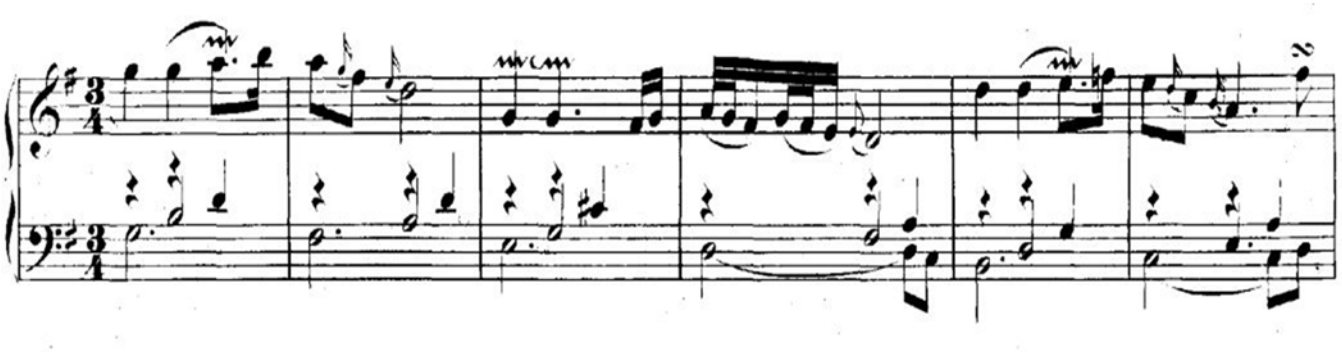
3 Alchemical free energy protocols for lipophilicity coefficients estimation **73**
- 3.1 Introduction . . . 73
- 3.2 Methods . . . 79
  - 3.2.1 Computing lipophilicity coefficients with alchemical free energy calculations . . . 79
  - 3.2.2 All-neutral and the two-species model for logD . . . 80
  - 3.2.3 logP simulations setup . . . 84
  - 3.2.4 logD simulations setup . . . 85

3.2.5 Alchemical free energies estimations 86
3.3 Results 87
3.3.1 logP results 87
3.3.2 logD results 90
3.4 Conclusion 98
3.5 Appendix 100

4 Alchemical free energy protocols for host-guest binding free energy predictions 103
4.1 Introduction 103
4.2 Theory: binding free energy models 105
4.2.1 Host-Guest simulation set-up 110
4.2.2 Alchemical free energy production simulations 110
4.2.3 Estimation of free energy changes for models A, B and C 111
4.2.4 Host-guest experimental setup 112
4.2.5 Statistical comparison with other submissions 114
4.3 Results 115
4.3.1 Host-Guest binding free energy predictions 115
4.3.2 Comparison with other participants submissions 120
4.4 Conclusions 124

5 Electrostatic finite size artifacts in alchemical free energy calculations 125
5.1 Introduction 126
5.2 Theory 127
5.2.1 Free energy correction terms for the BWRF approach . 127
5.2.2 Correction scheme for the PME approach 133
5.2.3 Correction scheme with alchemical counterions 135
5.3 Methods 137
5.3.1 Sodium ion 138
5.3.2 Acetate, methylammonium and functionalized fullerene 138
5.3.3 SAMPL5 host-guest systems: OA and CBC 139
5.3.4 Correction schemes implementations 140
5.4 Results and discussion 142
5.4.1 Sodium ion hydration free energy 142
5.4.2 Polyatomic molecules: acetate and methylammonium . 144
5.4.3 Functionalized fullerene correction terms 147
5.4.4 OA host-guest binding free energy 149
5.4.5 CBC host-guest binding free energy 151
5.5 Conclusions 155

6 Conclusions **158**



Chapter 1

Introduction

1.1 The drug discovery process

The history of modern drug discovery starts in 1856, when an English student, William Henry Perkins, tried to synthesize quinine from allytoluidine, an aniline derivative. Despite all his efforts, the experiment ended up in a dark substance that was able to turn fabrics purple: the tyrian purple [1], the very first synthetic dye. Perkins understood the importance of his discovery and he rapidly established his own factory on the following year. At the same time other companies were founded, trying to optimize Perkins' reaction to develop new possible synthetic dyes. As a result, the continuous research for synthetic dyes led to a rapid development and demand of organic chemists, who could discover novel reactions through the application of research. The rise of German economics enabled Germany to lead this field, investing capitals in dye industries. Thus, between 1858 and 1862 factories such as Bayer, Ciba, Sandoz and Farbenfabrikent Hoechst appeared. At that point many companies realized that their chemists not only could produce dyes but could also look for organic molecules that could be employed as possible drugs. As a matter of fact, dyes showed anti-bacterial properties, as Domagh [2] discovered with Protosil, a red dye which could be used to fight against streptococcal bacteria. Therefore, dyes were initially used as drug prototypes and, after a few years, drug research became a completely

independent process from its origin [3]. In Britain, France and Canada the dye-companies focused on the extraction of alkaloids from plants, which could have therapeutic effects on sick people. Seminal efforts for defining the very first drug-design process, were achieved by Ehrlich [4]. Ehrlich was able to synthesize the first synthetic drug, an arsenic compound to treat syphilis and with his research team he laid the foundations for the first reliable biological screening and evaluation procedures. The greatest result of this early drug discovery challenges arrived at the end of the 1930s, when a Scottish scientist, Alexander Fleming, building on works of Florey and Chain, discovered penicillin. Penicillin was the first synthetic drug, made available on a large scale, which helped soldiers during the Second World War against bacterial infections. In spite of this success, most of these drug design studies were carried out without knowing the chemical structure and properties of the biological targets and many times serendipity was the key of success [1].

After the Second World War medicinal chemistry experienced a radical change. Between 1950 and 1970, researchers focused their attention on understanding, *in vivo*, the mode of action of possible drug candidates. In the late 1960s, Beecham and Pfizer found new molecules with similar pharmacokinetic properties to penicillin, such as sulfactams or clavulanates [5]. Starting from these compounds and following different pathways, it was possible to develop new drugs with different therapeutic effects (hypoglycaemic agents, diuretics, antihypertensive etc). However, even if experiments on animals were giving promising results, many times scientists were struggling to understand and deal with the pharmacokinetics effects in human bodies. - absorption, distribution, metabolism and excretion (ADME).

During the 1980s continuous technological development brought the drug discovery process into a new era. In particular, the advent of computational chemistry gave beneficial effects to drug design. For the very first time it was possible to understand the atomic details of protein-ligand interactions, thanks to the seminal works by Martin Karplus and Andrew McCammon [6, 7]. Additionally, combinatorial chemistry made possible the creation of large libraries of possible drug candidates and Quantitative Structure-Activity Relationship(QSAR) methods facilitated the idealization of new compounds [8].

The endless progress of hardware resources led to nowadays, with the advent of graphical cards for scientific calculations and the creation of supercomputing facilities [9], which enable simulations of very large biomolecular systems [10, 11, 12] and accelerate the drug screening process. In parallel, the advent of more sophisticated biophysical essays has brought to a more efficient in vitro sampling of the chemical space. As a result, in the last 20 years, drug discovery has become an intersection among several scientific

disciplines, such as genomics, proteomics, physical and organic chemistry, computer science and pharmacology.

All these features led to the definition of a "critical path" of action [13] for the development of new drugs, shown in fig. **1**. The critical path divides the drug discovery process into two main stages: an initial basic research and subsequent clinical development and approval.

The basic research stage can be subdivided into three major steps:

- a pure biological research, where a biomolecular target is identified and, wherever possible, the target structure is solved by using crystallographic or nuclear magnetic resonance (NMR) techniques. After the identification of the biomolecular target a library with thousands of possible drug candidates is created;

- a molecular-prototype design and discovery phase. One of the leading approaches at this stage, which will be the central part of this thesis, performs an initial docking screening process of the molecular libraries. Then, free energy simulations are run to understand the binding affinities of drug candidates. These studies allow the definition of a lead compound and its optimization, through iterative computational procedures;

- the preclinical development,where a small batch of compounds are optimised to improved diverse drug-like properties (ADME)

All these steps require between 6 and 9 years of work [14], accounting for at least US $ 32 billion per year in the R&D pharmaceutical industry [15]. In spite of such a great financial investment, the pharmaceutical industry is facing the "Eroom's law" [16], shown in fig. **2**. An inverse proportional relationship is observed between the number of new drugs approved by the US Food and Drug Administration (FDA) per billion of US $ spent on the R&D pharmaceutical research. The total R&D spending was doubled to US $65 billion over the period 2000-2010, with very small number of approved drugs (less than 1 per year per US $ billion spending.). As a matter of fact, new drugs must be more effective than those currently on offer in order to be approved, which puts the R&D pharmaceutical sector under pressure to look for new methodologies, which could enhance the basic research stage.

To alleviate this problem, a significant amount of computational research is focused on the optimization stage of the *Structure Based Drug Design* process (SBDD) [17, 18, 19]. SBDD is a promising paradigm for drug-design, where ligands are created based on the structural information from experimental data. The most promising hit molecules, that typically come from a

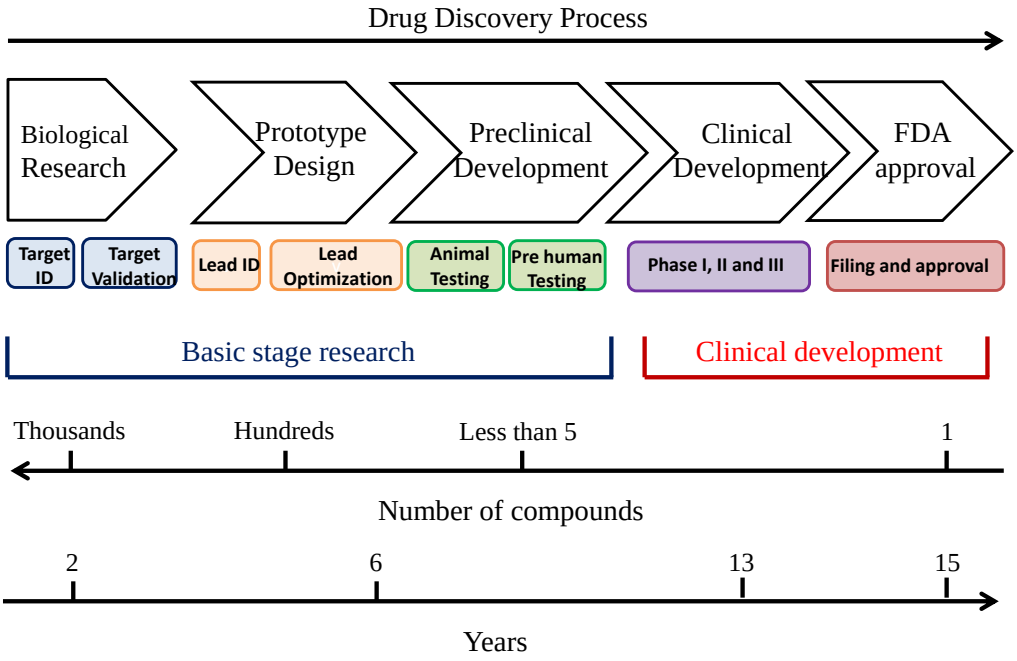


Figure 1: Critical path for the drug discovery process [13]. The process starts with the biological research, where a biological target is identified. Whenever possible a library of thousands of compounds against this target is screened. Then, the prototype design stage uses tools from physical sciences, such as computational chemistry, to forecast drug properties and to define lead compounds. In this stage, drugs undergo a lead optimization process, which should give a few lead compounds (around 5 molecules) ready for the clinical stage. The preclinical development is needed to understand the ADME properties of drug candidates, before authorizing the compounds for the human trials. The last part of the drug discovery process are devoted to the clinical development, which follow three phases where drugs are tested on larger and larger number of volunteers. The very last step is the approval by a drug administration organization, such as the FDA, and the launch on the market. The overall process takes approximately 15 years on average.

high throughput or biophysical assay, undergo a *hit to lead* stage, where a limited optimization is done to increase the drug-target affinity. Then, once potential drugs are found, molecules are extensively and iteratively optimized in the *lead optimization* stage, trying to enhance the ADME properties whilst

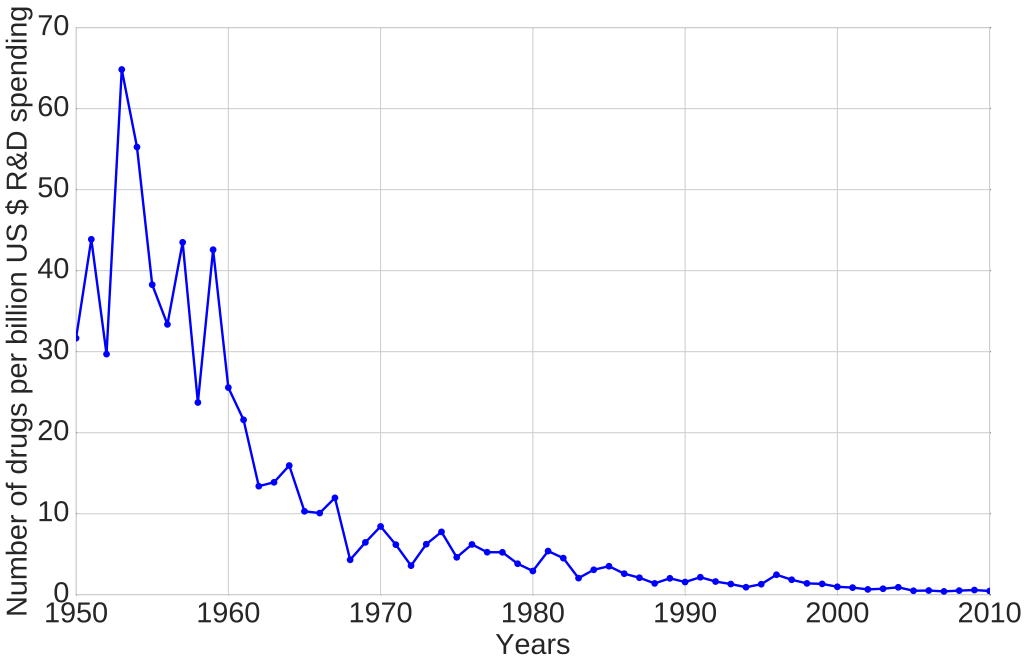


Figure 2: Eroom's Law in pharma R&D [16]. The plot shows how many drugs are approved per US $ 1 billion spending from the 1950s till 2010. All the figures were inflation-adjusted. Overall, the number of approved drugs by the US FDA has halved every 9 years, with less than 1 approved drug for 1 billion US $.

maintaining the potency. Correct predictions at the hit to lead and lead optimization stage would be beneficial for the basic research stage, leading to less expenses for the pre-clinical trial stage and less time consumed for the whole critical path. To reach this desired outcome, it is important to understand exactly the chemical problem involved and how computational methods can help in this process.

An important aspect in the lead optimization is the exact *in silico* estimation of the protein-ligand binding affinity, which can be retrieved from the experimental dissociation constant K_d:

$$\begin{cases} PL \rightleftarrows P + L \\ \\ K_d = \frac{[P][L]}{[PL]} \end{cases} \tag{1.1}$$

where P is the target protein, L the drug candidate, $[PL]$ is the equilibrium concentration of the complex species, $[P]$ and $[L]$ the equilibrium concentration of the protein and ligand respectively. K_d can be computationally estimated through the calculation of the Gibbs free energy of binding:

$$\Delta G = k_B T \ln \frac{K_d}{C_0} \tag{1.2}$$

where the ΔG denotes the change of free energy upon binding, k_B is the Boltzmann constant, T is the temperature and C_0 the standard state concentration (usually 1 M).

Many techniques have been developed to compute eq. 1.2. As an example Åqvist modelled the protein ligand binding by introducing the Linear Interaction Energy method (LIE) [20]. In this model, the linear response theory for electrostatics forces [21] is employed, to compute the free energy of binding of a ligand. A drawback of this technique is the difficulty in estimating empirical parameters which govern the overall protein-ligand binding prediction.

Another popular technique is the Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA) [22]. MM-PBSA exploits molecular dynamics and Poisson Boltzmann calculations to retrieve the free energy of binding of eq. 1.2. A problem of the MM-PBSA is the use of an implicit solvent, thus possible water mediated behavior in the protein-ligand association may be not detected.

Finally, a very popular and promising technique, which will be employed and studied in this thesis, is the alchemical free energy calculation. The alchemical approach exploits Zwanzig and Kirkwood ideas [23, 24], to compute the free energy of binding as a free energy for changing a molecule

from an initial to a final end state [25]. Aldeghi *et al.* [26] have employed alchemical free energy calculation to compute eq. 1.2 for a set of diverse inihibitors binding to bromodomain protein 4 (BRD4), achieving an accuracy of 0.6 kcal·mol^{-1}. Furthermore, the authors have shown how this level of accuracy ccan be achieved in pseudo-prospective approach. Chodera *et al.* [27] have proved the potential of alchemical free energy calculation for drug-discovery purposes. In particular, authors have predicted the affinities for kinase mutations binding to Abl, a major target in chronic myelogenous leukemia (CML). Additionally, these calculations have shown an accuracy of about 1 kcal·mol−1 in predicting the resistance for eight FDA-approved kinase inhibitors across 144 clinically identified kinase point mutations. Finally, by correctly mixing experimental bio asseys and alchemical free energy calculations, Jorgesen [28] have shown how alchemical free energy calculations can be pivotal in the lead optimization phase, helping in the synthesis of non-nucleoside inihibitors for HIV-1 reverse transcriptase (NNRTIs). Six compounds are reported with an association constant of 20 nM for protection of human MT-2 cells against the cytopathogenicity of HIV-1.

1.2 Statistical mechanics concepts: ensemble and free energy

Thermodynamic quantities, such as the Gibbs free energy in eq. 1.2, can be interpreted from an atomistic point of view, adopting the concept of *ensemble*. The ensemble idea was initially developed by Boltzmann and Gibbs [29, 30, 31] and it refers to a collection of systems (or particles), that share common macroscopic properties. Boltzmann and Gibbs stated that each thermodynamic property of a macroscopic system, like pressure, can be computed as an average from the mechanical property arising in each ensemble's particles [32, 33]. Indeed, if we consider an ensemble, at any moment in time, each of the total $\mathcal{N}$ particles has a given momentum $\mathbf{p}$ and position $\mathbf{q}$ and Hamiltonian function $\mathcal{H}(\mathbf{q}, \mathbf{p})$. The entire set of $\mathcal{N}$ positions and momenta defines a precise point in a $(\mathbf{q}^N, \mathbf{p}^N)$ multidimensional space $\Gamma(\mathbf{q}^N, \mathbf{p}^N)$, called *phase space*. The entire collection of particles adopting a particular conformation through time described an imaginary curve in the phase space, called *trajectory* [34]. Rather than focusing on the time evolution of the trajectory, thus on each single particle mechanics, statistics can be employed to extract the thermodynamic properties of the system. In this way, it is possible to describe the ensemble behavior by means of a distribution function $\mathcal{P}(\Gamma)$, which allows to compute the average properties of the

ensemble, namely the average macroscopic thermodynamic state. Mathematically, the *Liouville thereom* and *Birkhoff theorem* assure that along the trajectory the ensemble averages can be performed at any point in time and that these averages are ergodic, so the time average of a function along a trajectory is related to the ensemble average [35]. Hence, the aim of statistical mechanics is the correct description and computation of an ensemble probability function, in order to retrieve the macroscopic thermodynamic quantities of a system.

One of the most important ensemble is the *canonical ensemble*. This ensemble is characterized by a number of $\mathcal{N}$ of particles, it is coupled with a thermal bath to keep a constant temperature T and the entire system has a fixed volume V. The ensemble has a total energy $\mathcal{E}$, described by the Hamiltonian function $\mathcal{H}(\mathbf{q}^N, \mathbf{p}^N)$, given as the sum of all the $\mathcal{N}$ particles of the entire ensemble. The total energy of the system is not fixed, so all the possible energy states of the energy spectrum should be considered to make some evaluation on the macroscopic thermodynamic state of this ensemble. In the canonical ensemble it is possible to prove that the probability of a particle to be in a state j, described by a specific set of position $\mathbf{q}^N$ and momenta $\mathbf{p}^N$ and a total energy of $\mathcal{H}_j$ is:

$$\mathcal{P}_j(\Gamma) = \frac{e^{-\beta\mathcal{H}_j(\mathbf{q}^N,\mathbf{p}^N)}}{\int\int e^{-\beta\mathcal{H}(\mathbf{q}^N,\mathbf{p}^N)} d\mathbf{q}^N d\mathbf{p}^N} \tag{1.3}$$

where $\beta{=}\frac{1}{k_B T}$ and k_B is the Boltzmann constant ($1.38064852{\times}10^{-23}\ \frac{J}{K}$). The denominator of eq. 1.3 is called *partition function* and is often indicated as Q_{NVT}. The partition function is the sum over all the possible energy states of the ensemble. The knowledge of this function is pivotal in statistical physics, as from there any thermodynamic property of a system can be computed.As an example the average energy $\bar{E}$:

$$\bar{E} = k_B T^2 \left(\frac{\partial \ln Q_{NVT}}{\partial T}\right) \tag{1.4}$$

Moreover, Q_{NVT} is linked to the Hemlholtz free energy A_{NVT} as:

$$A_{NVT} = -k_B T \ln(Q_{NVT}) \tag{1.5}$$

Eq. 1.5 is known as the *fundamental equation* of statistical mechanics. The passage from the Hemlholtz free energy A_{NVT} to the Gibbs free energy G_{NPT}, described in eq. 1.2 can be done by translating the canonical ensemble concepts into the *isothermal-isobaric ensemble*. In this ensemble the pressure p

is fixed, rather than the volume V, and the partition function can be defined as:

$$\Delta_{NPT} = \int\int\int e^{-\beta\mathcal{H}(\mathbf{q}^N,\mathbf{p}^N)} e^{-\beta pV} V^N d\mathbf{q}^N d\mathbf{p}^N dV \tag{1.6}$$

From here, it is possible to derive the Gibbs free energy:

$$G_{NPT} = -k_B T \ln \Delta_{NPT} \tag{1.7}$$

Q_{NVT}, and Δ_{NPT}, can be further developed under the classical approximation [36][1]:

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \int\int e^{-\beta\mathcal{H}(\mathbf{q}^N,\mathbf{p}^N)} d\mathbf{q}^N d\mathbf{p}^N \tag{1.8}$$

where the term $\frac{1}{N!}$ takes into account that fermions (e.g. electrons) are indistinguishable, $\frac{1}{h^{3N}}$ described the quantum mechanical origin of each particle, where h is the Plank's constant (6.62×10^{-34} $\frac{J}{s}$). Decomposing the Hamiltonian of eq. 1.8 into potential energy term $U(\mathbf{q}^N)$ and kinetic energy $K(\mathbf{p}^N)$:

$$\begin{aligned} Q_{NVT} &= \frac{1}{N!}\frac{1}{h^{3N}} \int\int e^{-\beta(U(\mathbf{q}^N)+K(\mathbf{p}^N))} d\mathbf{q}^N d\mathbf{p}^N \\ &= \frac{1}{N!}\frac{1}{h^{3N}} \int e^{-\beta(U(\mathbf{q}^N)} d\mathbf{q}^N \int e^{-\beta(K(\mathbf{p}^N))} d\mathbf{p}^N \\ &= Q_{id} \cdot Z_{NVT} \end{aligned} \tag{1.9}$$

Q_{id} indicates the analytical integral evaluation of the kinetic part:

$$Q_{id} = \frac{V^N}{N!\Lambda^{3N}} \tag{1.10}$$

where $\Lambda = \frac{h^2}{(2\pi m k_B T)^{1/2}}$ is the *de Broglie wavelength*, m the mass of each particle. Z_{NVT} is the *configurational integral*:

$$Z_{NVT} = \int e^{-\beta U(\mathbf{q}^N)} d\mathbf{q}^N \tag{1.11}$$

For a isothermal-isobaric ensemble, eq. 1.11, can be written as:

$$Z_{NPT} = \int\int e^{-\beta U(\mathbf{q}^N)+pV} d\mathbf{q}^N dV \tag{1.12}$$

Thus, by determining the potential energy of the system it is possible to solve the configurational integral and retrieve the Gibbs free energy. To solve this problem numerically, *molecular dynamics* techniques are employed, which allow to *sample* the potential energy landscape of the molecular system.

[1]For the ease of notation only Q_{NVT} will be considered here

1.3 Classical potentials

The potential energy function of a system, $U(\mathbf{q}^N)$ is usually a high-multi dimensional function, very complicated to compute with quantum mechanical techniques (*ab-initio*). Help comes from the Born-Oppenheimer approximation. In this approximation, the electron density is supposed to be instantaneously fixed from changes in nuclei's positions, as the electrons move much faster than the nuclei [37, 38]. From the Born-Oppenheimer approximation, the Schrödinger equation can be re-written in terms of nuclei interactions. This allows to describe the nuclear system as a simple classical mechanics system, making use of the Hamiltonian function:

$$\frac{1}{2}m_i\left(\frac{\partial \mathbf{r}_i}{\partial t}\right)^2 + U_i(\mathbf{r}_i) = E_{tot}(\mathbf{r}_i) \tag{1.13}$$

where the subscript i-th denotes the i-th nucleus of the system, characterized by a mass m_i, a coordinates vector $\mathbf{r}_i$, a kinetic energy $\frac{1}{2}m_i\left(\frac{\partial \mathbf{r}_i}{\partial t}\right)^2$, and a nucleus-nucleus interaction potential energy function $U_i(\mathbf{r}_i)$.

As a consequence, classical mechanics can be employed to describe the nuclei's motions and interactions in a biomolecular system. However, appropriate energy functions $U(\mathbf{r}_i)$ parameters, called *force field*, must be chosen to describe the i-th nucleus interactions. The most employed force field separates molecular interactions into two factors: *bonded* terms and *non-bonded* terms. Bonded terms describe molecules' internal degree of freedom and can be subdivided into:

- bonded interactions (or stretching) between two atoms
- angle interactions (or scissoring) between three atoms
- dihedral interactions between four connected atoms

The stretching interaction is usually modelled as a harmonic potential approximation. In this approximation, two bonded atoms can be considered as two masses linked by a spring to each other. The spring gives rise to vibrational degrees of freedom, which are a function of the internuclear distance of the two atoms. Thus, the potential energy can be written as a function of the square of the displacement from the equilibrium length:

$$U_{bond}(\mathbf{r}) = \frac{1}{2}k_b(r_{eq} - r)^2 \tag{1.14}$$

where k_b is the spring constant in kcal mol^{-1}Å^{-2}, r_{eq} is the equilibrium bond length and r the current length.

A similar description can be achieved for the scissoring motion [39]:

$$U_{\text{angle}}(\theta) = k_\theta(\theta_{eq} - \theta)^2 \tag{1.15}$$

where k_θ is the angle force constant, defined in kcal·mol^{-1}rad^{-2}, θ_{eq} and θ are the equilibrium angle and the current angle.

Finally, the four atom interactions, are described by means of dihedral angles and modelled through a Fourier cosine series:

$$U_{\text{dihedral}}(\phi) = \sum_{n=1}^{n_{max}} k_\phi \left[1 + \cos(n\phi - \delta)\right] \tag{1.16}$$

where again k_ϕ described the stiffness of the dihedral harmonic oscillator and the term $[1 + \cos(n\phi - \delta)]$ comes from the Fourier cosine series, where n is the multiplicity and the $\sum_{n=1}^{n_{max}}$ is extended up to all the multiplicities employed, ϕ is the actual dihedral degree and δ is a shift term. Eq. 1.16 can be tuned to correctly model the energy barriers in molecules. In particular, for a carbon-carbon bond with sp^3 hybridization, it is possible to model the presence of minima in $\pi/3$, π and $-\pi/3$ and maxima in $2\pi/3$, 0 and $-2\pi/3$, by selecting k_ϕ=1, n=3 and δ=0.

The *non-bonded* interactions are usually modelled with functions inversely proportional to the distance between two atoms. In particular, two main terms can be described: the *dispersion interactions* and the *Coulombic interactions*. The former are known also as *van der Waals* or *London dispersions*, which are implemented with the *Lennard-Jones* potential, or *6-12* potential [40]:

$$U_{\text{vdW}}(\mathbf{r}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{1.17}$$

where ϵ_{ij} is called well depth, in kcal·mol^{-1} and it describes the interaction potential energy minimum, while σ_{ij} is the collision diameter, namely the minimum distance with the interaction potential equal to zero. The subscript ij denotes that σ and ϵ must be computed as a product of two different atoms, i and j, van der Waals potential terms. In practice, ϵ_{ij} and σ_{ij} are usually computed through the Lorentz-Berthelot mixing rules [40]. ϵ_{ij} is the geometric mean between the i-th atom and j-th one ϵ values ($\epsilon_{ij} = \sqrt{\epsilon_i \times \epsilon_j}$). Instead, for σ_{ij} an arithmetic mean is used: $\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2}$. The 6-12 potential finely describes the dispersion interactions : at very short distances a repulsive behavior is dominating between the two atoms, as their electronic clouds are overlapping one another. As atoms get farther and farther they experience a temporary attraction, since the electrons in the two adjacent

atoms occupy positions that allow the formation of a instantaneous dipole. Finally, for long distances the energy of interaction of two molecules will tend asymptotically to zero.

Finally, the electrostatic long-range interactions are modelled through the Coulombic potential:

$$U_{Coulomb}(\mathbf{r}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \tag{1.18}$$

where q_i and q_j are the partial charges of atoms i and j, ϵ_0 is the vacuum dielectric constant (8.85418×10^{-12} F$\cdot$m^{-1}), ϵ_r the medium relative permittivity and r_{ij} the distance between atoms. The calculation of eq. 1.18 involves a $\mathcal{O}(N^2)$ operations, which is computationally expensive. To alleviate this cost two main techniques are often employed: Reaction Field (RF) and Lattice Sum (LS) simulations [41, 42]. Both models rely on the use of *radii of cutoff*, namely the electrostatic interactions are computed within a cutoff length, in order to alleviate the computational cost. Specifically, the RF models the "long-range" Coulombic interactions as a static response of a uniform dielectric continuum. The very first treatment was done by Lars Onsager [43] and successively extended by Kirkwood [44]. Generally a RF Coulombic potential has the form:

$$U_{Coulomb-RF}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} H(R_C - r) \left(r^{-1} + \frac{\alpha r^2}{2R^3} - \frac{\alpha + 2}{2R} \right) \tag{1.19}$$

where $H(R_C - r)$ is the Heaviside function ($H(R_C - r) = 1$ if $r < R_C$, $H(r) = 0$ otherwise), ϵ_0 is the permittivity of vacuum,and R_C is a cutoff distance. This distance defines a cutoff sphere, within which the Coulombic interactions generated by solutes' charge are accounted (r^{-1} term). The polarization arising from the molecules beyond the cutoff are modeled through a dielectric response, described by the parameter α. α is determined by the relative dielectric permittivity ϵ_s of the medium surrounding the cutoff sphere for each particle through:

$$\alpha = \frac{2(\epsilon_s - 1)}{2\epsilon_s + 1} \tag{1.20}$$

The advantage of RF methods is a straightforward implementation and the decrease of the computational cost from $\mathcal{O}(N^2)$ to $\mathcal{O}(n^2)$ ($n << N$). LS techniques , like the particle-particle particle-mesh (P3M) and the particle-mesh Ewald (PME) [45]. relies on the use of the Fast Fourier Transform [46] for the long-range interactions and they reduce the Coulombic computational cost to $\mathcal{O}(N \log N)$. Generally, LS divides the Coulombic contribution into

three terms: *direct space sum*, *reciprocal space sum* and *self-energy term*:

$$\begin{cases} U_{dir} & = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=i+1}^{N} q_i q_j \frac{\mathrm{erfc}(\gamma r_{ij})}{r_{ij}} \quad \text{if} \quad r_{ij} < R_C \\ U_{rec} & = \frac{1}{2\pi V}\sum_{i=1}^{N}\sum_{j=i+1}^{N} q_i q_j \sum_{\mathbf{k}\neq 0} \frac{e^{(-(\pi\mathbf{k}/\gamma)^2 + 2\pi i \mathbf{k}\cdot(\mathbf{r}_i - \mathbf{r}_j))}}{\mathbf{m}^2} \quad \text{if} \quad r_{ij} > R_C \\ U_{self} & = -\frac{\gamma}{\sqrt{\pi}}\sum_{i=1}^{N} q_i \end{cases} \tag{1.21}$$

where i and j run over all the N particles of the system, $\mathbf{k}$ denotes the integer wave vectors, $\mathbf{r}_i$ is the position of the i-th particle, while r_{ij} is the distance between particles i and j, V is the volume of the periodic cell, $\gamma = \sqrt{-\log(2\delta)}/R_C$ where δ is a tolerance parameter and erfc denotes the error function. In the direct space sum all the pairs that are further apart than the cutoff distance R_C are ignored. The ignored terms are computed in the reciprocal space. The efficiency of LS methods is paid by a more complicated computational implementation.

The long range electrostatic and van der Waals interactions are often coupled with periodic boundary conditions (PBC). PBC are used to avoid problems with boundary effects, which strongly influences the properties of the whole system [47]. The aim of a molecular simulation is to study a macroscopic system, while computational techniques can only simulate a small box of few nanometers edges. Thus, PBC tries to mimic a macroscopic system by replicating the computational box in the x, y and z directions, with the same molecular feature for each box. Any change taking place in the reference box will be replicated in all the other boxes, generating a homogeneous systems, where there is no boundary or surface. If any molecule leaves the box along one direction (e.g. z), an image of that molecule will enter into the box from the opposite direction (e.g.$-z$), to maintain the number of molecules constant.

Hence, potential energy $U(\mathbf{r})$ can be computed from force field parameters as:

$$\begin{aligned} U(\mathbf{r}) = & \sum_{bonds} \frac{1}{2} k_b \left[r_{eq} - r\right]^2 + \sum_{angles} \frac{1}{2} k_\theta \left[\theta_{eq} - \theta\right]^2 + \\ & \sum_{dihedrals} k_\phi \left[1 + \cos(n\phi - \delta)\right]^2 + \\ & \sum_{i=1}^{N}\sum_{j=i+1}^{N} \left[\frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} + \frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right] \end{aligned} \tag{1.22}$$

1.4 Barker-Watts Reaction Field

Given the importance of the Barker-Watts Reaction Field for this work, a general derivation of eq. 1.19 is given. In 1936, Lars Onsager [43] developed a scheme to treat the electrostatic contributions for a dipole immersed in a dielectric, which is called *reaction field*. The derivation assumes a dipole of moment μ immersed in a spherical cavity with radius R_C, from now on called *radius of cutoff*, surrounded by a uniform dielectric material of constant ϵ_s. The dipole is supposed to be rigid and its shape does not influence the electrostatic potential. It follows that the total electrostatic potential $\psi(\mathbf{r})$, as a function of the position vector $\mathbf{r}$, must satisfy the Laplace equation:

$$\Delta\psi(\mathbf{r}) = 0 \tag{1.23}$$

where $\Delta = (\frac{\partial^2\psi(\mathbf{r})}{\partial x}^2, \frac{\partial^2\psi(\mathbf{r})}{\partial y}^2, \frac{\partial^2\psi(\mathbf{r})}{\partial z}^2)$ denotes the Laplacian operator.

The electrostatic potential $\psi(\mathbf{r})$ can be decomposed into two functions $R(\mathbf{r})$ and $\Theta(\theta)$. $R(\mathbf{r})$ denotes the radial electrostatic component, which depends on the vector position only, and $\Theta(\theta)$ is the angular contribution, which takes into account all the rotational degrees of freedom θ. A known solution for $R(\mathbf{r})$ and $\Theta(\theta)$ is based on the use of spherical harmonics [48] :

$$\psi(\mathbf{r},\theta) = \sum_{l=0}^{\infty}\left(A_l r^l + \frac{B_l}{r^{l+1}}\right)P_l(\cos\theta) \tag{1.24}$$

where the sum is extended to all the l-th terms, A_l and B_l are constants to be determined by imposing boundary conditions, $P_l(\cos\theta)$ denotes the l-th Legendre polynomial. Additionally, $\psi(\mathbf{r},\theta)$ is given by the sum of two contributions: $\psi_{in}(\mathbf{r},\theta)$ which described the electrostatic potential inside the solute and $\psi_{out}(\mathbf{r},\theta)$ which gives a description of the potential outside the solute. For these two elements the following conditions are present in the system:

$$\begin{cases} \lim_{r\to 0}\psi_{in}(\mathbf{r},\theta) = \frac{\mu\cos\theta}{4\pi\epsilon_0 r^2} \\ \lim_{r\to\infty}\psi_{out}(\mathbf{r},\theta) = 0 \end{cases} \tag{1.25}$$

where $\frac{\mu\cos\theta}{4\pi\epsilon_0 r^2}$ is the dipole electrostatic potential, ϵ_0 the vacuum permittivity. $\psi_{in}(\mathbf{r},\theta)$ and $\psi_{out}(\mathbf{r},\theta)$ can be expressed in a general form using eq. 1.24:

$$\begin{cases} \psi_{in}(\mathbf{r},\theta) = \frac{\mu\cos\theta}{4\pi\epsilon_0 r^2} + \sum_{l=0}^{\infty} A_l r^l P_l(\cos\theta) \\ \\ \psi_{out}(\mathbf{r},\theta) = \sum_{l=0}^{\infty}\frac{B_l}{r^{l+1}}P_l(\cos\theta) \end{cases} \tag{1.26}$$

Additionally, along the spherical cavity, the following conditions must be satisfied:

- The potential is continuous across the boundaries :

$$\psi_{in}(R_C, \theta) = \psi_{out}(R_C, \theta) \tag{1.27}$$

- The normal component of the dielectric displacement across the surface is continuous:

$$\frac{\partial \psi_{in}(R_C, \theta)}{\partial \mathbf{r}} = \epsilon_s \frac{\partial \psi_{out}(R_C, \theta)}{\partial \mathbf{r}} \tag{1.28}$$

Eq. 1.27 and 1.28 can be solved for all the spherical harmonic of the general electrostatic potential, eq. 1.26. For $l = 0$ a trivial solution is found, which coincides with the dipole electrostatic potential. For all the terms $l \geq 1$ it is possible to write from eq. 1.27:

$$\begin{cases} \frac{\mu \cos\theta}{4\pi\epsilon_0 R_C^2} + A_1 R_C \cos\theta = \frac{B_1}{R_C^2} \cos\theta \quad \text{if } l = 1 \\ \\ A_l R_C^l = \frac{B_l}{R_C^{l+1}} \quad \text{if } l \neq 1 \end{cases} \tag{1.29}$$

which gives B as a function of A:

$$\begin{cases} B_1 = \frac{\mu}{4\pi\epsilon_0} + A_1 R_C^3 \quad \text{if } l = 1 \\ \\ B_l = (R_C^{2l+1}) A_l \quad \text{if } l \neq 1 \end{cases} \tag{1.30}$$

Thus from condition 1.28, the electric displacement becomes:

$$\begin{cases} -\frac{\mu}{2\pi\epsilon_0 R_C^3} + A_1 = -\frac{2\epsilon_s}{R_C^3} B_1 \\ \\ A_l l R_C^{l-1} = -\epsilon_s (l+1) \frac{B_l}{R_C^{l+2}} \end{cases} \tag{1.31}$$

Substituting eq. 1.30 into eq. 1.31:

$$\begin{cases} A_1 = -\frac{\mu}{2\pi\epsilon_0 R_C^3} \left(\frac{\epsilon_s - 1}{1 + 2\epsilon_s} \right) \\ \\ A_l = B_l = 0 \end{cases} \tag{1.32}$$

From this solution, B_1 can be expressed as:

$$\begin{aligned} B_1 &= \frac{\mu}{4\pi\epsilon_0} - \frac{\mu}{2\pi\epsilon_0} \left(\frac{\epsilon_s - 1}{1 + 2\epsilon_s} \right) \\ &= \frac{\mu}{2\pi\epsilon_0} \left(\frac{1}{2} - \frac{\epsilon_s - 1}{1 + 2\epsilon_s} \right) \\ &= \frac{\mu}{4\pi\epsilon_0} \left(\frac{3}{1 + 2\epsilon_s} \right) \end{aligned} \tag{1.33}$$

Combining eq. 1.32 and eq. 1.33 $\psi_{in}(\mathbf{r},\theta)$ and $\psi_{out}(\mathbf{r},\theta)$ can be expressed as:

$$\begin{cases} \psi_{in}(\mathbf{r},\theta) = \frac{\mu\cos\theta}{4\pi\epsilon_0 r^2}\left\{1 + 2\frac{r^3}{R_C^3}\left(\frac{\epsilon_s - 1}{1+2\epsilon_s}\right)\right\} \\ \\ \psi_{out}(\mathbf{r},\theta) = \frac{\mu\cos\theta}{4\pi\epsilon_0 r^2}\left(\frac{3}{1+2\epsilon_s}\right) \end{cases} \tag{1.34}$$

Thus, combining the terms of eq. 1.34, the *reaction field* electrostatic potential $\psi(\mathbf{r},\theta)$ is:

$$\psi(\mathbf{r},\theta) = \frac{\mu\cos\theta}{4\pi\epsilon_0 r^2}\left\{1 + 2\frac{r^3}{R_C^3}\left(\frac{\epsilon_s - 1}{2\epsilon_s + 1}\right) + \frac{3}{2\epsilon_s + 1}\right\} \tag{1.35}$$

In particular, $\frac{\mu\cos\theta}{4\pi\epsilon_0 r^2}$ models the induced polarization due to the dipole, the term $\frac{\mu\cos\theta}{4\pi\epsilon_0 r^2} 2\frac{r^3}{R_C^3}\frac{\epsilon_s - 1}{2\epsilon_s + 1}$ is called *Onsager reaction field* and it measures the electric field which acts upon the dipole as a result of the electric displacements induced by its own presence and $\frac{\mu\cos\theta}{4\pi\epsilon_0 r^2}\frac{3}{2\epsilon_s+1}$ is the external moment of the immersed dipole and it determines the force which the dipole exerts upon a distant charge in the dielectric. Eq. 1.35 can be extended for polyatomic systems as:

$$\psi(\mathbf{r},\theta) = \sum_i^N \sum_{j>i}^N \frac{q_i q_j}{4\pi\epsilon_0 r^2}\left\{1 + 2\frac{r^3}{R_C^3}\left(\frac{\epsilon_s - 1}{2\epsilon_s + 1}\right) + \frac{3}{2\epsilon_s + 1}\right\} \tag{1.36}$$

where N is the total number of atoms in the system.

Barker and Watts [49] implemented the reaction field approach for the first time, to study dielectric properties of water models in Monte Carlo simulations. One of the advantages of the reaction field scheme is it is easiness of implementation in molecular dynamics routines. Furthermore, the inclusion in a cutoff-based Coulombic equation of the reaction field and dielectric polarization onto the dipole improves the simulated properties over straight-cutoff truncation methods and significantly reduces the straight-cutoff artefacts.

1.5 Molecular Dynamics

Molecular dynamics (MD) is an integration technique, which solves Newton's equations of motion, eq. 1.13, for each atom of the system, and it allows to sample the potential energy landscape of a biomolecule. A good definition for MD is given by Frenkel and Smith [40]

> "Molecular Dynamics simulations are in many respects very similar to real experiments. When we perform a real experiment [...]

> we prepare a sample of the material that we wish to study. We connect this sample to a measuring instrument, and we measure the property of interest during a certain time interval [...] In a Molecular dynamics simulation, we follow exactly the same approach. First, we prepare a sample: we select a model system consisting of N particles and we solve Newton's equations of motion for this system until the properties of the system no longer change with time"

The practical implementation of a molecular dynamics engine follows these steps:

1. Selection of initial random velocities for all i-th atoms of the system, drawn from the Maxwell-Boltzmann distribution [40] $p(\mathbf{v}_i) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} e^{-\beta\left(\frac{m_i \mathbf{v}_i^2}{2}\right)}$ where m_i is the mass of the i-th atom and $\mathbf{v}_i$ its velocity vector;
2. Evaluation of the potential energies $U(\mathbf{r})$ for the entire system;
3. Evaluation of the force acting on each atom, as $F_i = -\nabla_i U_i(\mathbf{r})$ where ∇_i denotes the gradient with respect to $i = x, y, z$;
4. Integration of Newton's equation of motion;
5. Update all the atomic positions and restart from step 1.

For the last step, a large class of algorithms is available. A popular one is the Verlet algorithm [50], which derives the positions at time $t + \Delta t$ without using the velocities, as:

$$r(t + \Delta t) \simeq 2r(t) - r(t - \Delta t) + \frac{f(t)}{m}\Delta t^2 \tag{1.37}$$

where $r(t)$ are the actual particles' positions, $r(t-\Delta t)$ the positions at a $t-\Delta t$ time and Δt is the timestep, typically about 1 or 2 fs. For all the simulations described in this work, the *velocity Verlet* algorithm, which provides both the atomic positions and velocities at the same instant of time:

$$\begin{cases} r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{f(t)\Delta t^2}{2m} + \mathcal{O}(\Delta t^3) \\ \\ v(t + \Delta t) = v(t) + \frac{\Delta t}{2m}(f(t + \Delta t) + f(t)) + \mathcal{O}(\Delta t^3) \end{cases} \tag{1.38}$$

where $\mathcal{O}(\Delta t^3)$ denotes all the series terms with index greater than 2. In practice, eq. 1.38 is split in three parts:

$$\begin{aligned} v(t+\frac{\Delta t}{2}) &= v(t) + \frac{f(t)\Delta t}{2m} \\ r(t+\Delta t) &= r(t) + r(t+\frac{\Delta t}{2})\Delta t \\ v(t+\Delta t) &= v(t+\frac{\Delta t}{2}) + \frac{f(t+\Delta t)\Delta t}{2m} \end{aligned} \tag{1.39}$$

where the first equation computes the half-step velocity, using the information retrieved from step at time t. This is allows the calculation of positions at time $t+\Delta t$. Finally, from the new atomic positions defined by $r(t+\Delta t)$ the half-step velocity is updated to the full velocity $v(t+\Delta t)$. Although this approach may seem tricky at a first glance, it guarantees a lower memory usage than eq. 1.37.

Because Newton's equation of motion conserves the total energy, MD simulations naturally form the NVE ensemble. Algorithms that connect the system to a thermostat or barostat allow the sampling of the NVT or NPT ensemble. For NVT ensembles is necessary to control the system's temperature. The temperature is related to the kinetic energy:

$$\sum_{i=1}^{N} \frac{1}{2} m_i v_i^2(t) = \frac{3}{2} N k_B T(t) \tag{1.40}$$

thus, to keep constant the temperature to a value T_{ref}, it is possible to scale the velocities by a factor ν:

$$\sum_{i=1}^{N} \frac{1}{2} m_i (\nu v_i(t))^2 = \frac{3}{2} N k_B T_{ref} \tag{1.41}$$

Thus, the scaling factor ν will be equal to:

$$\nu = \sqrt{\frac{T_{ref}}{T(t)}} \tag{1.42}$$

where $T(t)$ is the current temperature at the step at time t. However, this scheme would be not really useful, as there would be no kinetic energy fluctuations. Eq. 1.42 can be modified, by supposing the system is coupled to a thermal bath [51] with a temperature $T_{bath} = T_{ref}$. In this case the scaling factor ν is equal to:

$$\nu = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_{bath}}{T(t)} - 1\right)} \tag{1.43}$$

where τ is a parameter that describes how tightly the system is coupled to the bath, T_{bath} is the bath temperature, while Δt denotes the time difference between the steps. Another approach is the *Andersen thermostat* [52]. In this case the system is coupled to a thermal bath by randomly selecting a subset of particles at the beginning of each MD step, whose velocities are scaled. This thermostat reflects random collisions between the system's particles and the heat bath. The probability that a given particle is selected at time given a time step Δt is:

$$\mathcal{P}_{Andersen} = 1 - e^{-f\Delta t} \tag{1.44}$$

where f is the collision frequency. Then, each component of the particle velocity is set to:

$$\mathbf{v}_i(t) = \sqrt{\frac{k_B T}{m}} R \tag{1.45}$$

where R is a random number chosen from a normal distribution with mean of zero and variance one.

To keep the pressure constant the volume needs to be controlled. For a molecular system, the pressure can be monitored using the *virial thereom* [53]

$$P = \frac{2}{3V(t)} \langle \sum_{i=1}^{N} \frac{1}{2} m_i v_i(t)^2 \rangle - \langle \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{dU(\mathbf{r}_{ij}}{d\mathbf{r}_{ij}} \mathbf{r}_{ij} \rangle \tag{1.46}$$

where $V(t)$ is the system volume at time t, $\langle \sum_{i=1}^{N} \frac{1}{2} m_i v_i(t)^2 \rangle$ is the average kinetic energy of the system at time t,$\langle \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{dU(\mathbf{r}_{ij})}{d\mathbf{r}_{ij}} \mathbf{r}_{ij} \rangle$ is the *virial coefficient*, which expresses the average force acting on atoms i and j. The control onto the virial coefficient allows to keep constant the pressure to a value P_{ref}. As an example, the Berendsen barostat [54] controls the pressure P of the system starting from the pressure rate equation:

$$\frac{dP(t)}{dt} = \frac{P_{ref} - P(t)}{\tau} \tag{1.47}$$

where τ is the barostat relaxation time constant (usually 10 ps). For each MD step the volume is scaled by a factor η:

$$\eta(t) = 1 - \frac{\Delta t}{\tau} \gamma (P_{ref} - P(t)) \tag{1.48}$$

where $\gamma = \frac{1}{k_B T} \frac{\langle U(\mathbf{r})^2 \rangle - \langle U(\mathbf{r}) \rangle^2}{\langle U(\mathbf{r}) \rangle}$ is the isothermal compressibility of the system. The scaling factor η can be *isotropic*, thus the same scaling is applied for x, y and z directions of the unit cell, otherwise *anisotropic*. γ is usually a specific constant, which is often referred to the liquid water compressibility.

Another popular barostat is the Monte Carlo barostat [55, 56]. At regular intervals the box is scaled by a factor s:

$$s = \left(\frac{V(t) + \delta V}{V(t)} \right)^{1/3} \tag{1.49}$$

where $V(t) + \delta V$ is the proposed new volume and $V(t)$ the current volume size. The change in volume is accepted or rejected based on the weight function:

$$\Delta W = \Delta U(\mathbf{r}) + P(t)\delta V - N k_B T \ln \left(\frac{V + \delta V}{V} \right) \tag{1.50}$$

where $\Delta U(\mathbf{r})$ is the change in potential energy from the volume change, $P(t)$ is the system pressure, N is the number of molecules in the system. If $\Delta W \leq 0$ the change is accepted, if $\Delta V > 0$ the change is accepted with probability $e^{-\frac{\Delta W}{k_B T}}$. Monte Carlo barostat tends to be less computationally expensive than Berendsen barostat, since the virial coefficient is not computed. Other methodologies to keep temperature and pressure constant in MD simulations are the Parrinello-Rahman barostat [57] and extended systems thermostat and barostat [58].

Finally, it is often necessary to apply constraints on the solutes' atoms (e.g. SHAKE [59], SETTLE [60] and LINCS [61]), to allow the use of larger integration time step. As a matter of fact bond oscillations have a relatively high frequency and low amplitude, which would oblige the integration time step to be 1 fs or less. However, replacing the bond vibrations with holonomic constraints allow the integration time step to be at least 2 fs, halving the computational cost of the molecular dynamics.

1.6 Alchemical free energy calculations

According to eq. 1.7, the Gibbs free energy could written as:

$$\begin{aligned} G_{NPT} &= -k_B T \ln Z_{NPT} \\ &= k_B T \ln \frac{1}{Z_{NPT}} \\ &= k_B T \ln \frac{N! h^{3N}}{\int \int e^{-\beta(U(\mathbf{q}^N)+pV)} d\mathbf{q}^N dV} \end{aligned} \tag{1.51}$$

It follows that the computation of G_{NPT} would require the sampling of the phase space of a system using MD simulations. As a matter of fact eq. 1.51 can be computed as an ensemble average. Considering the equality:

$$1 = \frac{1}{(8\pi^2 V)^N} \int \int e^{-\beta(U(\mathbf{q}^N)+pV)} e^{\beta(U(\mathbf{q}^N)+pV)} d\mathbf{q}^N dV \tag{1.52}$$

where the constant factor $\frac{1}{(8\pi^2 V)^N}$ arises from a total integration on the translational and rotational degrees of freedom on the phase space, eq. 1.51, can be expressed as:

$$\begin{aligned} G_{NPT} &= k_B T \ln \frac{N! h^{3N}}{(8\pi^2 V)^N} \frac{\int\int e^{\beta(U(\mathbf{q}^N)+pV)} e^{-\beta(U(\mathbf{q}^N)+pV)} d\mathbf{q}^N dV}{\int\int e^{-\beta(U(\mathbf{q}^N)+pV)} d\mathbf{q}^N} \\ &= k_B T \ln \frac{N! h^{3N}}{(8\pi^2 V)^N} \int\int e^{\beta(U(\mathbf{q}^N)+pV)} \mathcal{P}(\Gamma) d\mathbf{q}^N dV \\ &= k_B T \ln \frac{N! h^{3N}}{(8\pi^2 V)^N} \langle e^{\beta(U(\mathbf{q}^N)+pV)} \rangle \end{aligned} \tag{1.53}$$

However, estimation of the free energy through the ensemble average in eq. 1.53 is not feasible with MD simulations. MD simulations are rarely able to sample high energy configurations, which make a large contribution to the ensemble average due to the sign of the potential energy in the exponential [62]. Furthermore, given the presence of many high energy states in a molecular system, the direct estimation of the free energy for a molecule using eq. 1.53 results in a vain task.

To solve this complicated task, eq. 1.53 can be reformulated, focusing on the calculation of the free energy differences between thermodynamic states. This can be done with the Zwanzig relation [23] also called *free energy perturbation* (FEP) [25]. Employing statistical mechanics, the free energy change for a thermodynamic system changing from a state A to a final state B, can be expressed as:

$$\Delta G = -k_B T \ln \frac{Z_{NPT}^{N+B}}{Z_{NPT}^{N+A}} \tag{1.54}$$

where Z_{NPT}^{N+X} is the isothermal-isobaric configurational integral for a system with N solvent molecules and, for instance here, a solute in a chemical state $X{=}A$ or B. Eq. 1.54 can be expanded as:

$$\Delta G = -k_B T \ln \frac{\int\int e^{-\beta(U_B(\mathbf{q}^N)+pV)} d\mathbf{q}^N dV}{\int\int e^{-\beta(U_A(\mathbf{q}^N)+pV)} d\mathbf{q}^N dV} \tag{1.55}$$

Since the potential energy function for both states can be written as:

$$\Delta U(\mathbf{q}^N) = U_B(\mathbf{q_A}^N) - U_A(\mathbf{q_B}^N) \tag{1.56}$$

where $U_A(\mathbf{q_A}^N)$ is the potential energy function for the system A (or *reference system*), $U_B(\mathbf{q_B})^N$ the potential energy for molecule B (or *target system*),

eq. 1.55 can be rewritten as:

$$\begin{aligned}
\Delta G &= -k_B T \ln \frac{\int\int e^{-\beta(U_B(\mathbf{q}^N)+PV)} d\mathbf{q}^N dV}{\int\int e^{-\beta(U_A(\mathbf{q}^N)+PV)} d\mathbf{q}^N dV} \\
&= -k_B T \ln \frac{\int\int e^{-\beta(\Delta U(\mathbf{q}^N)} e^{-\beta(U_A(\mathbf{q}^N)+PV)} d\mathbf{q}^N dV}{\int\int e^{-\beta(U_A(\mathbf{q}^N)+PV)} d\mathbf{q}^N dV} \\
&= -k_B T \ln \int\int e^{-\beta(\Delta U(\mathbf{q}^N)} \mathcal{P}_A(\mathbf{q}^N, V) d\mathbf{q}^N dV \\
&= -k_B T \ln \langle e^{-\beta(\Delta U(\mathbf{q}^N))} \rangle_A
\end{aligned} \tag{1.57}$$

where $\langle ... \rangle_A$ denotes the ensemble average over the configurations samples from the reference state system A. This approach was first introduced by Zwanzig [23]. Thus, any free energy ΔG (e.g. due to binding or hydration), between two molecular states A and B is the Boltzmann weighted probability of the difference $\Delta U(\mathbf{q}^N)$ of the potential energies for A and B. Computationally, FEP can be run in a molecular dynamic simulation for a system A, where at each step the difference in potential energy between the end states is computed and stored. Eq. 1.57 can be also reversed to compute the backward transformation, from B to A:

$$\Delta G = -k_B T \ln \langle e^{\beta(\Delta U(\mathbf{q}^N))} \rangle_B \tag{1.58}$$

However, this method suffers from numerical issues due to an asymmetry in the convergence properties [63] of eq. 1.57 and eq. 1.58. As described by Wu and Kofke [63, 64] eq. 1.57 has to show a decent overlap between the probable configurations in the phase space of molecule A and B to work properly.

Considering fig. **3** it is possible to have three scenarios:

1. A and B share common configurations with high probability;
2. A and B's have a partial overlap between configurations;
3. A and B's do not share any similar configuration;

In the first case, fig **3**A, the FEP equation can work properly for the transformation A→B, giving accurate results. However, when performing the reverse transformation, B→A, FEP could show some inconsistent results, as the important configurations of A may be not be a subset of the important configurations of B. In the case of partial overlap, fig **3**B, both transformations will result in a inconsistent free energy estimation. Depending on how much of the important configurations are shared in the overlap, in this case

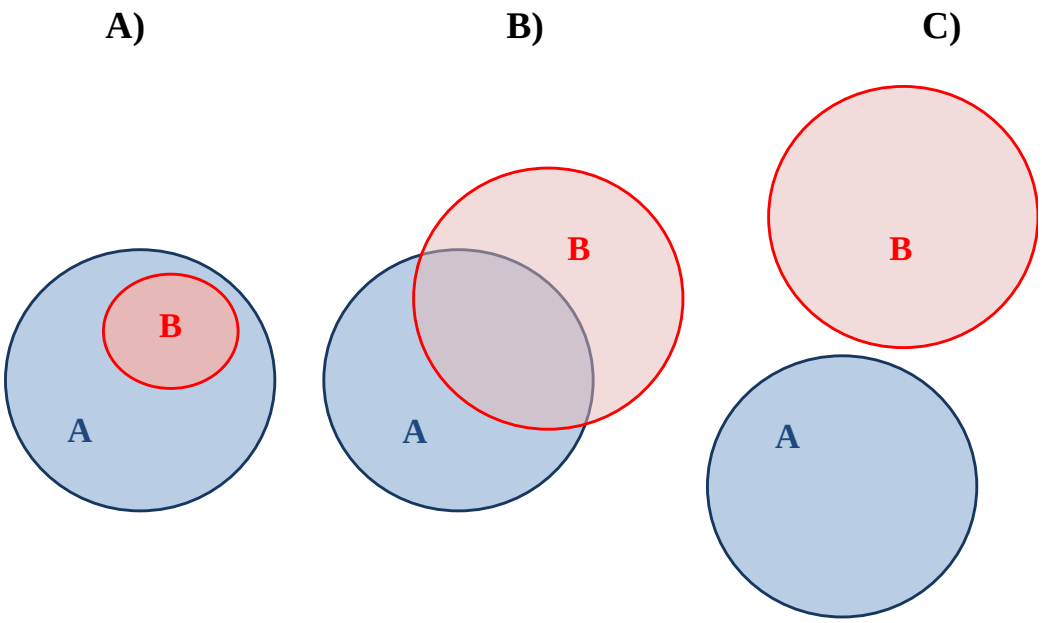


Figure 3: Free energy perturbation requires an overlap between the solutes'phase spaces. A) molecules A and B phase spaces overlap, in particular B is a subset of the phase space of A - meaning that both molecules can adopt similar configuration with high probability; B) molecules A and B shares a common subset; C) molecules A and B phase spaces do not overlap.

there could be a systematic offset in the free energy estimations. Finally, FEP fails to give reasonable results, within a reasonable amount of computing time, if there is no overlap between molecules configurations, as all the configuration sampled in one system are not important configurations in the other phase space system. However, it is important to remark that both in the case of partial overlap and no overlap for an ideal infinite sampling the free energy will tend to be asymptotically correct.

An alternative to equation 1.57 is the multistage calculation. In this case, the transformation A→B, or the backward B→A, is subdivided into intermediate simulations, where the force field parameters of the initial state are coupled to λ, defined in [0...1]. λ gradually modifies the initial state force field parameters into the final state one, passing through *alchemical intermediates*. These intermediates behave as a continuum between the two

end states, but they do not need to have a physical meaning.

The free energy changes can be computed with the *finite difference thermodynamic integration* [65]. For each MD step in a simulation carried on at λ_i, a difference between potential energies at λ_i and $\lambda_i + \Delta\lambda$ is evaluated, to collect two free energy terms as :

$$\begin{cases} \Delta G(\lambda_i \to \lambda_i + \Delta\lambda) = -k_B T \ln\langle e^{-\beta(U(\lambda_i+\Delta\lambda)-U(\lambda_i))}\rangle_{\lambda_i} \\ \\ \Delta G(\lambda_i \to \lambda_i - \Delta\lambda) = -k_B T \ln\langle e^{-\beta(U(\lambda_i-\Delta\lambda)-U(\lambda_i))}\rangle_{\lambda_i} \end{cases} \tag{1.59}$$

Then, a free energy gradient at λ_i is retrieved with finite differences:

$$\frac{\partial G}{\partial \lambda_i} = \frac{\Delta G_(\lambda_i \to \lambda_i + \Delta\lambda) - \Delta G(\lambda_i \to \lambda_i - \Delta\lambda)}{2\Delta\lambda} \tag{1.60}$$

where $\Delta\lambda$ is a control parameters for the gradient calculation (usually 0.01). Once all the MD simulation at each λ are done, all the gradients are collected and the final free energy change is computed by *thermodynamic integration*:

$$\Delta G_X = \int_{\lambda=0}^{\lambda=1} \frac{\partial G}{\partial \lambda} d\lambda \tag{1.61}$$

where the integral is performed across all the gradients values.

A further development of the multistage approach is the use of thermodynamic cycle, shown in fig. **4**. The thermodynamic cycle relies on the definition of free energy as a state function. Indeed, the free energy variation along a closed thermodynamic path must be equal to zero. Furthermore, thermodynamic cycles are reversible, which guarantees to consider forward and backward transformations. The use of free energy calculation with thermodynamic cycle allows the computation of free energy quantities that may be compared to experimental values. The transformation is performed between two phases p_1 and p_2, e.g. the hydration free energy will be computed between a vacuum and a solvated phase, giving a total free energy change ΔG_{p1} and ΔG_{p2} respectively. Thus, the free energy ΔG can be computed as the difference of the two legs of the thermodynamic cycle:

$$\Delta G = \Delta G_{p2} - \Delta G_{p1} \tag{1.62}$$

Computationally, this approach can be implemented with a *single topology* paradigm, which makes use of a common topology file that may describe the two molecules.

A common computational problem that arises in this type of simulation is the so-called "end-point catastrophes" [62]. At the end states some atoms

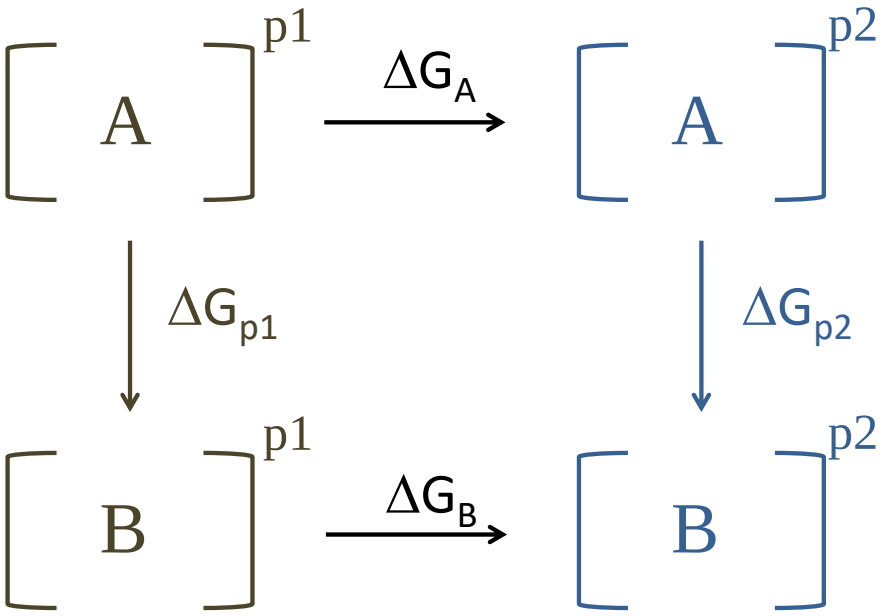


Figure 4: Thermodynamic cycle for a relative alchemical free energy calculation. Molecule end states are simulated between two phases, p_1 and p_2 (e.g. vacuum and solvated phases). Coupling the force field parameter of A to λ, the molecule can be transformed into its final state B. Each transformation is associated with a free energy change ΔG_{p1} and ΔG_{p2}. The desired free energy is computed from the difference of the two vertical legs of the cycle: $\Delta G = \Delta G_{p2} - \Delta G_{p1} = \Delta G_B - \Delta G_A$

may become "non-interacting", as for methane to ethanol transformation. In this case, the inter-atomic interactions may be unstable, due to the presence of steric clashes, leading to numerical instabilities. To prevent this situation, a soft core potential energy function is introduced in the MD engine:

$$U_{soft}(\lambda) = (1-\lambda)4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{(\lambda\delta\sigma_{ij}+r_{ij}^2)^6}\right) - \left(\frac{\sigma_{ij}^{6}}{(\lambda\delta\sigma_{ij}+r_{ij}^2)^3}\right)\right] + \frac{(1-\lambda)^n q_i q_j}{4\pi\epsilon_0\sqrt{(\lambda+r_{ij}^2)}} \qquad (1.63)$$

where δ and n are the soft core parameter for softening the Lennard-Jones and Coulombic interactions. It is clear that when the distance between atoms is zero (steric clashes) the potential is numerically computable, due to the presence of $\lambda\delta\sigma_{ij}$.

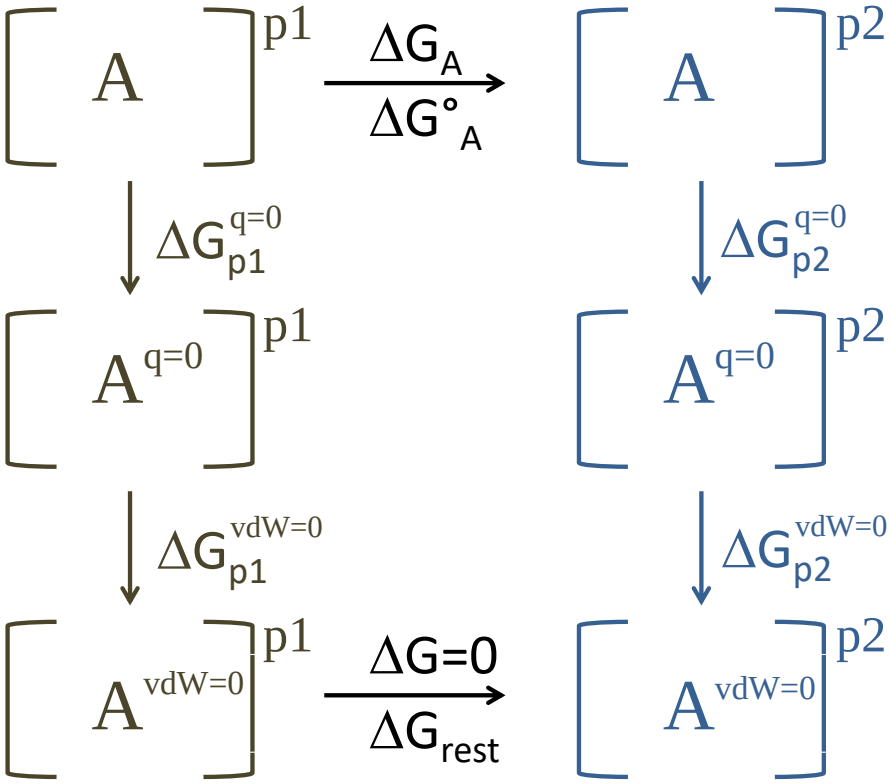


Figure 5: Thermodynamic cycle for absolute free energy calculation with the double annihilation scheme. The solute A is simulated in two phases p_1 and p_2 (e.g. solvated and complex phase). Initially, a discharging step is carried on by switching off A's partial charges, giving a total free energy changes $\Delta G_{p1}^{q=0}$ and $\Delta G_{p2}^{q=0}$. Subsequently, A's van der Waals parameters are turned off, retrieving free energy changes $\Delta G_{p1}^{vdW=0}$ and $\Delta G_{p2}^{vdW=0}$. The absolute free energy ΔG can be computed from the two legs of the cycle. In the case of host-guest system, a standard state correction may be necessary, ΔG_{rest}°, giving a standard free energy ΔG°.

A special case transformation is when a solute A is mutated into a "non-interacting" molecule, as if it were in an ideal thermodynamic state, as shown in fig. **5**, in which special considerations - binding or hydration - may apply (e.g. standard states if molality of process changes). In particular, this approach is called *double annihilation scheme*, which is often employed to compute absolute binding or hydration free energies [66, 67]. The double annihilation transforms the initial molecule A into an ideal thermodynamic state B, where the molecule does not have any interactions with the surrounding environment. Initially, molecule A's partial charges are switched off in a

discharging step, giving a free energy change $\Delta G_{p1}^{q=0}$ and $\Delta G_{p2}^{q=0}$ from phases p_1 and p_2 respectively. Following, the van der Waals terms of A's atoms are turned off in the *vanishing step*, resulting in free energy changes $\Delta G_{p1}^{vdW=0}$ and $\Delta G_{p2}^{vdW=0}$. The final free energy ΔG (e.g. free energy of binding) is computed as:

$$\Delta G = (\Delta G_{p2}^{q=0} + \Delta G_{p2}^{vdW=0}) - (\Delta G_{p1}^{q=0} + \Delta G_{p1}^{vdW=0}) \tag{1.64}$$

In the case of free energy of binding calculations for host-guest systems, eq. 1.64 can be decomposed into configurational integrals, giving:

$$\begin{aligned} \Delta G_{bind} &= -k_B T \ln \frac{Z_G^{q=0} Z_G^{vdW=0} Z_{HG} Z_{HG}^{q=0}}{Z_G Z_G^{q=0} Z_{HG}^{q=0} Z_{HG}^{vdW=0} Z_H} \\ &= -k_B T \ln \frac{Z_{HG}}{Z_H Z_G} \end{aligned} \tag{1.65}$$

where $Z_G^{q=0}$ and $Z_G^{vdW=0}$ denote the configurational integral of the guest in the solvated phase for the discharging and vanishing step respectively, while $Z_{HG}^{q=0}$ and $Z_{HG}^{vdW=0}$ are the isothermal-isobaric configurational integral for the guest in the complex phase for the discharging and vanishing step respectively. This expression for the binding free energy contrasts with standard binding free energy, usually defined from experimental measurements [67]:

$$\Delta G_{bind}^{\circ} = -k_B T \ln \frac{Z^{HG}}{Z^G Z^H} \frac{V}{V^{\circ}} \tag{1.66}$$

where V is the volume available to the guest in the bound phase and V° denotes the reference standard volume, typically 1 M (or 1660 ^{3}). Computationally, the standard state can be retrieved by applying a standard state correction. Usually, in the vanishing step the guest is restrained in the vicinity of the host. This procedure ensures that the guest would not drift off from the host's cavity, which could dramatically slow down the convergence of the computed free energy change. Depending on the type of restraint, it can be shown that the correction term is proportional to $\frac{V}{V^{\circ}}$, namely the ratio between the volume explored by the guest in the binding pocket, that can be computed numerically. Thus, the standard binding free energy can be eventually computed as:

$$\Delta G_{bind}^{\circ} = (\Delta G_{p2}^{q=0} + \Delta G_{p2}^{vdW=0}) - (\Delta G_{p1}^{q=0} + \Delta G_{p1}^{vdW=0}) + \Delta G_{restr} \tag{1.67}$$

1.6.1 Outline of the thesis

Alchemical free energy calculations rely on a solid mathematical and physical background, which is appealing for the development of robust computational

methods that may assist the drug discovery process. There is hope that alchemical methods can reduce the time and costs needed to carry out preclinical drug discovery research by reducing the number of compounds that have to be synthesized. However, it is currently hard to translate the "alchemical science" into a robust engineering tool, that can be used routinely to obtain accurate and reliable free energy estimations. The main goal of this thesis is to examine in depth the alchemical free energy calculation protocols, to make them better suited for high-throughput workflows.

Chapter 2 will discuss to what extent alchemical free energy calculations are reproducible between different simulation engines. As a result from a collaborative effort, SOMD [68, 69], GROMACS [70], CHARMM [71] and AMBER [72, 73] were employed to compute relative hydration free energies for a small dataset of organic molecules, which was considered as a reference to understand the limit of reproducibility of free energy calculations. Code specific implementation details in the free energy algorithm were studied and new protocols to enhance reproducibility were validated.

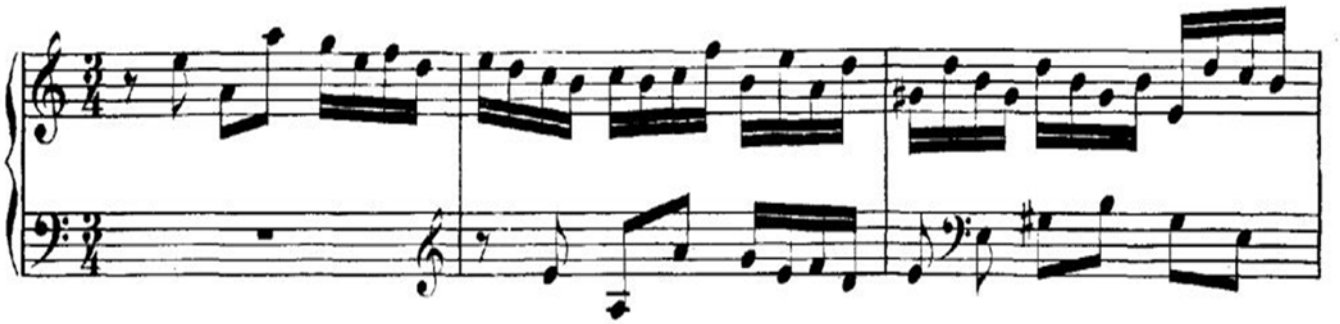
Chapter 3 will deal with the validation of protocols for lipophilicity coefficient measures. logP was computed with alchemical free energy calculations for 5 different molecules, with an increasing fluorination level. Computational models not only were able to retrieve the experimental trend, but they also gave structural insights into the solvation process. Then, logD coefficients were computed for 53 drug-like molecules in the context of the SAMPL5 [74] blind challenge. Two different models were considered in order to take into account the possible presence of protonated species in solvated phases.

As an extension of the SAMPL5 challange, chapter 4 will treat the validation of protocols for host-guest binding affinities predictions [75]. In this case, three different free energy models were compared with respect to experimental data, defining the best protocol of simulation for accurate binding affinities calculations.

Finally, chapter 5 will deal with finite size artefacts in alchemical free energy calculations. These artefacts arise when changes in the net-charge of a solute are performed during an alchemical simulation. This problem is well known since the early 1980s, when FEP were applied to polar solvent calculations [76, 77]. Further development for the sodium ion was done by Kastenhölz and Hünenberger [78, 79, 80]. Recently, Rocklin and Reif [81, 82] assessed protocols of simulation to correct for these problems. However, a full understanding and development for the Barker Watts reaction field approach [49] does not exist yet. This study will draw a comparison between all the various correction scheme, delineating the best procedures to avoid finite size artefacts in the context of binding free energy calculations.

Finally, chapter 6 will draw a summary and conclusions for all the findings

of the thesis, assessing to what extent alchemical free energy calculation are ready for molecular drug design.



Chapter 2

Reproducibility of Free Energy Calculations Across Different Molecular Simulation Software

This chapter explores the reproducibility problem in alchemical free energy calculations. Relative alchemical free energies of hydration for a dataset of small organic molecules were computed using AMBER, CHARMM, GROMACS and SOMD, to validate and understand to what extent alchemical free energy calculations are reproducible. Achieving an acceptable level of reproducibility requires considerable attention to package–specific details and simulation protocols. It is hoped that the work reported here will lead to new benchmarks for the computational community to validate new and future versions of software for free energy calculations.

2.1 Introduction

The calculation of free energies via molecular simulations [83, 84, 85, 86, 87] has been particularly attractive as it promises to circumvent certain limitations of experimental approaches. Specifically, biomolecular processes can be understood at the atomic level and computational techniques could lead to a cheaper and more time effective drug discovery process, especially if they

can predict the properties of new molecules before their synthesis. Thus, a multitude of methods have been devised to make reversible work estimates accessible through computation.

Among all the possible computational techniques, the alchemical free energy calculations have been applied in various forms for several decades now since the early days of computer simulation [88, 89, 90, 91, 92, 25]. The alchemical method has gained renewed attention in recent years both in academia and pharmaceutical industry. One reason for this success is that alchemical calculations are firmly rooted in statistical thermodynamics and they could give asymptotically correct free energy estimates, namely they are correct for a given potential energy function in the limit of sufficient simulation time [93, 94, 95, 83]. The name "alchemical" comes from the nonphysical intermediates that often need to be created to obtain reliable estimates of free energy differences between physical end states, and because parts or all of a molecule may effectively appear or disappear in a transformation.

As explained in chapter 1 section 1.5, the alchemical free energy simulations rely on the concept of thermodynamic cycles [92]. As the free energy is a state function, the sum of free energy changes computed around any closed cycle must be zero. This also implies that the reversible work can be computed along conveniently chosen legs of the cycle, even if the cycle is artificial. For example, fig. **1** shows that the relative free energy of hydration can be computed along the vertical legs of the cycle, that is, following a non-physical but computationally more efficient calculation rather than the horizontal legs of moving a molecule from the gas phase to the liquid phase.

In the context of drug design, one of the most popular alchemical technique in use do the relative alchemical free energy calculations (RAFE). In this case there is a mutation between two end states molecules as shown in fig. **1**. Computationally RAFEs are thought to be less demanding than absolute alchemical free energy calculations, as they typically require fewer shorter simulations to yield converged free energy changes. Therefore, it is without surprise that this approach has recently gained increased attraction in the computation of relative free energies of binding between small molecules, e.g. drug or lead like molecules and biomolecules [96, 97].

RAFEs can be calculated by making use of the so–called single topology method. Single topology means that there is only one connected representation of the molecule to be transformed into another molecule. Atoms of a given type are directly transformed, typically by linearly scaling the force field parameters, into atoms of a different type. The single topology method offers a straightforward route to implement RAFE calculations.[98, 62, 25, 99] In typical implementations, a certain number of non-interacting "dummy" atoms must hold the place of disappearing/appearing atoms in order to balance the

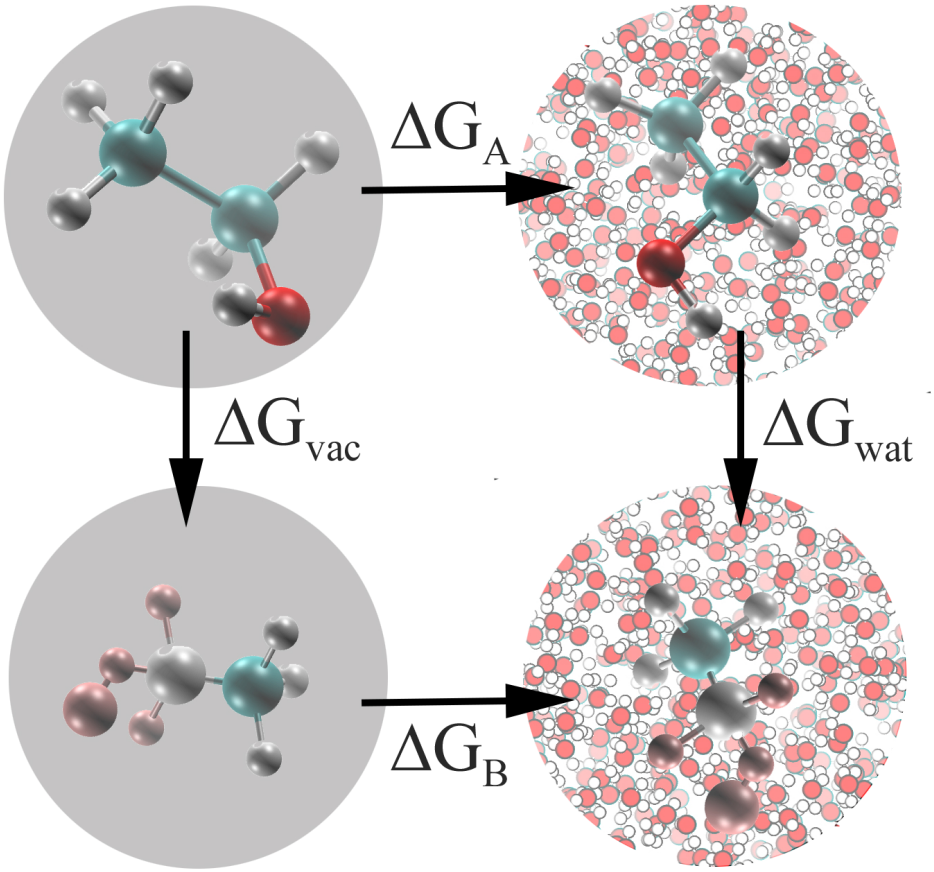


Figure 1: Thermodynamic cycle to compute the relative free energy of hydration $\Delta G_{hyd} = \Delta G_B - \Delta G_A = \Delta G_{wat} - \Delta G_{vac}$ for two molecules A and B. The example is for ethanol $\rightarrow$ methane transformation. Alchemical simulations are performed along the non-physical vertical legs, while the horizontal legs illustrate the physical process of moving a molecule from vacuum to solvated phase. For some transformation, as in ethanol$\rightarrow$methane, some atoms may become non-interacting (light red atoms in methane molecule).

number of atoms in both end states.

Generally, modern MD software (e.g. AMBER, [72] CHARMM, [71] GROMACS, [70] and SOMD. [68, 69]) support relative free energy calculation in a single topology paradigm [100]. However, in spite of this rigorous background, consistency and reproducibility are a matter of concern. The re-

producibility of free energy results among computer codes must be ensured, to allow a robust use of free energy calculations in molecular design problems. However, a systematic study of this problem in the alchemical context is missing. Recent related efforts have been reported by Shirts *et al.* [101] and by Hasse *et al.* [102]. The former study compares absolute free energy results for host-guest systems in the context of the SAMPL5 challange.This study is limited to compare the Coulombic potential differences that arise between AMBER [72], CHARMM [71], GROMACS [70], LAMMPS [103] and DESMOND [104] which, ultimately, lead to a different free energy value. Hasse's study went further, comparing liquid densities across DLPOLY [105], GROMACS [70] and NAMD [71]. Differently from this work, the aim of Hasse's paper is to highlight how achieving a good reproducibility of results can be quite a complicated and surprising task.

Given the importance for the drug discovery process, it is critical that free energy changes computed with different simulation software should be reproducible within statistical error, as this otherwise limits the transferability of potential energy functions and the relevance of properties computed from a molecular simulation to a given package. This is especially important as the community increasingly combines or swaps different simulation packages within workflows aimed at addressing challenging scientific problems [106, 107, 108, 109, 110].

This work assesses to what extent alchemical free energy calculations are reproducible among the packages AMBER, CHARMM, GROMACS and SOMD. The relative hydration free energies of a set of small organic molecules was taken as a reference to understand the reproducibility issue. A discussion about the obtained result with these packages will be reported, along with recommendation regarding simulation protocols, setup procedures and analysis techniques. Along with the predicted free energy values the experimental results are presented to allow the reader a comparison in terms of accuracy between the different codes. However, it is necessary that this study is focused onto the concept of reproducibility rather than precision and accuracy with respect to experimental data. This work is a result of a collaborative effort among the CCPBioSim (H. Loeffler), the University of California Irvine (D. Mobley and Guilherme Duarte Ramos Matos, the university of Chicago (B. Roux and D. Suh) and the university of Edinburgh. I personally contributed with SOMD simulations and to all the analyses.

2.2 Methods

One practical challenge is that the free energy methodologies used in one MD program are not always available in another package, or the same functionality is provided via different algorithms (e.g. algorithms for pressure and temperature scaling, integrators, etc). To circumvent some of these practical problems, relative free energies will be compared via three protocols:

1. In the "unified protocol" relative free energies are computed by scaling together all force field parameters i.e. partial charges, van der Waals parameters, and bonded parameters vary simultaneously along the alchemical path;

2. In the "split protocol" relative free energies are calculated by scaling separately the van der Waals parameters and the partial charges;

3. In the "absolute protocol" relative hydration free energies are retrieved as a the difference between two calculated absolute hydration free energies.

The following subsection will give initially an insight into the technical implementation of alchemical free energy calculations for AMBER, GROMACS, CHARMM and SOMD. Then, the relative free energy setup for each software will be discussed.

2.2.1 Alchemical Free Energy Implementations

One of the key difference in the free energy implementations among the four MD codes is the softcore potential [111, 112]. This can be generally written as a function of Coulombic and van der Waals terms:

$$V = 4\epsilon_{ij}(1-\lambda)\left[\left(\frac{\sigma_{ij}}{r_{LJ}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{LJ}}\right)^{6}\right] + (1-\lambda)^{n}\frac{q_i q_j}{4\pi\epsilon_0 r_{Coul}} \tag{2.1}$$

where ϵ_{ij} and σ_{ij} are the van der Waals parameters for the well-depth and equilibrium distance respectively, λ is the alchemical coupling parameter, r_{LJ} and r_{Coul} are modified distance parameters, which are defined differently for each software, as well as the Coulombic power n, q_i and q_j are the i-th and j-th atom partial charge, ϵ_0 the dielectric permittivity in vacuum. Details for r_{LJ}, r_{Coul} and n are given in table **1** for each software.

Another difference is how codes are scaling the force field parameters. It is possible to have two methodologies: "parameter scaling" and "energy

scaling" [99]. In the former case each parameter is scaled individually, e.g. in the case of a harmonic bond or angle term, the force constant and the equilibrium distance/angle are scaled individually. In the latter case, the total energy is scaled, all at once, or, equivalently for each individual force field contribution. While free energy is a state function that depends only on the end points, the pathways taken by the two methods through state space or alchemical space are different.

AMBER. This code uses a hybrid dual/single topology approach. All terms are energy scaled. The perturbed group must be entirely duplicated, i.e. for `sander` this means two topology files with one end state each, and for `pmemd` both end states in one topology file. The code loads two separate input topologies that describe the end states of interest and allows users to map atoms between the two end–states that will share the same coordinates for the free energy calculation. Evaluation of the interactions involving these atoms as a function of the coupling parameter is done by default via linear scaling of the energy and forces of the end–states. Alternatively the user can request the use of a softcore potential. The non–bonded interactions of atoms that are not paired between the end–states are handled with a softcore potential. In addition, bonded terms involving different unpaired atoms are ignored. This in effect amounts to defining unpaired atoms as dummy atoms in one of the end–states. This approach will be called "implicit dummy protocol" since the procedure is handled automatically by the software through analysis of the end–state topologies rather than via explicit definition of dummy atoms in an input topology.

The code cannot handle bond length changes involving a constraint. There is only one global λ for parameter transformation. Protocols that couple only some parameters (split protocol) must be emulated through careful construction of topologies. For instance one can keep the LJ and bonded terms fixed at the initial state for a charge transformation. The setup for the two end–states must therefore use identical atom types with only the charges varying.

Alternatively it is possible for the user to construct an input topology of a single molecule that explicitly contains dummy atoms such that the desired end–states can be simulated. This is a similar approach to that employed by SOMD and GROMACS, and this will be called "explicit dummy protocol".

CHARMM. The PERT module duplicates the topology similarly to `sander` but mapped atoms are given in the topology only once. The module requires balancing with explicit dummy atoms. All energy terms are linearly scaled

by the coupling parameter λ. The PSSP softcore potential is applied to all atoms in the perturbed group, whose parameters can be found in tab **1**. The code can handle constraints of changing bond lengths in the perturbed group but this may cause incorrect results with PSSP softcores.There is only one global λ for parameter transformation, however, the scripting facilities in CHARMM allow run time modification of topologies e.g. by setting charges or LJ parameters to arbitrary values.

GROMACS. This code uses a single topology description. Bonded terms are strictly parameter–scaled, which requires proper balancing of multi–term dihedrals, i.e. each individual term in the Fourier series must have an equivalent in both end states. If the term does not exist it must be created with parameters zeroing its energy. The softcore potential applies to dummy atoms only determined from atoms having zero LJ parameters in the end states. The code allows changing bond lengths involving constraints within the perturbed group but this can lead to instabilities and wrong results. There are separate λs for LJ, Coulomb and bonded parameters (and some other possible terms in the potential) which allows easy implementation of split protocols.

SOMD. SOMD is a software built by linking Sire and OpenMM molecular simulation libraries. [68, 69] This code uses a single topology description. The alchemical state is constructed at run time from an input topology together with a "patch" (list of force field parameters to be modified). All dummy atoms needed to describe the transformation must be present in the initial state. Bond and angle terms are parameter–scaled while the dihedral term is energy–scaled. The softcore potential applies to atoms that become dummy atoms in one end–state. Dummy atoms are specified by a keyword in the patch file. The code cannot handle constraints of changing bond lengths in the perturbed group. There is only one global λ for parameter scaling. Separated protocols (see below) must be emulated through careful construction of the patch file.

2.2.2 RAFE Setup

The initial setup for all relative free energy simulations has been carried out with the tool FESetup (version 1.2). [109]. FESetup is a perturbed topology writer for AMBER, CHARMM, GROMACS, SOMD and NAMD [113]. The tool makes use of a maximum common substructure search algorithm to automatically compute atoms that can be mapped, i.e. atoms that have

a direct relationship to an equivalent atom in the other state – atoms undergoing atom type conversion or modification. The only current limit is that rings are required to be preserved [114]. With this strategy, a single topology description is achieved: any atom that does not match is made a dummy atom. FESetup allows equilibration of the solvated simulation systems and ensures that "forward" and "backward" simulations have the same number of total atoms. With SOMD the mass of each perturbed atom is taken as the mass of the heavier end–state atom (e.g. a hydrogen atom that is perturbed to a carbon atom has an atomic mass of 12 amu at all lambda values). The other codes use the atom masses of the initial state (AMBER, CHARMM) or allow the user to define how masses vary as a function of lambda (GROMACS). The tool creates all input files with control parameters, topologies and coordinates as required for RAFE simulations. Full details on FESetup can be found in [109].

Figure **2** shows all 18 transformations considered in the present study. In the limit of sufficient sampling, RAFE simulations should not depend on the "forward" and "backward" direction of change with respect to the coupling parameter λ. However to test for possible discrepancies, both direction simulations have been run.

The ethane $\rightarrow$ methanol transformation is traditionally regarded as a standard test for RAFE simulations [25, 115]. The other transformations are centered around mutations from and to methane, and are meant to mimic components of typical transformations that could be attempted in the context of protein–ligand binding calculations. The 2–cyclopentanylindole to 7–cyclopentanylindole (2–CPI to 7–CPI) transformation has been added to include both deletion as well as insertion of sub–parts of the perturbed group in one transformation, an aspect not tested by the other transformations. For neopentane $\rightarrow$ methane two alternative mappings have been considered, see Figure **2**. One mapping has methane matched to a terminal methyl and the other one has the methane carbon matched with the central carbon in neopentane. The first approach will be called "terminally mapped" and the second one "centrally mapped". The choice of the current dataset was motivated by keeping the focus on probing for reproducibility among various MD packages, with rigid and neutral molecules to minimize statistical sampling errors, and avoid difficulties with charged particles [81, 116]. The force field was chosen to be GAFF (version 1.8), [117] utilizing AM1/BCC charges for the solute [118, 119] and TIP3P for the solvent. [120] Charges were computed with the `antechamber` program and missing bonded and vdW terms were generated with the `parmchk2` program, both from the AmberTools16 distribution.

While the MD packages principally allow a "one–step" transformation [121],

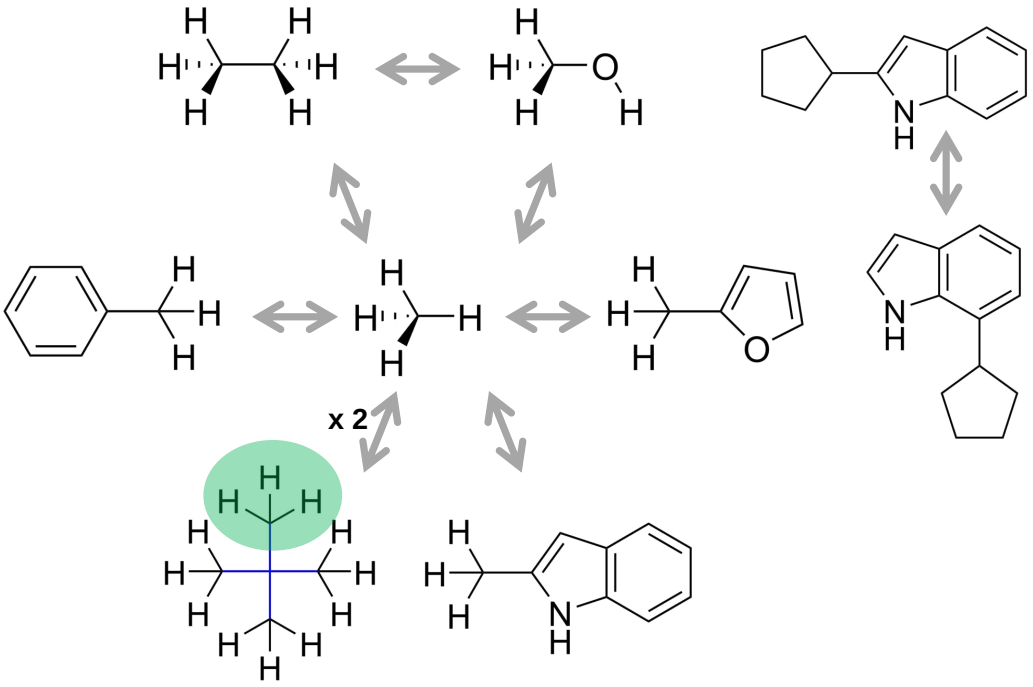


Figure 2: The thermodynamic cycles considered in this study. To compute the free energy of hydration, all pair–wise transformations have to be carried out once in solution and once in vacuum. Green and blue colours in neopentane show two alternative mappings for methane.

that is with both LJ and Coulombic parameters varied simultaneously (unified protocol), it has also been proposed that carrying out a split protocol may be more efficient. [122, 123, 124] In such a protocol the charges are transformed linearly between the end states followed by a mutation of the van der Waals parameters using a softcore potential on the LJ term only. [111, 112]. It is important to note that in the split protocol, charges have to be switched off before LJ parameters (and vice versa for the transformation in opposite direction) to avoid collapse of other atoms, e.g. solvents, onto a "naked" charge[125, 126, 121].

The production simulations were run at 298 K and 1.0 bar. Atomic masses were not changed along the alchemical transformations as this would affect only the kinetic energy, and would not contribute to the free energy change.

AMBER. The AMBER16 program was used for this set of free energy calculations. Typically 11 windows were used for charge mutations and 21 windows for vdW mutations. In some instances, steep variations in gradients were observed with this protocol and additional windows were added to obtain smoother integration profiles. The starting coordinates were usually taken directly from the pre–equilibrated setup step. In a very few cases it was necessary to use coordinates from the end of the simulation at a nearby λ state because of simulation instabilities. This happened in transformations with a larger number of dummy atoms. Absolute transformations were carried out using a one step protocol featuring 21 windows initially. For some perturbations additional windows were run in regions where the free energy gradients varied sharply. Each window was simulated for 2.5 ns, with the first 0.2 ns discarded prior analysis.

Water hydrogens (TIP3P) were constrained with SHAKE. None of the atoms in the perturbed group where constrained and hence the time step was set to 1 fs. An alternative protocol with SHAKE on bonds that do not change during transformation and a time step of 2 fs was also tested (see SOMD protocol below).

The temperature was controlled through a Langevin thermostat with a friction constant of $2.0\,\mathrm{ps}^{-1}$ and pressure rescaling through a Monte Carlo barostat with 100 steps between isotropic volume change attempts.

Long–range electrostatics in solution was handled with Particle Mesh Eward (PME) and an atom–based cutoff of 8.0 Å for the real-space Coulomb and vdW interactions. No cutoff was used for the vacuum simulations.

A Long Range Correction (LRC) term for truncated vdW interactions is applied during the MD simulations.

CHARMM. The version c40b1 was used for this set of free energy calculations. The PERT module was used to handle the alchemical transformations.

For the three protocols considered (split, unified and absolute), 21 evenly spaced windows were used and all windows were run for 1.5 ns with a timestep of 1 fs. Most windows used the same pre-equilibrated configuration. A few windows at the end-points (involving hydrogen being transformed to heavy atom or vice versa) were unstable due to steric clashes with starting coordinates and were equilibrated using 0.1 fs to 0.5 fs. Only water hydrogens (TIP3P) were constrained with SHAKE. Conditions of constant temperature and pressure control were maintained using the Berendsen weak coupling method, with a compressibility of $4.63 \times 10^{-5}\,\mathrm{atm}^{-1}$ and temperature and pressure coupling constants of $5.0\,\mathrm{ps}^{-1}$.

Long–range electrostatics in solution was handled with PME to order 6

with a cutoff of 12.0 Å for the real-space Coulomb and vdW interactions. No cutoff was used for the vacuum simulations. No LRC term was applied during the alchemical MD simulations but a solute-solvent LRC term was included in post-processing to calculate the final free energy.

The PSSP softcore potential function was used for the perturbed atoms. The PERT module currently does not currently support the force switching (option `VFSwitch`) for LJ potentials with softcores. The CHARMM PARAM27 force fields, however, is parameterized to use force switching [71]. Accordingly, we used the potential switching only (option `VSwitch`) with an inner cutoff of 10 Å and outer cutoff of 12 Å.

GROMACS. GROMACS version 4.6.7 was used to carry out this set of free energy calculations.

For the split and unified protocols 31 and 16 λ states were adopted, respectively, and simulated for 4.2 ns with time steps of 1.0 fs in water and vacuum. In the absolute protocol 20 alchemical steps were used both for the discharging and vanishing steps [127, 128] run for 5 ns. All the free energies were calculated from Langevin dynamics at 298 K. A friction coefficient of 1.0 ps/$\mathrm{m_{atom}}$ was used, where m_{atom} is the the mass of the atom. No holonomic bond or angle constraints were used. A Parrinello–Rahman barostat with $\tau_p = 10$ ps and compressibility equal to 4.5×10^{-5} bar^{-1} was used.

Two methods were used to calculate electrostatic interactions: Particle Mesh Ewald (PME) and charge group-based Reaction Field with a dielectric of 78.3, as implemented in the software. PME calculations were of order 6 and had a tolerance of 1.0×10^{-6}, with a grid spacing of 1.0 Å. We set the real-space electrostatic and vdW cutoffs to 10.0 Å; a switch was applied to the latter starting at 9.0 Å. A cutoff 50.0 Å was used for the vacuum simulations.

A Long Range Correction (LRC) term for truncated vdW interactions was applied during the MD simulations.All transformations required the use of softcore potentials to avoid numerical problems in the free energy calculation. We chose the 1–1–6 softcore potential for LJ terms (α=0.5 and σ=0.3) for atoms whose parameters were being perturbed and used the default softcore Coulomb implementation in paths where charges, LJ, and bonded terms were modified together, but no soft core potentials were applied to Coulomb interactions when electrostatic interactions were modified separately.

SOMD. This set of free energy calculations was carried out with SOMD from the Sire 2016.1 release. [68, 69]

Each alchemical transformation was divided into 17 evenly spaced win-

dows and simulated for 2 ns each both in water and in vacuum. The absolute hydration free energies were computed by annihilating non-bonded interactions of the solute in two steps. In the first step the free energy change for discharging the solute was computed. In the second step the free energy change for turning off the Lennard-Jones terms of the discharged solute was computed. Each step was carried out using 17 evenly spaced windows.

The starting coordinates for each window were obtained by energy minimization of the same pre-equilibrated configuration generated by FESetup.

A velocity-Verlet integrator was employed with a 2 fs time step. Only bonds involving hydrogens which are not alchemically transformed were constrained. This approach is referred as the "unperturbed H bond constraint protocol". Temperature control was achieved with the Andersen thermostat, [52] with a stochastic collision frequency of $10\,\mathrm{ps}^{-1}$. A Monte Carlo barostat assured pressure control, with isotropic box edge scaling moves attempted every 25 time steps.

A shifted atom–based Barker–Watts reaction field, [49] with a dielectric constant of 78.3 was adopted for the solution phase simulations with a cutoff of 10 Å. A similar cutoff was used for LJ interactions. The reaction field was not employed in the vacuum legs, where a Coulombic potential without cutoff was used. A protocol to account for the different treatment of intramolecular electrostatics in vacuum and solution is described in the supporting information.

The softcore parameters were set to default values for all the transformations, specifically $n = 0$ for Coulombic interactions and $\alpha = 2.0$ for the LJ potential [129]. Additionally, an end-point correction for truncated VdW potentials was applied by post-processing of end-state trajectories [130, 131].

Table 1: Summary of the technical details for the relative hydration free energy calculations carried out with the various codes.

	AMBER	CHARMM	GROMACS	SOMD
Version	AMBER16	c40b1	4.6.7	2016.1
Module	`pmemd, sander`	`PERT`	`gmx`	`somd-freenrg`
Protocol	Split protocol	Unified protocol	Split protocol	Unified protocol
Number of λ windows	11 (charge mutations) 21 (vdW mutations)	21 evenly spaced	31 (charge mutations) 31 (vdW mutations)	17 evenly spaced
Starting coordinates	FESetup pre-equilibration	FESetup pre-equilibration	FESetup pre-equilibration	FESetup pre-equilibration
Simulation length per window	2.5 ns	1.5 ns	4.2 ns	2 ns
Timestep	1 fs	1 fs	1 fs	2 fs
Electrostatic method	PME	PME	PME	atom based RF
Solvated phase cutoff	8	12	10	10
Vacuum phase cutoff	no cutoff	no cutoff	50	no cutoff
Constraint	none	none	none	H-bonds not perturbed
LRC corrections	during MD	Post-processing	during MD	Post-processing
Barostat	Monte Carlo	Berendsen	Parrinello-Rahman	Monte Carlo
Thermostat	Langevin	Berendsen	Langevin	Andersen
Soft core parameters	$r_{LJ} = (2\sigma_{ij}^6\lambda + r_{ij}^6)^{1/6}$ $r_{Coul} = (\beta\lambda + r_{ij}^p)^{1/p}$ $n = 1$	$r_{LJ} = (2\lambda + r_{ij}^2)^{1/2}$ $r_{Coul} = (\beta\lambda + r_{ij}^2)^{1/2}$ $n = 1$	$r_{LJ} = (2\sigma_{ij}^6\lambda + r_{ij}^6)^{1/6}$ $r_{coul} = r_{LJ}$ $n = 1$	$r_{LJ} = (2\sigma_{ij}\lambda + r_{ij}^2)^{1/2}$ $r_{Coul} = (\lambda + r_{ij}^2)^{1/2}$ $n = 1$

2.2.3 Free Energy Estimations

The thermodynamic integration method (chapter 1, section 1.5) was used to compute relative hydration free energies, as this is supported by all the tested MD packages "out–of–the–box". Equation 2.2 computes the free energy as

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \frac{\partial G}{\partial \lambda} d\lambda \tag{2.2}$$

where $\frac{\partial G}{\partial \lambda}$ denotes a collected free energy gradient for each simulated phase. For all the softwares the `alchemical analysis tools` [132] is employed to solve eq. 2.2, performing a cubic spline interpolation across all the gradients values.

All RAFE simulations were run in triplicate in forward as well as backward direction for a total of 6 simulations per mutation. The final hydration free energy ΔG_{HYD} was computed as the average for each direction separately. For comparison the absolute (standard) hydration free energies for all molecules in Figure **2** was also computed.

To estimate the reliability and convergence of the results, the standard error of the mean (SEM) has been calculated. The SEM is defined as

$$SEM_{\Delta G_{HYD}} = \frac{\sigma}{\sqrt{n}} \tag{2.3}$$

where σ is the sample standard deviation and $n = 3$ is the number independent simulations run.

The mean unsigned error (MUE) was also employed to compare data sets.

$$\text{MUE} = \frac{1}{N} \sum_{\text{i}=1}^{N} |y_{\text{i}} - x_{\text{i}}| \tag{2.4}$$

where N is the total number of samples, y_{i} and x_{i} are the i–th data to be compared.

2.3 Results

Absolute hydration free energies have been considered as the standard reference of comparison, since for the present dataset they can be calculated with high precision [128]. Prior work has successfully compared calculated absolute hydration free energies across GROMACS and DESMOND codes. [133]

Table **2** summarizes results for the absolute hydration free energies. The precision of the calculated free energies is similar between AMBER, CHARMM and GROMACS, whereas the SOMD free energies are less precise. This may reflect differences in the lambda schedules and length of trajectories between the different codes. Nonetheless the standard errors are typically well under 0.1 kcal·mol^{-1}, thus it becomes meaningful to investigate small differences of a few tenths of kcal·mol^{-1} between codes.

The ΔG_{HYD} obtained with the various MD packages in this way agree quite well given statistical errors, although some larger deviations are apparent as well. GROMACS predicts a smaller ΔG_{HYD} for methanol by about 0.2 kcal·mol^{-1}. Similarly, there are some small discrepancies in the toluene, 2–methylfuran and 2–methylindole cases, where CHARMM produces slightly smaller ΔG_{HYD} . These small discrepancies may be due to the differences in calculated water densities between CHARMM and other codes (typically smaller by ca. 0.01 g·cm^{-3}). The largest deviation can be found for one of the largest molecules (7–CPI) with the AMBER result being less negative than with the other MD packages by 0.4–0.8 kcal·mol^{-1}. This particular discrepancy does not correlate with significant variations in density between AMBER and other codes.

As an additional check densities in the fully decoupled states have been computed and compared with a pure TIP3P water box density (0.980 g·cm^{-3}) [134]. The average densities across all simulations are 0.980±0.002 g·cm^{-3}, 0.973±0.002 g·cm^{-3}, 0.970±0.002 g·cm^{-3}, 0.976±0.003 g·cm^{-3} for AMBER, CHARMM, GROMACS and SOMD respectively. AMBER and GROMACS show higher densities presumably because a LRC term was applied during the MD simulations, whereas LRC terms for SOMD and CHARMM are only applied via post-processing of trajectories.

Table **3** shows the MUE between SOMD, GROMACS, AMBER and CHARMM. CHARMM produces figures that agree the most with other MD packages. The largest difference reaches 0.2 kcal·mol^{-1} for SOMD and GROMACS. Variabilities between the codes may be partly explained by differences in densities due to different treatments of long range electrostatics and vdW interactions.

Having established the predictive value from absolute transformations

Table 2: Absolute hydration free energies obtained from AFE calculations, in kcal·mol^{-1}, end-state densities, in g·cm^{-3}, and experimental absolute hydration free energy, extract from [127]. Uncertainties on the last decimal are given in parenthesis.

Solute	AMBER		CHARMM		GROMACS		SOMD		ΔG_{HYD}^{exp}
	Free energy (kcal·mol^{-1})	Density (g·cm^{-3})	Free energy (kcal·mol^{-1})	Density (g·cm^{-3})	Free energy (kcal·mol^{-1})	Density (g·cm^{-3})	Free energy (kcal·mol^{-1})	Density (g·cm^{-3})	kcal·mol^{-1}
methane	2.47±0.01	0.986±0.001	2.48±0.01	0.977±0.001	2.44±0.01	0.987±0.001	2.52±0.02	0.982±0.001	2.00±0.60
methanol	-3.73±0.01	0.988±0.001	-3.72±0.01	0.980±0.001	-3.51±0.01	0.988±0.001	-3.70±0.05	0.987±0.001	-5.10±0.60
ethane	2.50±0.01	0.988±0.001	2.50±0.01	0.979±0.001	2.48±0.01	0.988±0.001	2.56±0.01	0.984±0.001	1.83±0.60
toluene	-0.72±0.01	0.991±0.001	-0.64±0.01	0.983±0.001	-0.72±0.01	0.991±0.001	-0.55±0.02	0.989±0.001	-0.90±0.20
neopentane	2.61±0.01	0.990±0.001	2.58±0.02	0.981±0.001	2.58±0.01	0.990±0.001	2.71±0.06	0.987±0.001	2.51±0.60
2-methylfuran	-0.49±0.02	0.991±0.001	-0.42±0.01	0.983±0.001	-0.51±0.01	0.991±0.001	-0.39±0.02	0.989±0.001	NA
2-methylindole	-6.24±0.01	0.993±0.001	-6.06±0.01	0.984±0.001	-6.35±0.01	0.993±0.001	-6.06±0.04	0.990±0.001	NA
2-CPI	-6.05±0.02	0.995±0.001	-6.18±0.04	0.992±0.001	-6.54±0.01	0.994±0.001	-6.14±0.09	0.991±0.001	NA
7-CPI	-5.66±0.03	0.995±0.001	-6.28±0.03	0.982±0.001	-6.52±0.02	0.995±0.001	-6.10±0.01	0.992±0.001	NA

Table 3: MUE, in kcal·mol^{-1}, between relative free energies obtained with the absolute protocol for the SOMD, GROMACS, AMBER and CHARMM packages.

Package	GROMACS	AMBER	CHARMM
SOMD	0.20±0.03	0.13±0.04	0.08±0.02
GROMACS		0.19±0.01	0.15±0.01
AMBER			0.12±0.01

the ΔG_{HYD} from relative mutations were computed. Table **4** summarizes the results for the four MD packages. Again the data is from the recommended protocol for each package (see detailed discussions in the following subsections).

Firstly, the internal consistency of the different codes with the computed absolute hydration free energies was computed. For each software, the number of times a calculated relative free energy deviates from the difference in reference absolute hydration free energies by more than 0.1 kcal·mol^{-1} was counted. According to this criterion, the AMBER explicit implementation is the least consistent (10 deviations), followed by AMBER implicit (6 deviations), SOMD (6 deviations), CHARMM (5 deviations), GROMACS (5 deviations). The perturbations that give a discrepancy are not the same across codes, for instance methane->toluene with AMBER explicit deviates from the reference absolute hydration free energies by 0.33 kcal·mol^{-1}, but at most 0.04 kcal·mol^{-1} with other codes. SOMD and GROMACS show deviations of ca. 0.25 kcal·mol^{-1} for methanol->methane but this is not the case for AMBER (implicit or explicit) or CHARMM.

Next, the consistency between forward and backward transformations in relative hydration free energies was compared. Again counting the number of deviations that exceed 0.1 kcal·mol^{-1} indicates that AMBER explicit is the least consistent (3 deviations), followed by AMBER implicit (2 deviations), CHARMM (2 deviations), GROMACS (1 deviation), SOMD (1 deviation). The largest deviation is observed with AMBER implicit for 2-methylindole <-> methane (0.36 kcal·mol^{-1}).

Then, a comparison between the relative free energies predicted by all the packages was drawn. CHARMM tends to show relative free energies with smaller values for a number of transformations: neopentane, 2–methylfuran and 2–methylindole. SOMD displays smaller values ΔG_{HYD} for the methanol and toluene transformations. The largest discrepancy, however, is in the neopentane transformation with central mapping where AMBER with implicit dummy atoms is about 0.5 kcal·mol^{-1} higher and CHARMM about 0.2 kcal·mol^{-1} lower than the other two codes. The terminal mapped neopentane case reveals AMBER to be in line with GROMACS and SOMD while CHARMM's results deviate further. AMBER deviates also quite strongly from the other codes in the cyclopentanylindole cases.

The MUEs of the relative free energy simulations are presented in Table **5**. They are on average slightly larger than the MUEs from the absolute simulations (Table **3**) and reach 0.26 kcal·mol^{-1} for AMBER compared with CHARMM.

To further understand the correctness of the transformations, table **6** shows the computed cycle closure errors for the cycle ethane→ methanol

Table 4: Comparison of relative free energies of hydration for various MD packages as obtained from absolute (AFE) and relative (RAFE) transformations via unified or split protocols. The values deduced from AFE transformations (given in the first row) were taken from Table 1. Similarly the experimental values can be computed as a difference from absolute hydration free energies.

Transformation[1]			AMBER[2]		CHARMM[3]	GROMACS[2]	SOMD[3]	ΔG^{exp}_{HYD}
			implicit[4]	explicit[4]				
AFE	ethane	methane		-0.02±0.01	-0.03±0.01	-0.04±0.01	-0.05±0.02	
RAFE	ethane	methane	+0.02±0.01	-0.13±0.02	+0.09±0.02	-0.04±0.02	0.04±0.02	0.16
RAFE	methane	ethane	+0.00±0.03	+0.19±0.03	+0.04±0.01	+0.02±0.01	-0.01±0.06	-0.16
AFE	methanol	methane		+6.20±0.01	+6.20±0.02	+5.95±0.01	+6.21±0.06	
RAFE	methanol	methane	+6.19±0.01	+6.20±0.02	+6.18±0.01	+6.20±0.01	+5.99±0.05	7.09
RAFE	methane	methanol	-6.20±0.03	-6.15±0.01	-6.21±0.01	-6.20±0.01	-5.97±0.04	-7.09
AFE	ethane	methanol		-6.22±0.01	-6.22±0.02	-5.98±0.01	-6.26±0.05	
RAFE	ethane	methanol	-6.20±0.01	-6.27±0.01	-6.25±0.01	-6.19±0.01	-6.09±0.03	-6.93
RAFE	methanol	ethane	+6.20±0.01	+6.25±0.01	+6.28±0.01	+6.19±0.01	+6.09±0.02	6.93
AFE	toluene	methane		+3.19±0.01	+3.12±0.01	+3.16±0.01	+3.07±0.03	
RAFE	toluene	methane	+3.24±0.02	+3.39±0.02	+3.04±0.02	+3.21±0.01	+2.89±0.09	2.88
RAFE	methane	toluene	-3.42±0.03	-3.52±0.03	-3.09±0.02	-3.20±0.01	-3.06±0.02	-2.88
AFE	neopentane	methane		-0.13±0.02	-0.11±0.02	-0.14±0.01	-0.19±0.06	
RAFE	neopentane[5]	methane	0.32±0.04	-0.03±0.06	-0.35±0.01	-0.15±0.02	-0.20±0.05	-0.52
RAFE	methane[5]	neopentane	-0.25±0.03	-0.07±0.03	+0.24±0.02	+0.16±0.05	+0.13±0.05	0.52
RAFE	neopentane[6]	methane	-0.13±0.01	-0.12±0.02	-0.56±0.02	-0.14±0.01	-0.11±0.01	-0.52
RAFE	methane[6]	neopentane	+0.13±0.03	+0.12±0.03	+0.40±0.02	+0.18±0.03	+0.10±0.06	0.52
AFE	2–methylfuran	methane		+2.96±0.02	+2.90±0.01	+2.95±0.01	+2.90±0.03	
RAFE	2–methylfuran	methane	+3.09±0.01	+3.10±0.01	+2.84±0.03	+2.93±0.05	+2.92±0.05	NA
RAFE	methane	2-methyfuran	-3.10±0.03	-3.15±0.03	-2.84±0.02	-2.96±0.01	-2.83±0.03	NA
AFE	2–methylindole	methane		+8.72±0.01	+8.53±0.02	+8.79±0.02	+8.57±0.03	
RAFE	2–methylindole	methane	+8.78±0.03	+8.78±0.04	+8.49±0.01	+8.73±0.03	+8.64±0.06	NA
RAFE	methane	2-methylindole	-9.14±0.02	-9.13±0.03	-8.56±0.02	-8.74±0.01	-8.67±0.08	NA
AFE	2–CPI	7–CPI		+0.39±0.04	-0.11±0.04	+0.02±0.05	+0.08±0.14	
RAFE	2–CPI[7]	7–CPI	+0.36±0.03	+0.63±0.06	-0.01±0.01	-0.01±0.03	-0.11±0.07	NA
RAFE	7–CPI[7]	2–CPI	-0.34±0.05	-0.50±0.03	-0.04±0.01	+0.20±0.04	+0.01±0.08	NA

[1] The values deduced from the AFE absolute of Table 1 are given first.
[2] split protocol.
[3] unified protocol.
[4] using either the implicit or the explicit dummy atom approach.
[5] central mapping.
[6] terminal mapping.
[7] partial re/discharge i.e.
only the charges of the appearingand the disappearing rings are switched

Table 5: MUE, in kcal·mol^{-1}, comparing relative free energies from relative simulations between SOMD, GROMACS, AMBER and CHARMM.

Package	GROMACS	AMBER	CHARMM
SOMD	0.11±0.01	0.23±0.01	0.15±0.01
GROMACS		0.16±0.01	0.13±0.01
AMBER			0.26±0.01

Table 6: Cycle closure errors, in kcal·mol^{-1}, for ethane→ methanol → methane → ethane

Package and Protocol	Closure Error
AMBER implicit	0.07 ± 0.04
AMBER explicit	0.02 ± 0.05
GROMACS split reaction field	0.05 ± 0.02
GROMACS unified reaction field	0.13 ± 0.03
GROMACS split PME	0.04 ± 0.01
GROMACS unified PME	0.18 ± 0.03
CHARMM	0.01 ± 0.03
SOMD	-0.11 ± 0.08

→ methane → ethane. Uncertainties were estimated by propagating uncertainties from the individual perturbations. AMBER explicit dummy and CHARMM are the only protocol consistent within uncertainty estimates, but the deviations observed with the other protocols are small. The largest discrepancy is observed with the GROMACS unified PME protocol, with the error just under 0.2 kcal·mol^{-1}.

Finally, table **7** reports whether the codes reproduced consistent changes in mean box volumes between forward and backward transformations. Generally codes present a consistency among volumes values. GROMACS gives the most precise volume changes, whereas SOMD gives the least precie volume changes. This could indicate the barostats used by the different simulation packages relax volume fluctuations with different efficiency, or that they sample different volume fluctuations.

Table 7: Changes in volumes, in Å^3, for selected perturbations across packages

transformation		AMBER	CHARMM	GROMACS	SOMD
methane	ethane	38±6	24±3	31±3	61±25
ethane	methane	-29±7	-30±4	-28±3	-50±30
ethane	methanol	-33±9	-36±2	-36±5	-66±36
methanol	ethane	38±8	37±3	36±5	46±31
methane	methanol	-5±11	-9±2	-3±5	11±20
methanol	methane	8±10	1±2	4±3	27±16
methane	toluene	93±11	89±3	110±2	120±56
toluene	methane	-106±8	-89±8	-113±3	-145±23
methane	2-methylindole	164±11	64±3	142±3	140±22
2-methylindole	methane	-138±8	-120±7	-139±5	-166±45
methane	neopentane	99±9	90±2	115±4	117±50
neopentane	methane	-105±11	-100±2	-114±4	-68±43

2.3.1 AMBER

Using AMBER for RAFE simulations has revealed several problems with the implementation. Some bugs were identified and the developers have fixed those for AMBER16, e.g. energy minimization in `sander` led to diverged coordinates for mapped atoms. For a single topology description, however, it is necessary to have the same coordinates. Other issues are that vacuum simulations can only be carried out with the `sander` program because `pmemd` cannot handle AFE simulations in vacuum as of this writing. A disadvantage of `sander` is that it cannot be used to simulate the λ end points, [135] such that the TI gradients need to be extrapolated (minimum and maximum allowed λs are 0.005 and 0.995). Also, `sander` considers the whole system as the perturbed region while `pmemd` restricts this to a user chosen atom selection. This has obvious implications for performance [135].

In contrast to the other three codes, AMBER does not yield correct relative free energies with the unified protocol, i.e when all force field parameters are scaled simultaneously as shown in Table **8**.

The issue becomes apparent when more than a few dummy atoms are involved, while the unified protocol works for the smaller transformations.The split RAFE protocol and absolute free energies, however, are very close to the other MD packages as demonstrated in Table **9**.

End point geometries appear to be another issue with AMBER simulations in both solution and vacuum. This is most obvious in the neopentane

Table 8: Comparison between split and unified protocol in AMBER. The values for the split protocol highlights inconsistencies in the code in bold font. All the values are in kcal·mol^{-1}.

transformation		split protocol ΔG_{HYD}	unified protocol ΔG_{HYD}	ΔG^{exp}_{HYD}
ethane	methane	0.02	-0.07	0.16
methane	ethane	0.01	-0.01	-0.16
methanol	methane	6.19	6.23	7.09
methane	methanol	-6.19	-6.31	-7.09
ethane	methanol	-6.20	-6.24	-6.93
methanol	ethane	6.19	6.19	6.93
toluene	methane	3.24	**5.64**	2.88
methane	toluene	-3.42	**−5.61**	-2.88
neopentane[1]	methane	0.31	**7.26**	-0.52
methane[1]	neopentane	-0.25	**−7.04**	0.52
neopentane[2]	methane	-0.13	**0.94**	-0.52
methane[2]	neopentane	0.12	**−0.96**	0.52
2-methylfuran	methane	3.08	**3.14**	NA
methane	2-methylfuran	-3.10	**−2.15**	NA
2-methylindole	methane	8.77	**11.13**	NA
methane	2-methylindole	-9.13	-8.98	NA

[1] central mapping.
[2] terminal mapping.

Table 9: Comparing AMBER results for simulations with various split protocols.The emphasis is here on the data with SHAKE enabled and a time step of 2 fs (last column). Implicit, explicit and absolute protocols had SHAKE disabled and a time step of 1 fs. All the values are in kcal·mol^{-1}

transformation		implicit ΔG_{HYD}	explicit ΔG_{HYD}	absolute ΔG_{HYD}	SHAKE[1] ΔG_{HYD}
ethane	methanol	-6.20±0.01	-6.27±0.01	-6.22±0.01	-6.18±0.01
methanol	ethane	-6.20±0.01	-6.25±0.01		
toluene	methane	3.24±0.02	3.39±0.02	3.19±0.01	3.27±0.03
methane	toluene	3.42±0.03	3.52±0.03		
neopentane	methane[2]	0.32±0.04	-0.03±0.06		0.35±0.02
methane	neopentane[2]	0.25±0.03	-0.07±0.03	-0.13±0.02	
neopentane	methane[3]	0.13±0.01	-0.12±0.02		
methane	neopentane[3]	-0.13±0.03	-0.12±0.03		

[1] implicit dummy atom protocol with δt =2 fs and SHAKE on all H–bonds except perturbed bonds.
[2] central mapping.
[3] terminal mapping.

→ methane test case with central mapping. As a matter of fact, the methane end state exhibits incorrect distances between the carbon and the four attached hydrogens of approximately 1.23 Å. This value is about 1.12 Å for the terminal dummy atoms in the other test cases but still higher than the expected 1.09 Å on average. Such an error depends on the number of dummy atoms immediately surrounding the central atom.

Furthermore, a constraint check was carried out. SHAKE was explicitly deactivated for all bonds in the perturbed region in these protocols. Table **9** shows selected results for transformations with SHAKE enabled for all bonds to hydrogens except those bonds that change bond length during transformation. The time step has been increased from 1 fs as used in the other three protocols to 2 fs. As the results are essentially the same as the non–SHAKE simulations, this SHAKE protocol appears to be a viable solution to increase the performance of RAFE simulations. We have repeated this protocol with AMBER in response to the results obtained with SOMD using this implementation. From a practical point of view, AMBER uses an *atom*-based mask for bond SHAKEs such that the mask must be set for the hydrogens in question while the same is not possible for their non–H counter–part in the other state because *all* bonds emanating from this atom would be affected. In general,

the free energies computed with each approach are in good agreement with each other and with the results of the other MD packages There are, however, a few notable deviations. Neopentane → methane with central mapping differs from the result with terminal mapping by about 0.4 kcal·mol^{-1}. The terminal mapping and the free energies from the explicit dummy simulations are, however, consistent with the absolute transformations (Table **2**).

2.3.2 CHARMM

A few bugs not previously reported in CHARMM c40b1 were found and careful alchemical free energy (AFE) setup was needed to produce robust and accurate results. The PERT module does not allow a hydrogen bond constraint (SHAKE) to be applied on the perturbed region, and this requires end point λ to be equilibrated carefully. These windows at end-point λ were started with their own equilibration using timesteps of 0.1 fs to 0.5 fs before the production run. The VSwitch option was used to apply a switching function to the potential since that option cannot be applied to forces for calculations run with the PERT module.

The PSSP softcore potential function cannot handle Long-Range Correction (LRC) correctly. This effect is not clearly shown when the initial and final states are comparable in size, but the deviation becomes larger for perturbations that involve large changes in solute size, or for absolute alchemical free energy calculations. It is necessary to disable the LRC to obtain consistent free energies from relative and absolute alchemical free energy calculation protocols.

Table **10** shows the relative free energies obtained from CHARMM simulations. While results from all three protocols (split, unified, absolute) seem to be in good agreement with each other, the split-protocol results are more precise due to the additional amount of data generated. It is notable that the split-protocol results are more similar to the ones obtained by other MD packages (i.e. neopentane and toluene), but the relative-unified results are more consistent with the CHARMM absolute simulations (e.g. 2-methylindole). Overall, the relative free energies obtained by these three different protocols are in good agreement with those reported for the other MD packages (Tables 1 and 3).

Table 10: Results for CHARMM simulations with split and unified protocols for relative alchemical free energy and unified protocol for absolute values. All the values are in kcal·mol^{-1}.

transformation		split ΔG_{HYD}	unified ΔG_{HYD}	absolute(unified) ΔG_{HYD}
ethane	methane	-0.09±0.01	-0.09±0.02	-0.03±0.01
methane	ethane	0.04±0.01	0.04±0.01	
methanol	methane	6.20±0.01	6.18±0.01	6.20±0.01
methane	methanol	-6.30±0.01	-6.21±0.01	
ethane	methanol	-6.21±0.01	-6.25±0.01	-6.22±0.02
methanol	ethane	6.25±0.01	6.28±0.01	
toluene	methane	3.22±0.01	3.04±0.02	3.12±0.01
methane	toluene	-3.28±0.01	-3.09±0.02	
neopentane	methane[1]	-0.29±0.01	-0.35±0.01	
methane	neopentane[1]	0.15±0.01	0.24±0.02	-0.11±0.02
neopentane	methane[2]	-0.42±0.01	-0.56±0.02	
methane	neopentane[2]	0.31±0.01	0.40±0.02	
2-methylfuran	methane	2.87±0.01	2.84±0.03	2.90±0.01
methane	2-methylfuran	-2.93±0.01	-2.84±0.02	
2-methylindole	methane	8.88±0.01	8.49±0.01	8.53±0.02
methane	2-methylindole	-8.81±0.01	-8.56±0.02	
2-CPI	7-CPI	-0.08±0.01	-0.01±0.01	-0.11±0.04
7-CPI	2-CPI	-0.01±0.01	-0.04±0.01	

[1] central mapping.
[2] terminal mapping.

Table 11: Relative hydration free energies obtained from GROMACS simulations, in kcal·mol^{-1}.

transformation		split[1] RF ΔG_{HYD}	split[1] PME ΔG_{HYD}	unified[2] RF ΔG_{HYD}	unified[2] PME ΔG_{HYD}	absolute[3] RF ΔG_{HYD}	absolute[3] PME ΔG_{HYD}
ethane	methane	-0.02±0.01	-0.03±0.02	-0.02±0.01	-0.03±0.01	-0.06±0.01	-0.04±0.01
methane	ethane	0.01±0.02	0.02±0.01	0.05±0.02	0.01±0.02		
methanol	methane	6.16±0.01	6.20±0.01	7.30±0.02	7.38±0.01	5.77±0.01	5.95±0.01
methane	methanol	-6.17±0.01	-6.20±0.01	-7.09±0.02	-7.17±0.02		
ethane	methanol	-6.12±0.01	-6.18±0.01	-7.11±0.01	-7.21±0.02	-5.83±0.01	-5.98±0.01
methanol	ethane	6.12±0.01	6.19±0.01	7.34±0.01	7.40±0.01		
toluene	methane	3.22±0.01	3.21±0.01	3.23±0.01	3.22±0.01	2.97±0.01	3.16±0.01
methane	toluene	-3.25±0.01	-3.20±0.01	-3.22±0.01	-3.21±0.01		
neopentane	methane[4]	-0.10±0.01	-0.15±0.02	-0.08±0.02	-0.18±0.03	-0.18±0.01	-0.14±0.01
methane	neopentane[4]	0.11±0.02	0.16±0.05	0.00±0.03	0.18±0.03		
neopentane	methane[5]	-0.11±0.01	-0.13±0.01	-0.14±0.01	-0.14±0.01		
methane	neopentane[5]	0.10±0.03	0.18±0.03	0.09±0.01	0.15±0.02		
2-methylfuran	methane	2.99±0.01	2.93±0.05	3.05±0.01	3.00±0.01	2.87±0.01	2.95±0.01
methane	2-methylfuran	-3.00±0.01	-2.96±0.01	-3.05±0.01	-3.01±0.01		
2-methylindole	methane	8.71±0.02	8.73±0.03	8.73±0.01	8.80±0.03	8.44±0.02	8.79±0.02
methane	2-methylindole	-8.73±0.03	-8.74±0.01	-8.30±0.02	-8.77±0.04		
2-CPI	7-CPI	-0.07±0.02	-0.03±0.03	-0.10±0.05	-0.20±0.10	-0.02±0.05	0.02±0.02
7-CPI	2-CPI	0.12±0.06	0.20±0.04	0.04±0.06	0.14±0.09		

[1] results obtained from alchemical transformations with electrostatic and bonded scaling separate from vdW parameter change.
[2] results obtained from alchemical transformation with all parameters scaling together.
[3] results obtained from absolute free energy calculations.
[4] central mapping.
[5] terminal mapping.

2.3.3 GROMACS

Table **11** lists the relative free energies obtained from GROMACS simulations. Relative free energies are in good agreement with each other and with ΔG_{HYD} obtained from the other software used in this study (Tables **2** and **4**). A noteworthy exception is the difference between the unified and split results of methane → methanol and its reverse process. This was investigated further with additional split protocol simulations using Coulomb softcore potentials (Table **12**).

There is a difference of approximately 1.5 kcal·mol^{-1} between the split protocol without Coulomb softcore potentials and both protocols that use it. Deeper analyses suggest that softening of the electrostatic interactions

Table 12: ΔG_{HYD} results in different scenarios with or without Coulomb softcore potentials, in kcal·mol^{-1}.

		without Coulomb softcore		with Coulomb softcore		absolute	
		RF	PME	RF	PME	RF	PME
Transformations		ΔG_{HYD}	ΔG_{HYD}	ΔG_{HYD}	ΔG_{HYD}	ΔG_{HYD}	ΔG_{HYD}
ethane	methane	-0.02±0.01	-0.04±0.02	-0.03±0.04	-0.02±0.04	-0.06±0.01	-0.04±0.01
methane	ethane	0.01±0.02	0.02±0.01	0.01±0.04	0.02±0.04		
methanol	methane	6.16±0.01	6.20±0.01	7.32±0.03	7.42±0.04	5.77±0.01	5.95±0.01
methane	methanol	-6.17±0.01	-6.20±0.01	-7.14±0.03	-7.21±0.03		
ethane	methanol	-6.12±0.01	-6.18±0.01	-6.15±0.02	-6.21±0.02	-5.83±0.01	-5.98±0.01
methanol	ethane	6.12±0.01	6.19±0.01	6.15±0.02	6.21±0.02		
toluene	methane	3.22±0.01	3.21±0.01	3.22±0.04	3.21±0.04	2.97±0.01	3.16±0.01
methane	toluene	-3.25±0.01	-3.20±0.01	-3.27±0.04	-3.22±0.04		
neopentane	methane[1]	-0.10±0.01	-0.15±0.02	-0.13±0.08	-0.13±0.08	-0.18±0.01	-0.14±0.01
methane	neopentane[1]	0.11±0.02	0.16±0.05	0.12±0.08	0.15±0.08		
neopentane	methane[2]	-0.12±0.01	-0.13±0.01	-0.10±0.04	-0.13±0.04		
methane	neopentane2[2]	0.10±0.03	0.18±0.03	0.08±0.06	-0.15±0.06		
2-methylfuran	methane	2.99±0.01	2.93±0.05	3.07±0.03	3.02±0.04	2.87±0.01	2.95±0.01
methane	2-methylfuran	-3.00±0.01	-2.96±0.01	-3.08±0.03	-3.02±0.04		
2-methylindole	methane	8.71±0.02	8.73±0.03	8.79±0.04	8.82±0.05	8.44±0.02	8.79±0.02
methane	2-methylindole	-8.73±0.03	-8.74±0.01	-8.79±0.05	-8.81±0.06		
2-cyclopentanylindole	7-cyclopentanylindole	-0.07±0.02	-0.03±0.03	-0.12±0.03	-0.14±0.05	-0.02±0.05	0.02±0.02
7-cyclopentanylindole	2-cyclopentanylindole	0.12±0.06	0.20±0.04	-1.20±0.20	-1.50±0.10		

[1] central mapping.
[2] terminal mapping.

requires adjustments in the λ-distance between states in the rapidly varying part of the $\partial G/\partial\lambda$. A variant that combined the bonded terms with the vdW transformation did not change this result. Thus, the split protocol without Coulomb softcore potentials was found to be the most effective way to calculate relative free energies with the current GROMACS implementation.

Additionally it is worth mentioning is that relative free energy simulations that feature alchemical transformations of a hydrogen atom into a heavy atom will crash if the bond involving the hydrogen atom is constrained with algorithms such as SHAKE or LINCS. Successful simulations require turning off the bond constraint and decreasing the time step to 1 fs. Alternative protocols that require some scripting and changes in the topology file could be pursued in the future. For instance 2 fs constraints protocols similar to those used in SOMD or AMBER in this study could be implemented via the definition of a new atom type for alchemically perturbed hydrogen atoms.

Table 13: Final relative free energy of hydration ΔG_{HYD} estimations and standard error for SOMD unperturbed hydrogen bonds protocol, *RAFE*, compared with relative free energy of hydration computed from absolute free energy simulations, *RAFE-absolute*. All the values are in kcal·mol^{-1}

transformation		RAFE ΔG_{HYD}	RAFE-absolute ΔG_{HYD}
ethane	methane	0.01±0.05	0.04±0.02
methane	ethane	-0.04±0.02	0.04±0.02
methanol	methane	5.99±0.05	6.21±0.05
methane	methanol	-5.97±0.04	-6.21±0.05
ethane	methanol	-6.09±0.03	-6.26±0.05
methanol	ethane	6.09±0.02	6.26±0.05
toluene	methane	2.89±0.09	3.06±0.03
methane	toluene	-3.06±0.02	-3.06±0.03
neopentane	methane[1]	-0.20±0.05	-0.19±0.06
methane	neopentane[1]	0.13±0.05	0.19±0.06
neopentane	methane[2]	-0.11±0.01	-0.19±0.06
methane	neopentane[2]	0.10±0.06	0.19±0.06
2-methylfuran	methane	2.92±0.05	2.90±0.03
methane	2-methyfuran	-2.83±0.03	-2.90±0.03
2-methylindole	methane	8.64±0.06	8.57±0.03
methane	2-methylindole	-8.67±0.08	-8.57±0.03
2–cyclopentanylindole	7–cyclopentanylindole	0.11±0.07	0.08±0.14
7–cyclopentanylindole	2–cyclopentanylindole	0.01±0.08	0.08±0.14

[1] central mapping.
[2] terminal mapping.

2.3.4 SOMD

Table **13** compares relative free energy of hydration $\Delta\ G_{HYD}$ according to the protocol with unperturbed H bond constraints, with relative ΔG_{HYD} obtained from two absolute free energy calculations.

Overall for the unperturbed H bonds constraint a very good agreement is observed between both absolute and relative calculations (R^2=0.99±0.01 and MUE = 0.10±0.03 kcal·mol^{-1}), highlighting internal consistency within SOMD.

To achieve this level of reproducibility within SOMD it was crucial to pay close attention to constraints. Specifically, bonds that involve unperturbed hydrogen atoms are constrained. Bonds involving hydrogen atoms that are

Table 14: Relative free energy of hydration ΔG_{HYD} computed with all bond constraints, *All bonds*, no constraints, *None*, and unperturbed hydrogen bond constraint, *unpert H bonds*. All the values are in kcal·mol^{-1}

transformation		*All bonds* ΔG_{HYD}	*None* ΔG_{HYD}	*unpert H bonds* ΔG_{HYD}
ethane	methane	-0.48±0.01	-0.18±0.04	0.01±0.05
methane	ethane	0.49±0.01	0.01±0.02	-0.04±0.02
methanol	methane	6.06±0.01	6.49±0.01	5.99±0.05
methane	methanol	-6.08±0.01	-6.15±0.01	-5.97±0.04
ethane	methanol	-6.22±0.01	-6.14±0.03	-6.09±0.03
methanol	ethane	6.23±0.01	6.09±0.01	6.09±0.02
toluene	methane	3.73±0.27	3.09±0.06	2.89±0.09
methane	toluene	-3.79±0.03	-3.07±0.06	-3.06±0.02
neopentane	methane[1]	-2.09±0.01	-0.14±0.14	-0.20±0.05
methane	neopentane[1]	2.04±0.01	0.01±0.06	0.13±0.05
neopentane	methane[2]	-0.48±0.01	-0.14±0.06	-0.11±0.01
methane	neopentane[2]	0.59±0.02	0.14±0.06	0.10±0.06
2-methylfuran	methane	3.38±0.02	2.81±0.03	2.92±0.05
methane	2-methyfuran	-3.40±0.03	-2.89±0.06	-2.83±0.03
2-methylindole	methane	9.29±0.06	8.72±0.05	8.63±0.06
methane	2-methylindole	-9.10±0.04	-8.61±0.04	-8.67±0.08

[1] central mapping.
[2] terminal mapping.

perturbed to a heavy element are unconstrained. Additionally the atomic mass of the perturbed hydrogen atom is set to the mass of the heavy atom it is perturbed to. Bonds involving hydrogen atoms that are perturbed to another hydrogen atom type are constrained. We stress that it is acceptable to artificially increase the atomic mass of hydrogen atoms because the calculated excess free energy changes do not depend on atomic masses.

This protocol suppresses high frequency vibrations in flexible bonds involving hydrogen atoms, thus enabling a time step of 2 fs, whilst giving essentially negligible errors due to the use of constraints for perturbed bonds. This is apparent from the comparison with the absolute hydration free energy calculations. Additionally, the protocol yields relative hydration free energy very similar (MAE = 0.09 kcal·mol^{-1}) to those computed from simulations where no constraints are applied on the solutes and a timestep of 1 fs is used, table **14**

By contrast, a protocol that constrains all bonds in a solute leads to significant differences with the absolute hydration free energies. For instance neopentane → methane (centrally mapped) gives a RAFE ΔG_{HYD}=2.04±0.01 kcal·mol^{-1} whereas the absolute hydration free energy calculations give ΔG_{HYD}=-0.19±0.06 kcal·mol^{-1}.

This discrepancy occurs because in the SOMD implementation, the energies of constrained bonds are not evaluated, but the calculation of the energies of the solute at perturbed λ values is carried out using the coordinates of the reference λ trajectory. This leads to a neglect of contributions of the bonded term (and associated coupled terms) to the free energy change. The effect is more pronounced for perturbations that feature a large change in equilibrium bond lengths, such as those where a hydrogen atom is perturbed to/from a heavy atom.

The reaction fields implemented in SOMD and GROMACS differ somewhat (atom-based shifted Barker Watts, [49] vs group based switched Barker Watts), but nevertheless SOMD and GROMACS RF produce comparable results with a MUE of 0.18 kcal·mol^{-1}. Overall, the SOMD free energy estimations are in good agreement with the other MD packages, as the MUE suggests (see Table **5**). For the methane → neopentane transformations SOMD yields consistent results between central and terminal mappings, as shown in Table **13**. Reaction field and GROMACS PME results are in good agreement. All SOMD RAFE simulations were carried out with simultaneous transformation of Lennard-Jones, charges, and bonded terms. This suggests that the failure of the GROMACS "unified protocol" in some instances may be due to differences in the softcore Coulomb implementations.

2.4 Discussion and Conclusions

This study addressed whether contemporary MD packages such as AMBER, CHARMM, GROMACS and SOMD are able to reproduce relative alchemical free energies of hydration for a set of neutral small organic molecules, given a pre–defined force field. Establishing a simulation protocol that leads to consistent results across codes has been cumbersome due to technical difficulties encountered with every code. The MD codes have a wide range of options and setup features which makes it difficult for the inexperienced user to decide on the most appropriate ones. All the dataset input files, setup, simulations and analyses shown in this work are freely available at https://github.com/halx/relative-solvation-inputs .

The computed free energies appear to be in reasonably good agreement with each other (see Tables **2** and **4**). The average MAE between all codes 0.14 kcal·mol^{-1} for absolute free energies and 0.17 kcal·mol^{-1} for relative free energies. This can be interpreted as the current "limit of reproducibility" for the alchemical relative hydration free energy for a small dataset of neutral molecules. There is some doubt, however, over the AMBER results because the particular version of the tested software cannot reproduce the correct endpoint geometries. This is particularly evident in the neopentane to methane case with central mapping where also the relative free energies are clearly different from the other packages. These issues may reflect a bug in the AMBER package but it was impossible to isolate it.

It was not possible to define a *universal* protocol that could be recommended for use with all four codes. Unified protocols do not appear to work with AMBER and GROMACS, while SOMD and CHARMM had no problem in this regard. It was not clear where the problem may lie, e.g. only with the vacuum leg of the thermodynamic cycle. In the case of AMBER the vacuum simulation has currently been done with the separately developed `sander` module. The problem may be a consequence of the different softcore functions used in these MD packages but further investigations are needed to resolve this issue.

The unperturbed H bond protocol is an interesting alternative which applies constraints to all non–transforming bonds and thus allowed us to increase the time step to 2 fs. The split protocol was found to work well for all codes. It appears to be the most efficient approach for GROMACS as shown with the methanol to methane case because the unified protocol produces a less smooth function [136]. A complete separation of λs may not be necessary though as a certain degree of overlap between vdW and Coulumb λ may be a viable solution [137] for equilibrium AFEs.

Comparison between codes is hampered by several factors. Firstly, the codes use different simulation algorithms e.g. electrostatics are handled differently in vacuum i.e. infinite cutoff vs. reaction field. Temperature and pressure control, time step integrators, etc. are other examples. But the data here suggest that, if there are any systematic errors introduced through these algorithms, then they are small. It is reassuring that AFEs for the systems tested here show only a small dependence on MD protocol decisions (provided a correct implementation).

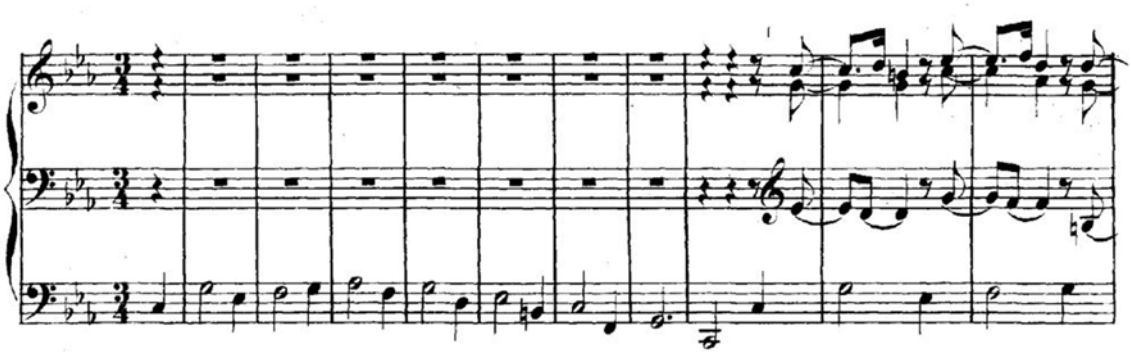
More specifically, various issues with current code bases have been revealed through this work. Constraints in connection with varying bond length can cause errors with GROMACS, just as masses must not be allowed to vary in RAFE simulations, both to avoid crashes and incorrect results from the software. CHARMM has issues with constraints and the PSSP softcores, and the PERT module cannot make use of the force switch as it is now standard for CHARMM force fields. Care must be taken when using the LRC long range correction keyword to avoid producing inconsistent results. AMBER's problem with end point geometries and unified protocols has been pointed out above.

Another practical issue is the complex setup associated with the split protocol. For instance in GROMACS it is necessary to carry out two separate simulations per λ because discharging and recharging groups cannot be selected separately. λ paths as implemented in GROMACS could also be beneficial for other codes as they make the setup of split protocols easier. The alternative we have used in codes lacking this feature is to mimic this protocol through careful constructions of topologies via scripting.

The primary focus of this work was to achieve low statistical errors to establish if codes are able to reproduce free energies. The protocols efficiency for each software was not investigated in details, as this would require further, complicated studies. For absolute calculations the most demanding protocol and most precise protocol is GROMACS (200 million aggregate time–steps per solute, average SEM 0.01 kcal·mol^{-1}), the least demanding protocol is CHARMM (31.5 million time–steps per solute, average SEM 0.015 kcal·mol^{-1}). SOMD's aggregate time–steps is comparable to CHARMM (34 million time–steps) but the free energies are less precise (average SEM 0.045 kcal·mol^{-1}). For relative calculations, the least demanding protocol is SOMD (17 million time–steps), but this is also the least precise (average SEM 0.05 kcal·mol^{-1}). Remarkably the most demanding protocol (GROMACS 197.4 million time–steps, average SEM 0.02 kcal·mol^{-1}) is less precise than CHARMM that used fewer time–steps (31.5 million time-steps, average SEM 0.01 kcal·mol^{-1}). Further work should be pursued to understand what algorithmic details in the various implementations are important for

the efficiency of the free energy calculations.

Beyond careful protocol validation, further automation of alchemical free energy studies will also decrease user errors, and thus increases reproducibility. Various attempts in this direction are currently underway for both absolute and relative setups [138, 139, 140, 108, 141, 142, 109]. To conclude, it is hoped that this study will contribute to improve best practices in the computational field, to improve the transferability of alchemical free energy calculation protocols across software. Reproducibility is crucial to enable robust use of alchemical free energy methods in molecular design.



Chapter 3

Alchemical free energy protocols for lipophilicity coefficients estimation

The work presented in this chapter was inspired by two publications [143, 144], where alchemical free energy calculations were used to predict the lipophilicity coefficients of drug-like molecules. In [143], AFE were employed to predict changes in logP for the increased fluorniation of cyclohexyl ring. The decrease in logP was correctly predicted by AFE and, unlike empirical logP estimators, atomistic details from molecular simulations could explain the causes of such a trend. In [144] alchemical free energy calculations were employed to estimate distribution coefficients, logD, which are a natural extension of the logP. In particular, logD requires considerations about the protomeric and tautomeric states of a molecule in different aqueous and organic phases. This chapter indicates how alchemical free energy methods may be used to predict the lipophilicity of drug-like molecules.

3.1 Introduction

Absorption and distribution are two pivotal indicators in medicinal chemistry, assessing the effectiveness of a drug candidate molecule. The former

gives an indication on the drug's bioavailability and how the compound can travel through the bloodstream, before being taken up by the target cells. The distribution points out how the drug can pass through different cells compartments and how the molecule is distributed through the body. Only molecules with favourable absorption and distribution profiles may be progressed towards clinical studies of drug efficacy. Thus, knowing in advance these features is pivotal in the decision-making of pre-clinical drug discovery.

A cheap experimental test, to predict absorption and distribution properties for drug candidates, is the logP measurement [145, 146]. The logP is the logarithm based 10 of the ratio of un-ionized species between an organic phase, e.g. octanol, and an aqueous, calculated as [147, 148, 149, 150]:

$$\log P = \log \frac{[A]_o}{[A]_w}, \tag{3.1}$$

where $[A]_o$ is the concentration of the solute in the organic phase, and $[A]_w$ the concentration in the water phase. Classically, logP is determined by using the shake-flask method [147]. A drug candidate is dissolved in a flask containing organic solvent and water by shaking. Eventually, the concentration of the solute in each solvent is measured with UV/Vis spectroscopy [151]. Another methodology makes use of a high-performance liquid chromatography [152]. The logP is determined by measuring the retention time with similar compounds with known logP valus [153].

The logP neglects the presence of possible mixtures of protomeric and tautomeric states for a molecule A. Therefore, the distribution coefficient, or logD, can be defined:

$$\log D = \log \left(\frac{\sum_i^{N_q} \sum_j^{N_{taut}} [A_j]_i^o}{\sum_k^{N_q} \sum_l^{N_{taut}} [A_l]_k^w} \right) \tag{3.2}$$

where the sums are extended over all the possible protonation (N_q) and tautomeric state (N_{taut}) i and j in the organic phase, and k and l in water phase, for a molecule A.

Given the fundamental role of logP and logD measurements, it is without surprise that the medicinal chemistry community is trying to achieve such predictions by means of computational techniques. Remarkable results have been obtained by QSAR, such as ChemAxon [154], MolInspiration [155], ALogPs [156], that are readily available to the community via web-server. However, such methodologies do not give insights about the possible interactions that drug-like molecules establish with the surrounding environment. For this reason, "trajectory-based" methods, such as alchemical free energy (AFE) molecular dynamics, can be valuable to forecast logP and logD values.

Unfortunately, logP predictions by molecular simulations can be fairly difficult. Indeed, simulations may be strongly influenced by the correct force field model of the organic solvent [157, 158, 159]. logD predictions are more complex, since protonated and tautomerization states of a molecules should be consider. Additionally, a drug like molecule may present multiple distinct protomeric and tautomeric states at the same pH, creating ambiguity for the selection of a correct model of each species. Finally, the presence of charged species could give rise to the well-known finite size artifacts [80, 82, 81], which should be corrected to retrieve an accurate logD estimation.

Thus, the aim of this chapter is to devise a suitable protocol for logP and logD estimations, along with an assessment of current limitations in the context of AFE calculations. logP were computed for a set of five molecules, shown in fig. **1**,which have shown interesting peculiarities in logP trends [143, 160].

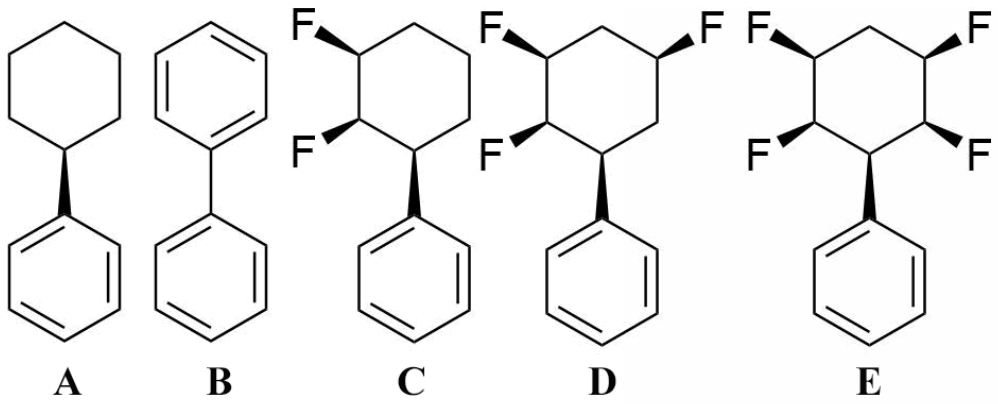


Figure 1: Series of bicyclic rings studied for logP calculations. A: cyclohexylbenzene, B: biphenyl, C:2,3-difluoro-1,1'-biphenyl, D: 2,3,5-difluoro-1,1'-biphenyl and E: 2,3,5,6-difluoro-1,1'-biphenyl.

logD estimations were considered as part of the 5th Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL). The experimental measurements were carried out at Genentech, according to Lin and Pease [161, 162] for 53 drug-like molecules, shown in fig. **2**, **3**, **4**. The choice of using trajectory based alchemical methods was partially motivated by the previously reported success with simple molecules such as caffeine (**080**) that were treated with general molecular mechanics force fields [163]. To push the limit of molecular simulation protocols the dataset also included large and chemically complex molecules such as rifampicin (**083**) or reserpine (**065**), fig. **4**.

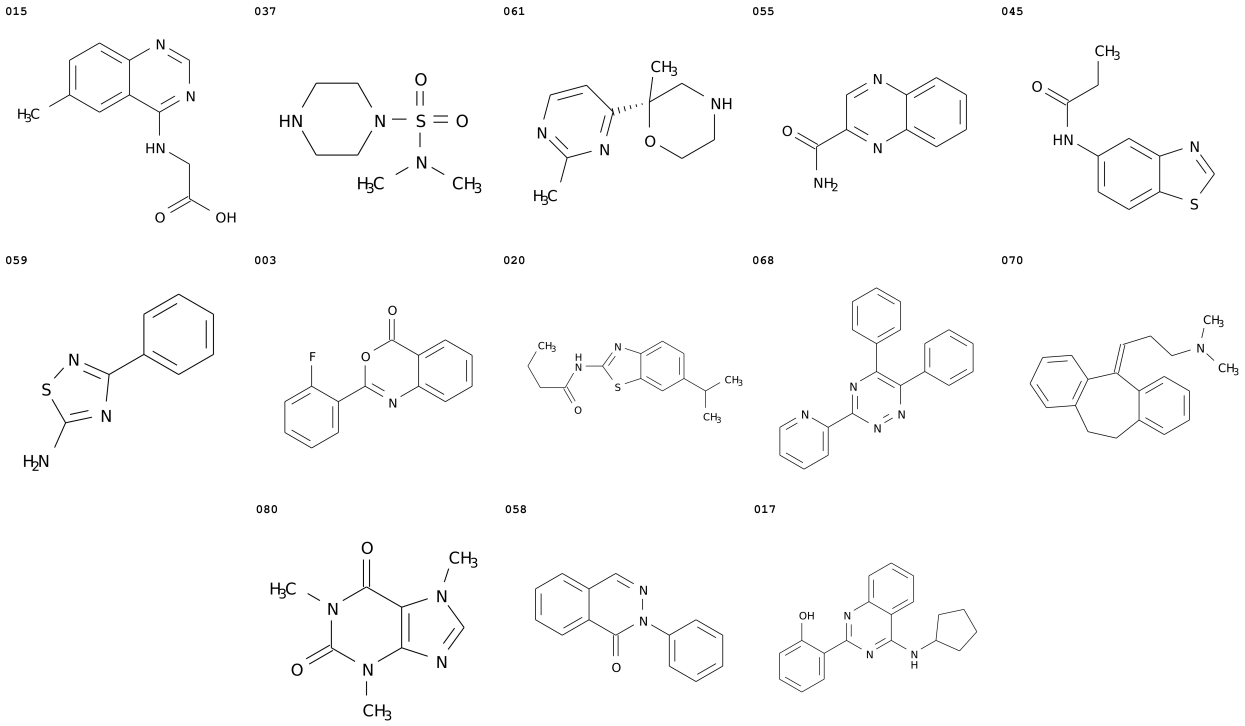


Figure 2: SAMPL5 distribution coefficient molecules batch 0. The chemical structures were created with OBabel [164].

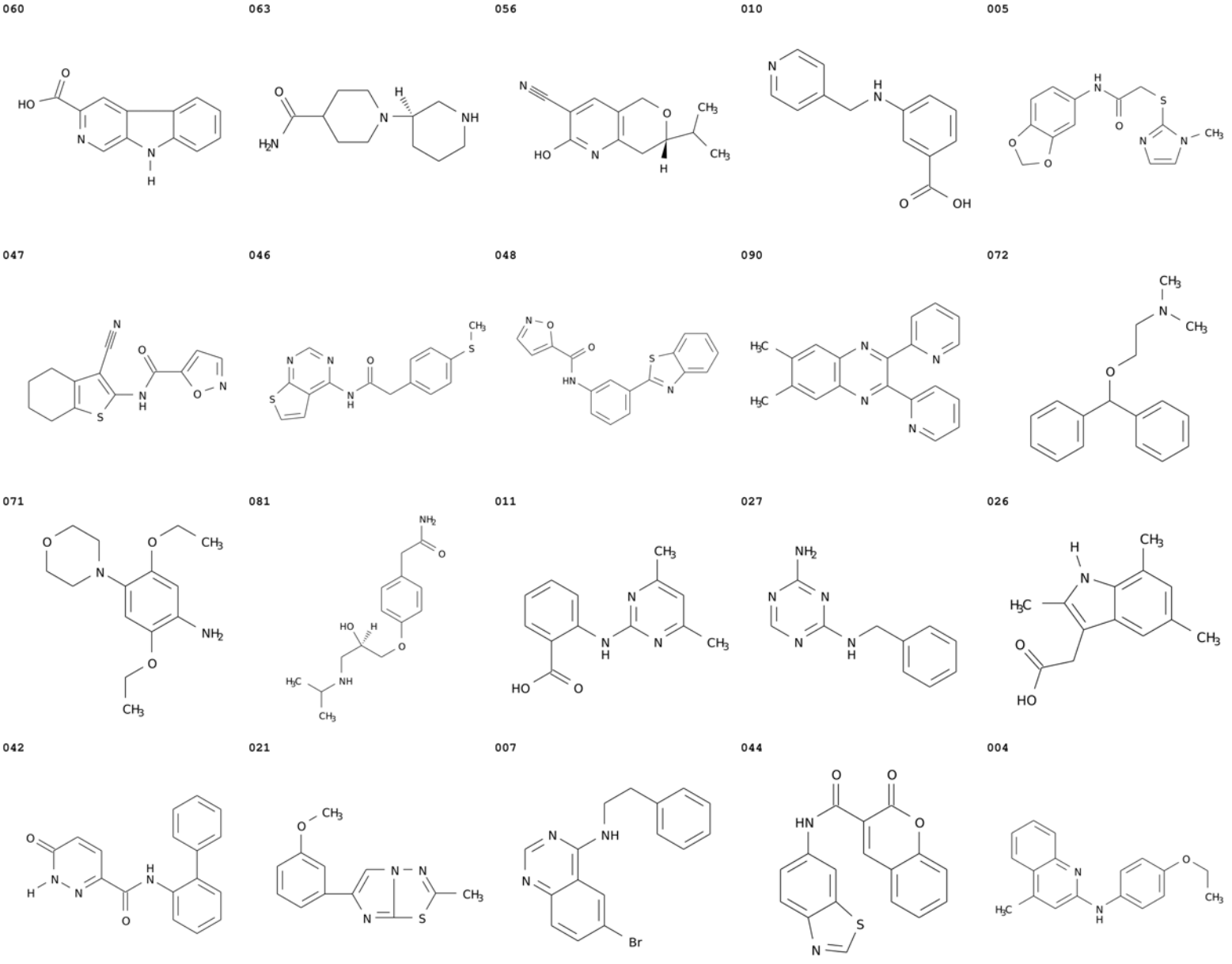


Figure 3: SAMPL5 distribution coefficient molecules batch 1. The chemical structures were created with OBabel [164].

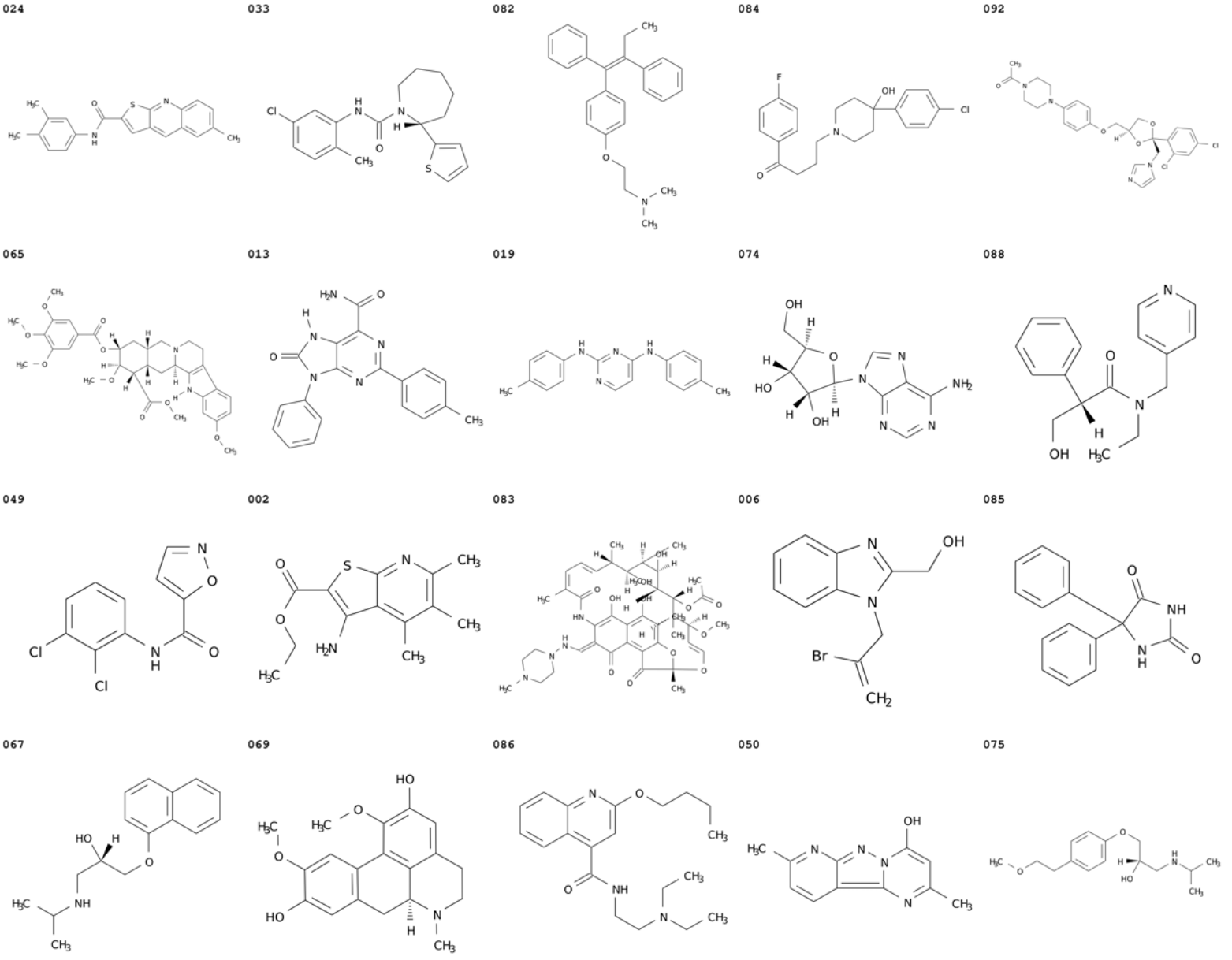


Figure 4: SAMPL5 distribution coefficient molecules batch 2. The chemical structures were created with OBabel [164].

3.2 Methods

3.2.1 Computing lipophilicity coefficients with alchemical free energy calculations

The partition coefficient logP and the distribution coefficient logD are defined by eqs. 3.1 and 3.2 respectively. Considering only the partition coefficient for sake of simplicity, it is possible to show a relationship between solvation free energies and lipophilicity values. Referring to a standard concentration of 1 M, both solvation free energy in organic phase, ΔG_o, and in water phase, ΔG_w, can be written as:

$$\begin{aligned} \Delta G_o &= -k_B T \ln \frac{[A]^o}{1M} \\ \Delta G_w &= -k_B T \ln \frac{[A]^w}{1M} \end{aligned} \tag{3.3}$$

where k_B is the Boltzmann constant, T the temperature, $[A]^o$ the concentration of a generic molecule A in the organic phase and $[A]^w$ is the concentration of molecule A in water phase. By isolating the single concentrations, eq. 3.3 can be written as:

$$\begin{aligned} e^{-\beta \Delta G_o} &= [A]^o \\ e^{-\beta \Delta G_w} &= [A]^w \end{aligned} \tag{3.4}$$

where $\beta = \frac{1}{k_B T}$. Substituting eq. 3.4 into eq. 3.1, the logP for molecule A can be expressed as:

$$\log P = \log e^{-\beta(\Delta G_o - \Delta G_w)} \tag{3.5}$$

Employing the logarithmic properties, the relation between solvation free energies and logP can be retrieved:

$$\log P = -\frac{\Delta G_o - \Delta G_w}{k_B T 2.303} \tag{3.6}$$

where the factor 2.303 comes from the $\log e$. A similar derivation can be done for logD calculations, which it is the key equation for using alchemical free energy calculations in lipophilicity calculations.

To compute each solvation free energy in eq. 3.6, the double annihilation method [66, 67], shown in fig. **2**, was employed. Each individual solvation free energies can be computed as:

$$\Delta G_{solv} = (\Delta G^{elec}_{solv} + \Delta G^{vdW}_{solv}) - (\Delta G^{elec}_{vac} + \Delta G^{vdW}_{vac}) + \Delta G_{FUNC}, \tag{3.7}$$

where *solv* is either the organic or water phase. ΔG^{elec}_{solv} and ΔG^{elec}_{vac} are the discharging step excess free energy change, in either solvent and vacuum respectively. The vanishing free energies in solvent and vacuum are given by ΔG^{vdW}_{solv} and ΔG^{vdW}_{vac}. The correction term ΔG_{FUNC} is used to account for using Barker-Watts reaction field (BWRF) electrostatics in the water and cyclohexane phase. Indeed, in both solvated phases, the system's Coulombic interactions are calculated based on BWRF. Thus, two different dielectric constant are adopted for water and organic phase simulations. However, for simulations in vacuum a reaction-field is inappropriate and instead a Coulombic potential without cutoffs was employed. Because a reaction-field is applied to all intra and intermolecular pairwise interactions, this leads to an inconsistent description of the intramolecular electrostatic interactions of the solute in the solvated and vacuum simulations.

Therefore, to enable meaningful comparisons, ΔG_{FUNC} corrects the intramolecular Coulombic interactions consistently between solvated and vacuum legs of the thermodynamic cycle depicted in fig. **2**. The ΔG_{FUNC} term is obtained via post-processing the $\lambda = 0.0$ trajectories of the discharging step of a solvated simulation and use of the Zwanzig relation [23]:

$$\Delta G_{FUNC} = -\beta^{-1} \ln\langle \exp[-\beta(U_{ic,nc}(\mathbf{r}) - U_{ic,sim}(\mathbf{r}))]\rangle_{sim}, \tag{3.8}$$

where $U_{ic,nc}(\mathbf{r})$ is the solute intramolecular electrostatic potential that depends on the coordinates $\mathbf{r}$ of the solute and Coulomb law, $U_{ic,sim}(\mathbf{r})$ is the intramolecular electrostatic potential term as computed during the simulation with a BWRF cutoff.

Finally, although logP for molecules in fig. **1** were carried out in octanol as organic phase, cyclohexane was chosen as a solvent for these simulations. Firstly, cyclohexane highly simplifies the problem, as it prevents for the presence of water molecules in the organic phase. Furthermore, analysing 85 molecules from the Minnesota database [165], it is possible to highlight an almost linear relationship between logP values calculated using octanol and logP values calculated with cyclohexane, as fig. **6** shows.

3.2.2 All-neutral and the two-species model for logD

logD values may be also estimated with the double annihilation technique. Here, two models were devised to compare logD calculations. The first model, *all-neutral*, neglects all the possible protonated species and reduces the calculation of logD to a logP simulation, where free energy of solvation are retrieved from eq. 3.7. Although the *all-neutral* model can be thought as a rough approximation, given the small number of ionisable species in the

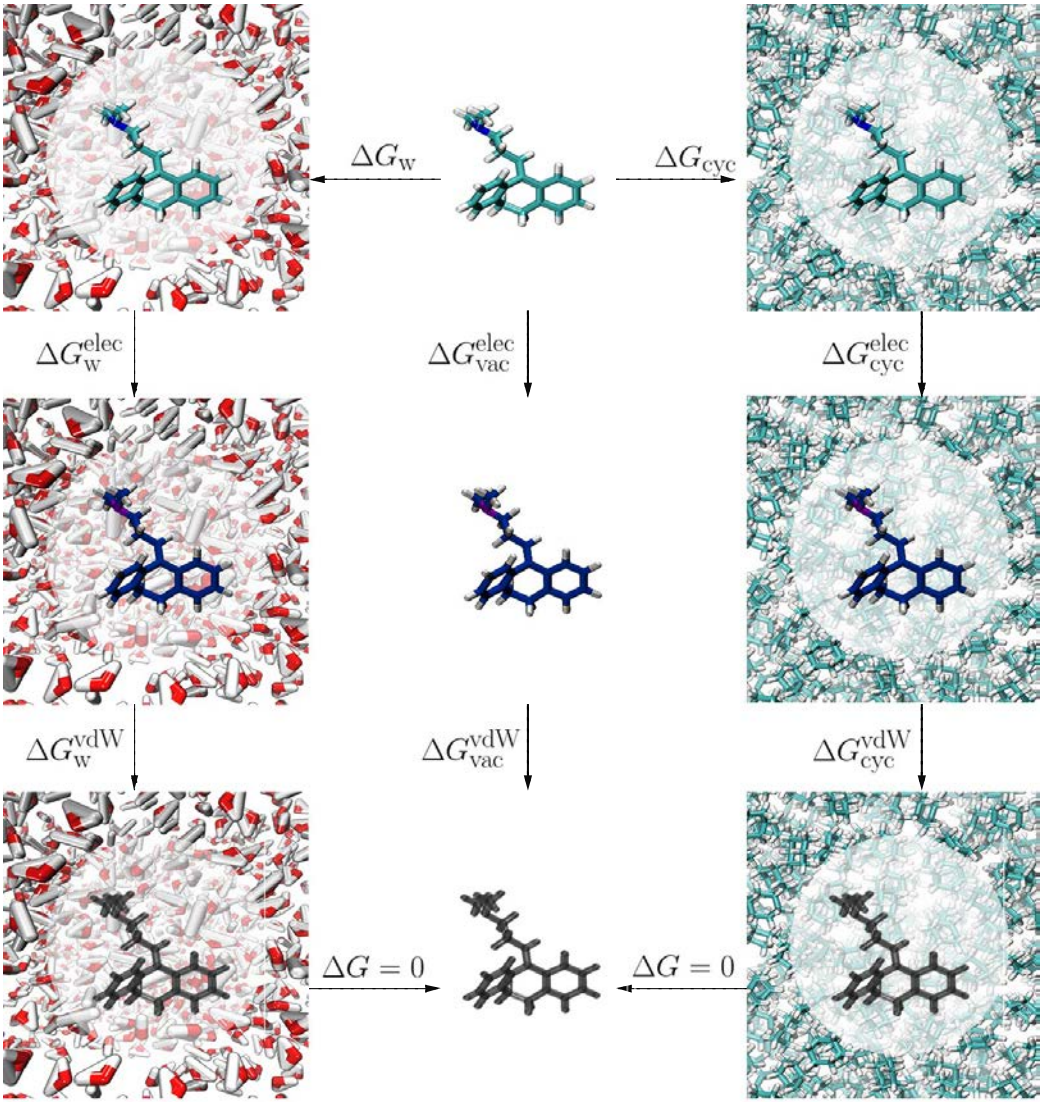


Figure 5: Thermodynamic cycle for log D calculation. First the atoms' partial charges are turned off retrieving ΔG_{w}^{elec}, ΔG_{vac}^{elec} and ΔG_{cyc}^{elec} in water, vacuum and cyclohexane phase respectively. Then, van der Waals terms are switched off and ΔG_{w}^{vdW}, ΔG_{vac}^{vdW} and ΔG_{cyc}^{vdW} are calculated in each phase.

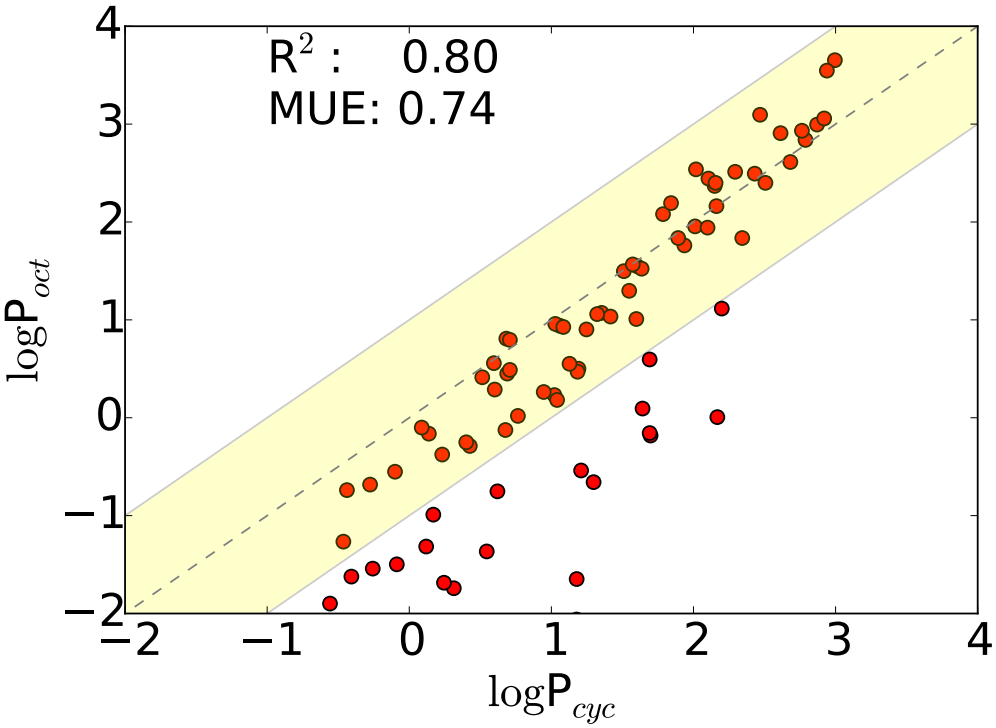


Figure 6: Relationship between logP values experimentally computed using octanol as organic phase (logP$_{oct}$) and cyclohexane (logP$_{cyc}$). The grey line assumes a perfect correlation and the yellow shaded interval corresponds to an error of 1 logP.

SAMPL5 dataset, it is expected that logP and logD would coincide for the majority of molecules. The second model, *two-species*, was derived to consider the presence of charged species at pH 7.4 for ionisable species. In particular, due to the difficult task of modelling correctly all the possible protonated state of a single molecule, only the two major species were considered. In this way it is possible to split the protonation state of a molecule into a dominant charged species, which reflects the majority of charged species at pH 7.4, and a neutral one. Generally, assuming all activity coefficients to be unity, the distribution coefficient logD is given by eq. 3.2. The concentration of the most populated species in water phase at pH 7.4 is given by:

$$[A]_w^{dom} = f^{chemicalize}(A^{dom}) \times [A]_{tot}, \tag{3.9}$$

where $f^{chemicalize}(A^{dom})$ is the fraction of the dominant species $[A]_w^{dom}$ predicted by ChemAxon [154] at pH 7.4 and $[A]_{tot}$ is set to 1 M by convention. Note that the fraction of dominant species is determined by considering potentially multiple equilibria between different charged states and tautomers. For each $[A]_w^{dom}$ it is assumed that there exists only another one species in solution, which is the $[A]_w^{dom}$ conjugate pair $[A]_w^{con}$. If there are multiple ionisable sites, $[A]_w^{con}$ is taken to be the conjugate pair that is expected to have the highest population on the basis of the pKa values of each ionisable site. Thus:

$$[A]_w^{con} = 1 - [A]_w^{dom} \tag{3.10}$$

Considering only these two species, the logD can be simplified as:

$$\log D = \log \left(\frac{[A]_o^{con} + [A]_o^{dom}}{[A]_w^{con} + [A]_w^{dom}} \right) \tag{3.11}$$

With the knowledge of $[A]_w^{con}$ and $[A]_w^{dom}$ an effective pKa^{eff} can be defined as:

$$pKa^{eff} = pH - \log \frac{[A]_w^{dom}}{[A]_w^{con}} \tag{3.12}$$

where for simplicity it was assumed that the dominant form is the base and the conjugate form the acid. Although $[A]_w^{dom}$ and $[A]_w^{con}$ are conjugate pairs, the term pKa^{eff} is used since the term $f^{chemicalize}(A_w^{dom})$ was derived by considering the co-existence of more than two species. Thus rearranging eq. 3.12 it is possible to express $[A]_w^{dom}$ and $[A]_w^{con}$ as:

$$\begin{cases} [A]_w^{dom} &= (10^{-pKa^{eff}+pH}) \\ \\ [A]_w^{con} &= (10^{-pH+pKa^{eff}}) \end{cases} \tag{3.13}$$

Eq. 3.13 allows to define the logD, eq. 3.11 as:

$$\begin{cases} \log D & = \log\left(\frac{[A]_o^{con}+[A]_o^{dom}}{[A]_w^{con}+[A]_w^{dom}}\right) \\ & = \log\left(\frac{[A]_o^{con}}{[A]_w^{con}+[A]_w^{dom}} - \frac{[A]_o^{dom}}{[A]_w^{con}+[A]_w^{dom}}\right) \\ & = \log\left(\frac{[A]_o^{con}}{[A]_w^{con}+10^{-pKa^{eff}+pH}[A]_w^{con}} + \frac{[A]_o^{dom}}{[A]_w^{dom}+10^{-pH+pKa^{eff}}[A]_w^{dom}}\right) \end{cases} \tag{3.14}$$

Now, recognizing that $10^{-pKa^{eff}+pH} = \frac{10^{-pKa^{eff}}}{10^{-pH}}$ and $10^{-pH+pKa^{eff}} = \frac{10^{-pH}}{10^{-pKa^{eff}}}$, eq. 3.14 can be simplified as:

$$\log D = \log\left(P_{A^{con}}\left(\frac{1}{1+\frac{10^{-pka^{eff}}}{10^{-pH}}}\right) + P_{A^{dom}}\left(\frac{1}{1+\frac{10^{-pH}}{10^{-pKa^{eff}}}}\right)\right) \tag{3.15}$$

where $P_{A^{con}} = \frac{[A]_o^{con}}{[A]_w^{con}}$ and $P_{A^{dom}} = \frac{[A]_o^{dom}}{[A]_w^{dom}}$. For molecules with a single ionisable site and no alternative tautomeric forms $pKa^{eff} = pKa$ and if $P_{A^{con}} >> P_{A^{dom}}$ eq. 3.15 is further simplified to a more commonly used approximation [166]:

$$\log D = \log\left(P_{A^{con}}\left(1+\frac{10^{-pKa}}{10^{-pH}}\right)^{-1}\right) \tag{3.16}$$

3.2.3 log**P** simulations setup

logP were estimated for molecules shown in fig. **1**. Starting from the SMILE string [167, 168], each molecule was converted into a pdb format with obabel [164]. Then, a GAFF AM1-BCC parametrization was performed with `Antechamber`, module of AmberTools16 [73]. Experimental logP measurements were carried on with 1-octanol as organic phase. As organic solvent cyclohexane was chosen to study the logP trend computationally. Employing `tleap` package, each compound was solvated in GAFF cyclohexane and TIP3P water [120] cubic boxes, setting the minimum distance between solute and box walls to 20.0 Å. Then, all the systems were energy minimized with 100 steps of steepest descent algorithm and equilibrated. Firstly, the water molecules were equilibrated in a NVT ensemble for 200 ps at 298 K, followed by a NPT simulation for further 200 ps at 1 atm with AmberTools module Sander. Finally, a molecular dynamics simulation in NPT ensemble was run, to reach the final desired density ($0.7\ g \cdot cm^{-3}$ for cyclohexane and $1 g \cdot cm^{-3}$ for water).

Alchemical free energy calculations were run for 11 equidistant λ windows, both for the discharging and vanishing steps. Each λ window lasted for 2 ns,

employing a velocity-Verlet integrator in the MD, with a time step of 2 fs. Simulations were performed in a NPT ensemble and temperature control was achieved with an Andersen thermostat with a coupling constant of 10 ps^{-1}. Pressure control was maintained by a Monte Carlo barostat, with isotropic scaling every 100 fs. Periodic boundary conditions were imposed with a 12 Å cutoff for the non-bonded interactions, using a shifted atom-based Barker Watter reaction field, with a dielectric constant of 82.0 for TIP3P water and 1.0 for GAFF cyclohexane. All solute's bonds were constrained.

3.2.4 log**D simulations setup**

All molecules were parameterized with the general Amber force field (GAFF) [117], solvated in cubic boxes of TIP3P water molecules [120] and GAFF cyclohexane. Solutes were initially energy minimized for 100 cycles by using the steepest descent method with harmonic positional restraints and a restraint force constant of 10 kcal·mol^{-1} $Å^{-2}$ applied to the whole water molecules or cyclohexane molecules respectively. Secondly, a NVT equilibration of 200 ps at 298 K, following a NPT equilibration at 1 atm with Amber module Sander[73] were carried out. Finally, a 2 ns simulation in the NPT ensemble was run with SOMD (rev 15.1) [68, 69], to reach a final density of 1 g/cc and 0.7 g/cc for water and cyclohexane respectively. Then, coordinate files were retrieved with CPPTRAJ [169]. From the mol2 file the topology and the coordinates for vacuum simulations were created with the help of `tleap`. For the *two-species model*, the pKa and protonated states of each molecule was computed with ChemAxons [154] at pH 7.4. Then, Antechamber 14 [73] was run to obtain AM1-BCC charges [118]. In the case of charged species the molecules were then re-solvated and underwent the same procedure as described above for the uncharged species.

Each discharging step was divided into nine equidistant λ windows. For the vanishing step, 11 equidistant λ windows were used, and an additional window was added at 0.950, to capture large fluctuations in the free energy changes towards the end of the decoupling process. Each λ window was run for 2 ns in the organic and aqueous phase, except molecules **007**, **013**, **019**, **024**, **042**, **056**, **065**, **071**, **088**, and **092**, whose vanishing step lasted for 6 ns, to improve the precision of the computed free energy changes. Additionally, for vacuum simulation each λ window was run for 0.8 ns. A velocity-Verlet integrator was employed with a time step of 4 fs using a hydrogen mass repartitioning scheme (HMR) [170] by constraining all bonds. All simulations were performed at 298 K and 1 atm in a NPT ensemble, using an

atom-based Barker Watts reaction field[49] with a dielectric constant of 78 for the water phase and a dielectric constant of 1.0 for the cyclohexane phase. The non-bonded interactions cutoff was set to 12 Å and periodic boundary conditions were imposed. Andersen thermostat with a coupling constant of 10 ps^{-1}[52] assured the temperature control, while a Monte Carlo barostat was used for pressure control, attempting isotropic box edge scaling every 25 time steps.

3.2.5 Alchemical free energies estimations

To ensure accuracy and consistency of logP and logD calculations, all solvation free energies in water and cyclohexane were computed twice assigning different initial molecular velocities, drawn from the Maxwell-Boltzmann distribution. All solvation free energies were estimated by using MBAR [171]. Thus, the computed lipophilicity coefficients are reported as the average of the two independent simulations and statistical uncertainties were computed as:

$$err(\Delta G) = \frac{\sigma}{\sqrt{n}}, \tag{3.17}$$

where σ is the standard deviation of both runs and n=2, unless otherwise stated. These are the error bars reported in the results section.

Insights into the statistical distribution of the logD estimations of the correlation coefficient, R^2, mean unsigned error, MUE and Kendall τ were retrieved with a bootstrapping scheme. Each data point was modelled with a normal distribution with its mean given by the computed free energy and σ the associated computed error. Ten thousand samples were then drawn from the artificial normal distribution for each data point and correlated with the experimental values, giving a distribution of R^2, MUE and Kendall τ. This final statistical distributions may not be symmetric around the mean and uncertainties in the dataset metrics are reported with a 95% confidence interval written in the follow way $-z < \mu < z$, where $-z$ is the lower bound and z the upper bound of the confidence interval and μ the mean of the distribution. All simulation input files and post processing scripts needed for reproducing the results as well results files can be found in a github repository `https://github.com/michellab/Sire-SAMPL5`.

3.3 Results

3.3.1 logP results

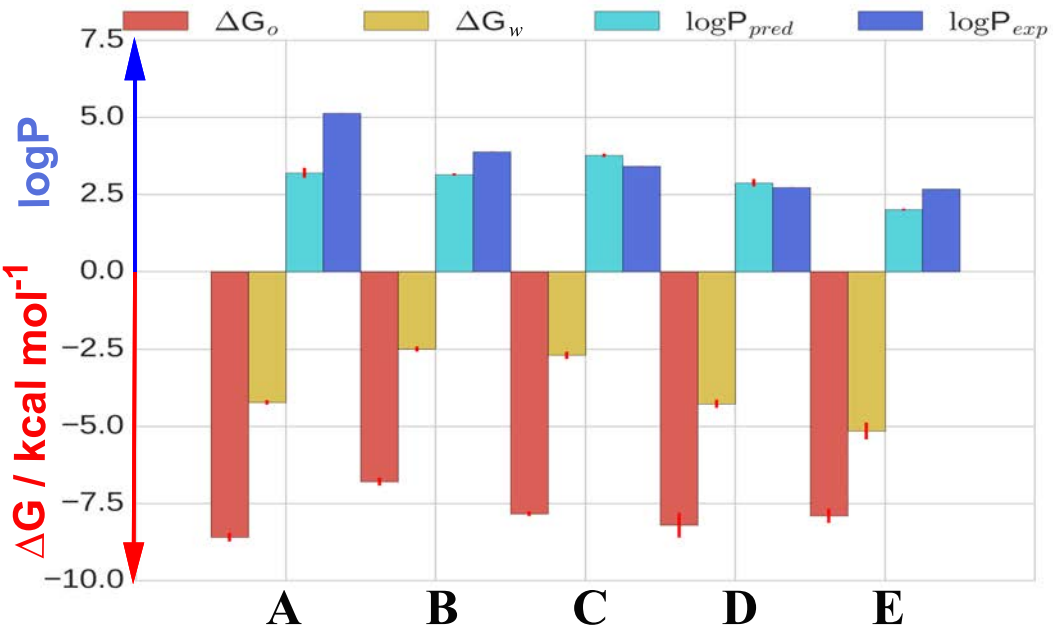


Figure 7: The positive y-axis depicts a comparison between the calculated (cyan) and measured (blue) logP for compounds **A**, **B**, **C**, **D** and **E**, depicted in fig. **1**. The negative y-axis shows calculated solvation free energy in cyclohexane, ΔG_o (red), and aqueous, ΔG_w (yellow) phases.

Fig. **7** and tab. **1** show a comparison of the calculated lipophilicity coefficient, $logP_{pred}$, and the experimental measure, $logP_{exp}$, as well as the estimated solvation free energy in water phase, ΔG_w and in the organic-cyclohexane phase, ΔG_o. Overall, the logP calculations are in good agreement with the experimental data trend, with a Kendall tau of 0.5±0.1 and mean unsigned error of 0.77±0.07 logP units. Increased fluorination tends to decrease logP values. The use of trajectory-based simulations allows to investigate the origin of this decrement. Such a trend is mainly due to a more rapid decrease in the hydration free energies, passing from -2.49±0.08 kcal·mol^{-1} for molecule B to -5.15±0.27 kcal·mol^{-1} for molecule **E**. There are less variations in solvation free energy in cyclohexane, where all the

fluorinated compounds (C, D and E) show an average value of about -7.5 kcal·mol^{-1}.

Therefore, further insights were pursued to help rationalise the calculated differences in hydration free energies, by using grid-cell theory (GCT) analyses of the MD simulation trajectories [172, 173, 174, 175]. GCT is an MD trajectory post-processing method that spatially resolves the water contribution to enthalpies, entropies and free energies of hydration for small molecules, host/guests and protein-ligands complexes.

Fig. **8** depicts the spatially resolved hydration thermodynamics around the non-fluorinated cyclohexane **A** and the tetrafluorinated cyclohexane **E**. Comparison of the water density contours show water structuring above and below the π-cloud of the phenyl ring due to the expected weak hydrogen bonding interactions in this region. In addition, the fluorine atoms in **E** induce further structuring of water around the cyclohexyl moiety, with a more pronounced effect around the hydrogen face of the cyclohexane (panels 1 and 2). Owing to the different polarities of the cyclohexyl ring in **E**, water near the fluorine-face preferentially orients hydrogen atoms towards the ring, whereas water near the hydrogen-face preferentially orients oxygen atoms towards the ring. Water near the hydrogen-face is more enthalpically stabilised and entropically destabilised with respect to bulk, whereas the energertics are not significantly different from the bulk in the vicinity of the fluorine face (panels 3-4 and 5-6). Overall, favourable enthalpic contributions offset unfavourable entropic contributions for water near the hydrogen face and water in this region makes additional favourable contributions to the hydrtaion free energy (panels 7-8). Therefore, the decreased lipophilicity of **A** with respect to **E** is attributed to enhanced hydrogen bonding interactions between water and the hydrogen face of the all-*cis* tetrafluorcyclohexane ring.

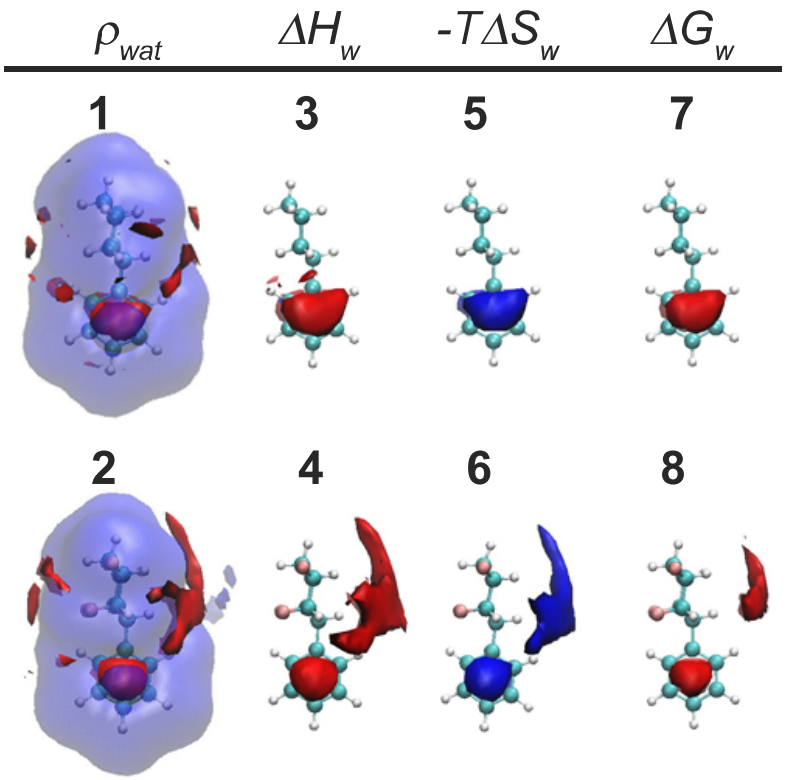


Figure 8: Spatial resolution of hydration thermodynamics around molecules **A** and **E**. Panels **1** and **2** show isocontours for density (red: $\rho_{wat} > 2.33$ bulk density, blue: $\rho_{wat} < 0.5$ bulk density). Panels **3** and **4** show isocontours for regions where water is enthalpically stabilised with respect to bulk water (red: $\Delta H_w < 0.0055$ kcal·mol^{-1} Å^{-3}). Panels **5** and **6** show isocontours for regions where water is entropically destabilised with respect to bulk water (blue: $-T\Delta S_w > 0.0033$ kcal·mol^{-1} Å^{-3}). Panels **7** and **8** show isocontours for regions where water is more stable than bulk water (red: $\Delta G_w < 0.0055$ kcal·mol^{-1} Å^{-3}).

Table 1: Comparison between the predicted $\log P_{pred}$ and the experimental measures, $\log P_{exp}$, along with their standard errors, σ_{err} for molecules **A**, **B**, **C**, **D** and **E**, shown in fig. **1**. ΔG_o and ΔG_w are the solvation free energy in cyclohexane and the hydration free energy respectively, along with their standard errors, expressed in kcal·mol^{-1}.

Compound	$\Delta G_o \pm \sigma_{err}$	$\Delta G_w \pm \sigma_{err}$	$\log P_{pred} \pm \sigma_{err}$	$\log P_{exp} \pm \sigma_{err}$
A	-8.59±0.14	-4.22±0.07	3.20±0.16	5.14±0.01
B	-6.79±0.13	-2.49±0.08	3.15±0.04	3.88±0.01
C	-7.83±0.08	-2.69±0.12	3.77±0.05	3.42±0.01
D	-8.20±0.04	-4.27±0.13	2.88±0.12	2.73±0.01
E	-7.90±0.23	-5.15±0.27	2.01±0.03	2.68±0.01

3.3.2 log**D results**

Fig. **9** and **10** show the performance of *all neutral* and *two species* models respectively with respect to the experimental values, along with the determination coefficient R^2, MUE and Kendall τ. Solvation free energies for the whole dataset of fig. **2**, **3**, **4** are reported in the appendix in tab. **2** and tab. **3**.

The *all-neutral* model was applied for all the 53 molecules of the SAMPL5 dataset. There is an acceptable agreement with the experimental trend, as highlighted by the Kendall τ = 0.41±0.03, with a MUE of 2.20±0.08 logD units. The high MUE is due to 19 ionisable species, like molecules **075** and **083**. As an example, the *all-neutral* model predicts a logD for molecule **075** of 3.00±0.71, which contrasts the experimental value of -2.80±0.30, and for molecule **083** a logD 7.95±1.20 (-1.90±0.40 experimentally). This last molecules is particularly challenging, since it likely adopts multiple protonation states at pH 7.4, thus the high discrepancy with respect to the experimental value may be due to the large size and number of functional groups present in **083**. Considering only the 19 ionisable molecules, the *all-neutral* shows a high MUE (2.66±0.13 logD units), highlighting the limits of this approach to the logD calculation.

Additionally, the *all-neutral* model consistency and reproducibility was compared with *all-neutral* results run by the Mobley group (UCI) [74]. The same input files were used, but free energy calculations were performed with software GROMACS [176]. Fig. **11**, **12** and **13** compares SOMD and GROMACS logD, hydration free energies and solvation free energies in cyclohexane respectively. Comparing the logD predictions, a fair agree-

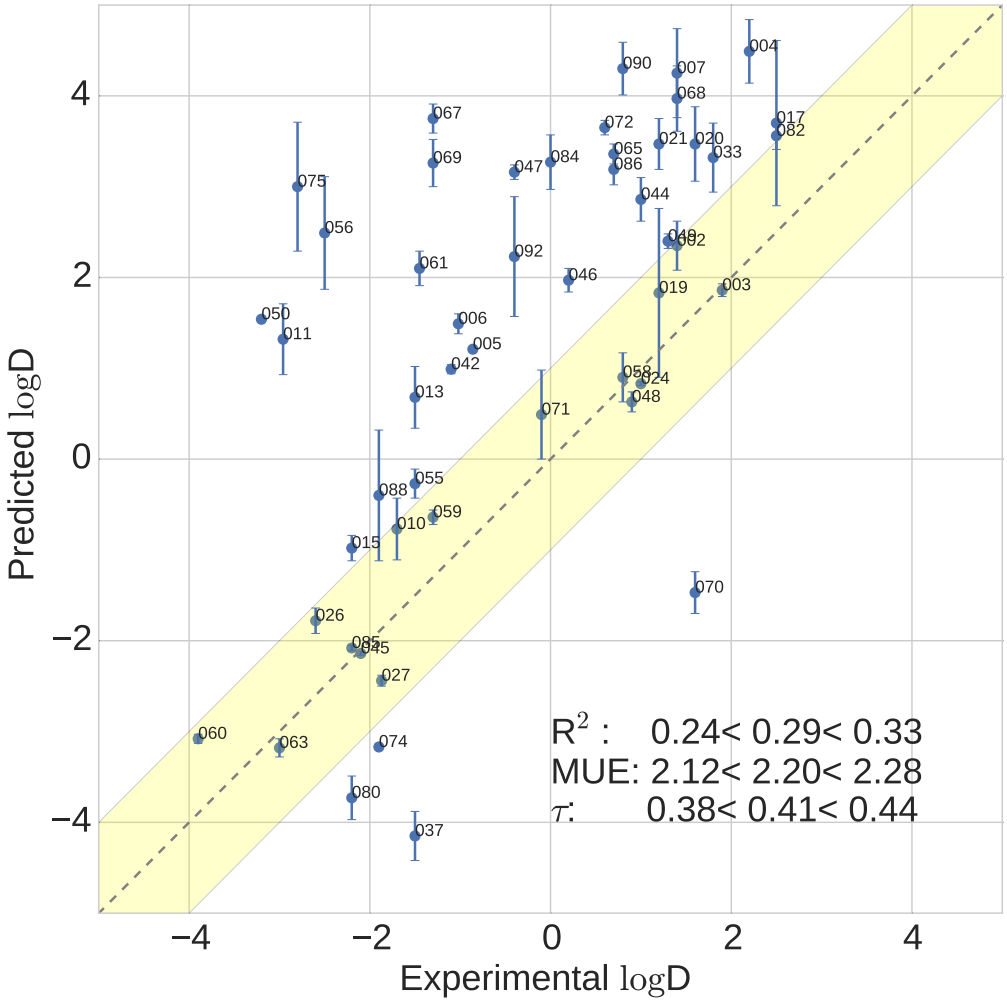


Figure 9: Scatter plot of the computed logD for molecules modelled as neutral, *all-neutral*, in water and in cyclohexane, R^2, MUE (in logD units) and Kendall τ values are given with 95 % confidence intervals. The grey line assumes a perfect correlation and the yellow shaded interval corresponds to an error of 1 logD.

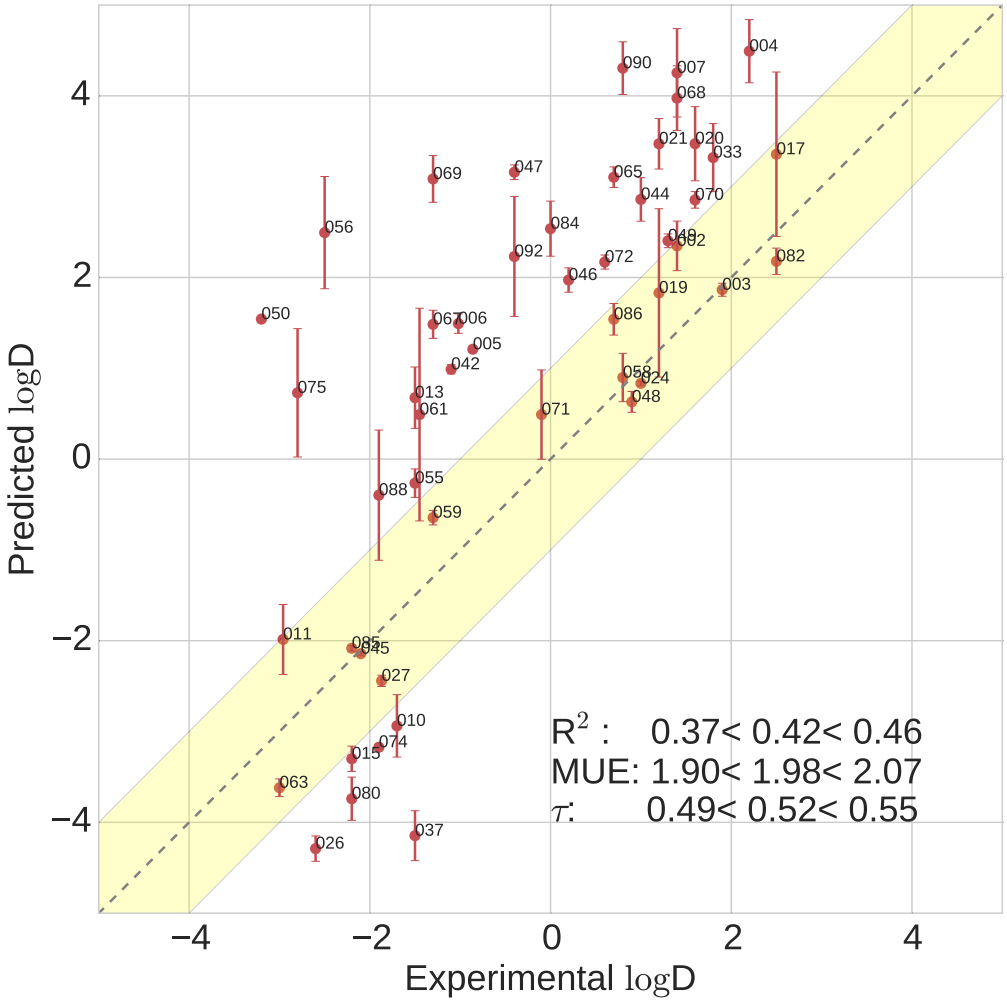


Figure 10: Scatter plot of the computed logD for molecules modelled as neutral, *two-species*, in water and in cyclohexane, R^2, MUE (in logD units) and Kendall τ values are given with 95 % confidence intervals. The grey line assumes a perfect correlation and the yellow shaded interval corresponds to an error of 1 logD.

ment is observed with R^2=0.61±0.06 and the mean unsigned deviation MUD = 0.85±0.09 logD units. As explained before, molecule **083** is an outlier in SOMD predictions, however for GROMACS the predicted logD was 1.21±0.09, which is closer to the experimental value. The next outlier is molecule **017**, which shows SOMD logD = 3.7± 0.9 and GROMACS logD = 6.25± 0.04, followed by **082**, SOMD logD = 3.6± 0.1 and GROMACS logD = 6.56±0.05. Additionally, comparing the solvation free energy between SOMD and GROMACS, differences in cyclohexane solvation free energy for **082** and **017** are present. In particular, **082** is the largest outlier, with an absolute difference between SOMD and GROMACS predictions of 4.2 kcal·mol^{-1}, while **017** shows a difference of 3.3 kcal·mol^{-1}. Nonetheless, the free energy predictions are overall in better agreement, with R^2=0.94±0.02 and MUE=0.75±0.09 kcal·mol^{-1} of hydration free energy and R^2=0.85±0.01 and MUE= 1.01±0.09 kcal·mol^{-1} for solvation free energy in cyclohexane. In the GROMACS protocol, the alchemical free energies were computed with 20 λ windows, both for the discharging and vanishing step, and using PME [177] for electrostatics calculations. In contrast, SOMD uses 9 λ windows for the discharging step and 12 for the vanishing one, along with a Barker-Watts atom based reaction field [49]. These protocol differences may be the source of variability.

Thus, the *two-species* model was employed for the prediction of actual logD values. Considering the co-existence of two states for ionisable molecules, gives a remarkable improvement in the logD predictions with respect to the *all-neutral* model. All the statistical estimators improve, R^2 = 0.42±0.06, Kendall τ= 0.52±0.03 and MUE improves by 0.22 logD units. Considering ionisable molecules only, MUE is improved by 0.66 logD units, equivalent to 0.90 kcal·mol^{-1} in the difference of solvation free energies, along with a R^2= 0.52±0.07 and Kendall τ= 0.57±0.06. As in the *all neutral* model, molecule **083** is an outlier, with a predicted logD of 7.94±0.4. However unlike for the *all-neutral* model, the logD value for molecule **075** is more accurate (0.72±0.70 vs 3.00±0.71). Again, the possible presence of multiple ionisable sites makes difficult to model rigorously the partitioning process with AFE calculations.

To test the utility of using an effective pKa in the *two species* model, estimations for charged molecules were compared by application of eq. 3.16. For the 19 protonated molecules considered, the *two-species* model obtained a MUE = 2.1 logD units, while eq. 3.16 gives a MUE = 2.3 logD units. This is due to 5 molecules that have a different pKa and effective pKa values, owing to the co-existence of multiple protomers and tautomers at pH 7.4 (**010, 011, 015, 060, 063**). For these 5 molecules the *two-species* model performs well with a MUE = 1.0 logD units, which is significantly better

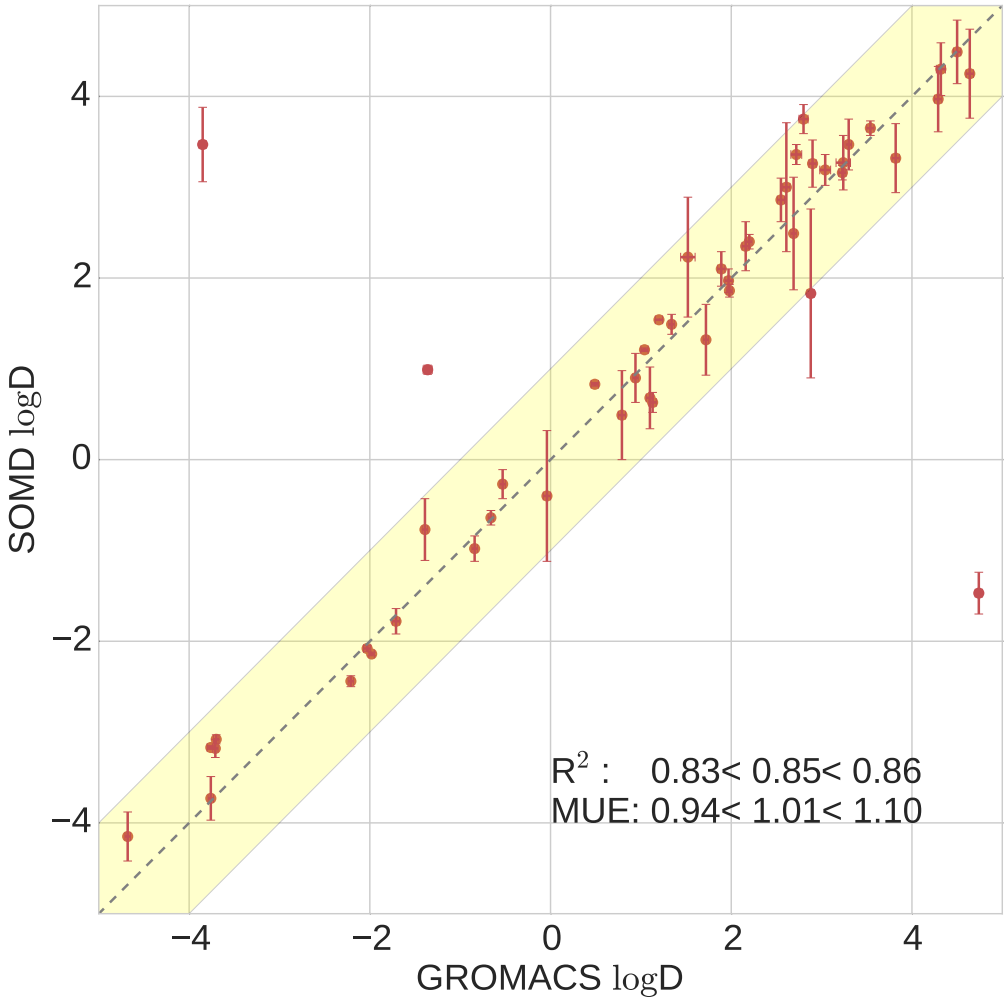


Figure 11: Comparison between SOMD and GROMACS logD estimations, both based on considering all the 53 molecules as neutral. The MUE is given in logD units, while the shaded yellow area is ± 1 logD deviation bound.

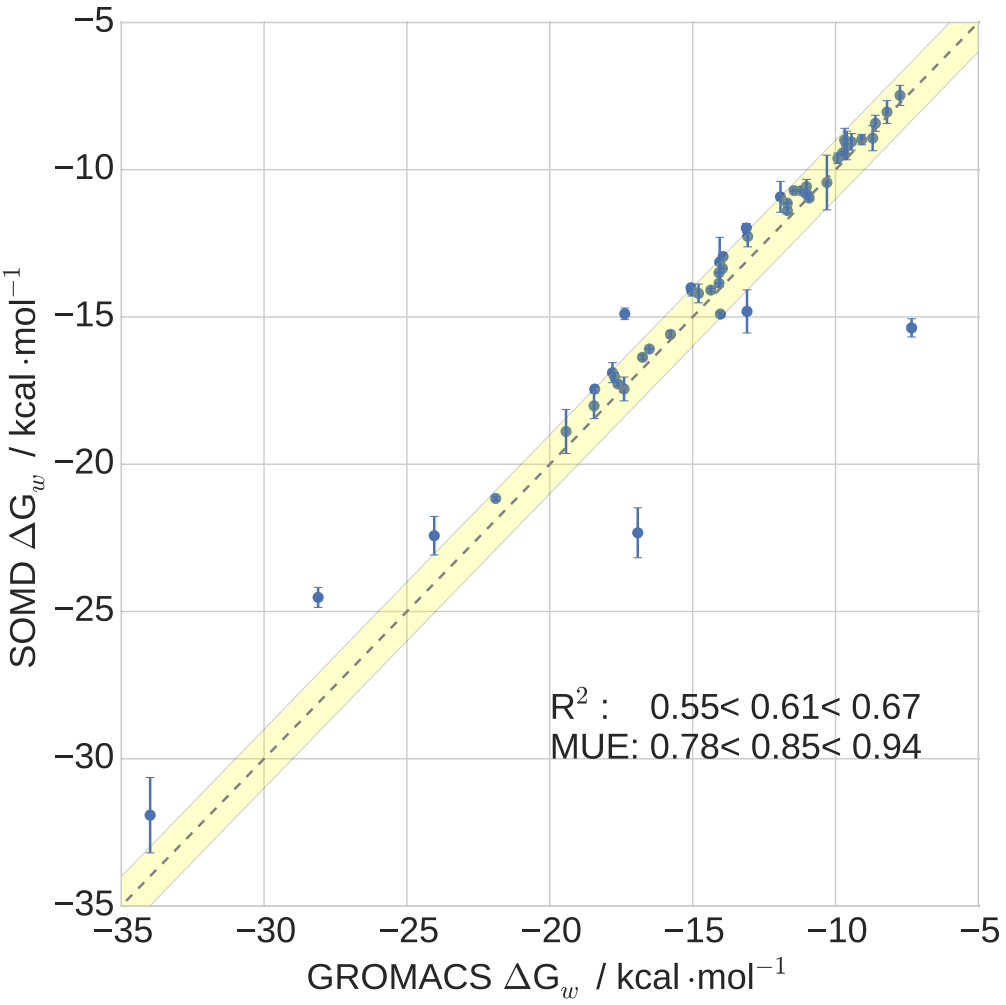


Figure 12: Comparison between SOMD and GROMACS ΔG_w estimations, both based on considering all the 53 molecules as neutral. MUE is given in kcal·mol^{-1} while the shaded yellow area is $\pm$ 1 kcal·mol^{-1} deviation bound.

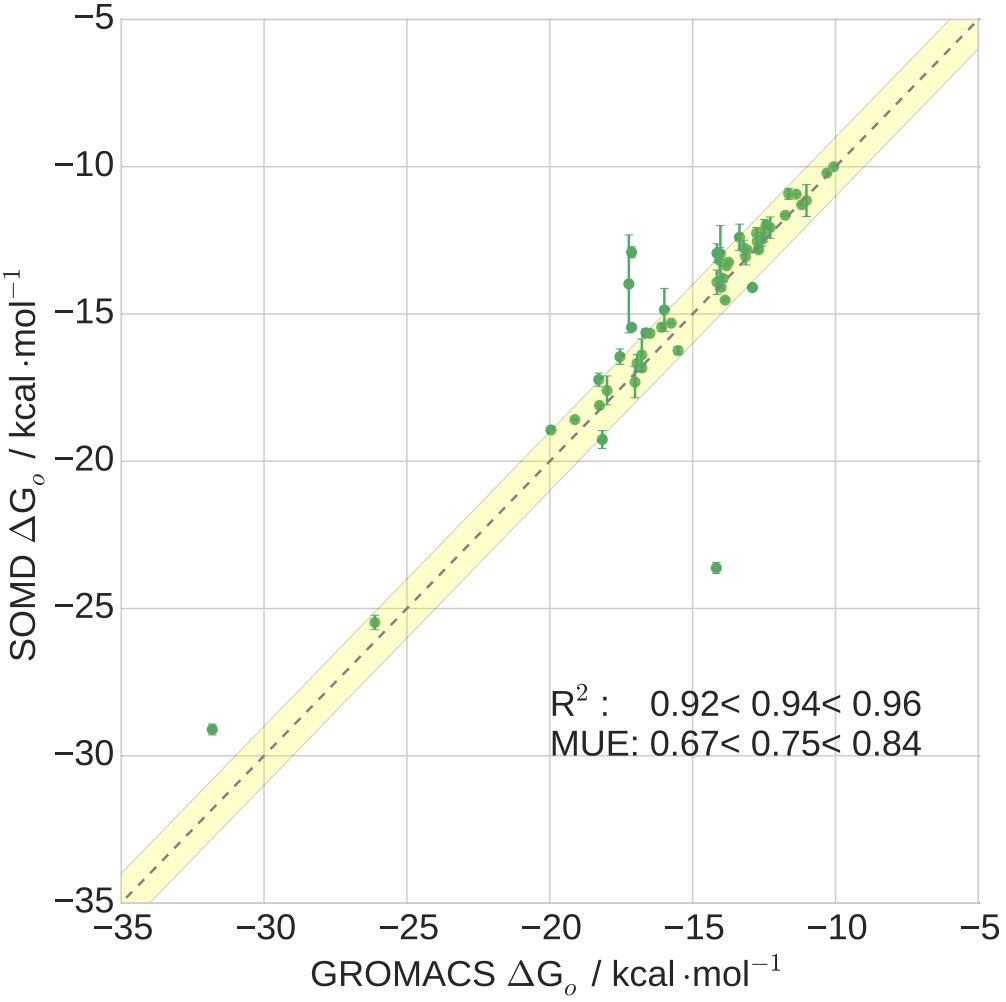


Figure 13: Comparison between SOMD and GROMACS ΔG_o estimations, both based on considering all the 53 molecules as neutral. MUE is given in kcal·mol^{-1} while the shaded yellow area is $\pm$ 1 kcal·mol^{-1} deviation bound.

than the MUE = 2.4 given by eq. 3.16. However, given the small size of the dataset, it is not possible to assert whether the improvements are statistically significant.

3.4 Conclusion

AFE were employed to estimate lipophilicity coefficients for a dataset of drug like molecules, as shown in fig. **2**, **3** and **4**. logP were estimated for 5 bicyclic rings rings, which show an increase in hydrophilicity as fluorination increases. AFE were able to reproduce the experimental trend, by modelling the logP process as a transfer from aqueous phase to cyclohexane-organic phase, using a double annihilation technique. Furthermore, MD analyses suggest that the decreased lipophilicity of molecules were due to enhanced hydrogen bonding interactions of water molecules with the hydrogen face of the cyclohexane ring, with respect to bulk water. Additionally, the orientation of the water molecules near the cyclohexane face of the ring was consistent with the hydrogen bond donor ability of the polar hydrogens on the cyclohexane ring. This contrasts with the energetics of water near the fluorine face of the ring which are comparable to bulk water. Finally, this study shows how AFE can improve logP estimations with respect to QSAR techniques, giving insights about the atomistic behavior and interactions.

Modelling more complex molecules, such as those in SAMPL5 (fig. **2**, **3** and **4**) highlighted more difficulties for AFE methods. Two models were considered: *all-neutral* model, which do not take into account the presence of ionized molecules at pH 7.4, and *two-species*, which considers the co-existence of at least of two species for ionisable species. Simulated logD values show a low correlation with experimental data for the *all neutral* model, especially for highly complicated system such as molecules **083** and **075**. A similar statistical trend was found by Bannan *et al* [74], which approximate all the molecules as neutral as well. An improvement is shown for the *two species* model, achieving a R^2 of 0.42±0.04 and lowering the MUE to 1.98±0.19 logD units. Comparing the *two-species* model to other participants submissions, results are encouraging, since the model is among the top ranked submissions in terms of MUE. The best predictions were achieved by COSMO approach [178, 179] and the 3D-RISM [180, 181]. This further suggests that quantum-based methods are more refined and tuned to correctly forecast lipophilicity coefficients with respect to trajectory-based models. Further inspection of the results demonstrated that the contribution of charged species ($P_{A^{dom}}$) to the predicted logD values was negligible. While this suggests that the evaluation of vacuum to cyclohexane transfer free energies of charged species are unnecessary, it will be interesting to evaluate this assertion in more complex scenarios where for instance charged solutes partition into cyclohexane together with clusters of water molecules, as tested by Bannan *et al.* [74]. The approach could be further generalised to handle more com-

plex molecules that adopt multiple charge states, but a drawback is that the results depend on the values of ionisation and tautomerisation equilibrium constants. Consequently, robust predictions will require accurate computation of vacuum to solvent transfer free energies, and also pKa constants.

A major problem for the MD models is the pKa estimation. Indeed, many of the molecules in fig. **2**, **3** and **4** could adopt multiple protonation states in aqueous solution at pH 7.4. A third source of error was introduced by the present of finite size electrostatics artefacts, which arises when a net charges are perturbed between the two end states of alchemical free energy calculations. A test to correct for these errors was done for the 19 ionisable species, but not considered due to the preliminary stage of this approach. On one side, the correction terms were meaningful for hydration free energy estimations. On the other side, cyclohexane correction terms were underestimated. A possible explanation is that GAFF force field as used here is unbalanced and favors solvation of solutes in a non-polar solvent.

In conclusion, predictions of lipophilicity coefficients can be done by using AFE techniques. However, efforts should be devoted to improve models for logD estimations. In particular, attention should be focused on pKa corrections and polarisable force-fields for modelling transfer between polar/non-polar solvents. It would be useful to devise a new dataset, to enable testing and training the accuracy and robustness of molecular simulation protocols. This could be carried out by separating existing datasets into compounds predicted to adopt a single protomer/tautomer form in aqueous and organic phases, and ionisable compounds that may adopt multiple charged states. In the first case, logD and logP would be equivalent and their evaluation would not require pKa considerations. Ideally, force fields validated on this category of compounds could be then combined with pKa estimators, to address the more challenging situation where multiple species contribute to a logD value.

3.5 Appendix

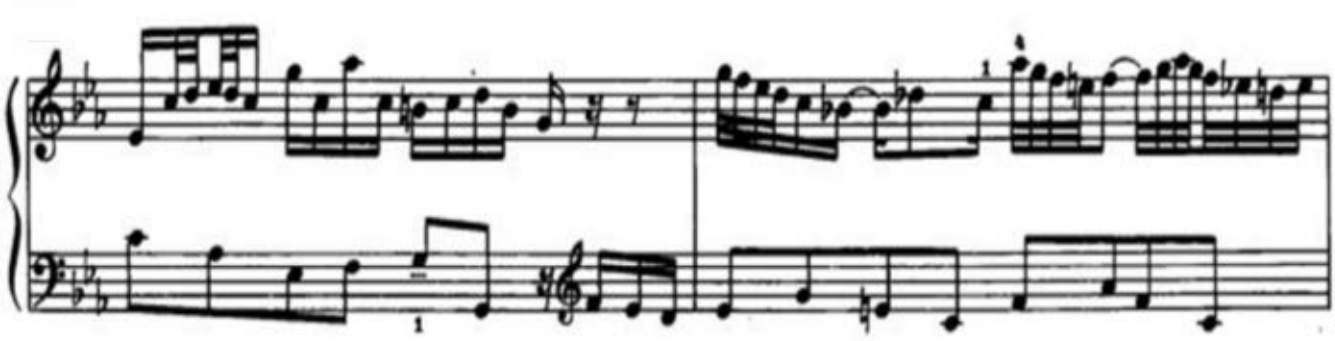
Table 2: *all-neutral* hydration free energies and logD values. *molecule* denotes the molecule number (fig. **2**, **3** and **4**, ΔG_w is the hydration free energy, ΔG_o the solvation free energy in cyclohexane, logD the predicted distribution coefficient while $\log D_{exp}$ is the experimental value. σ denotes the standard error for each quantity. All the free energy values are in kcal·mol^{-1}.

molecule	ΔG_w	σ_w	ΔG_o	σ_o	logD	$\sigma_{\log D}$	$\log D_{exp}$	$\sigma_{\log D_{exp}}$
002	-10.72	0.04	-13.92	0.41	2.35	0.43	1.40	0.30
003	-9.44	0.04	-11.98	0.14	1.86	0.14	1.90	0.10
004	-9.18	0.49	-15.31	0.01	4.49	0.01	2.20	0.30
005	-14.01	0.05	-15.66	0.08	1.21	0.08	-0.86	0.09
006	-9.61	0.18	-11.65	0.03	1.49	0.03	-1.02	0.09
007	-9.05	0.07	-14.86	0.73	4.25	0.76	1.40	0.30
010	-14.10	0.19	-13.05	0.28	-0.77	0.29	-1.70	0.40
011	-11.39	0.09	-13.18	0.43	1.32	0.45	-2.96	0.08
013	-18.02	0.43	-18.94	0.03	0.68	0.03	-1.50	0.40
015	-14.09	0.06	-12.75	0.25	-0.98	0.26	-2.20	0.30
017	-8.93	0.43	-13.98	1.66	3.70	1.72	2.50	0.30
019	-14.81	0.73	-17.31	0.53	1.83	0.55	1.20	0.40
020	-18.89	0.75	-23.62	0.19	3.47	0.20	1.60	0.30
021	-9.05	0.28	-13.79	0.10	3.47	0.10	1.20	0.30
024	-17.45	0.06	-18.59	0.02	0.83	0.02	1.00	0.40
026	-13.35	0.00	-10.92	0.19	-1.78	0.20	-2.60	0.10
027	-15.59	0.11	-12.26	0.20	-2.44	0.20	-1.87	0.07
033	-10.92	0.53	-15.46	0.01	3.32	0.01	1.80	0.20
037	-17.44	0.41	-11.15	0.54	-4.61	0.56	-1.50	0.10
042	-14.89	0.20	-16.24	0.13	0.99	0.13	-1.10	0.30
044	-14.20	0.32	-18.11	0.01	2.86	0.01	1.00	0.40
045	-13.86	0.01	-10.93	0.02	-2.14	0.02	-2.10	0.20
046	-12.95	0.03	-15.64	0.15	1.97	0.16	0.20	0.30
047	-11.13	0.02	-15.45	0.13	3.16	0.14	-0.40	0.30
048	-16.37	0.07	-17.23	0.22	0.63	0.23	0.90	0.40
049	-9.49	0.03	-12.77	0.08	2.40	0.08	1.30	0.10
050	-10.71	0.03	-12.82	0.02	1.54	0.03	-3.20	0.60
055	-10.58	0.25	-10.21	0.03	-0.27	0.03	-1.50	0.10
056	-8.99	0.40	-12.40	0.45	2.49	0.46	-2.50	0.10
058	-10.84	0.00	-12.07	0.36	0.90	0.38	0.80	0.10

059	-10.88	0.11	-10.00	0.00	-0.64	0.00	-1.30	0.30
060	-17.03	0.14	-12.82	0.08	-3.08	0.08	-3.90	0.20
061	-8.43	0.27	-11.29	0.01	2.10	0.01	-1.45	0.09
063	-16.89	0.34	-12.55	0.21	-3.18	0.22	-3.00	0.40
065	-24.52	0.34	-29.11	0.19	3.36	0.19	0.70	0.20
067	-8.98	0.18	-14.10	0.03	3.75	0.04	-1.30	0.30
068	-10.96	0.06	-16.39	0.54	3.97	0.56	1.40	0.30
068	-10.96	0.06	-16.39	0.54	3.97	0.56	1.40	0.30
069	-13.14	0.84	-17.60	0.49	3.26	0.51	-1.30	0.30
070	-15.37	0.31	-13.37	0.00	-1.47	0.00	1.60	0.30
071	-12.27	0.35	-12.94	0.32	0.49	0.33	-0.10	0.50
072	-7.48	0.34	-12.46	0.24	3.65	0.24	0.60	0.30
074	-21.16	0.07	-16.83	0.09	-3.17	0.10	-1.90	0.30
075	-10.44	0.93	-14.53	0.04	3.00	0.04	-2.80	0.30
080	-17.27	0.06	-12.17	0.38	-3.73	0.40	-2.20	0.20
081	-22.33	0.85	-14.10	0.05	-6.02	0.05	-2.20	0.30
082	-8.04	0.39	-12.90	0.19	3.56	0.20	2.50	0.40
083	-31.92	1.28	-42.77	0.35	7.95	0.37	-1.90	0.40
084	-11.98	0.15	-16.45	0.26	3.27	0.27	-0.00	0.20
085	-16.09	0.00	-13.24	0.05	-2.08	0.05	-2.20	0.40
086	-14.90	0.07	-19.26	0.31	3.19	0.32	0.70	0.20
088	-13.50	0.02	-12.96	0.96	-0.40	0.99	-1.90	0.30
090	-10.80	0.10	-16.67	0.30	4.30	0.31	0.80	0.20
092	-22.43	0.66	-25.47	0.25	2.23	0.25	-0.40	0.30

Table 3: *two-species* hydration free energies and logD values. *molecule* denotes the molecule number (fig. **2**, **3** and **4**, *concentration* is the concentration of the "charged" species and the neutral counterpart, ΔG_w is the hydration free energy, ΔG_o the solvation free energy in cyclohexane, logP is the logP for the specific species, pKa^{eff} is the effective pKa (eq. 3.12), logD-pKa is the distribution coefficient computed with eq. 3.16, $two-species$ denotes the *two-species* model logD prediction, $\log D_{exp}$ is the experimental distribution coefficient. σ denotes the standard error for each quantity. All the free energy values are in kcal·mol^{-1}.

molecule	concentration	ΔG_w	σ_w	ΔG_o	σ_o	logP	pKa	pKa^{eff}	logD-pKa	two-species	$\log D_{exp}$	$\sigma_{\log D_{exp}}$
010	0.68	−14.10	0.19	−13.05	0.28	−0.77	4.60	5.24	−3.57	−2.94	−1.70	0.40
010 charged	99.32	−79.57	0.20	−52.96	0.32	−19.49						
011	0.05	−11.39	0.09	−13.18	0.43	1.32	3.31	4.10	−2.78	−1.99	−2.96	0.08
011 charged	99.95	−70.30	0.19	−49.27	0.41	−15.40						
015	0.48	−14.09	0.06	−12.75	0.25	−0.98	3.73	5.08	−4.65	−3.30	−2.20	0.30
015 charged	99.52	−69.54	0.06	−47.25	0.09	−16.32						
017	45.78	−8.93	0.43	−13.98	1.66	3.70	7.30	7.33	3.34	3.36	2.50	0.30
017 charged	54.22	−63.54	0.54	−47.28	0.74	−11.91						
026	0.31	−13.35	0.00	−10.92	0.19	−1.78	4.90	4.89	−4.28	−4.29	−2.60	0.10
026 charged	99.69	−79.68	0.05	−50.74	0.71	−21.19						
037	9.87	−16.50	0.13	−10.84	0.25	−4.15	8.36	8.36	−5.15	−5.15	−1.50	0.10
037 charged	90.13	−110.99	0.52	−42.56	1.01	−50.11						
060	0.83	−17.03	0.14	−12.82	0.08	−3.08	1.36	5.32	−9.12	−5.16	−3.90	0.20
060 charged	99.17	−78.51	0.04	−52.11	0.04	−19.34						
061	23.25	−8.43	0.27	−11.29	0.01	2.10	7.92	7.92	1.47	1.47	−1.45	0.09
061 charged	76.75	−62.49	0.00	−38.08	0.04	−17.88						
063	36.58	−16.89	0.34	−12.55	0.21	−3.18	9.90	7.64	−5.68	−3.62	−3.00	0.40
063 charged	63.42	−70.08	0.20	−39.12	2.35	−22.67						
065	55.73	−24.52	0.34	−29.11	0.19	3.36	7.30	7.30	3.10	3.10	0.70	0.20
065 charged	44.27	−61.08	0.88	−52.47	0.88	−6.30						
067	0.54	−8.98	0.18	−14.10	0.03	3.75	9.67	9.67	1.48	1.48	−1.30	0.30
067 charged	99.46	−52.83	0.05	−34.40	0.37	−13.50						
069	66.16	−13.14	0.84	−17.60	0.49	3.27	7.11	7.11	3.09	3.09	−1.30	0.30
069 charged	33.84	−53.61	0.35	−41.17	1.11	−9.11						
070	0.43	−6.24	0.13	−13.37	0.00	5.22	9.76	9.77	2.86	2.86	1.60	0.30
070 charged	99.57	−51.37	0.47	−32.91	0.16	−13.52						
072	3.30	−7.48	0.34	−12.46	0.24	3.65	8.87	8.87	2.17	2.17	0.60	0.30
072 charged	96.70	−52.38	0.20	−33.37	0.01	−13.91						
075	0.54	−10.44	0.93	−14.53	0.04	3.00	9.60	9.67	0.80	0.73	−2.80	0.30
075 charged	99.46	−54.26	0.35	−34.66	0.60	−14.35						
081	0.54	−22.33	0.85	−14.10	0.05	−6.02	9.60	9.67	−8.23	−8.29	−2.20	0.30
081 charged	99.46	−63.27	0.20	−35.91	0.46	−20.03						
082	4.14	−8.04	0.39	−12.90	0.19	3.56	8.76	8.77	2.18	2.18	2.50	0.40
082 charged	95.86	−50.83	0.44	−33.39	0.82	−12.77						
084	18.45	−11.98	0.15	−16.45	0.26	3.27	8.05	8.05	2.53	2.54	0.00	0.20
084 charged	81.55	−48.96	0.20	−36.16	0.25	−9.37						
086	2.22	−14.90	0.07	−19.26	0.31	3.19	9.00	9.04	1.58	1.54	0.70	0.20
086 charged	97.78	−58.93	0.51	−40.54	0.72	−13.47						



Chapter 4

Alchemical free energy protocols for host-guest binding free energy predictions

This chapter is mainly adapted from the paper: "Blinded prediction of host-guest standard free energies of binding in the SAMPL5 challenge" [131]. As part of the SAMPL5 contest, 22 host guest systems were studied and three alchemical free energy models,*model A*, *model B* and *model C*, were evaluated. Model A yields a free energy of binding based on computed free energy changes in complex and solvated phases; model B adds long range dispersion corrections to the previous protocol; model C takes into account standard state correction. Results for each host-guest systems based on models A, B and C are analyzed and compared with other SAMPL5 submissions.

4.1 Introduction

Accurate and reliable predictions of binding affinity for drug-like molecules, interacting with target-protein, is one of the major objective for computer-aided drug design (CADD). Robust and consistent predictions could enhance and accelerate the early stage drug discovery, reducing costs and time of experimental assays on candidate ligands [182]. Unfortunately, despite sev-

eral efforts to improve the reliability of protein-ligand binding predictions, CADD methodologies still suffer from systematic errors, such as sampling error [183], inaccurate potential energy functions [184] and presence of finite size effects [81].

Before proceeding with accurate protein-ligand binding affinity calculations, alchemical free energy calculations are tested and validated with host-guest case studies. Host-guest systems can be regarded as toy model for protein-ligand binding, where there are fewer degrees of freedom with respect to protein-ligand case [185, 186, 187, 188, 173], which could allow a systematic study of alchemical free energy parameters.

For these reasons host-guest simulations were included in the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) [189, 190, 191] project, a community-wide blinded predictions challenge to evaluate the reliability of computational methods used by the computational chemistry community.

In 2015 we took part in the fifth edition of SAMPL (SAMPL5). In this context the binding free energies for a dataset of 22 guests and three different host molecules were requested. The hosts were two octa-acid hosts, OA and TEMOA, and a cucurbituril clip, CBC, as shown in Fig. **1**. Specifically, OA [192] has four flexible propionate side chains bearing two rotatable single bonds each, while, TEMOA contains four additional methyl groups, which alter the shape and depth of the hydrophobic cavity. Both system are rigid basket shaped hosts. CBC [193] is an acyclic molecular clip, chemically related to the cucurbiturils. It is made of two glycoluril units, each with four sulfonate solubilizing groups. Unlike OA and TEMOA, CBC is a more flexible host, which has shown a high binding affinity for ferrocene, adamantane and bicyclooctane guests [194].

As regards the free energy techniques adopted for previous SAMPL editions, various computational methods have been tested to predict host-guest binding free energy, ranging from quantum chemical [195, 196, 197] to molecular mechanical approaches [198, 199]. Molecular dynamics (MD) or Monte Carlo (MC) simulations are frequently carried out to estimate the ensemble averages that yield standard free energies of binding and different approximations lead to various ways of estimating free energies of binding from molecular simulations trajectories, e.g. free energy perturbations (FEP) [25], finite difference thermodynamic integration (FDTI) [65], or end-states only variants such as MM-PBSA [200].

To predict standard free energies of binding for these 22 host-guest complexes a trajectory based alchemical free energy approach was employed. The aim of this study was to examine robust alchemical free energy protocols which make use of the general Amber forcefield (GAFF) [117]. The

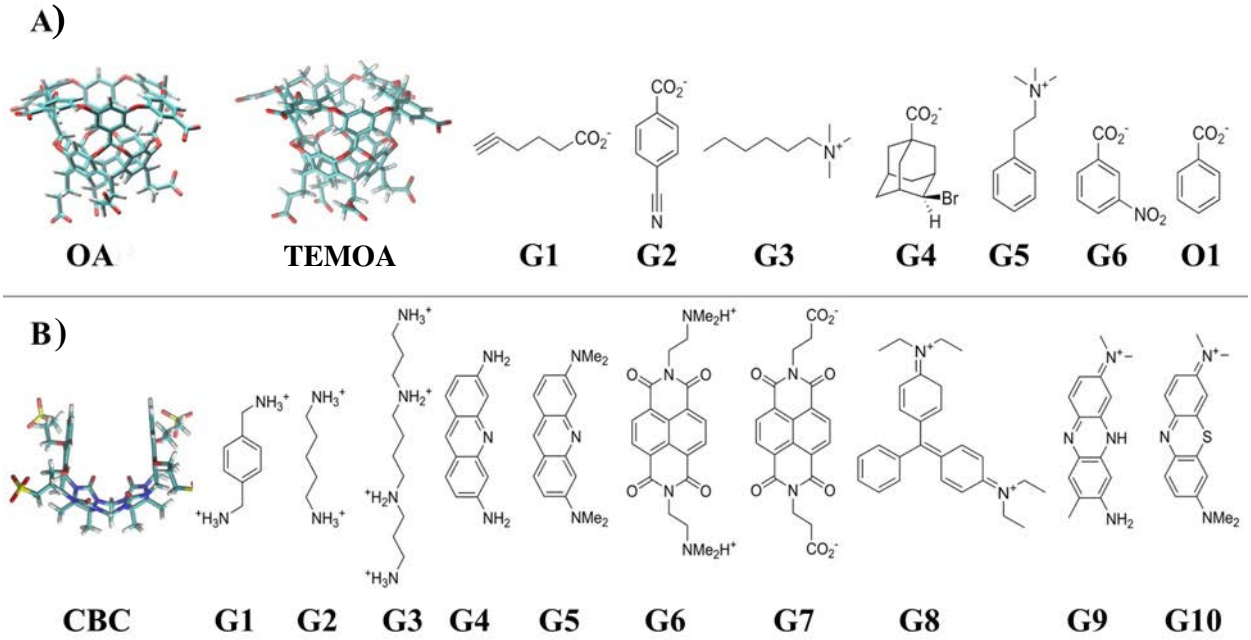


Figure 1: A): Octa-acid systems OA and TEMOA and their respective six guests, plus SAMPL 4 guest O1. B): Cucurbituril clip CBC and its ten guests.

validation is carried out by comparing computed binding energies with experimental data, for three different variants of a double annihilation methodology for binding free energy predictions. Additionally, these three models were compared to other SAMPL5 submissions. The choice of GAFF was motived by recent successes obtained by Mishra [201] and Aldeghi [26]. GAFF is not guaranteed to be the most accurate force-field, but it remains an attractive choice due to the ease of parameters generation, especially given the limited time available in SAMPL between datasets release and deadline for predictions submissions.

4.2 Theory: binding free energy models

As already mentioned in chapter 1 section 1.5, the free energy of binding ΔG_{bind} can be computed from alchemical free energy calculations as:

$$\Delta G_{bind} = -k_B T \ln \frac{Z^{HG}}{Z^G Z^H}, \tag{4.1}$$

where k_B is the Boltzmann constant, T the temperature, Z^{HG}, Z^G and Z^H are the isothermal-isobaric configurational integral of the complex, the guest

and host respectively, while the solvent component has been taken out from eq. 4.1 for the sake of simplicity. To evaluate the expression of eq. 4.1, the double annihilation technique [66, 67] can be used by means of a thermodynamic cycle, as depicted in fig. **2**. First the guest's partial charges are turned off both in water and in the host-guest-complex phase (discharging step), giving the discharging free energy change ΔG^{solv}_{elec} and ΔG^{host}_{elec} respectively. Secondly, the guest is fully decoupled from the solvent or host, switching off the van der Waals terms (vanishing step), ΔG^{solv}_{vdW} and ΔG^{host}_{vdW}. The discharging and vanishing steps are usually performed with a series of intermediate simulations that depend on a coupling parameter $\lambda \in [0, 1]$. Closure of the thermodynamic cycle in Fig 2 shows that in the double annihilation technique the free energy of binding ΔG_{bind} is given as:[62]

$$\Delta G_{bind} = (\Delta G^{solv}_{elec} + \Delta G^{solv}_{vdW}) - (\Delta G^{host}_{elec} + \Delta G^{host}_{vdW}). \tag{4.2}$$

In the above calculations an empirical distance-restraint term is added to the potential energy function. This is done to prevent the non-interacting guest from drifting out of the host cavity, which leads to slow convergence of free energy changes computed via molecular simulations. A flat-bottom restraining potential is used between one atom of the guest, chosen to be the one closest to the center of mass, and four equivalent carbon atoms of the host. The restraint potential for atom j of the guest is based on work presented in [173] and takes the following form:

$$U^{restr}(d_{ij}) = \sum_{i=1}^{N_{host}} \begin{cases} 0 & if|d_{ij} - R_{ij}| \leq D_{ij} \\ \kappa_{ij} \left(|d_{ij} - R_{ij}| - D_{ij}\right)^2 & if|d_{ij} - R_{ij}| > D_{ij} \end{cases}, \tag{4.3}$$

where $U^{restr}(d_{ij})$ is the potential energy of the restraint as a function of the distance between the i-th guest atom and j host atoms $d_{ij} = ||\mathbf{r}_j - \mathbf{r}_i||$ where $|| \circ ||$ denotes a 2-norm, D_{ij} is the restraint deviation tolerance, R_{ij} the reference distance between host and guest atom, κ_{ij} the restraint force constant, and N_{host} the number of host atoms that contribute to the restraint. Free energies of binding computed according to eq. 4.2, will be referred to as *model A*.

Model A neglects the contribution of long range dispersions interactions, since the potential energy function is evaluated with non-bonded cutoffs. Following work from Shirts et al. [130], it is possible to introduce a long range dispersion correction term to the free energy of binding as a post processing step of the simulation trajectories. This leads to a corrected free energy of

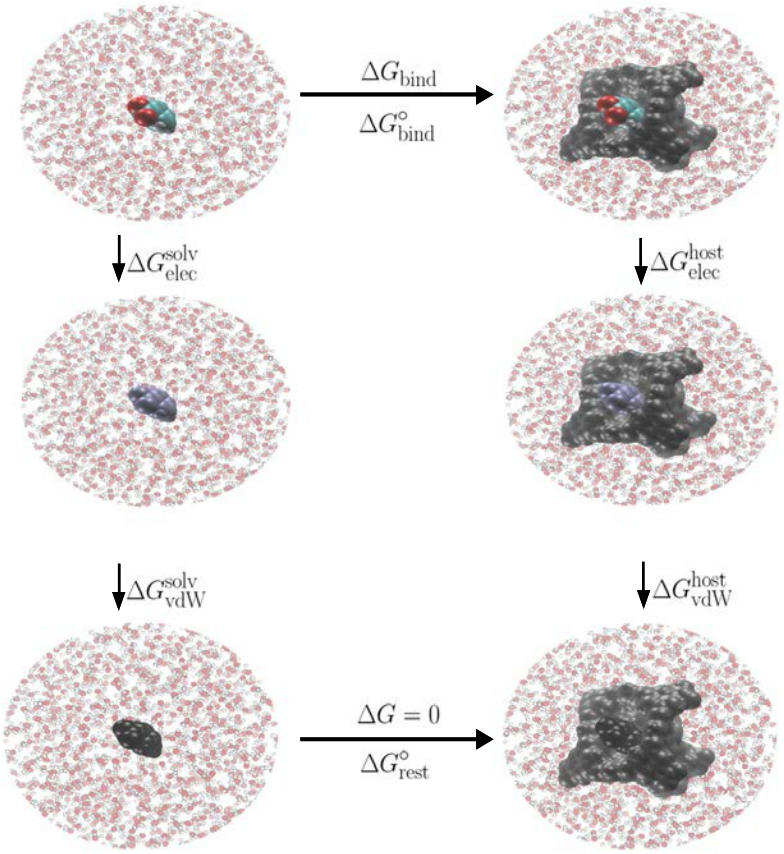


Figure 2: Thermodynamic cycle for free energy of binding calculation. First, a fully interacting ligand is simulated in water phase (top left), then charges and Lennard-Jones terms are switched off sequentially, resulting in a fully non interacting ligand in water (bottom left). The same transformation is performed in the complex, passing from a fully interacting ligand (top right) to a dummy ligand (bottom right). The bottom middle step is concerned with the evaluation of a free energy change associated with introduction of a host-guest restraining potential in the complex simulations. This term is neglected in *model A* and *model B*, and numerically evaluated with respect to standard state conditions in *model C*. Consequently *model A* and *model B* yield a free energy of binding ΔG_{bind}, whereas *model C* yields a standard free energy of binding ΔG^{0}_{bind}.

binding $\Delta G_{bind,LJRC}$, or *model B*, given by:

$$\begin{aligned}\Delta G_{bind,LJLRC} = &(\Delta G^{solv}_{elec} + \Delta G^{solv}_{vdW}) \\ &- (\Delta G^{host}_{elec} + \Delta G^{host}_{vdW}) \\ &+ (\Delta G^{host}_{LJLRC} - \Delta G^{solv}_{LJLRC}).\end{aligned} \tag{4.4}$$

The Lennard-Jones dispersion correction ΔG^{X}_{LJLRC} can be computed making use of the Zwanzig relation [23] in the following way:

$$\Delta G^{X}_{LJLRC} = -\beta^{-1} \ln\langle \exp[-\beta(U_{LJ,long}(\mathbf{r}) - U_{LJ,sim}(\mathbf{r})]\rangle_X + U_{LJ,ana}, \tag{4.5}$$

where X= *host* or *solv*, $\beta = \frac{1}{k_B T}$, $U_{LJ,long}$ is the Lennard-Jones energy computed for a "long" cutoff and $U_{LJ,ana}$ is an analytical correction for extending the long cutoff to infinity. The long range correction $U_{LJ,long}$ is estimated in a post processing step of the vanishing step $\lambda = 0$ and $\lambda = 1$ trajectories, by extending the domain of the typical Lennard-Jones cutoff radius in the simulations to a radius which covers the entire box. To define this long cutoff, the minimum box length in all directions $\min(L_x, L_y, L_z)$ in the input coordinates is calculated, and the new cutoff radius is set to $r_c, long = 0.95 \min(L_x, L_y, L_z)/2$ to allow for some fluctuations in box size during the simulation. This allows an averaging over the whole trajectory of the additional contribution of the long range potential $U_{LJ,long}$, with respect to the simulated Lennard-Jones term $U_{LJ,sim}$. This correction, however, does not account for an infinitely large box size giving rise to an analytical correction over an infinite domain, which is an additive constant given below:

$$U_{LJ,ana} = 8\pi\rho \sum_{i=1}^{N} \epsilon_i \sigma_i^3 \left[\frac{1}{9} \left(\frac{\sigma_i}{R_C} \right)^9 - \frac{1}{3} \left(\frac{\sigma_i}{R_C} \right)^3 \right] \tag{4.6}$$

where N is the number of solute's atoms, ρ is the solvent density in mol·Å^{-3}, assuming an isotropic solvent, $\epsilon_i = \sqrt{\epsilon_i \epsilon_{solv}}$ is the resulting Lennard Jones well depth for the i-th solute's atom and the solvent ϵ_{solv} value, in kcal·mol^{-1} and similarly $\sigma_i = 0.5 \times (\sigma_i + \sigma_{solv})$ is the Lennard-Jones distance, in Å, for the i-th solute's atom and the solvent σ_{solv} value. Lennard-Jones parameters for the solvent are those of the oxygen atom of the TIP3P water model [120]. It is implicitly assumed that the radial distribution function $g(\mathbf{r})$=1 for $R_C > r$.

Both *model A* and *model B* lack a well defined reference state in their definition of the free energy change upon binding of the guest molecules. Therefore a third model is proposed to enable a standard state definition. For this purpose the standard state correction is added to the free energy of

binding given by eq. 4.4. The standard free energy of binding is given by:

$$\begin{aligned}\Delta G^{\circ}_{bind} = &(\Delta G^{solv}_{elec} + \Delta G^{solv}_{vdW}) \\ &- (\Delta G^{host}_{elec} + \Delta G^{host}_{vdW}) \\ &+ (\Delta G^{host}_{LJLRC} + \Delta G^{solv}_{LJLRC}) + \Delta G^{\circ}_{restr},\end{aligned} \tag{4.7}$$

where ΔG°_{restr} accounts for the introduced flat-bottom restraint. Considering the cycle in fig. **2**, the restraint free energy change can be computed as:

$$\Delta G^{\circ}_{restr} = -k_B T \ln \left(\frac{Z_{H \circ\circ G^{ideal},solv}}{Z_{H,solv} Z_{G_{gas}}} \right), \tag{4.8}$$

where $Z_{H \circ\circ G^{ideal},solv}$ is the configuration integral for the restrained decoupled guest bound to the host, $Z_{H,solv}$ is the configuration integral for the solvated host and $Z_{G_{gas}}$ is the configuration integral for the guest in an ideal thermodynamic state. Assuming that the restraint potential is independent from the solvent and host degrees of freedom, the contribution of the all degrees of freedom but those of the guest cancel out and eq.4.8 simplifies to:

$$\Delta G^{\circ}_{restr} = -k_B T \ln \left(\frac{Z_{\circ\circ G^{ideal},solv}}{Z_{G_{gas}}} \right), \tag{4.9}$$

where $Z_{\circ\circ G^{ideal},solv}$ is the configuration integral for the decoupled guest. Because the guest has no intermolecular interactions in both thermodynamic states defined in eq. 4.9, and because the restraint does not hinder rotational motions, internal and rotational contributions to the configuration integrals cancel out and the only term left is the translational contribution to the configuration integral. For $Z_{G_{gas}}$ a standard volume V° =1660 Å^3·molecule^{-1} (1 M)is used. Hence, eq. 4.9 simplifies to:

$$\Delta G^{\circ}_{restr} = -k_B T \ln \left(\frac{V^{restr}}{V^{\circ}} \right), \tag{4.10}$$

where the restraint volume V^{restr} is given by:

$$V^{restr} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} dx_j \, dy_j \, dz_j \exp(-\beta U(d_{i1}, \ldots, d_{iN_{host}})). \tag{4.11}$$

V^{restr} can be calculated by numerically integrating eq. 4.11. In this contex, initially the coordinates of the host-guest complex in the generated trajectory at $\lambda = 1$ of the vanishing step are aligned onto the first frame of the trajectory. Then, the average coordinates of each of the four host atoms

used for the restraint were computed. Next, a grid spacing and an integration domain are defined. The grid spacing is set to 0.1 Å and the integration domain is defined by the rectangular cuboid that is given by the minimum and maximum coordinates of the four defined host atoms with an additional buffer around the bounding domain of ± 5 Å. Numerical integration is then performed via multidimensional trapezoidal rule.

4.2.1 Host-Guest simulation set-up

Guests input files were created using `tleap`. Force field parameters and molecules' library files were extracted from the provided topologies using the python module `parmed` [73]. Thus, guests were solvated in a rectangular cuboid box of TIP3P water molecule [120], with a minimum distance between the solute and the box of 12 Å. Ions were added to neutralize the overall charge of the box. Then, solutes were energy minimized with a maximum of 100 steps of steepest descent gradient. Next, solutes were restrained with a force constant of 10 kcal·mol^{-1}·Å^{-2}, to allow water to equilibrate in an NVT ensemble at 298 K. Finally, a NPT equilibration for 200 ps at 1 atm was done using Amber MD module Sander [73], followed by a 2 ns NPT simulation run with SOMD (rev. 2015.0.0) [68, 69], to reach a final water density of about 1 g/cm^3. The final coordinate files were retrieved with CPPTRAJ [169]. The same protocol for preparation and equilibration was used for the host-guest system files.
Additionally a reference system, OA-O1 was taken from the SAMPL4 [202] challenge and set-up from scratch. Guest O1 was obtained from the modification of compound G6 using Maestro (v.10.1.012, rel 2015-1, Schrödinger) [203] and further parametrized using AM1-BCC charges [118] using Antechamber 14 [73]. Complex and water phase systems were created with *tleap*, according to the above protocol.

4.2.2 Alchemical free energy production simulations

For the discharging steps nine equidistant λ windows were selected for the host-guest complex and the guest in water phases. For the vanishing step 12 and 18 equidistant windows were used for octa-acid guests and CBC guests respectively. The reasoning behind these different choices was that the CBC guests were larger and more flexible systems than octa-acids, therefore a denser number of λ windows was deemed necessary to guarantee a good overlap of the potential energy distributions of neighbouring λ windows [204]. All

simulations were run for 8 ns. A velocity-Verlet integrator was used with a time step of 4 fs using a hydrogen mass repartitioning (HMR) scheme [170] and all bonds were constrained. Every simulation was performed in an NPT ensemble and temperature control was achieved with an Andersen Thermostat with a coupling constant of 10 ps^{-1} [52]. Pressure was maintained by a Monte Carlo barostat that attempted isotropic box edge scaling every 100 fs. Periodic boundary conditions were imposed with a 12 Å atom-based cut-off distance for the non-bonded interactions, using a shifted Barker Watts reaction field [49] with dielectric constant of 78.3. In the host-guest complex the guest molecules were restrained according to eq. 4.3. The parameters were $R_{ij} = 5$ Å, $D_{ij} = 2$Å and $\kappa_{ij} = 10$ kcal·mol^{-1}· $Å^{-2}$.

4.2.3 Estimation of free energy changes for models A, B and C

Individual free energy contributions from the discharging and vanishing steps were estimated by using multistate Bennett's acceptance ratio (MBAR) [171]. To estimate accuracy and consistency of the computed binding free energy from eq. 4.2, each simulation was repeated twice from the same initial coordinates, using different initial assignments of velocities drawn from the Maxwell-Boltzmann distribution. Final binding free energies are reported as the average of both runs and statistical uncertainties were calculated according:

$$err(\Delta G) = \frac{\sigma}{\sqrt{n}}, \tag{4.12}$$

where σ is the standard deviation of both runs and n=2 unless otherwise mentioned.

The computed binding free energies with each model are then compared to experimental values considering three different measures: the determination coefficient R^2, mean unsigned error (MUE) and Kendall τ. To gain insight into the distribution of the three different measures a bootstrapping scheme is used in which each computational free energy point is considered to be a normal distribution with its mean given by the computed free energy and σ the associated computed error. One million samples are then drawn from the artificial normal distributions for each data point and correlated with the experimental values, giving rise to a distribution of R^2, MUE and Kendall τ. The resulting distributions are typically not symmetric around the mean and uncertainties in the dataset metrics are reported with a 95% confidence interval.

4.2.4 Host-guest experimental setup

The hosts octa-acid (OA) and the tetra-endo-methyl octa-acid (TEMOA) fig. **1**, are water soluble deep-cavity cavitands, containing a hydrophobic binding pocket and an exterior coating of eight carboxylic acid groups arranged in a square anti-prism array [205]. Both octa-acid systems have a well-defined binding pockets formed by eight aromatic rings, whose hydrogen atoms are projected into the midsection of the cavity and act as weak hydrogen bond donors [206]. Experimental binding affinities were measured by Dr. Bruce Gibb, from Gibb laboratory, with NMR for guests G1, G2, G3 and G5, while for G4 and G6 ITC was used. All the NMR spectra were recorded at 25°C, with the exception of TEMOA-G4 which was carried out at 5°C. The NMR titrations were done in 10 mM phosphate buffer (pH=11.3) and the corresponding association constant K_a were estimated from the standard titration and curve fitting process. ITC measurements for OA-G4, OA-G6 and TEMOA-G6, were performed in 50 mM phosphate buffer (pH=11.5) at 25°C. From curve fitting the binding isotherms were processed and a binding free energy estimated. The high basic pH = 11.3 - 11.5 was chosen in order to have a good host solubility.

CBC is composed of a central methylene-bridged glycoluril dimer, two dialkoxnaphthalene sidewall and four sodium sulfonate solubilizing groups, fig. **1**. By virtue of the conformational constraints of the polycyclic ring system, comprising seven membered rings, CBC is constrained to adopt a C-shape with nearly parallel aromatic walls, separted by $\sim$7Å [207]. In this case, NMR and UV/VIS and fluorescence spectroscopy titrations were carried on by the Isaacs laboratory. In particular CBC-G1, CBC-G2 and CBC-G3 were studied with NMR, in a sodium phosphate buffer of 20 mM. CBC-G4, CBC-G5, CBC-G6, CBC-G7, CBC-G8 and CBC-G9 were measured with UV/VIS, since the association constant K_a exceeded the ^{1}H NMR range ($K_a \leq 10^4 M^{-1}$). CBC-G10 was instead measured with fluorenscence spectroscopy which provide a reliable measure of the affinity of the 1-methylene blue (G10) complex. Additional tests were made with UV/VIS titration, which displayed a well defined 1-methylene blue complex of 1:1 stoichiometry.

All the experimental standard data, standard free energy of binding $\Delta G°$, its uncertainty σ and the association constant K_a for each host-guest system, are reported in table **1**.

Table 1: Experimental measurements for the three host-guest systems, OA, TEMOA and CBClip. For each host the ligand name, mol, is reported along with the technique used, the buffer concentration (mM), the standard free energy of binding ΔG°_{bind} in kcal·mol^{-1} and the uncertainty σ, and the association constant K_a in M^{-1}.

		OA		
mol	Technique	buffer	$\Delta G^{\circ}_{bind} \pm \sigma$	K_a
G1	NMR	10	-5.39±0.01	$5.0\pm0.1\times10^{3}$
G2	NMR	10	-4.73±0.01	$1.3\pm0.1\times10^{3}$
G3	NMR	10	-4.49±0.01	$5.1\pm0.1\times10^{3}$
G4	ITC	50	-9.36±0.01	$7.4\pm0.1\times10^{6}$
G5	NMR	10	-4.50±0.01	$1.9\pm0.1\times10^{3}$
G6	ITC	50	-5.33±0.01	$8.1\pm0.1\times10^{3}$

		TEMOA		
mol	Technique	buffer	$\Delta G^{\circ}_{bind} \pm \sigma$	K_a
G1	NMR	10	-5.47±0.01	$6.9\pm0.7\times10^{3}$
G2	NMR	10	-5.25±0.01	$4.9\pm0.3\times10^{3}$
G3	NMR	10	-5.73±0.01	$2.3\pm0.6\times10^{4}$
G4	NMR	10	-2.40±0.01	$5.5\pm0.2\times10$
G5	NMR	10	-3.90±0.01	$7.3\pm0.3\times10^{2}$
G6	ITC	50	-4.52±0.01	$2.0\pm0.1\times10^{3}$

		CBC		
mol	Technique	buffer	$\Delta G^{\circ}_{bind} \pm \sigma$	K_a
G1	NMR	20	-5.84±0.01	$1.9\pm0.1\times10^{4}$
G2	NMR	20	-2.52±0.01	70±8.0
G3	NMR	20	-4.02±0.01	$8.8\pm0.5\times10^{2}$
G4	UV/VIS	20	-7.25±0.01	$2.0\pm0.1\times10^{5}$
G5	UV/VIS	20	-8.54±0.01	$1.8\pm0.2\times10^{6}$
G6	UV/VIS	20	-8.68±0.01	$2.2\pm0.2\times10^{6}$
G7	UV/VIS	20	-5.18±0.01	$6.2\pm0.2\times10^{3}$
G8	UV/VIS	20	-6.18±0.01	$3.3\pm0.2\times10^{4}$
G9	UV/VIS	20	-7.40±0.01	$2.6\pm0.1\times10^{5}$
G10	Fluorescence	20	-10.38±0.1	$3.9\pm0.2\times10^{7}$

4.2.5 Statistical comparison with other submissions

A statistical comparison based on correlation coefficient, R^2, and root-mean-square error, RMSE, was made between SOMD and other participants predictions for OA, TEMOA and CBC. The statistical analysis was carried on by Yin *et al.* [75]. To determine how each calculation method perform an uncertainty in RMSE and R^2 was computed with bootstrap resampling with replacement, as described above.

Table 2: Comparison of the binding free energy ΔG_{bind} of *models A, B* and standard binding free energy of ΔG°_{bind} of *model C* with respect to experimental data, *Experimental*, for the SAMPL4 OA-O1 complex.

OA-G1			
model A	*model B*	*model C*	ΔG°_{bind}
-6.1±0.5	-6.1±0.5	-4.4±0.5	-3.7

4.3 Results

4.3.1 Host-Guest binding free energy predictions

To test the precision and accuracy of the *models A,B, C*, the free energy of binding of guest O1 to host OA, adopted from SAMPL4, was retrospectively predicted. Table **1** compares the results with experimental data [202]. Both *models A,B* yield a similar free energy of binding ΔG_{bind} = -6.1 ± 0.5 kcal·mol^{-1}. This is because the long-range corrections for Lennard-Jones interactions implemented in *model B* produce a negligible correction term of 0.03 kcal·mol^{-1}. By contrast, the addition of a standard state correction in *model C* leads to standard free energy of binding of ΔG°_{bind}=-4.4 ± 0.5 kcal·mol^{-1} which is in good agreement with the experimental data of ΔG°_{bind} = -3.7 kcal·mol^{-1}.

Thus, blinded predictions were performed for each SAMPL5 host-guest. Fig. **3** contrasts the predictive power of the different models against the experimental data, released after submission of the predictions, while fig. **4** compares the statistical significancy of all the three modes, in terms of R^2, MUE and Kendall τ. For the complete dataset, all three models yield a similar R^2 value of ca. 0.64< 0.70< 0.75. Furthermore, *Models A,B* have a similar MUE 4.3< 4.5< 4.7 kcal·mol^{-1}, whereas *model C* is statistically more accurate, with a MUE of 3.20< 3.39< 3.57 kcal·mol^{-1}. The accuracy of the predictions for the three different hosts was also considered individually and summarised in table **3**. As judged by the MUE measures, the models perform better across the octa-acid systems than for CBC. In particular, *model C* gives the best predictions compared to *A* and *B* for octa-acid systems, with a MUE of 1.79< 2.16< 2.52 kcal·mol^{-1} and 1.42< 1.72< 2.02 kcal·mol^{-1} for OA and TEMOA respectively. R^2 is on average slightly higher for both octa-acid hosts (R^2: 0.77< 0.87< 0.93. *model C* for OA and R^2: 0.52< 0.74< 0.93, *model C* for TEMOA) than CBC (R^2=0.69< 0.76< 0.82) for models

A, B and C, but the trend is not strong given statistical uncertainties.

Next, attention focused on the guests whose predictions exhibited the largest discrepancy with experimental data. For octa acid systems, guest G5 is the greatest outlier, with an underestimation in ΔG°_{bind} of about -3.2 kcal·mol^{-1} and -2.8 kcal·mol^{-1} for OA and TEMOA respectively. A similar trend was found also by other participants, which make use of GAFF and MD approaches [75, 208]. Yin *et al.* found a flipping motion in the four benzoic acid around the OA cavity entrance toward the guest, which characterize part of the all OA conformations in complex phase,. This, however, was not found analyzing the $\lambda = 0.0$ discharging step SOMD trajectories.

Additionally, the validity of the G5 binding mode was evaluated. For this purpose, G5 was rotated by approximately 180° degrees about its centre of mass, such that the amine group pointed towards the bottom of the host cavities.Then, calculations were repeated using these new coordinates after solvent equilibration. Binding free energy predictions from *model C* obtained for this alternative binding mode were poor: ΔG°_{bind}=1.8±0.1 kcal·mol^{-1} and 16.7±0.1 kcal·mol^{-1} for OA and TEMOA respectively, so that the original binding mode is more likely. Finally, it is worth mentioning that Sderhjelm *et al.* [208] have achieved a good agreement with experimental data for G5 molecules, but only after analysis of 70 ns of molecular dynamics and adding an empirical correction for the presence of finite size artefacts. This suggests that the presence of finite size artefacts and sampling could have biased G5 simulations results.

An intersting guest is G4, which shows a high binding affinity for OA but very low for TEMOA, regardless the two host molecules are quite similar. Predictions reflect the experimental trend, recognizing G4 as a nM compound for OA and a mM compound for TEMOA. Furthermore, as revealed in the experimental study [205], G4 is freer to tumble in the OA host cavity. In this way, the G4 molecule may explore the binding cavity of OA, till the bromine is able to form halogen-hydrogen interactions at the midsection of the OA host molecule, where benzoic protons are present. Differently, for TEMOA-G4 systems, the presence of methyl branches at the rim of the host cavity may prevent the guest to tumble freely with an entropic penalty for the guest.

For the CBC host, the MUE is about 4.8$<$ 5.1$<$ 5.4 kcal·mol^{-1} for all models, despite of a reasonable correlation coefficient R^2= 0.65$<$ 0.76$<$ 0.81. In particular, *model C* performs better than A and B, but large errors are present for a series of guests. As depicted in table **3**, the guests G2 and G3 are predicted to bind substantially worse than what is observed in experiments. As a matter of fact, these two molecules are made of linear flexible alkyl chains, containing several positively charged ammonium groups

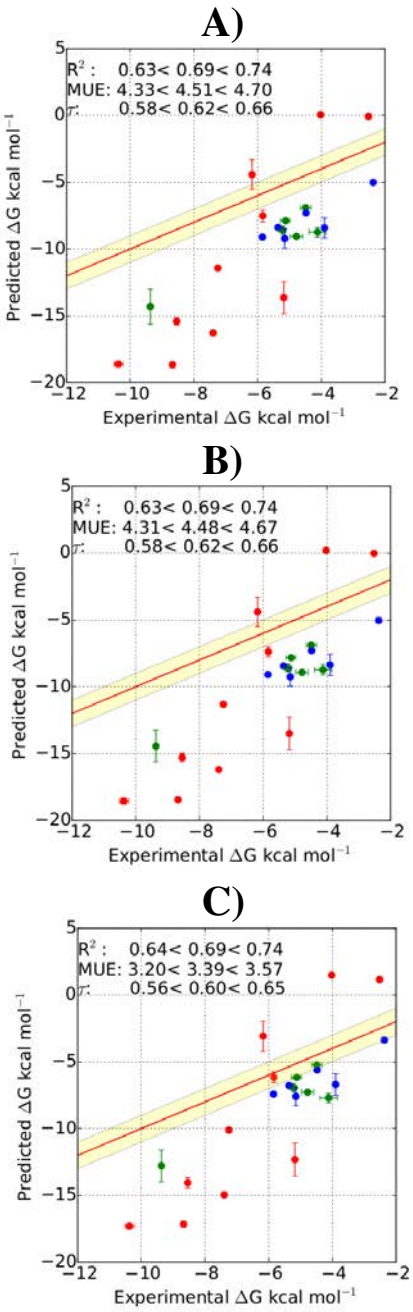


Figure 3: Comparison of the binding free energy ΔG_{bind} of A): *model A*, B): *model B* and C): *model C* for all the host-guest systems with respect to experimental data. The red line indicates ideal correlation between experiments and computational results and the yellow shaded region gives a binding free energy error bound of 1 kcal·mol−1. OA systems are colored in blue, TEMOA in green and CBC in red. Error bars denote ± eq 4.12.

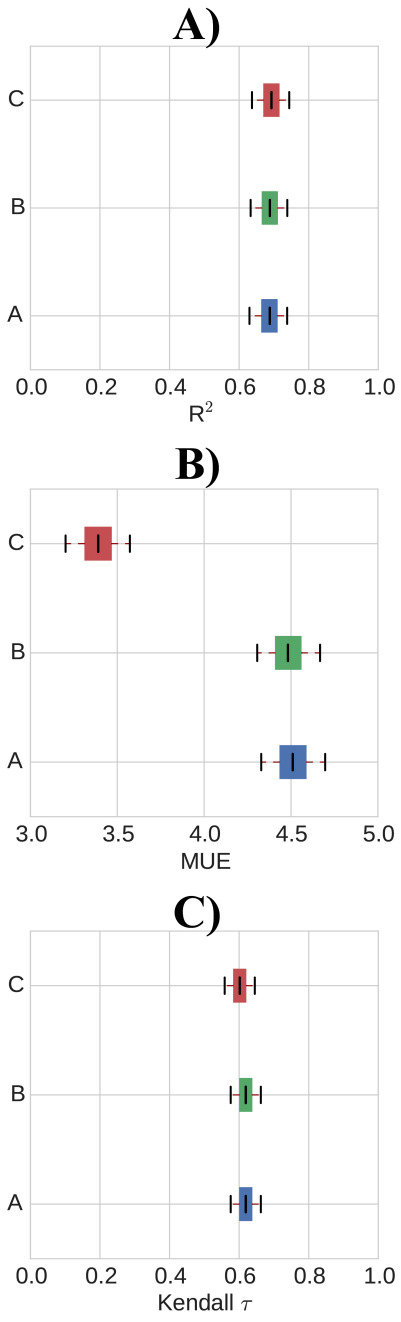


Figure 4: Statistical significance for *model A*, *model B*, *model C* for A): R^2; B): MUE and C): Kendall τ, for all the dataset of host-guest systems. The overlap between notches points out a statistical similarity for *model A* and *model B*, while *model C* is significantly better for R^2 and MUE with respect to these two previous models .

Table 3: Calculated binding free energy, in kcal·mol^{-1},experimental binding free energy ΔG°_{exp} determination coefficient R^2, mean unsigned error (MUE), in kcal·mol^{-1}, *model A*, *B* and *C*. Uncertainties in the calculated individual binding free energies are the standard error of the mean. Dataset metrics are given with a 95% confidence interval.

	OA			
	ΔG_{bind} *A*	ΔG_{bind} *B*	ΔG°_{bind} *C*	ΔG°_{exp}
G1	-8.58±0.31	-8.60±0.35	-6.94±0.30	-5.39±0.01
G2	-6.91±0.22	-6.89±0.15	-5.23±0.15	-4.37±0.01
G3	-9.05±0.19	-8.93±0.18	-7.28±0.14	-4.49±0.01
G4	-14.31±1.32	-14.45±1.18	-12.79±1.18	-9.36±0.01
G5	-8.73±0.36	-8.73±0.37	-7.71±0.37	-4.50±0.01
G6	-7.88±0.10	-7.82±0.10	-6.16±0.10	-5.33±0.01
MUE	3.33< 3.73<4.01	3.37<3.72<4.07	1.79<2.16<2.52	
R^2	0.83<0.91<0.96	0.83<0.91<0.96	0.77<0.87<0.93	

	TEMOA			
	ΔG_{bind} *A*	ΔG_{bind} *B*	ΔG°_{bind} *C*	ΔG°_{exp}
G1	-8.38±0.11	-8.44±0.03	-6.78±0.03	-5.47±0.01
G2	-9.21±0.74	-9.26±0.68	-7.59±0.67	-5.25±0.01
G3	-9.10±0.03	-9.09±0.01	-7.42±0.01	-5.73±0.01
G4	-5.01±0.01	-5.04±0.21	-3.36±0.21	-2.40±0.01
G5	-8.41±0.74	-8.36±0.81	-6.69±0.81	-3.90±0.01
G6	-7.30±0.17	-7.29±0.16	-5.61±0.16	-4.52±0.01
MUE	3.08<3.38<3.67	3.09<3.39<3.69	1.42<1.72<2.02	
R^2	0.51<0.73<0.93	0.52<0.74<0.92	0.52<0.74<0.93	

	CBC			
	ΔG_{bind} *A*	ΔG_{bind} *B*	ΔG°_{bind} *C*	ΔG°_{exp}
G1	-7.52±0.44	-7.37±0.40	-6.16±0.41	-5.84±0.01
G2	-0.08±0.05	-0.02±0.05	1.17±0.06	-2.52±0.01
G3	0.06±0.02	0.21±0.04	1.49±0.03	-4.02±0.01
G4	-11.42±0.15	-11.32±0.19	-10.11±0.19	-7.25±0.01
G5	-15.41±0.28	-15.30±0.34	-14.05±0.37	-8.54±0.01
G6	-18.65±0.23	-18.46±0.20	-17.16±0.23	-8.68±0.01
G7	-13.64±1.21	-13.50±1.22	-12.32±1.23	-5.18±0.01
G8	-4.43±1.10	-4.40±1.09	-3.07±1.11	-6.18±0.01
G9	-16.28±0.08	-16.20±0.05	-14.96±0.05	-7.40±0.01
G10	-18.61±0.21	-18.56±0.22	-17.30±0.23	-10.38±0.01
MUE	5.38<5.65<5.94	5.32<5.60<5.88	4.82<5.12<5.41	
R^2	0.70<0.76<0.81	0.70<0.76<0.82	0.69<0.76<0.82	

($+2e$ for G2, $+4e$ for G3), which likely results in electrostatics artefacts for the binding free energy estimation. By contrast, G4, G5, G6, G7, G9 and G10 are predicted to bind significantly better than what observed in reality. These compounds are made of conjugated aromatic rings. At pH 7.4 they could adopt multi-protonation states, as suggested from pKa calculations [154]. Thus, it is unclear whether the discrepancies are due to force field parametrization and modelling errors or finite-size effects.

4.3.2 Comparison with other participants submissions

As a separate issue, the reproducibility of standard free energies of binding was evaluated by comparing the results from *model C* with those reported by the Gilson lab (UCSD) for the octa-acid hosts [75]. The same input files were used, but the free energy calculations were performed with the *pmemd.cuda* program from AMBER 14 [209], and a different potential of mean force based 'attach-pull-release' (APR) methodology. [210] Figure **5** shows that a good agreement is observed between both OA and TEMOA hosts, with a mean unsigned differences of about 0.4 kcal·mol^{-1} in the former case and 0.6 kcal·mol^{-1} for the latter. At first glance this level of variability seems reasonable given the typical statistical uncertainties of each methodology. Nonetheless closer inspection indicates that OA-G5, TEMOA-G5 and TEMOA-G4 show significant discrepancies. Since the *model C* standard free energies of binding were only estimated from two repeats a concern was that the error estimates were not reliable. To test this two additional repeats were performed for these systems. The standard free energies of binding obtained from four repeats of *model C* are: ΔG°_{bind}(SOMD, OA-G5)= -6.9±0.1 , ΔG°_{bind}(SOMD, TEMOA-G4)= -3.4±0.2 , ΔG°_{bind}(SOMD, TEMOA-G5)= -6.5±0.3 kcal·mol^{-1} respectively. The results were statistically identical to those obtained from two repeats (table **1**) for TEMOA-G4 and TEMOA-G5, but not OA-G5. Personal discussions with the Gilson lab prompted additional APR calculations which produced revised values for ΔG°_{bind}(APR, TEMOA-G4)= -4.3±0.3 kcal·mol^{-1}, and ΔG°_{bind}(APR, OA-G5)= -4.5±0.5 kcal·mol^{-1}. Additionally, fig. **6** compares *model A*, *model B* and *model C* with other SAMPL5 submissions. Overall, explicit solvent based methods show a greater reliability over electronic structures schemes. In particular, APR attains on average a correlation coefficient $R^2 \sim 0.8\pm0.2$ for octa acid systems, with a RMSE $\sim$ 0.8±0.2 kcal·mol^{-1} for OA and 1.6±0.6 kcal·mol^{-1} for TEMOA. Another explicit solvent methodology, employing GAFF parametrization and TIP3P water model was based on metadynamic approach [211, 208],, achieved a good $R^2 \sim 0.7\pm0.3$ for OA along

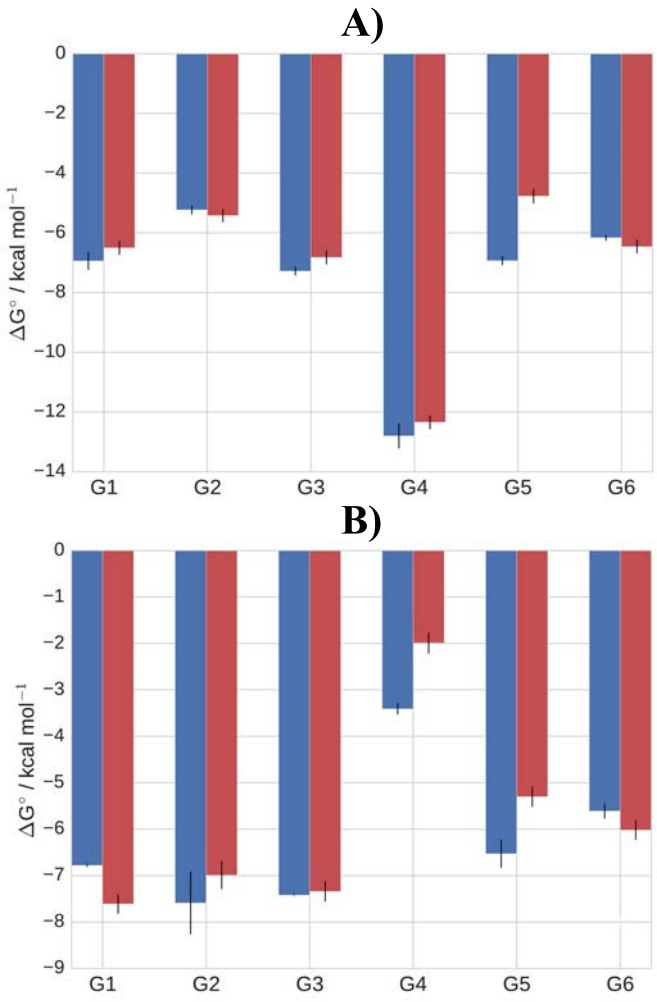


Figure 5: **A** Comparison between standard binding free energies computed with *model C* (blue) and with the attach-pull-release method (red) for **A**) OA and **B**) for TEMOA.

with RMSE$\sim$0.9$\pm$0.3 kcal·mol^{-1}. Due to the narrow range of experimental results in TEMOA, metadynamics, as APR, show a performance deterioration, with R^2 $\sim$0.3$\pm$0.3 and RMSE$\sim$1.6$\pm$0.5. Finally, fair results were reached by TI/HBAR methods, based on CGENFF force field parametrization [212] and MD simulations with TI [24] and HBAR [213] free energy estimations. Overall, the correlation ranges in $0.4<R^2<0.7$ and the $1.3<$ RMSE$<$ 1.9 kcal·mol^{-1} for OA, while for TEMOA the R^2 is about 0.4$\pm$0.3 and 2.0$<$ RMSE$<$ 3.1 kcal·mol^{-1}. An encouraging results was gained by the structure based "Movable type" protocols MovTyp-1 and MovTyp-2 [214]. MovTyp simulates local partition functions using Monte Carlo integration and use a statistical potential, KECSA [215], to calculate each atom pairwise probability distribution. On average MovTyp attains a good $R^2 \sim 0.7\pm0.3$ and RMSE $\sim$ 1.0$\pm$0.2 kcal·mol^{-1}, while for TEMOA R^2 decays to 0.5$\pm$0.3 and RMSE$\sim$3.1$\pm$1.0 kcal·mol^{-1}. On the other side, pure electronic structure based methods, fail to achieve a good correlation and RMSE for octa acid system. The three set of predictions – DLPNO-CCSD(T), DFT/TPSS-n and DFT/TPSS-c [216] – are generated with a dispersion-corrected density functional theory calculation, in conjuction with the COSMO-RS continuum solvation model [217]. DLPNO-CCSD(T) adopts the DLPNO-CCSD(T) level of theory, combined with COSMO-RS. Both DLPNO and DFT methods show a very low correlation for octa acid system, R^2 $\sim$0.4$\pm$0.3, and a high RMSE, which is on average about 5.0 $\pm$1.0 kcal·mol^{-1}.

For the CBC, fewer submission were done. In this case there is no a clear comparison for GAFF TIP3P based methods. however, *model A*, *model B* and *model C* perform the best in terms of R^2 (0.7 $\pm$0.2), while the RMSE is high, ranging from 5.7 to 6.4 kcal·mol^{-1}. Better results in terms of RMSE were retrieved only by the MovTyp-1 algorithm, 3.5$\pm$0.7 kcal·mol^{-1}, and TI-dock methods [218], RMSE$\sim$3.6$\pm$0.7 kcal·mol^{-1}, although a null correlation $R^2 \sim 0.1$ $\pm$0.1 was obtained. The worse performance of free energy methods in CBC simulation could be due to the acyclic aspect of this host, which confers more flexibility, which could slow down the convergence of the calculated free energy changes.

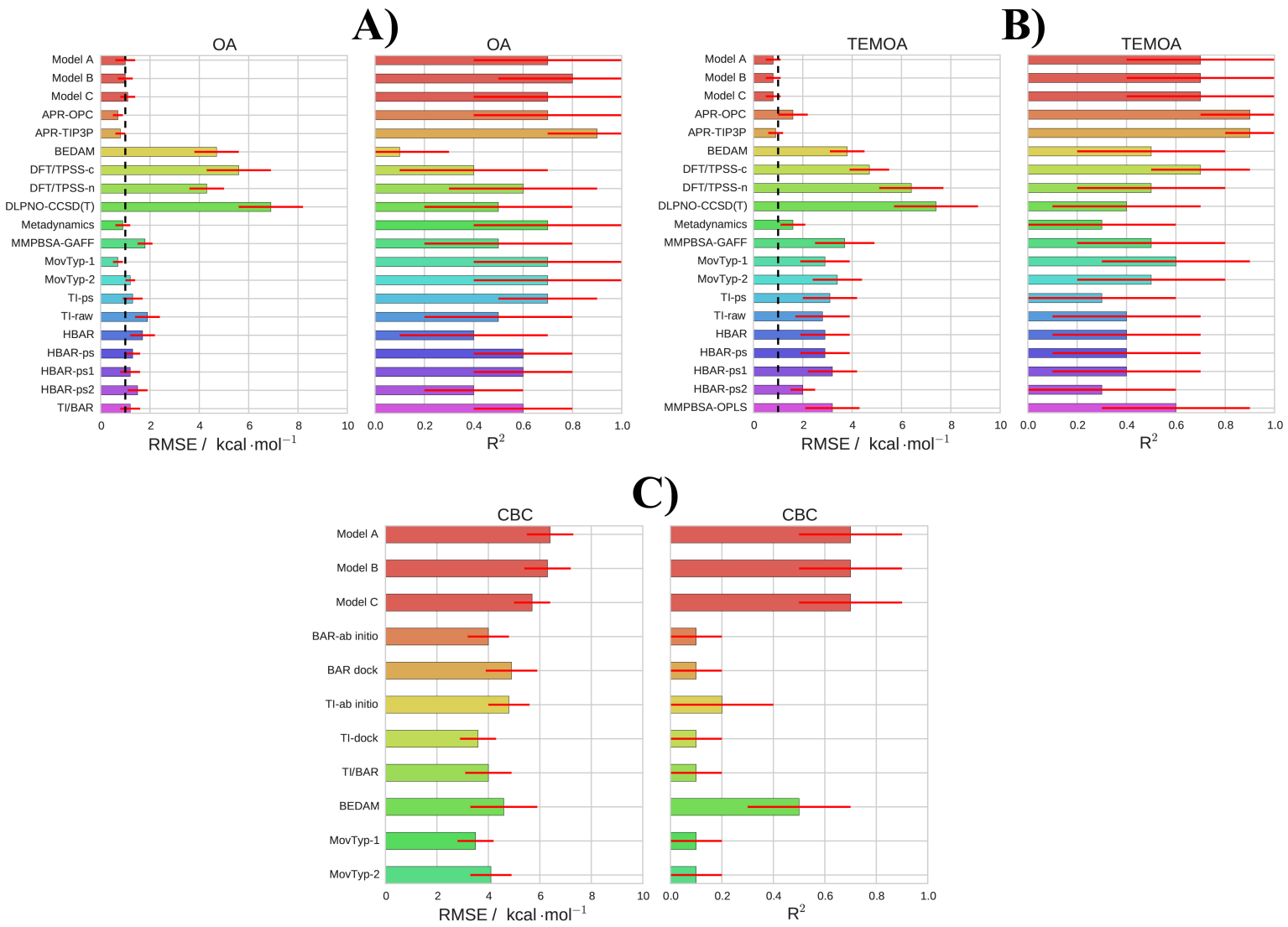
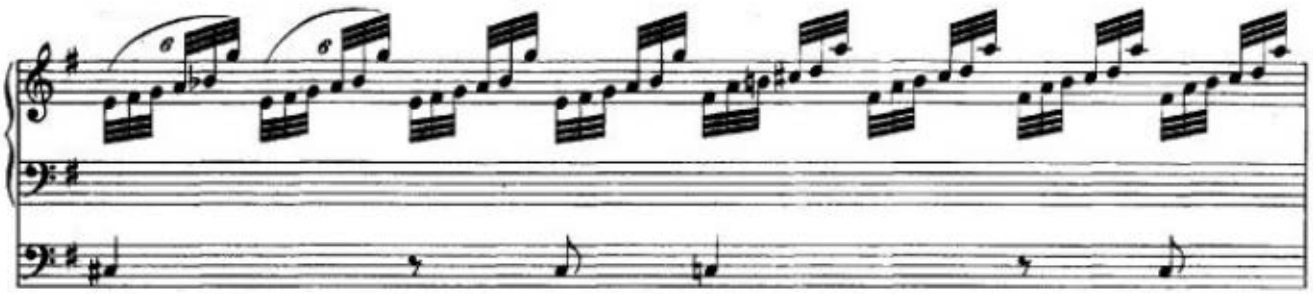


Figure 6: Comparison of SOMD *model A*, *model B*, *model C* submission with respect to other SAMPL5 participants estimations, in terms of RMSE, in kcal·mol^{-1}on the left , and R^2 coefficient,on the right, for all the entire OA dataset. In red *model A*, *model B* and *model C*, the dashed black line on the RMSE plot denotes 1 kcal·mol^{-1}

4.4 Conclusions

In the context of the SAMPL5 challenge, the binding free energies for 22 host-guest systems, OA, TEMOA and CBC were computed using variation of alchemical free energy models (*model A*, *B* and *C*). SOMD predictions were among the top-ranked SAMPL5 submissions with respect to R^2 metric (fig. **6**.) This result was unexpected given the size of the guests that was deemed large for an absolute binding free energy calculation, suggesting that longer per-λ simulation time than what was used here would be necessary. Factors that may have contributed to this outcome include the relative rigidity of some of the guests (e.g. OA and TEMOA), the rapid relaxation of the hosts upon guest decoupling. Encouragingly the results were also reasonably predictive, at least as judged by correlation with experiment (R^2 $0.65 < 0.70 < 0.74$). Nevertheless, systematic errors are present and the same models do not fare as well when ranked according to a mean unsigned error metric. *Model B* yields results that are identical to *model A* since the Long-range correction for missing dispersion interactions is essentially negligible. This was surprising, given previous reports, that showed significant contribution to standard free energies of binding [130]. For the systems considered here it seems that the simulation cutoffs used were sufficient to include most of the guest-host dispersion interactions. Satisfactorily, addition of a standard state correction term in *model C* systematically improves agreement with experimental data. In addition the computed standard free energies of binding for *model C* agree well with those produced independently by members of the Gilson lab (UCSD) using a different code and methodology. It is well known that the computation of free energies of solvation of charged solutes via molecular simulations is typically affected by significant finite-size effects [79, 80, 82, 81]. Given the broad range of net charges in the guests considered here it is perhaps surprising that encouraging R^2 values were obtained. For the host-guest binding energies reported here errors due to finite-size electrostatics is mitigated since partial error compensation occur between the simulations of the solvated guest and the host-guest complex. However it seems reasonable to anticipate that the significant MUE values would be decreased with the use of suitable schemes to reduce or eliminate finite-size errors. Other areas were further improvement could be sought for this dataset include the explicit consideration of multiple tautomeric forms of the guests, as well as a more systematic evaluation of alternative potential energy functions.



Chapter 5

Electrostatic finite size artifacts in alchemical free energy calculations

It is well known that absolute alchemical free energy calculations of charged molecules and relative alchemical free energy calculations that involve a change of net charge are affected by finite size artefacts. These artefacts are mainly due to effective Coulombic potentials and their associated parameters (e.g. cutoff, box size and periodic boundary conditions). This chapter explores correction protocols for these artefacts. A first protocol, adapted from works of Rocklin and Reif [81, 82], was developed for Barker Watts Reaction Field (BWRF) atomistic cutoff based simulations. A second approach makes use of alchemical counterions to keep a net-charge balance during the simulations. The hydration free energy of small organic charged species was computed to validate the former protocol. Then, both approaches were tested in binding free energy calculations using the OA and CBClip host-guest dataset of SAMPL5.

5.1 Introduction

Accurate and reliable protein-ligand binding free energy estimations would be beneficial for the drug discovery process. This would allow to enhance the screening of drug candidates and reduce the cost of the basic research stage. Over the past decades many free energy techniques have been devised [83, 84, 85, 86]. In particular the alchemical free energy calculations have shown significant success and a great potential in host-guest binding affinities studies [201, 219, 131]. However, the accuracy of the alchemical free energy calculations is always limited by force field errors [95, 184], sampling [183, 220] and methodological problems, such as the use of constraints [99, 221].

Beyond these issues, finite size artefacts [78, 81, 222, 82] constitute another major source of error. These artefacts affect both absolute and relative alchemical free energy calculations, when a charged solute is alchemically perturbed to a different net charge end state. The use of effective Coulombic potentials, along with cutoff lengths, periodic boundary condition and approximated solvent models introduces an offset potential in the simulation cell, which makes the simulated model deviate from a realistic conditions.

Historically, these artefacts are well known since the 1970s, thanks to seminal works of Neumann and Steinhauser [76, 223, 224], which stated the influence of boundary conditions and fictitious Coulombic potentials for dielectric constant calculations. Neumann's formulations were further developed by Hummer [225, 226, 227] focusing the attention on the problem of ionic species hydration free energy calculation. Although the computation of hydration free energy for ionic species seems an easy task, alchemical techniques fail to give a reasonable answer with errors ranging between 20 and 40 kcal·mol^{-1}. In 2006 these works were revisited by Kastenholz and Hünenberger [79, 80], which designed a correction protocol for ionic hydration free energies calculations run with Lattice Sum and Reaction Field methods. Later on, Rocklin [81] and Reif [82] sought to extend Kastenholz and Hünenberger's protocol to protein-ligand binding calculations by devising an analytical-numerical correction scheme.

These last two contributions fully defined the finite size artefacts as arising from distinct problems:

- Errors arising from Coulombic potential approximations [49, 41], such as cutoff-scheme, along with the presence of boundary conditions. The use of these approximation induces a fictitious polarization around the solutes, which are not identical to ideal Coulombic polarizations;

- The solvent model does not reflect the real macroscopic bulk proper-

ties (e.g. difference in dielectric constant). Thus, each solvent model would introduce a different polarization around the net-charged solute, different from the real case scenario. Furthermore, for lattice sum scheme (e.g. Particle Mesh Ewald (PME) methods), artificial interactions between solutes and theirs periodic copies are present, as well as for solvent molecules, resulting in an *undersolvation* effect;

- *charge sorting* errors due to the specific convention for summing up the contributions of solvent charges to the Coulombic potential at the cutoff. This error generates a further polarization to the solute, which is not present in a real system.

- The presence of effective Coulombic potentials give rise to spurious interactions between host and guest atoms.

In this chapter, these artefacts will be discussed in the context of the Barker-Watts Reaction Field atom-based scheme (BWRF), which is used by SOMD. In the section 5.2 the theoretical derivation of the BWRF scheme and its Poisson-Boltzmann correction scheme will be presented. To further validate the BWRF correction approach, the Rocklin methodology [81] for PME-based simulation will be introduced. A third scheme, employed only for host-guest binding free energy calculations, which makes use of alchemical counterions to balance the total net-charge along the alchemical transformation will be discussed. Section 5.4 will show the validation of the results for charged species solvation free energies. Thus, BWRF and Rocklin correction schemes will be extended to the case of host-guest binding free energy simulations and compared with the alchemical counterions approach.

5.2 Theory

5.2.1 Free energy correction terms for the BWRF approach

Fig. **1** shows the three correction terms, according to the BWRF framework, to deal with the finite size artefacts, to correct a periodic and effective Coulombic simulation box to a non-periodic and infinite electrostatic system. ΔG_{POL} corrects for the polarization effects due to the use of effective Coulombic potentials and periodic boundary conditions. Furthermore, ΔG_{POL} corrects also for the incorrect dielectric properties of the solvent model employed (e.g. TIP3P water [120] has a dielectric constant of 82 [228], while the experimental value is 78.3). ΔG_{PSUM} takes into account the charge sorting

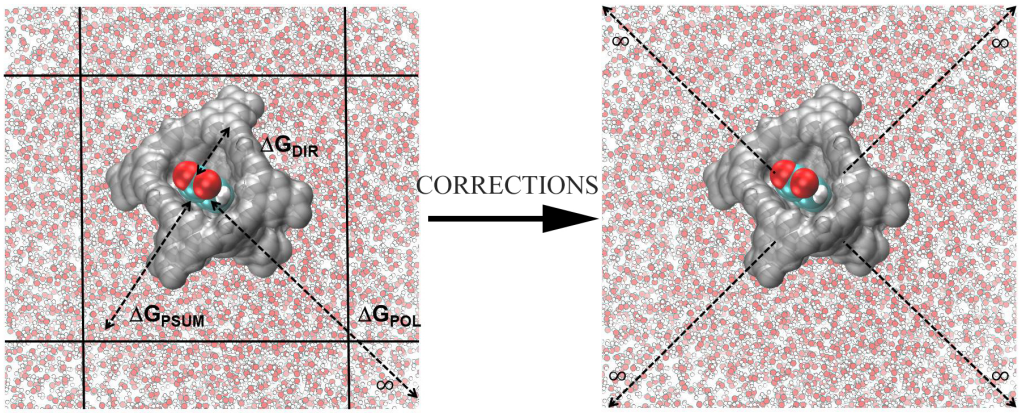


Figure 1: The three correction terms ΔG_{POL}, ΔG_{PSUM} and ΔG_{DIR} allows to pass from a periodic and effective Coulombic simulation box to a non-periodic Coulombic and infinite extended system. ΔG_{POL} accounts for polarization effects, considering an infinitely long system; ΔG_{PSUM} corrects for solvent polarization artefacts arising at the cutoff; ΔG_{DIR} correct for host-guest electrostatic interactions.

problem, due to solvent molecules at the cutoff sphere. ΔG_{DIR} corrects for spurious host-guest interactions, due to the use of effective Coulombic potential.

ΔG_{POL} is computed as:

$$\Delta G_{POL} = \Delta G_{NP} - \Delta G_{RF} \tag{5.1}$$

where ΔG_{NP} is the electrostatic charging free energy due to Coulombic interactions under non-periodic conditions, ΔG_{RF} is the electrostatic charging free energy for BWRF simulations under PBC. Both ΔG_{NP} and ΔG_{RF} are computed from complex phase and solvated phase with implicit solvent calculations. ΔG_{NP} is estimated using *APBS* software [229]. The implicit solvent linear Poisson-Boltzmann (PB) equation:

$$\Delta\epsilon(\mathbf{r})\Phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) + 4\pi\frac{\sum_i c_i^\infty z_i^2 q^2 \Phi(\mathbf{r})}{k_B T}\lambda(\mathbf{r}) \tag{5.2}$$

is solved through an adaptive multi-level finite element approach both for the complex and solvated phases. In eq. 5.2, $\epsilon(\mathbf{r})$ denotes the dielectric field as a function of the vector position $\mathbf{r}$, $\rho(\mathbf{r})$ the charge density in space, c_i^∞ is the concentration of an ion i at an infinite distance from the solute, z_i the i-th ion valency, q the proton charge (1.60217×10^{19} C), k_B the Boltzmann constant, T the temperature and $\lambda(\mathbf{r})$ is the accessibility to ions at point $\mathbf{r}$. The use

of the linearized PB equation is justified by studies made by Fogolari *et al.* [230, 231], which have shown that the electrostatic potential $\Phi(\mathbf{r})$ from the linearized PB equation is close to the solution obtained from solving the pure non linear equation for most of the biological relevant cases. As described in [82], there is a self Coulombic term included in the potential $\Phi(\mathbf{r})$. To cancel out this contribution, $\Phi(\mathbf{r})$ is computed as a difference between the potential generated by the solutes in presence of a dielectric and a vacuum reference potential, whose dielectric constant ϵ is set to 1. Once $\Phi(\mathbf{r})$ is estimated, the electrostatic free energy at $\lambda = 0.0$, $\Delta G_{NP-\lambda=0.0}$ can be estimated as:

$$\Delta G_{NP-\lambda=0.0} = \frac{1}{2}\sum_{i=1}^{N} q_i \Phi(\mathbf{r}_i) \tag{5.3}$$

where the sum is extended to all the N solutes' atoms, q_i is the i-th atom's partial charge and $\Phi(\mathbf{r}_i)$ refers to the electrostatic potential at the atom positions. Similarly, by setting guest's atoms charges to zero, it is possible to compute $\Delta G_{NP-\lambda=1.0}$. Considering the complex phase, *HG*, and the solvated phase, *G*, simulations, the final ΔG_{NP} is equal to:

$$\Delta G_{NP} = (\Delta G^{G}_{NP-\lambda=0.0} + \Delta G^{G}_{NP-\lambda=1.0}) - (\Delta G^{HG}_{NP-\lambda=0.0} - \Delta G^{HG}_{NP-\lambda=1.0}) \tag{5.4}$$

where $\Delta G^{G}_{NP-\lambda=0.0}$ and $\Delta G^{G}_{NP-\lambda=1.0}$ are the electrostatic free energies for the solvated guest at $\lambda = 0.0$ and 1.0, while $\Delta G^{HG}_{NP-\lambda=0.0}$ and $\Delta G^{HG}_{NP-\lambda=1.0}$ are the electrostatic free energies for the host-guest system at $\lambda = 0.0$ and 1.0.

ΔG_{RF} is computed using a custom code kindly provided by P. Hünenberger [232, 233] and modified accordingly. The code computes the electrostatic potential $\Phi_{RF}(\mathbf{r})$ mimicking implicit solvent BWRF simulation conditions under periodic boundary condition for the complex and solvated phase. The final potential is retrieved by solving the fundamental electrostatic equations [234]:

$$\begin{cases} \mathbf{E}(\mathbf{r}) & = \mathbf{V}(\mathbf{r}) + \int\int\int_{\Re^3} \mathbf{T}(\mathbf{r})\mathbf{P}(\mathbf{r})d\mathbf{r} \\ \\ \mathbf{P}(\mathbf{r}) & = \epsilon_0 \left[\epsilon(\mathbf{r}) - 1\right] \mathbf{E}(\mathbf{r}) = \epsilon_0(\epsilon_s - 1)\mathcal{H}(\mathbf{r})\mathbf{E}(\mathbf{r}) \end{cases} \tag{5.5}$$

where $\mathbf{E}(\mathbf{r})$ is the electric field, as a function of the atoms' coordinates $\mathbf{r}$, $\mathbf{V}(\mathbf{r})$ is the vacuum field, namely the field generated by the solute's atoms partial charges under the BWRF, $\int\int\int_{\Re^3}$ is a volume integral, $\mathbf{P}(\mathbf{r})$ is the polarization field. $\mathbf{T}(\mathbf{r}) = (3\mathbf{r}_{ij,\alpha} \times \mathbf{r}_{ij,\beta} - \mathbf{r}^2_{ij}\delta_{\alpha,\beta})\mathbf{r}^{-5}_{ij}$ is the dipole-dipole interaction tensor, where α and $\beta = x, y, z$, $\mathbf{r}_{ij}$ is the distance between dipole i and dipole j and $\delta_{\alpha,\beta}$ is the Kronecker delta (. The second equation describes the polarization field as a function of ϵ_0, the vacuum permittivity, $\epsilon(\mathbf{r})$

dielectric vector and $\mathcal{H}$ is the Heaviside function, which evaluates 0 inside the reference solute molecule and 1 outside. To account for the presence of PBC in the simulation, eq. 5.5 are solved in the reciprocal space using the Fast Fourier Transform (FFT) [46]. Since, neither $\mathbf{E}(\mathbf{r})$ nor $\mathbf{P}(\mathbf{r})$ are known initially, the code computes the overall electric field in the space, by iteratively calculating the vacuum field inside the solute in such a way that the overall electric field inside the solute becomes zero. To obtain an accurate description of the electrostatic potential, the entire space is subdivided into a grid domain of $N_x \cdot N_y \cdot N_z$ points. Once the electrostatic field is computed, it is possible to retrieve the polarization field and, eventually, the electrostatic potential as:

$$\Phi_{RF}(\mathbf{r}) = \int\int\int_{\Re^3} \mathbf{t}(\mathbf{r})\mathbf{P}(\mathbf{r})d\mathbf{r} \tag{5.6}$$

where $\mathbf{t}(\mathbf{r}) = -\mathbf{r}_{ij,\alpha} \times \mathbf{r}_{ij}^{-}3$ is the charge-dipole interaction tensor, evaluated for all the solute charges. As per the non-periodic electrostatic potential, $\Phi_{RF}(\mathbf{r})$ contains a self Coulombic term, which is subtracted by a vacuum reference potential. Finally the solvation free energy contribution at $\lambda = 0.0$, given by a BWRF potential under PBC, is computed as:

$$\Delta G_{RF-\lambda=0.0} = \frac{1}{2}\sum_{i}^{N_{atoms}} q_i \Phi_{RF}(\mathbf{r}_i) \tag{5.7}$$

where N is the total number of solute's atoms (N_{atoms}) and $\Phi_{RF}(\mathbf{r}_i)$ refers to the electrostatic potential at the atoms'site $\mathbf{r}_i$. Similarly for $\Delta G_{NP-\lambda=1.0}$, $\Delta G_{RF-\lambda=1.0}$ is computed by setting guest'atoms charges to zero. ΔG_{RF} is thus equal to:

$$\Delta G_{RF} = (\Delta G^{G}_{RF-\lambda=0.0} + \Delta G^{G}_{RF-\lambda=1.0}) - (\Delta G^{HG}_{RF-\lambda=0.0} - \Delta G^{HG}_{RF-\lambda=1.0}) \tag{5.8}$$

where $\Delta G^{G}_{RF-\lambda=0.0}$ and $\Delta G^{G}_{RF-\lambda=1.0}$ are the electrostatic free energies for the solvated guest at $\lambda = 0.0$ and 1.0, while $\Delta G^{HG}_{RF-\lambda=0.0}$ and $\Delta G^{HG}_{RF-\lambda=1.0}$ are the electrostatic free energies for the host-guest system at $\lambda = 0.0$ and 1.0.

The radius of cutoff present in the BWRF equation, eq. 1.36, defines a *cutoff sphere*. At the boundaries of the sphere it is possible to compute the Coulombic contribution with two different approach: an atom-based or *P-summation* approach and a group-based or *M-summation* method [235]. As fig. **2** shows, the former approach computes the charges contribution consider only those atoms within the radius of cutoff, which may cause a split of solvent molecules at the boundaries. The second method computes charges' contributions considering the molecules in the vicinity of the cutoff as a unique entity, even if some atoms may lie beyond the cutoff radius. Both

summation schemes may give raise to fictitious polarization onto the solute's atoms. In SOMD the BWRF computes charges based on the P-summation scheme, thus a correction ΔG_{PSUM} is devised.

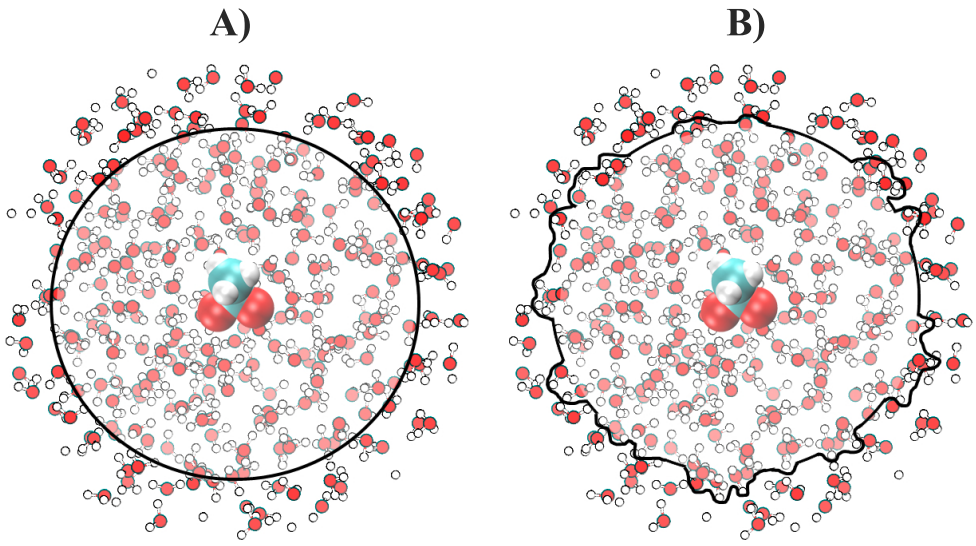


Figure 2: Difference between *P-summation* and *M-summation*. *P-summation* sums up electrostatic contributions for all the single atoms within the cutoff radius; *M-summation* scheme considers the electrostatic contributions for entire molecules, whose atoms may lie at the boundary of the cutoff length.

In literature studies [226, 227] the solvent molecules at the cutoff are assumed to be in an orientational disorder limit (ODL) distribution. The combination of the electrostatic description of a solvent molecule and the statistical properties of the ODL distribution allows to devise ΔG_{PSUM}. Solvent molecules can be described by a charge distribution $\tau(\xi)$, where ξ denotes the distance from the center of the distribution within a sphere of radius R defined by the solvent radius. $\tau(\xi)$ is overall neutral over the spatial extent and it vanishes beyond a distance R:

$$\begin{cases} 4\pi \int_0^\infty d\xi \xi^2 \tau(\xi) = 0 \\ \\ \tau(\xi) = 0 \quad \text{for} \quad \xi > R \end{cases} \tag{5.9}$$

The charge distribution $\tau(\xi)$ can be described as an isotropic quadrupole (IQ) [234] with magnitude:

$$\gamma = 4\pi \int_0^\infty \xi\xi^4\tau(\xi) \tag{5.10}$$

where γ coincides with the quadrupole moment trace of the charge distribution. Thus, the electrostatic potential generated by this charge distribution within the solvent radius can be computed as:

$$\Phi(r) = \int_r^\infty dr\mathbf{E}(\mathbf{r}) \tag{5.11}$$

where $\mathbf{E}(\mathbf{r}) = \epsilon_0^{-1}r^{-2}\int_0^r d\xi\xi^2\tau(\xi)$ is the radial electric field, as a function of the radial coordinate r. Considering an integration over the spherical volume $\frac{4}{3}\pi R^3$ around the center of the distribution $\tau(\xi)$, the electrostatic potential generated by a solvent molecule at R can be computed as a function of the quadrupole moment trace, eq. 5.10:

$$\Phi(R) = (-6\epsilon_0)^{-1}\gamma \tag{5.12}$$

Now, eq. 5.12 can be coupled with the configurational canonical distribution $\mathcal{P}_{ODL}$ for ODL molecules. $\mathcal{P}_{ODL}$ can be divided in a positional contribution $\mathbf{S}$ and orientational one Θ:

$$\mathcal{P}_{ODL} = p_S(\mathbf{S})p_\Theta(\Theta) \tag{5.13}$$

where $p_S(\mathbf{S}) \sim V^{-1}$, where V is the volume available to the water molecules, and $p_\Theta(\Theta) \sim (4\pi)^{-1}$ in a spherical domain. As a consequence, it is now possible to compute the isotropically average charge distribution of solvent molecules in ODL configuration as:

$$\tau(\xi) = (4\pi)^{-1}\sum_{i=1}^{n} q_i r_i^{-2}\delta(\xi - r_i) \tag{5.14}$$

where the sum is extended up to all the n solvent molecules at the interface, $\delta(\xi - r_i)$ is the Kronecker delta. This distribution satisfies the electroneutrality condition, eq. 5.9 and its magnitude is given by the quadrupole moment trace γ. Thus, the polarization issues due to summation scheme at the cutoff sphere gives a contribution equal to $\Phi_{ODL} = (6\epsilon_0)^{-1}\gamma\rho$, where ρ is the solvent number density.

The ODL potential Φ_{ODL} can be used to compute the correction term ΔG_{PSUM} in presence of a BWRF, which, for a single ion, has the form:

$$\Delta G_{PSUM} = -q_I\frac{2(\epsilon_s - 1)}{2\epsilon_s + 1}\left(1 - \frac{R_I^3}{R_C^3}\right)\Phi_{ODL} - q_I\frac{3}{2\epsilon_{BW} + 1}\Phi_{ODL} \tag{5.15}$$

where q_I is the ionic charge, ϵ_s is the solvent dielectric constant, the term $-q_I \frac{2(\epsilon_s-1)}{2\epsilon_s+1}\left(1-\frac{R_I^3}{R_C^3}\right)\Phi_{ODL}$ represents the interaction between the ionic charge and the ODL potential, due to the reaction field, and the additional factor $\frac{R_I^3}{R_C^3}$ takes into account the presence of the ion in the cutoff sphere. The term $-q_I \frac{3}{2\epsilon_{BW}+1}\Phi_{ODL}$ denotes the interaction between the ionic charge and the ODL potential, due to the forces exerted upon the dipole form the continuum electrostatic environment. Eq. 5.15 can be extended to polyatomic systems [82]:

$$\Delta G_{PSUM} = -Q_G \frac{2(\epsilon_s-1)}{2\epsilon_s+1}\Phi_{ODL}\frac{\langle N_w(R_c)\rangle}{\frac{4}{3}\pi R_C^3} + Q_G \frac{3}{2\epsilon_s+1}\Phi_{ODL} \tag{5.16}$$

where $\langle N_w(R_C)\rangle$ is the average number of water molecules within the cutoff sphere and Q_G the solute's net-charge.

Finally, ΔG_{DIR} treats the host-guest direct interactions only, which are affected by an offset due to the effective Coulombic potential in use:

$$\Delta G_{DIR} = U_{NP}^{dir} - U_{RF}^{dir} \tag{5.17}$$

U_{NP}^{dir} is the pure Coulombic potential between host's and guest's atoms, while U_{RF}^{dir} the reaction field Coulombic potential of interaction.

Hence, considering all the correction terms, the final charging correction free energy contribution is:

$$\Delta G_{COR} = (\Delta G_{POL}^{HG} + \Delta G_{PSUM}^{HG} + \Delta G_{DIR}^{HG}) - (\Delta G_{POL}^{G} + \Delta G_{PSUM}^{G}) \tag{5.18}$$

where HG denotes the calculations in the complex phase, G in the solvated phase.

5.2.2 Correction scheme for the PME approach

Lattice sum methods [41, 40, 46], like particle-particle particle mesh (P3M) or Particle Mesh Ewald (PME), are widely used nowadays in computer simulations. The advantage of these methodologies is the use of the reciprocal Fourier transform to compute the long range electrostatic interactions, enhancing the convergence of calculations with respect to direct calculations. However, the reciprocal space treatment makes the computational implementation more complicated and further finite size artefacts are introduced. Rocklin *et al.* [81] developed an analytical and numerical scheme for the correction of finite size artefacts for lattice-sum based molecular simulations. This correction approach will be named from now on as the PME correction scheme.

Rocklin corrections can be enlisted as:

- Correction for *periodicity-induced net-charge interactions*. When lattice-sum schemes are coupled with PBC conditions, spurious interactions arise between the host-guest system reference computational box when a solutes present a net-charge. The correction term ΔG_{NET} can be estimated as:

$$\Delta G_{NET}(L) = -\frac{\psi_{LS}}{8\pi\epsilon_0}\left[(Q_H + Q_G)^2 - Q_P^2\right]\frac{1}{L} \tag{5.19}$$

  where L is the is the dimension of the unit cell, supposing to be the same length in x,y and z directions, $\psi_{LS} = -2.837297$ is the Wigner integration constant [236], Q_H is the protein/host net charge, Q_G the ligand/guest net charge. When simulations are carried out with neutralizing counterions $Q_H = 0$ and the same is done for the solvated guest simulations

- Correction for *periodicity-induced net-charge undersolvation*. In lattice-sum simulations the solvent in the periodic replicas of the computational box is perturbed by the periodic copies of the solute, being unavailable for the solvation of the solute itself in the reference computational box. In this case, a correction term ΔG_{USV} can be retrieved as :

$$\Delta G_{USV}(L) = \frac{\psi_{LS}}{8\pi\epsilon_0}\left(1 - \frac{1}{\epsilon_s}\right)\left[(Q_H + Q_G)^2 - Q_P^2\right]\frac{1}{L} \tag{5.20}$$

  where ϵ_s is the solvent dielectric permittivity. ΔG_{USV} differs from ΔG_{NET} by a factor ϵ_s. For polar solvent, such as water, the contribution given by $\Delta G_{USV} + \Delta G_{NET}$ could be negligible, while this is not possible for organic solvent simulations.

- The charge sorting correction term for PME is ΔG_{DSC}:

$$\Delta G_{DSC}(L) = -\frac{\gamma Q_G}{6\epsilon_0}\frac{N_s}{L^3} \tag{5.21}$$

  where γ is the quadrupole moment trace of the solvent model, Q_G the ligand charge and N_s the number of solvent molecules in the simulation box.

- Correction for the *residual integrated potential effects*. This term reflects similar polarization issues encountered in BWRF and it takes into account the offset potential introduced by the presence of net-charges of

the solutes. The treatment is again related to the quadrupole-moment trace of the solute, giving a term ΔG_{RIP}:

$$\Delta G_{RIP}(L) = [(I_H + I_G)(Q_H + Q_G) - I_H Q_H]\frac{1}{L^3} \tag{5.22}$$

where I_H and I_G are the spatial integral of the electrostatic potential under non-periodic boundary conditions due to the protein/host and ligand/guest charges respectively, evaluated using APBS [229]. In the case where no protein is present in the simulation box $I_P = 0$.

5.2.3 Correction scheme with alchemical counterions

Fig. **3** shows a third approach to correct for the finite size artefacts. This scheme relies on the use of alchemical counter ions, in order to conserve the system electro-neutrality during the alchemical transformation of the solute and it was used for BWRF simulations only. Once the molecule is fully decoupled, after the vanishing step, each counter ion is switched on, going through a reverse vanishing and a reverse discharging step. The correction term $\Delta G_{COR}^{counterions}$ can be retrieved from the thermodynamic cycle:

$$\begin{aligned}\Delta G_{COR}^{counterions} &= \sum_{i=1}^{n}(\Delta G_{ion_i}^{wat} - \Delta G_{ion_i}^{host}) \\ &= \sum_{i=1}^{n}(\Delta G_{ion_i-elec}^{solv} + \Delta G_{ion_i-vdW}^{solv}) - (\Delta G_{ion_i-elec}^{host} + \Delta G_{ion_i-vdW}^{host})\end{aligned} \tag{5.23}$$

where the sum is extended up to all the n counter-ions, $\Delta G_{ion_i}^{wat}$ is the free energy change for turning on the van der Waals and charges of the i-th ion in the solvated phase, $\Delta G_{ion_i}^{host}$ is the free energy change for the i-th ion in the bound phase. $\Delta G_{ion_i-elec}^{solv}$ and $\Delta G_{ion_i-vdW}^{solv}$ are the reverse-discharging and reverse-vanishing free energy changes in solvated phase respectively, while $\Delta G_{ion_i-elec}^{host}$ and $\Delta G_{ion_i-vdW}^{host}$ are the reverse-discharging and reverse-vanishing free energy changes in the complex phase.

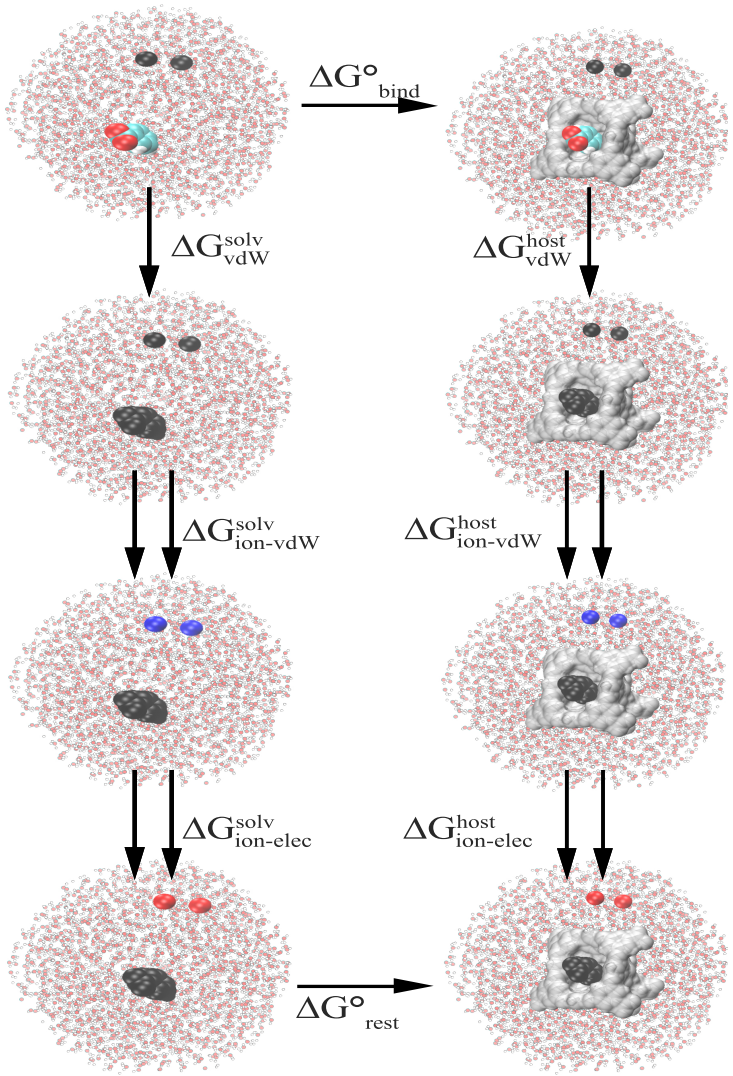


Figure 3: Thermodynamic cycle for the alchemical counterions approach. Once the guest's van der Waals parameters are switched off, ΔG^{solv}_{vdW} in the solvated phase and ΔG^{host}_{vdW} in the complex phase, n ions' interaction parameters are turned on. In this case, the cycle shows the transformation of two ions to keep the net-charge balance in the simulated box. Firstly, in a de-vanishing step two ions van der Waals terms are turned on, giving two free energy changes $\Delta G^{solv}_{ion-vdW}$ and $\Delta G^{host}_{ion-vdW}$ in the solvated and complex phase respectively. Then, ions' partial charges are switched on, giving $\Delta G^{sol}_{ion-elec}$ and $\Delta G^{host}_{ion-elec}$ free energy changes in solvated and complex phase respectively. The cycle is closed by computing the standard state term ΔG°_{rest}.

5.3 Methods

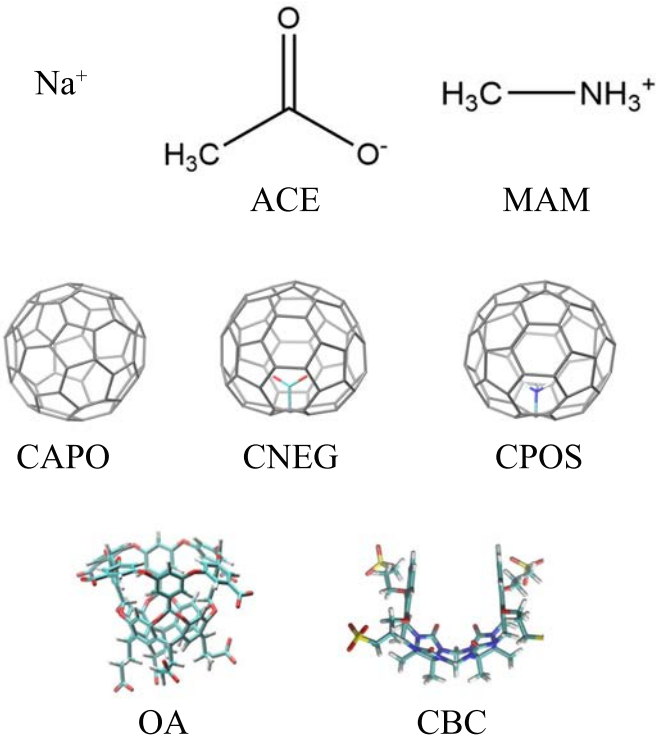


Figure 4: Dataset employed to study the finite size artefacts. Sodium ion, Na^+, hydration free energy was computed as an initial benchmark for BWRF and PME correction schemes. Acetate (ACE) and methylammonium (MAM) molecules solvation discharging free energy were computed both for BWRF and PME approaches. The BWRF correction terms were compared with Reif and Oostenbrink results [82], for the complex of ACE and MAM with CAPO, CPOS and CNEG. Finally, BWRF and PME correction schemes were extended to deal with SAMPL5 OA and CBC dataset (refer to fig. 1 chapter 4 for guest structures).

5.3.1 Sodium ion

The sodium ion hydration free energy was computed as a benchmark test to to validate and assess the accuracy of BWRF and PME correction schemes. This test was done for different box sizes and cutoff lengths. Initially, sodium ion was solvated into a "small" cubic box, with an average edge size of 30Å. Next, sodium ion was placed in a "medium" size box, with an average dimension of 43Å. Finally, a third box of 61Å edge was created.

For each system the `tleap` [73] module of AmberTool16 was employed. Joung and Cheatham [237] parameters were used to parametrize the sodium ion charges and van der Waals terms. TIP3P water molecules [120] were selected along with the *iso* option to have the most similar edge lengths in each direction. Then an energy minimization of 100 steps with the steepest descent method was done using AmberTools module `sander`. Following a NVT equilibration at 298 K and a NPT equilibration at 1 atm both for 200 ps. Finally, 2 ns simulations were run in the NPT ensemble using SOMD (rev. 15.1) [69] to reach a final density of 1 g/cc and to obtain a fully equilibrated simulation box. The final coordinate files were retrieved with CPPTRAJ [169].

The hydration free energy ΔG_{hyd} of each system was computed for two non-bonded interaction cutoff lengths: one at 10 Å and another one at 12 Å both with BWRF and PME framework. Each simulation was 2 ns long and 11 equidistant λ windows were employed for both the discharging and the vanishing step. All the water molecules' bonds were constrained, a Monte Carlo barostat was adopted to keep the pressure constant with moves attempted every 50 fs, along with an Andersen thermostat to keep constant the temperature to 298 K. The water reaction field dielectric constant was set to 82.0. Each free energy change was estimated with MBAR [171]. Two independent runs were done and the ΔG_{HYD} was taken as an average of both runs. The standard error was computed as:

$$err_{\Delta G_{hyd}} = \frac{\sigma}{\sqrt{2}} \tag{5.24}$$

where σ is the standard deviation between the two calculated hydration free energies.

5.3.2 Acetate, methylammonium and functionalized fullerene

Following Reif and Oostenbrink algorithm [82], acetate (ACE) and methylammonium (MAM) solvation discharging free energies were computed and

corrected with BWRF and PME correction schemes, in order to have a meaningful comparison with literature results. For consistency with the Reif paper, the GROMOS 53A6 force field parameters [238] were used to parametrized the ACE molecule. Starting from the pdb format file of acetate, created with Maestro (v.10.1.012, rel 2015-1, Schrödinger) [203], charges were assigned using AmberTools 16 module Antechamber based on the AM1-BCC method [118], creating a mol2 file of the ACE molecule. Then, the mol2 file was loaded in `tleap` and placed in a TIP3P water box [120]. The same protocol was done for the MAM molecule, however the GAFF force field parameters [117] were employed. The same equilibration protocol used in sec. 5.3.1 was followed for both molecules. The initial, ACE box had a size of 37.48×37.79×37.70 Å edge and MAM a box size of 38.71×37.09×37.22 Å. The solvation discharging free energies were computed as a difference between a solvated phase and vacuum phase discharging free energy step. An average value for the solvation discharging free energy was computed from two independent simulations and the standard error computed according to eq. 5.24. Thus, the same free energy parameters of sec. 5.3.1 were selected both for ACE and MAM calculations.

Then, the BWRF correction terms were compared with Reif [82] results for three functionalized fullerene cavities: a neutral one (CAPO), a positively charge one (CPOS) and a negatively charged (CNEG). CAPO fullerene was created starting from a mol2 file, parametrized with GAFF atom type and neutral carbon atoms using Antechamber. Thus, employing AmberTool 16 module `parmchk` and `tleap` the topology and the coordinates file of the neutral cavity were created. Using Maestro, a pdb format file of the bound host-guest system was saved. Thus, the topology and coordinates file were written using `tleap`, loading the new pdb file along with the force field information of the guest molecule and CAPO fullerene. The same approach was used for CPOS and CNEG. The parameters for the functionalized moieties of CPOS and CNEG, were retrieve from MAM and ACE molecules respectively. Due to the absence of hydrogen atoms in CPOS and CNEG moieties, the atoms' partial charges were re-scaled accordingly, in order to have a total net charge of +1e and -1e respectively. The correction terms were computed as an average between two replicates simulations and the standard error was estimated with eq. 5.24. The free energy parameters were the same of sec. 5.3.1, but with a non-bonded cutoff set to 14 Å.

5.3.3 SAMPL5 host-guest systems: OA and CBC

Correction schemes were applied to correct for host-guest binding free energy cases. The CBC and OA host-guest dataset from SAMPL5 were selected as

test cases for this study (fig. 1.1 chapter 4). The CBC system files were created from scratch, in order to have the same box size between complex phase and solvated phase simulations. Solvated phase simulation files were created from complex phase topology and coordinate files. Thus, the host molecule was removed from the system and CBC heavy atoms were substituted with water molecules. Then, the new solvated guest was re-equilibrated according to the protocol of sec. 5.3.1. OA dataset was considered as a further extension of this study and the same input files of the SAMPL5 challenge were used. In this case complex phase box edges differ from the solvated ones.

The choice of these two datasets was motived by three reasons. First, the OA host-guest system is a very rigid host system interacting with guests whose charges are between -1 e and +1 e. This aspect reflects some of the systems used in literature [81, 82]. Secondly, the CBClip was thought to be a more complex case, due to its high flexibility and highly charged guests. As a matter of fact, CBC guest show a variety of charges ranging from -2 e (G7) up to +4 (G3)-4e (G3). These features make the CBC dataset an interesting case for the validation and application of the correction schemes. Third, the different box sizes between complex and solvated phases in OA dataset could shed light onto size effects in the BWRF and PME correction schemes.

5.3.4 Correction schemes implementations

The BWRF correction free energies can be computed by using the Sire command `somd-coultailcorrection`. This script loads the $\lambda = 0.0$ discharging step trajectory of solvated guest and complex phase with *mdtraj* [239] module. For each frame the atomic coordinates are read and copied into a Sire molecular system. Firstly the BWRF Poisson calculation is performed. In the first section of the Poisson solver input file the number of grid points (N_x, N_y and N_z) is specified. Then, the maximum number of iterations for the electric field to converge is set to 200, with a convergence limit set to 10^{-3}kJ·mol^{-1}·nm^{-1}·e^{-1}. In the second section of the input file, the solute atoms' coordinates, radius and atomic partial charge are written. For each atom the radius is computed as:

$$r_{LJ} = \frac{1}{2}(\sigma^2)^{\frac{1}{6}} + s_{rad} \tag{5.25}$$

where σ is the Lennard-Jones distance parameter and s_{rad} is the solvent probe radius, fixed to 1.4Å supposing water molecules as solvent.

Following, the PB calculation is performed with APBS. In the APBS input file the solute's atoms coordinates are written in XML format, along with their partial charges and radius (computed according to eq. 5.25). For

each PB calculation, the same grid dimension and grid spacing as in the BWRF Poisson routine are employed. The grid is centered onto the solute coordinates and the linear Poisson Boltzmann equation option is selected (`lpbe`). The equation is solved under multiple Debye-Hückel boundary conditions (`mdh`). The solute dielectric constant `pdie` was set to 2.0 according to literature [81] values, while the solvent dielectric `sdie` was set to 82 for solvated phase calculations. A cubic B-spline discretization was selected, to have the biomolecular charges mapped onto the nearest and next-nearest neighbor grid point.

Thus, the term ΔG_{PSUM} is computed according to eq. 5.16. The ΔG_{DIR} term is computed only for host-guest systems. For each trajectory frame, the difference between the host-guest Coulombic potential and the BWRF electrostatic potential is computed.

Finally, all the free energy terms employed for the calculation of ΔG_{COR} are written onto a text file, to further check if any problem has arisen during the calculation. The final correction term is computed from the average value of all the ΔG_{POL}, ΔG_{PSUM} and ΔG_{DIR} values.

For the PME simulation OpenMM 7.0 [69] was employed. The python module `parmed` was used to load the topology and coordinate file of each system. Then, a reference OpenMM system was created, by imposing constraints to all bonds, rigid water molecules and non bonded interactions cutoff. Then, for any λ window the solute atoms partial charges were scaled accordingly. For each iteration, a selected number of steps of MD were run onto the system. Following, the current system coordinates were copied onto all the other thermodynamic λ state and their potential energy values were retrieved and saved in a text file. After all the λ windows have been run, `pymbar` was used to estimate the electrostatic free energy change.

The counterions approach was run with SOMD only. Each studied host-guest system was re-created, preserving the original box size dimension but introducing as many non-interacting ions as the total solute net-charge. Then, since it is currently not technically possible to run multiple alchemical transformations at the same time in SOMD, two different sets of simulations were carried out. The first set was done for decoupling the guest's atoms, both in complex and solvated phase. In the second set of simulations, as many simulations as the total number of non-interacting counterions were run. Here, each single ion was perturbed from a non-interacting ideal state to a fully interacting state through a de-vanishing and charging step, with an ad-hoc morphing file.

5.4 Results and discussion

5.4.1 Sodium ion hydration free energy

The sodium ion "small", "medium" and "big" box sizes created in `tleap` are depicted in tab. **1**, while the best protocol results for the hydration free energy under BWRF and PME correction schemes are shown in tab. **2** and **3** respectively. In particular, for the PME approach a grid of 257 points in each direction was employed as done by Rocklin [81], while for the sodium ion BWRF correction a grid of 65 points gave a reasonable free energy correction.

The BWRF electrostatic potential shows a strong dependence on box sizes and cutoff lengths, as reflected by the raw hydration free energies ΔG_{hyd}^{raw}. Indeed, raw values vary from -62.81±0.01 kcal·mol^{-1} to -67.84±0.02 kcal·mol^{-1}. This increment is counter balanced by the correction term ΔG_{POL} which gives a contribution due to polarization errors between -25 and -19 kcal·mol^{-1}. The final corrected hydration free energy estimations more consistent with each other, within fluctuations of maximum 1.66 kcal·mol^{-1} . The correction scheme brought an improvement in terms of standard deviation, passing from 2.18 kcal·mol^{-1} to 1.61 kcal·mol^{-1}

PME replicates the trend observed with BWRF results. PME raw hydration free energies exhibit a low standard deviation, 0.49 kcal·mol^{-1}. The free energy difference between the small box at 10 Å cutoff and the big box at 12 Å cutoff is about 0.7 kcal·mol^{-1} with an underestimation against the experimental value of about 10 kcal·mol^{-1} [240, 241, 242]. In this case, the charge sorting problem is the major source of errors, with an average correction ΔG_{dsc} of -16.4 kcal·mol^{-1} for each system. The final hydration free energy ΔG_{hyd}^{corr} is -104.87±0.39 kcal·mol^{-1} on average. Both for BWRF and PME schemes the ΔG_{hyd}^{corr} is in better agreement with the experimental value (-99 kcal·mol^{-1}) than the raw simulation results. kcal·mol^{-1} [242, 241, 240].

The grid size and dielectric constant sensitivity of the PB and Poisson solvers were tested both for BWRF and PME correction schemes. Overall, results are consistent one another with differences less than 1 k_BT and values were not reported in this thesis.

Table 1: `tleap` box edges length L_x, L_y and L_z in Å for the three sodium ion systems called "small", "medium" and "big".

	L_x	L_y	L_z
small	25.29	25.58	25.79
medium	43.44	43.65	43.77
big	61.49	61.99	61.62

Table 2: Sodium ion hydration free energy results the BWRF approach corrected with the Poisson scheme (sec. 5.2.1). *box size* denotes the edges length of each simulated box, *cutoff* is the cutoff length employed for the non-bonded interactions, ΔG^{raw}_{hyd} is the raw hydration free energy from the PME discharging and vanishing simulation, ΔG_{POL}, ΔG_{DIR} and ΔG_{PSUM} are the correction terms explained in sec. 5.2.1. ΔG^{corr}_{hyd} is the final corrected hydration free energy. $\mu \pm \sigma$ indicates the average and standard deviation of ΔG^{raw}_{hyd} and ΔG^{corr}_{hyd} values. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

Sodium ion BWRF						
box size	cutoff	ΔG^{raw}_{hyd}	ΔG_{POL}	ΔG_{DIR}	ΔG_{PSUM}	ΔG^{corr}_{hyd}
small	10A	-62.81±0.01	-25.78±0.01	0.00±0.00	-15.58±0.01	-104.16±0.02
	12A	-66.24±0.01	-22.32±0.01	0.00±0.00	-15.73±0.03	-104.33±0.01
medium	10A	-63.64±0.01	-26.85±0.01	0.00±0.00	-15.68±0.01	-106.17±0.01
	12A	-67.57±0.04	-23.02±0.02	0.00±0.00	-15.72±0.06	-106.31±0.07
big	10A	-63.73±0.02	-23.05±0.14	0.00±0.00	-15.73±0.03	-102.50±0.20
	12A	-67.84±0.02	-19.28±0.03	0.00±0.00	-15.70±0.04	-102.80±0.04
$\mu \pm \sigma$		-65.30±2.18				-104.73±1.61

Table 3: Sodium ion hydration free energy results the PME approach with Rocklin correction scheme. *box size* denotes the edges length of each simulated box (refer to tab. **1**), $cutoff$ is the cutoff length employed for the Coulombic non-bonded interactions, ΔG^{raw}_{hyd} is the raw hydration free energy from the PME discharging and vanishing simulation, ΔG_{net}, ΔG_{usv}, ΔG_{dsc} are the Rocklin correction terms, as explained in sec. 5.2.2, and ΔG^{corr}_{hyd} is the final corrected hydration free energy. $\mu \pm \sigma$ indicates the average and standard deviation of ΔG^{raw}_{hyd} and ΔG^{corr}_{hyd} values. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

Sodium ion PME							
box size	cutoff	ΔG^{raw}_{hyd}	ΔG_{net}	ΔG_{usv}	ΔG_{rip}	ΔG_{dsc}	ΔG^{corr}_{hyd}
small	10A	-87.98±0.26	0.01±0.01	-0.01±0.01	0.08±0.01	-16.54±0.03	-104.58±0.23
	12A	-87.66±0.02	0.01±0.01	-0.01±0.01	0.08±0.01	-16.52±0.02	-104.25±0.04
medium	10A	-88.98±0.04	0.01±0.01	-0.01±0.01	0.02±0.01	-16.25±0.03	-105.36±0.01
	12A	-88.68±0.31	0.01±0.01	-0.01±0.01	0.02±0.01	-16.31±0.02	-105.11±0.29
big	10A	-88.60±0.03	0.01±0.01	-0.01±0.01	0.02±0.01	-16.24±0.01	-104.98±0.02
	12A	-88.60±0.09	0.01±0.01	-0.01±0.01	0.02±0.01	-16.21±0.01	-104.97±0.06
$\mu \pm \sigma$		-88.41±0.49					-104.87±0.39

5.4.2 Polyatomic molecules: acetate and methylammonium

In this case the solvated phase discharging free energy change was compared between BWRF and PME for different cutoff lengths, because it was technically challenging to implement a soft-core vanishing step into the PME code.

Tab. **4** shows BWRF solvated phase discharging free energy estimations and the correction terms. As for the sodium ion case, the solvated phase discharging free energy ΔG^{raw}_{chg} is influenced by the cutoff length, with an average values between the three systems of -76.76±3.58 kcal·mol^{-1} and -41.78±3.57 kcal·mol^{-1} for ACE and MAM respectively. The correction terms show an improvement in the discharging free energy, reducing the standard deviation to 0.56 kcal·mol^{-1} for ACE and to 0.43 kcal·mol^{-1} for MAM. In the end, ACE solvated phase discharging free energy have an average value of about -80 kcal·mol^{-1}, while MAM a ΔG^{corr}_{chg} of -78.35±0.43 kcal·mol^{-1}.

Tab. **5** reports the solvated phase discharging free energy for the PME approach with Rocklin correction scheme. As before PME seems to be cutoff independent with an average ΔG^{raw}_{chg} of -97.19±0.48 kcal·mol^{-1} for ACE and -62.12±0.27 kcal·mol^{-1} for MAM. The correction scheme reduces further the

standard deviation from of ΔG_{chg}^{raw} for ACE, passing from 0.48 kcal·mol^{-1} to 0.34 kcal·mol^{-1} for ACE, while MAM deviation remains stationary. The final solvated phase discharging free energies are in line one another and with the BWRF results.

To check further the consistency and sensitivity of the BWRF correction scheme, ACE and MAM solvated phase discharging free energy corrections were evaluated for three different number of points per grid, whose values were suggested by APBS developers: 65, 97 and 129 (tab. **6**). Overall, no appreciable differences can be highlighted for finer grid spacing in APBS and Poisson solvers, which proves the consistency and of reliability of the correction scheme. In particular, both ACE and MAM have a small difference of about 1 kcal·mol^{-1} between the corrected discharging free energy at 10 Å cutoff and 65 grid points and simulations at 14 Å cutoff and 129 grid points, with values ranging between -80.26±0.06 and -81.45±0.09 kcal·mol^{-1} for ACE and -77.87±0.04 and -78.64±0.09 kcal·mol^{-1} for MAM.

Table 4: ACE and MAM solvated phase discharging free energy estimations for the BWRF approach. *mol* indicates the investigated system, cutoff the cutoff length for the system, ΔG^{raw}_{chg} is the solvated phase raw discharging free energy change, ΔG_{POL}, ΔG_{PSUM}, ΔG_{DIR}, are the BWRF correction terms, while ΔG^{corr}_{chg} is the corrected solvated discharging free energy. $\mu \pm \sigma$ indicates the average and standard deviation of ΔG^{raw}_{chg} and ΔG^{corr}_{chg} values. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

BWRF solvated discharging free energy						
mol	cutoff	ΔG^{raw}_{chg}	ΔG_{POL}	ΔG_{PSUM}	ΔG_{DIR}	ΔG^{corr}_{chg}
ACE	10A	-73.04±0.05	-23.05±0.03	15.83±0.01	0.00±0.00	-80.26±0.03
ACE	12A	-77.08±0.01	-19.50±0.02	15.68±0.02	0.00±0.00	-80.90±0.01
ACE	14A	-80.18±0.03	-16.92±0.02	15.71±0.08	0.00±0.00	-81.38±0.12
$\mu \pm \sigma$		-76.76±3.58				-80.84±0.56
MAM	10A	-38.06±0.02	-24.14±0.01	-15.68±0.04	0.00±0.00	-77.87±0.07
MAM	12A	-42.10±0.05	-20.59±0.03	-15.70±0.11	0.00±0.00	-78.38±0.03
MAM	14A	-45.19±0.03	-18.04±0.04	-15.60±0.09	0.00±0.00	-78.82±0.15
$\mu \pm \sigma$		-41.78±3.57				-78.35±0.43

Table 5: ACE and MAM solvated phase discharging free energy estimations for the PME approach. *mol* indicates the investigated system, cutoff the cutoff length for the system, ΔG^{raw}_{chg} is the solvated phase raw discharging free energy change, ΔG_{NET}, ΔG_{USV}, ΔG_{RIP}, ΔG_{DSC} are the Rocklin correction terms, while ΔG^{corr}_{chg} is the corrected solvated discharging free energy. $\mu \pm \sigma$ indicates the average and standard deviation of ΔG^{raw}_{chg} and ΔG^{corr}_{chg} values. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

PME solvated discharging free energy							
mol	cutoff	ΔG^{raw}_{chg}	ΔG_{NET}	ΔG_{USV}	ΔG_{RIP}	ΔG_{DSC}	ΔG^{corr}_{chg}
ACE	10A	-96.97±0.59	-0.01±0.01	0.01±0.01	-2.64±0.01	16.24±0.60	-83.36±0.60
ACE	12A	-97.28±0.29	-0.01±0.01	0.01±0.01	-2.64±0.01	16.29±0.28	-83.64±0.28
ACE	14A	-97.62±0.07	-0.01±0.01	0.01±0.01	-2.63±0.01	16.22±0.07	-84.04±0.08
$\mu \pm \sigma$		-97.19±0.48					-83.68±0.34
MAM	10A	-61.83±0.18	-0.01±0.01	0.01±0.01	-2.41±0.01	-14.77±0.18	-79.03±0.19
MAM	12A	-62.37±0.11	-0.01±0.01	0.01±0.01	-2.41±0.01	-14.78±0.11	-79.57±0.11
MAM	14A	-62.17±0.01	-0.01±0.01	0.01±0.01	-2.41±0.01	-14.74±0.04	-79.32±0.04
$\mu \pm \sigma$		-62.12±0.27					-79.30±0.27

Table 6: Solvated phase discharging free energy for ACE and MAM molecules with BWRF correction approach for a different number of points per grid. *mol* denotes the system under investigation, *Grid size* is the number of points of the grid for the x, y and z direction, *Spacing* is the spacing between grid points in Å, *cutoff* is the cutoff length employed in the simulation, ΔG_{chg}^{raw} the raw solvated discharging free energy, ΔG_{POL}, ΔG_{PSUM} and ΔG_{DIR} are the correction terms in the BWRF scheme and ΔG_{chg}^{corr} is the corrected solvated discharging free energy.

mol	Grid size	Spacing	Cutoff	ΔG_{chg}^{raw}	ΔG_{POL}	ΔG_{PSUM}	ΔG_{DIR}	ΔG_{chg}^{corr}
ACE	65	0.58×0.58×0.58	10Å	-73.04±0.05	-23.05±0.03	15.83±0.01	0.00±0.00	-80.26±0.06
	97	0.39×0.39×0.39		-73.04±0.05	-23.28±0.01	15.83±0.01	0.00±0.00	-80.49±0.05
	129	0.29×0.29×0.29		-73.04±0.05	-23.15±0.01	15.83±0.01	0.00±0.00	-80.36±0.05
ACE	65	0.58×0.58×0.58	12Å	-77.08±0.01	-19.50±0.02	15.68±0.02	0.00±0.00	-80.90±0.03
	97	0.39×0.39×0.39		-77.08±0.01	-19.61±0.09	15.75±0.10	0.00±0.00	-80.94±0.13
	129	0.29×0.29×0.29		-77.08±0.01	-19.59±0.03	15.75±0.10	0.00±0.00	-80.92±0.10
ACE	65	0.58×0.58×0.58	14Å	-80.18±0.03	-16.92±0.02	15.71±0.08	0.00±0.00	-81.39±0.09
	97	0.39×0.39×0.39		-80.18±0.03	-17.10±0.01	15.71±0.08	0.00±0.00	-81.57±0.09
	129	0.29×0.29×0.29		-80.18±0.03	-16.98±0.01	15.71±0.08	0.00±0.00	-81.45±0.09
MAM	65	0.60×0.57×0.57	10Å	-38.06±0.02	-24.14±0.01	-15.68±0.04	0.00±0.00	-77.87±0.04
	97	0.40×0.38×0.38		-38.06±0.02	-23.94±0.01	-15.68±0.04	0.00±0.00	-77.67±0.04
	129	0.30×0.29×0.29		-38.06±0.02	-23.97±0.01	-15.68±0.04	0.00±0.00	-77.70±0.04
MAM	65	0.60×0.57×0.57	12Å	-42.10±0.05	-20.59±0.03	-15.70±0.11	0.00±0.00	-78.38±0.12
	97	0.40×0.38×0.38		-42.10±0.05	-20.48±0.08	-15.70±0.11	0.00±0.00	-78.27±0.14
	129	0.30×0.29×0.29		-42.10±0.05	-20.48±0.08	-15.70±0.11	0.00±0.00	-78.27±0.14
MAM	65	0.60×0.57×0.57	14Å	-45.19±0.03	-18.04±0.04	-15.60±0.09	0.00±0.00	-78.82±0.10
	97	0.40×0.38×0.38		-45.19±0.03	-17.82±0.01	-15.60±0.09	0.00±0.00	-78.61±0.10
	129	0.30×0.29×0.29		-45.19±0.03	-17.85±0.01	-15.60±0.09	0.00±0.00	-78.64±0.09

5.4.3 Functionalized fullerene correction terms

To further test the consistency of the BWRF correction terms, ΔG_{POL}, ΔG_{PSUM} and ΔG_{DIR} values were compared with Reif and Oostenbrink results [82] for the functionalized fullerene system, as shown in tab. **7** and **8**. All the free energy correction terms are in line with Reif calcualtions. ΔG_{POL} is consistent for all the host-guest systems complex and solvated phases, with a maximum discrepancy of 0.8 kcal·mol^{-1} with respect to Reif's results. ΔG_{PSUM} differs of about 1 kcal·mol^{-1} with respect to Reif results, but this is likely due to the use of a different water model (TIP3P in BWRF simulations and SPC in Reif simulations). Finally, ΔG_{DIR} values are in

agreement with Reif calculations for all the charged host-guest systems.

Table 7: Functionalized fullerene charging corrections results for the BWRF scheme of SOMD. *Guest* denotes the guest molecule, ACE or MAM, *Host* if the simulation refers to the complex phase simulation with a fullerene (CAPO), negatively charged fullerene (CNEG) and positively charged fullerene (CNEG) host, ΔG_{POL}, ΔG_{PSUM} and ΔG_{DIR} are the SOMD BWRF correction values. All the free energy terms are in kcal·mol^{-1} and errors were estimated according to eq. 5.24

Guest	Host	ΔG_{POL}	ΔG_{PSUM}	ΔG_{DIR}
ACE	CAPO	-16.37±0.01	15.03±0.03	-0.01±0.01
	CNEG	-49.34±0.11	15.29±0.04	34.18±0.01
	CPOS	15.71±0.13	14.99±0.01	-35.09±0.01
MAM	CAPO	-16.32±0.01	-14.81±0.05	0.00±0.00
	CNEG	16.52±0.01	-15.08±0.09	-34.78±0.01
	CPOS	-49.94±0.18	-14.88±0.14	34.05±0.02

Table 8: Functionalized fullerene charging corrections results for the BWRF scheme computed by Reif and Oostenbrink [82]. *Guest* denotes the guest molecule, ACE or MAM, *Host* if the simulation refers to the complex phase simulation with a fullerene (CAPO), negatively charged fullerene (CNEG) and positively charged fullerene (CNEG) host, ΔG_{POL}, ΔG_{PSUM} and ΔG_{DIR} are the BWRF correction values. All the free energy terms are in kcal·mol^{-1} while errors were not provided in the original paper

Guest	Host	ΔG_{POL}	ΔG_{PSUM}	ΔG_{DIR}
ACE	CAPO	-16.90	16.11	0.00
	CNEG	-50.22	16.18	34.23
	CPOS	16.68	16.23	-34.63
MAM	CAPO	-17.04	-16.20	0.00
	CNEG	16.68	-16.18	-34.68
	CPOS	-50.17	-16.30	34.03

5.4.4 OA host-guest binding free energy

Given the consistency shown by the BWRF correction scheme, the OA host-guest system was studied. Even though the host is relatively rigid, the BWRF correction scheme fails to give accurate free energy estimations (tab. **9**). If on one side negatively charged guests (G1, G2, G4 and G6) present a correction term of about -1 kcal·mol^{-1}, positively charged molecules (G3 and G5) are subjected to an underestimation in binding of about 10 kcal·mol^{-1}. The statistical analysis clearly indicates that corrected standard free energies of binding $\Delta G^{\circ}_{corr-bind}$ have an increment of 3 kcal·mol^{-1} in the MUE with respect to the raw standard binding free energy results, along with a low correlation with R^2=0.63±0.07 and τ=0.58±0.05

Rocklin corrections, tab. **10** gives a useful comparison, to understand whether the BWRF is wrong. In this case, the vanishing free energy change was taken from the BWRF simulations, which seems to be a reasonable approximation as the electrostatic term in this case is reduced to zero. However, also in this case the correction scheme fails to give an accurate estimation. For all the guest systems and overestimation is present in the correction terms, with a maximum binding free energy of -25.29±0.40 kcal·mol^{-1} for G3 and 13.54±0.18 kcal·mol^{-1} for G6. Statistically, PME results are worse than BWRF ones, with a MUE of 14.17±0.12 kcal·mol^{-1} and lower R^2 (0.30±0.06) and τ (-0.19±0.10).

Differently from the BWRF and PME schemes, the alchemical counterions corrections range between -1 and +1 kcal·mol^{-1}, as depicted in tab. **11**. This results in a similar statistical behavior to the raw predictions, with a R^2 0.85±0.05, MUE = 2.63±0.11 kcal·mol^{-1} and τ=0.48±0.12.

Table 9: BWRF binding free energies for the OA dataset of SAMPL5 (fig 1.1 chapter 4). *mol* indicated the host-guest complex, *phase* the complex or solvated simulated phase, ΔG^{raw}_{chg} is the raw discharging free energy, ΔG_{POL}, ΔG_{PSUM} and ΔG_{DIR} are the BWRF correction terms, ΔG_{LJ} is the vanishing step free energy change, $\Delta G^{\circ}_{raw-bind}$ is the standard raw binding free energy, $\Delta G^{\circ}_{corr-bind}$ is the standard binding free energy corrected for the finite size artefacts. Standard state correction magnitude are 1.65 kcal·mol^{-1} on average. ΔG°_{exp} is the experimental binding free energy. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

BWRF OA binding affinities										
mol	phase	net charge	ΔG^{raw}_{chg}	ΔG_{POL}	ΔG_{PSUM}	ΔG_{DIR}	ΔG_{LJ}	$\Delta G^{\circ}_{raw-bind}$	$\Delta G^{\circ}_{corr-bind}$	ΔG°_{exp}
OA-G1	complex	−1	58.20±0.12	-240.27±1.29	12.17±0.01	191.68±0.64	12.79±0.39	-6.92±0.44	-7.89±0.11	-5.39±0.01
	solvated	−1	65.04±0.01	-18.94±0.04	15.46±0.03	-31.98±0.01	-2.63±0.11			
OA-G2	complex	−1	19.92±0.02	-237.05±0.35	12.19±0.05	192.29±0.03	8.79±0.23	-5.26±0.32	-6.30±0.92	-4.73±0.01
	solvated	−1	28.00±0.04	-18.76±0.21	15.50±0.02	-28.27±0.01	-6.20±0.04			
OA-G3	complex	1	32.71±0.01	200.10±0.62	-12.15±0.04	-198.71±0.79	6.60±0.09	-7.68±0.12	3.03±0.52	-4.49±0.01
	solvated	1	37.34±0.28	-19.17±0.01	-14.97±0.04	12.67±0.01	-7.37±0.19			
OA-G4	complex	−1	46.48±0.93	-233.99±0.52	12.01±0.08	200.33±0.33	5.74±0.44	-12.67±1.89	-16.80±1.21	-9.36±0.01
	solvated	−1	51.80±0.01	-18.31±0.04	15.21±0.02	-14.41±0.01	-13.90±0.13			
OA-G5	complex	1	8.06±0.03	203.41±0.01	-12.08±0.03	-204.15±2.31	6.40±0.35	-7.08±0.52	4.59±1.96	-4.50±0.01
	solvated	1	14.03±0.02	-18.93±0.01	-15.07±0.01	9.51±0.01	-8.31±0.17			
OA-G6	complex	−1	16.77±0.09	-237.45±0.73	12.23±0.01	181.80±0.69	6.59±0.06	-6.23±0.15	-7.64±0.23	-5.33±0.01
	solvated	−1	23.23±0.01	-19.54±0.91	15.60±0.18	-38.84±0.01	-7.75±0.09			
R^2								0.80< 0.85< 0.91	0.56< 0.63< 0.70	
MUE								2.00< 2.12< 2.24	5.04< 5.15< 5.27	
τ								0.46< 0.60< 0.73	0.46< 0.58< 0.73	

Table 10: PME binding free energies for the OA dataset of SAMPL5 (fig 1.1 chapter 4). *mol* indicated the host-guest complex, *phase* the complex or solvated simulated phase, ΔG^{raw}_{chg} is the raw discharging free energy, ΔG_{NET}, ΔG_{USV}, ΔG_{RIP} and ΔG_{DSC} are the PME correction terms, $\Delta G^{\circ}_{raw-bind}$ is the standard raw binding free energy, $\Delta G^{\circ}_{corr-bind}$ is the corrected standard binding. ΔG°_{exp} is the experimental binding free energy. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

PME OA binding affinities									
mol	phase	ΔG^{raw}_{chg}	ΔG_{NET}	ΔG_{USV}	ΔG_{RIP}	ΔG_{DSC}	$\Delta G^{\circ}_{raw-bind}$	$\Delta G^{\circ}_{corr-bind}$	ΔG°_{exp}
OA-G1	complex	72.87±0.89	0.01±0.01	-0.01±0.01	14.68±0.04	15.62±0.03	-9.39±0.08	12.21±0.01	-5.39±0.01
	solvated	78.90±0.14	0.01±0.01	-0.01±0.01	-4.11±0.01	12.81±0.01			
OA-G2	complex	84.64±0.42	0.01±0.01	-0.01±0.01	15.95±1.92	15.59±0.01	-7.82±0.33	12.40±2.27	-4.73±0.01
	solvated	91.81±0.39	0.01±0.01	-0.01±0.01	-4.89±0.01	16.22±0.02			
OA-G3	complex	-62.54±0.34	0.01±0.01	-0.01±0.01	-14.45±0.05	-15.58±0.05	-9.17±0.69	-25.29±0.40	-4.49±0.01
	solvated	-57.75±1.03	0.01±0.01	-0.01±0.01	-3.14±0.01	-10.78±0.01			
OA-G4	complex	68.28±0.17	0.01±0.01	-0.01±0.01	14.60±0.10	15.57±0.05	-14.17±0.06	5.52±0.14	-9.36±0.01
	solvated	73.73±0.61	0.01±0.01	-0.01±0.01	-4.17±0.02	14.64±0.04			
OA-G5	complex	-34.54±0.81	0.01±0.01	-0.01±0.01	-14.42±0.18	-15.65±0.07	-9.34±0.24	-20.21±1.54	-4.50±0.01
	solvated	-29.18±1.35	0.01±0.01	-0.01±0.01	-3.84±0.01	-15.36±1.90			
OA-G6	complex	111.48±0.99	0.01±0.01	-0.01±0.01	14.64±0.09	15.63±0.01	-8.11±0.13	13.54±0.18	-5.33±0.01
	solvated	117.70±0.87	0.01±0.01	-0.01±0.01	-3.89±0.02	12.50±0.05			
R^2							0.81< 0.86< 0.91	0.23< 0.29< 0.35	
MUE							4.03< 4.14< 4.26	14.05< 14.17< 14.29	
τ							0.33< 0.48< 0.60	-0.30< -0.20< -0.01	

Table 11: BWRF simulations for the OA Clip dataset of SAMPL5 (fig 1.1 chapter 4) with the alchemical counterions approach. *mol* indicates the host-guest complex, $\Delta G^{\circ}_{raw-bind}$ is the raw standard binding free energy, $\Delta G^{\circ}_{corr-bind}$ is the standard binding free energy corrected for the finite size artefacts. ΔG°_{exp} is the experimental binding free energy. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

mol	$\Delta G^{\circ}_{raw-bind}$	$\Delta G^{\circ}_{corr-bind}$	ΔG°_{exp}
OA-G1	−7.15±0.41	−6.82±0.33	−5.39±0.01
OA-G2	−6.44±0.26	−6.67±0.51	−4.73±0.01
OA-G3	−8.12±0.09	−7.83±0.13	−4.49±0.01
OA-G4	−12.12±0.19	−12.25±0.22	−9.36±0.01
OA-G5	−8.17±0.34	−7.89±0.37	−4.50±0.01
OA-G6	−7.34±0.29	−7.43±0.19	−5.33±0.01
R^2	0.80< 0.85< 0.91	0.79< 0.85< 0.90	
MUE	2.00< 2.12< 2.24	2.51< 2.63< 2.74	
τ	0.46< 0.60< 0.73	0.33< 0.48< 0.60	

5.4.5 CBC host-guest binding free energy

As for the OA systems, the BWRF correction gives overestimated correction terms, tab. **12**. On average the total correction is about -7 kcal·mol^{-1} for most of the system, while it is +13 kcal·mol^{-1} for G7 guest. Statistically, the corrected binding free energies $\Delta G^{\circ}_{corr-bind}$ are less precise and accurate than the raw standard ones $\Delta G^{\circ}_{raw-bind}$, with an increment of about 4 kcal·mol^{-1} in the MUE (9.40±0.50 kcal·mol^{-1}).

The statistical accuracy of predictions is even worse for the case of PME predictions, as tab. **13** shows. Also in this set of simulations the vanishing step was run with SOMD under BWRF. Although the $\Delta G^{\circ}_{raw-bind}$ statistics is roughly similar to the BWRF one, again the corrected $\Delta G^{\circ}_{corr-bind}$ gets less precise with an increment of 5 kcal·mol^{-1} in the MUE (9.38±0.50 kcal·mol^{-1}). Also for PME simulations the correction schemes gives unreasonable free energies corrections ranging between -7 to +7 kcal·mol^{-1}.

Table 12: BWRF binding free energies for the CBC dataset of SAMPL5 (fig 1.1 chapter 4). *mol* indicates the host-guest complex, *phase* the complex or solvated phase, ΔG^{raw}_{chg} is the raw discharging free energy, ΔG_{POL}, ΔG_{PSUM} and ΔG_{DIR} are the BWRF correction terms, ΔG_{LJ} is the vanishing step free energy change, $\Delta G^{\circ}_{raw-bind}$ is the standard raw binding free energy, $\Delta G^{\circ}_{corr-bind}$ is the standard binding free energy corrected for the finite size artefacts. ΔG°_{exp} is the experimental binding free energy. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

BWRF CB7Clip binding affinities										
mol	phase	net charge	ΔG^{raw}_{chg}	ΔG_{POL}	ΔG_{PSUM}	ΔG_{DIR}	ΔG_{LJ}	$\Delta G^{\circ}_{raw-bind}$	$\Delta G^{\circ}_{corr-bind}$	ΔG°_{exp}
CBC-G1	complex	2	26.64±0.01	170.10±0.19	-25.58±0.02	-260.97±0.50	2.39±0.01	-8.50±0.05	-17.79±0.58	-5.84±0.01
	solvated	2	24.68±0.02	-75.69±0.05	-31.48±0.01	0.00±0.00	-5.40±0.01			
CBC-G2	complex	2	34.92±0.12	171.17±1.02	-25.81±0.10	-262.27±2.36	2.19±0.34	-1.90±0.18	-8.37±2.40	-2.52±0.01
	solvated	2	36.40±0.01	-78.61±0.15	-31.84±0.11	0.00±0.00	-2.42±0.01			
CBC-G3	complex	4	17.67±0.28	216.50±0.84	-51.35±0.05	-513.28±0.51	5.64±0.08	-6.06±0.45	-13.03±0.25	-4.02±0.01
	solvated	4	19.13±0.02	-279.00±0.19	-62.17±0.20	0.00±0.00	-3.10±0.02			
CBC-G4	complex	1	105.90±0.12	101.66±0.29	-12.61±0.01	-127.30±0.16	-2.46±0.67	-11.19±0.82	-16.58±0.49	-7.25±0.01
	solvated	1	102.99±0.02	-17.80±0.02	-15.06±0.14	0.00±0.00	-12.04±0.04			
CBC-G5	complex	2	11.31±0.01	101.16±0.18	-12.59±0.07	-125.91±0.05	-12.92±0.23	-13.80±0.18	-18.88±0.18	-8.54±0.01
	solvated	2	7.90±0.02	-17.34±0.04	-14.92±0.07	0.00±0.00	-24.54±0.04			
CBC-G6	complex	2	-2.95±0.01	169.22±0.44	-25.59±0.03	-244.88±0.09	-8.60±0.28	-17.47±0.10	-24.34±0.20	-8.68±0.01
	solvated	2	-6.67±0.01	-63.88±0.11	-30.50±0.09	0.00±0.00	-23.56±0.16			
CBC-G7	complex	−2	171.78±0.05	-278.21±0.67	26.78±0.04	234.21±0.71	-9.36±1.08	-15.37±1.23	-2.08±0.64	-5.18±0.01
	solvated	−2	170.40±0.13	-60.74±0.36	30.23±0.06	0.00±0.00	-24.64±0.10			
CBC-G8	complex	1	-38.98±0.02	97.25±0.14	-12.39±0.01	-119.35±0.04	-39.69±1.20	-4.42±1.21	-7.80±0.85	-6.18±0.01
	solvated	1	-41.98±0.02	-16.22±0.01	-14.90±0.05	0.00±0.00	-42.33±0.06			
CBC-G9	complex	1	-17.54±0.01	102.31±0.06	-12.70±0.02	-126.81±0.24	-6.33±0.72	-15.47±0.78	-20.03±0.18	-7.40±0.01
	solvated	1	-21.48±0.01	-17.45±0.01	-15.20±0.09	0.00±0.00	-19.09±0.06			
CBC-G10	complex	1	-6.14±0.01	99.84±0.29	-12.52±0.02	-126.72±0.11	-9.65±0.04	-17.86±0.05	-24.70±0.25	-10.38±0.01
	solvated	1	-14.12±0.01	-17.42±0.01	-15.15±0.09	0.00±0.00	-20.85±0.01			
R^2								0.55< 0.64< 0.73	0.47< 0.57< 0.66	
MUE								4.75< 5.09< 5.43	8.93< 9.38< 9.84	
τ								0.68< 0.76< 0.82	0.73< 0.78< 0.82	

As for the OA dataset, the alchemical counterions approach, tab. **14**, gives much better results than the Poisson based calculations. In this scheme, corrections are ranging between -2 and -1 kcal·mol^{-1}. In particular, the counterions calculations perform very well for highly charged guests molecules, as for G2 (+2e), whose $\Delta G^{\circ}_{corr-bind} = -1.84 \pm 0.01$ kcal·mol^{-1} is in line with respect to the experimental value of -2.50 ± 0.07 kcal·mol^{-1}, and for G3 (+4e) ($\Delta G^{\circ}_{corr-bind} = -4.52 \pm 0.24$ kcal·mol^{-1} and -4.02±0.03 kcal·mol^{-1} experimentally). Overestimation in $\Delta G^{\circ}_{corr-bind}$ are still present for G9 and G10 guests (-15.16±0.46 kcal·mol^{-1} and -17.57±0.05 kcal·mol^{-1} respectively). The alchemical counterions approach achieves a lower MUE with respect to the raw estimations(4.70±0.26 kcal·mol^{-1}) and is overall superior to the other correction approaches.

Table 13: PME binding free energies for the CBC dataset of SAMPL5 (fig 1.1 chapter 4). *mol* indicated the host-guest complex, *phase* the complex or solvated simulated phase, ΔG^{raw}_{chg} is the raw discharging free energy, ΔG_{NET}, ΔG_{USV}, ΔG_{RIP} and ΔG_{DSC} are the PME correction terms, $\Delta G^{\circ}_{raw-bind}$ is the standard raw binding free energy, $\Delta G^{\circ}_{corr-bind}$ is the corrected standard binding. ΔG°_{exp} is the experimental binding free energy. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24

PME CB7Clip binding affinities									
mol	phase	ΔG^{raw}_{chg}	ΔG_{NET}	ΔG_{USV}	ΔG_{RIP}	ΔG_{DSC}	$\Delta G^{\circ}_{raw-bind}$	$\Delta G^{\circ}_{corr-bind}$	ΔG°_{exp}
CBC-G1	complex	116.99±0.01	0.01±0.01	-0.01±0.01	-17.04±0.02	-31.4±0.02	-7.62±0.02	-15.43±0.02	-5.84±0.01
CBC-G1	solvated	117.16±0.01	0.01±0.01	-0.01±0.01	-8.68±0.0	-31.95±0.02			
CBC-G2	complex	102.92±0.01	0.01±0.01	-0.01±0.01	-14.27±0.16	-31.4±0.06	-1.86±0.34	-6.76±0.52	-2.52±0.01
CBC-G2	solvated	105.42±0.01	0.01±0.01	-0.01±0.01	-8.69±0.03	-32.08±0.02			
CBC-G3	complex	170.32±0.01	0.01±0.01	-0.01±0.01	-28.03±0.03	-62.64±0.09	-4.38±0.08	3.41±0.32	-4.02±0.01
CBC-G3	solvated	174.60±0.01	0.01±0.01	-0.01±0.01	-34.61±0.16	-63.85±0.03			
CBC-G4	complex	-50.98±0.01	0.01±0.01	-0.01±0.01	-8.63±0.01	-15.71±0.01	-11.85±0.67	-18.03±0.71	-7.25±0.01
CBC-G4	solvated	-53.70±0.01	0.01±0.01	-0.01±0.01	-2.19±0.01	-15.97±0.01			
CBC-G5	complex	9.58±0.01	0.01±0.01	-0.01±0.01	-8.64±0.01	-15.66±0.01	-14.88±0.23	-21.06±0.24	-8.52±0.01
CBC-G5	solvated	6.45±0.01	0.01±0.01	-0.01±0.01	-2.19±0.01	-15.93±0.01			
CBC-G6	complex	6.52±0.01	0.01±0.01	-0.01±0.01	-16.90±0.08	-31.29±0.02	-18.80±0.28	-26.60±0.76	-8.68±0.01
CBC-G6	solvated	2.59±0.01	0.01±0.01	-0.01±0.01	-8.55±0.01	-31.84±0.03			
CBC-G7	complex	6.21±0.01	0.01±0.01	-0.01±0.01	-17.18±0.01	-31.32±0.02	-17.74±1.08	-25.87±1.60	-5.18±0.01
CBC-G7	solvated	3.06±0.01	0.01±0.01	-0.01±0.01	-8.55±0.02	-31.83±0.01			
CBC-G8	complex	6.67±0.01	0.01±0.01	-0.01±0.01	-17.18±0.01	-31.33±0.01	-7.91±1.20	-16.07±1.26	-6.18±0.01
CBC-G8	solvated	2.29±0.01	0.01±0.01	-0.01±0.01	-8.55±0.01	-31.80±0.03			
CBC-G9	complex	6.91±0.01	0.01±0.01	-0.01±0.01	-16.88±0.10	-31.25±0.01	-17.28±0.72	-25.02±1.06	-7.40±0.01
CBC-G9	solvated	2.94±0.01	0.01±0.01	-0.01±0.01	-8.55±0.01	-31.84±0.01			
CBC-G10	complex	42.57±0.01	0.01±0.01	-0.01±0.01	-8.69±0.01	-15.69±0.01	-12.16±0.04	-18.44±0.19	-10.38±0.01
CBC-G10	solvated	34.90±0.01	0.01±0.01	-0.01±0.01	-2.18±0.01	-15.92±0.01			
R^2							0.38 < 0.44< 0.50	0.35 < 0.41< 0.48	
MUE							4.68 < 4.99< 5.30	10.76< 11.19< 11.62	
τ							0.51 < 0.59< 0.64	0.46< 0.53< 0.60	

Table 14: BWRF simulations for the CB7 Clip dataset of SAMPL5 (fig 1.1 chapter 4) with the alchemical counterions approach. *mol* indicates the host-guest complex, $\Delta G^{\circ}_{raw-bind}$ is the raw standard binding free energy, $\Delta G^{\circ}_{corr-bind}$ is the standard binding free energy corrected for the finite size artefacts. ΔG°_{exp} is the experimental binding free energy. All the values are in kcal·mol^{-1} and standard errors are computed with eq. 5.24.

mol	$\Delta G^{\circ}_{raw-bind}$	$\Delta G^{\circ}_{corr-bind}$	ΔG°_{exp}
CBC-G1	−8.52±0.02	−7.99±0.04	−5.84±0.01
CBC-G2	−1.93±0.13	−1.84±0.01	−2.52±0.01
CBC-G3	−6.00±0.35	−4.52±0.24	−4.02±0.01
CBC-G4	−11.28±0.59	−11.25±0.62	−7.25±0.01
CBC-G5	−13.78±0.14	−13.57±0.15	−8.54±0.01
CBC-G6	−17.37±0.08	−16.67±0.41	−8.68±0.01
CBC-G7	−15.45±0.92	−14.79±1.10	−5.18±0.01
CBC-G8	−4.32±0.81	−4.10±0.74	−6.18±0.01
CBC-G9	−15.46±0.52	−15.16±0.46	−7.40±0.01
CBC-G10	−17.93±0.04	−17.57±0.05	−10.36±0.01
R^2	0.55< 0.64< 0.73	0.59< 0.66< 0.74	
MUE	4.75< 5.09< 5.43	4.43< 4.70< 4.96	
τ	0.68< 0.76< 0.82	0.73< 0.77< 0.82	

5.5 Conclusions

Free energy calculations of charged species are a frequent task in the drug design process. These calculations are usually carried on with periodic boundary conditions and effective Coulombic potentials. Furthermore, solvent models may not reflect exactly the physical properties of the bulk solvent. All these approximations could affect the free energy calculation with the introduction of a bias potential which hinders a proper comparison with experiments. Throughout this chapter a correction scheme was introduced for BWRF based simulations. This scheme was then compared with Rocklin approach [81] for PME simulations. Finally, for host-guest systems, simulations with alchemical counterions were run to retrieve corrected free energies and establish a comparison with the BWRF and PME corrections.

Results were encouraging for hydration free energy calculations. Tests on sodium ion assured the consistency and reliability of BWRF and PME correction schemes. Overall, the estimated hydration free energies were in line with experimental values ($\Delta G^{hyd}_{corr} = -104$ kcal·mol^{-1} on average). The solvation discharging free energy for acetate (ACE) and methylammonium (MAM) gave a further validation for BWRF and PME schemes. Furthermore, ACE and MAM systems were used to check the BWRF correction terms sensitivity to the number of grid points. It was found that variation in grid points do not influence the total free energy correction and all the values converge towards similar values.

The extension of these two schemes to the binding free energy for OA and CBC systems from SAMPL5 was problematic. In both cases the BWRF and PME scheme fail to give reasonable binding free energies. In particular, the BWRF attained a worse MUE for both host-guest systems (MUE 5.2±0.1 kcal·mol^{-1} for OA and MUE 9.4±0.5 kcal·mol^{-1} for CBC). PME simulations showed a even more serious discharging free energy correction.

On the contrary, the BWRF alchemical counterions simulations gave more promising results to correct finite size artefacts in host-guest systems. For the OA dataset results were in line with the raw binding free energy estimations (MUE = 2.6±0.1 kcal·mol^{-1}, R^2 = 0.8±0.1 and τ=0.5±0.1) and a sensible improvement of about 0.5 kcal·mol^{-1} was attained for the CBC with respect to the raw binding vaules. Although these results show a better trend with experimental data, further testing should be carried out to check the convergence properties of this approach.

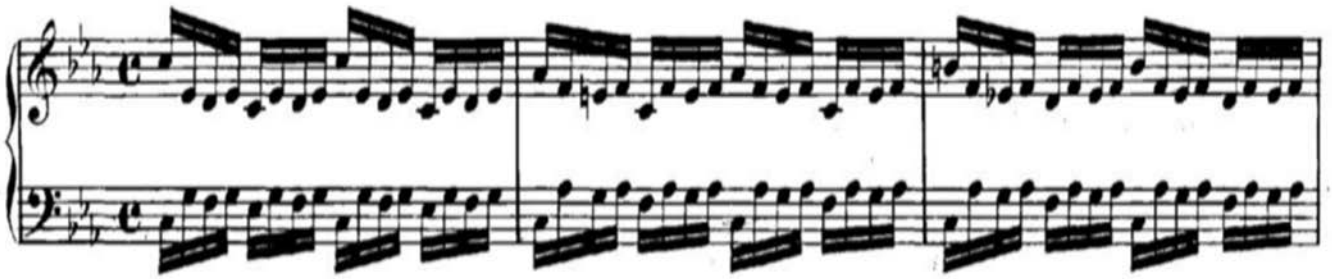
The aim of this work was to show the reliability of the current correction schemes for alchemical free energy simulations of charged species. For the very first time a BWRF atom-based cutoff correction scheme was imple-

mented and tested. Although a good agreement with literature studies, the BWRF correction approach failed to give reasonable values for the host-guest binding affinity calculations. This outcome could highlight the presence of a flaw in the theoretical framework. As a matter of fact, literature studies [81, 26, 243] lack of tests on highly charged host-guest systems. Guests usually bear a net-charge ranging from -1 e to +1 e and host has a rigid structure. Additionally, as regards the BWRF group-based cutoff correction scheme of Reif and Oostenbrink [82], the cutoff dependence of the Poisson solver was not investigated and the study is limited to artificial host-guest systems, which do not have experimental data. These aspects make the investigation upon SAMPL5 host-guest system complicated and impossible to rely on model validation with experimental data. Therefore further work should be carried out on new host-guest datasets to delineate a proper comparison between correction schemes, achieving a complete evaluation of the efficiency of Poisson-Boltzmann solvers. As a first step, neutral and rigid host with highly charged guests should be studied and compared with experimental value, in order to shed light on the limits of these correction schemes. After this step, the correction approach may be extended to highly charged host-guest system, allowing a classification of the all possible study cases.

Then, attention should be devoted to the force field parametrization. As a matter of fact, this study was focused on the use of GAFF [117] for host-guest systems parametrization, differently from literature studies. Since GAFF tends to be highly transferable between molecular models, the general force field may be already empirically account for a correction term for the finite size artefacts. As an example, the OA dataset shows a hydrophobic binding cavity. For all the guest molecules the raw binding free energy is in line with the experimental trend, while the PB correction results in an underestimation in binding.

Finally, this work could be further developed in terms of software development. Initially, the OpenMM PME script should be extended with the implementation of the soft-core potential, to compute the vanishing free energy changes. This would give a more meaningful comparison with BWRF raw free energy estimations. Furthermore, these simulations could help to understand the possible GAFF parametrization effect on host-guest binding calculations, comparing Rocklin correction scheme with the BWRF atom-based cutoff corrections. Secondly, Sire core libraries should be extended to deal with multi-alchemical systems at the same time. This approach would be greatly beneficial for the alchemical counterions scheme, avoiding the creation of multiple simulations files and ad-hoc morphing files. This would reduce the presence of errors introduced by the user in the creation of the input files. Lastly, further test should be carried out on the BWRF-Poisson

solver code and Sire BWRF correction code, to isolate possible bug which may affect the final correction estimation.



Chapter 6

Conclusions

Alchemical free energy calculations (AFE) are a promising technique that could be beneficial for the hit-to-lead and lead optimization stage in the drug discovery process. Although AFE have known a great success in pharmaceutical industries in the last decades, the reproducibility of results, the presence of finite size artefacts and the lack of robust protocols make these calculations far form being a reliable engineering tool that could be routinely used. Thus, this work presented a deep analysis of specific AFE hurdles with the hope that protocols and methods could be further extended in the future, and cover more relevant case studies for the medicinal chemistry.

Initially, chapter 2 showed to what extent AFE calculations are reproducible among different molecular simulation codes. Through a large scale collaborative effort with H. Loeffler (CCPBiosim), D. Mobley and G. Ratos (UCI), B. Roux and D. Suh (UCHI) and our group in Edinburgh, the relative hydration free energies for a small dataset of organic molecules was computed to assess the reproducibility between AMBER, GROMACS, CHARMM and SOMD. The simulation package results proved that it was impossible to define a universal simulation protocol, but specific protocols for each software were devised, achieving a reproducibility-limit of 0.2 kcal·mol^{-1}. The reproducibility was mainly hindered by specific details of the AFE implementations,. For instance the electrostatic contribution in vacuum phase, or the different way the reaction field was implemented (e.g. SOMD and GROMACS). Furthermore, for SOMD the use of all-bonds constraints mainly

affected the free energy predictions, with discrepancies between absolute and relative free energy calculations. This suggested a new constraint scheme, where only the perturbed covalent hydrogen bonds were constrained, to maintain an integration time step of 2 fs. Beyond the protocol validation and code analysis, this work could be employed as a new validation test for existing and future versions of molecular simulation codes, in order to improve the transferability of AFE protocols among different codes. This work could be further extended to investigate the reproducibility of solvation free energies for charged molecules. It would also be desirable to proceed with a study of reproducibility for host-guest systems. All these efforts would facilitate the creation of a reproducible alchemical free energy workflows.

Then, chapter 3 introduced new simulation protocols for lipophilicity coefficients estimations employing AFE techniques. In the first part, AFE were used to compute the experimental log P of five molecules with an increasing fluorine content. As a result, AFE were able to mimic the experimental trend (τ 0.50$\pm$0.10 and MUE 0.77$\pm$0.07 log P units), and trajectory analysis allowed to understand the molecular interactions that drove the partitioning process. As an extension of the log P calculations, two AFE protocols were devised to predict log D values for 53 small drug-like molecules for the blinded challenge SAMPL5. The first protocol, *all-neutral* model, considered all the molecules as neutral species. This model achieved a reasonable MUE 2.20$\pm$0.08 log D units, but it failed to predict log D for ionisable species. Thus, the *two-species* model was devised, to assume the co-existence of two species in organic and aqueous phases for ionisable molecules. This model lowered the MUE to 1.98$\pm$0.19 log D units and retrospectively it was among the top-ranked submission of SAMPL5. In spite of this success, the *two-species* protocol could be further tuned to achieve more reliable and accurate estimations to match QSAR accuracy. Future work and efforts should be devoted to consider all the possible protonated and tautomeria states of a molecule. Secondly, tests should be run to understand the influence on the log D values for cases where a water molecule migrate into the organic phase together with the solute and viceversa. Finally, a detailed analysis should be carried out to understand the influence of the finite size artefacts, and of the force field parametrization onto the solvation free energy estimates. As part of an experimental study, it would be extremelly useful to devise datasets that enable testing of all these sources of errors. As rough sketch such a dataset should consists of a batch of non-ionisable compounds where the predictions of log D = log P , which could be used to ensure any forcefield. Once this is established more complicated systems with multiple equilibria of different species can then be tackled.

Another validation of AFE protocols was presented in chapter 4, where

three different simulation strategies for host-guest binding affinities predictions were used. *Model A* computed the free energy of binding as a difference of free energy changes between complex and solvated phases; *Model B* added the long-range dispersion interactions corrections to *model A*; *model C* added a standard state for each host-guest system to the previous model. All the three protocols were among the top-ranked submissions of SAMPL5 for R^2 metrics. *Model C* was the best predictor with a MUE= 3.39±0.18 kcal·mol^{-1} and R^2 =0.69±0.05. In addition, the standard binding free energy predictions were reproducible with results produced by Gilson lab (UCSD) using a different code and methodology. This study has already led to further developments, thanks to the SAMPL6 challenge in 2018. *Model C* values were improved through a regression protocol, that could correct the free energy estimations from systematic errors such as errors for finite size artefacts and force field parametrization. Results showed a more accurate binding free energy estimations with MUE $<$ 1.0 kcal·mol^{-1} for OA and TEMOA. There is hope that this protocol could be further enhanced by considering the presence of a phosphate buffer in the experimental setup, studying its influence on the free energy estimations.

Finally, chapter 5 presented the finite size artefacts problem in AFE calculations. For the first time a correction scheme for BWRF atom-based simulations was devised and compared with literature protocols, which makes use of PME and BWRF group-based approaches. Initially, the BWRF was validated for hydration free energy estimations of the sodium ion, with analyses carried out for different box sizes and cutoff lengths. Once the consistency of the BWRF results was tested, the correction scheme was applied to polyatomic species, such as acetate and methylammonium, in order to correct the solvation discharging free energies. Results were consistent with one another and in line with literature values. Hence, the BWRF was tested for host-guest binding free energy predictions, using the OA and CBC dataset of SAMPL5 as case study. In this occasion the BWRF scheme failed to give a reasonable result, with overestimations in binding and inconsistency with experimental results. To further understand the finite size magnitude in host-guest systems a protocol, which makes use of alchemical counterions to keep the net-charge constant in the simulated box, was tested. The alchemical counterion approach gave better results than the raw free energy estimation and the BWRF correction protocols, showing a good agreement with experimental results in both dataset. Although the BWRF scheme was unsuccessful, this study indicated several paths to follow to achieve a complete understanding of the finite size problem. First, this work presented new tests for finite size correction schemes and it pointed out a limit of literature correction schemes for host-guest cases. The work used a dataset with pecu-

liarities still unexplored in literature, such as the presence of a very flexible host molecule and highly charged guests. There is hope that the alchemical counterion approach could also be extended and implemented in the SOMD code, to deal with more alchemical transformation at the same time, in order to have an easier simulation setup and maybe more consistent results with respect to the BWRF correction approach. Finally, the influence of GAFF force field on the finite size artefacts should be investigated both in the host-guest case and in the solvation free energy simulations, to understand to what extent GAFF can correct these errors and how the BWRF correction scheme could be modified.

Finally, the reader may wonder why musical scores are present at the beginning of each chapter. These are all Bach compositions, which guided me throughout the writing-up of this thesis. Each musical score has a peculiarity which reflects the subject presented and discussed in each chapter. The introduction was combined with the first Aria of the Goldberg variations. This piece reflects a smooth introduction into the AFE problems and it presents also technical difficulties, which could reflect the (sometimes) complicated nature of the statistical mechanics. The second chapter is described by the English Suite no. 2. This piece uses the same notes for the right hand and left hand and it was felt appropriate to introduce the concept of reproducibility. The Passacaglia in C minor was chosen for the third chapter, to show the complicated nature of the log D estimations. Beyond the two hands, this score has a part for the pedal, which is necessary to give the right emphasis and power to the composition. Thus, the pedal cannot be neglected when the composition is played, as the protonated and tautomeric states of a molecule could not be neglected to predict the correct log D value. Chapter 4 was described by the Prelude in C minor from the Partita no. 2, as this is an example of the *standard* use of the counterpoint, widely used during Bach's era, which reflects the use of *standard state* correction for host-guest binding affinities. Given the more complicated discussion of chapter 5, the third part of the Fantasia BWV 542 was selected, to give the idea of a complicated subject and study such as the finite size problem. Finally, the conclusions are described by the Prelude no. 2 in C minor BWV 847, which is a versatile composition, used countless times to play variations onto Bach music. This symbolizes the variations and future studies that can be taken starting from the protocols presented throughout this thesis.

Bibliography

[1] Thomas A Ban. The role of serendipity in drug discovery. *Dialogues in clinical neuroscience*, 8(3):335, 2006.

[2] Domagk Gerhard. Ein beitrag zur chemotherapie der bakteriellen infektionen. *DMW-Deutsche Medizinische Wochenschrift*, 61(07):250–253, 1935.

[3] Walter Sneader. *Drug discovery: a history*. John Wiley & Sons, 2005.

[4] Paul Ehrlich. Croonian lecture. on immunity with special reference to cell life. *Proceedings of the Royal Society of London*, 66(424-433):424–448, 1900.

[5] Ana Cecília A Roque. Ligand-macromolecular interactions in drug discovery. *Clifton*, 572, 2010.

[6] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[7] Andrew J McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585, 1977.

[8] Hugo Kubinyi. *3D QSAR in drug design*, volume 1. Springer Science & Business Media, 1993.

[9] David E Shaw, Martin M Deneroff, Ron O Dror, Jeffrey S Kuskin, Richard H Larson, John K Salmon, Cliff Young, Brannon Batson, Kevin J Bowers, Jack C Chao, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

[10] Joshua A Anderson, Chris D Lorenz, and Alex Travesset. General purpose molecular dynamics simulations fully implemented on graphics processing units. *Journal of Computational Physics*, 227(10):5342–5359, 2008.

[11] Mark S Friedrichs, Peter Eastman, Vishal Vaidyanathan, Mike Houston, Scott Legrand, Adam L Beberg, Daniel L Ensign, Christopher M Bruns, and Vijay S Pande. Accelerating molecular dynamic simulation on graphics processing units. *Journal of Computational Chemistry*, 30(6):864–872, 2009.

[12] John E Stone, David J Hardy, Ivan S Ufimtsev, and Klaus Schulten. Gpu-accelerated molecular modeling coming of age. *Journal of Molecular Graphics and Modelling*, 29(2):116–125, 2010.

[13] Janet Woodcock and Raymond Woosley. The fda critical path initiative and its influence on new drug development. *Annual Review of Medicine*, 59:1–12, 2008.

[14] Serge Mignani, Scot Huber, Helena Tomas, Joao Rodrigues, and Jean-Pierre Majoral. Why and how have drug discovery strategies in pharma changed? what are the new mindsets? *Drug Discovery Today*, 21(2):239–249, 2016.

[15] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug discovery*, 9(3):203, 2010.

[16] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature Reviews Drug discovery*, 11(3):191, 2012.

[17] Peter M Colman. Structure-based drug design. *Current opinion in Structural Biology*, 4(6):868–874, 1994.

[18] Tom L Blundell. Structure-based drug design. *Nature*, 384(6604):23, 1996.

[19] Mark A Murcko, Paul R Caron, and Paul S Charifson. Structure-based drug design. In *Annual Reports in Medicinal Chemistry*, volume 34, pages 297–306. Elsevier, 1999.

[20] Tomas Hansson, John Marelius, and Johan Åqvist. Ligand binding affinity prediction by linear interaction energy methods. *Journal of Computer-Aided Molecular Design*, 12(1):27–35, 1998.

[21] Frederick S Lee, Zhen-Tao Chu, Michael B Bolger, and Arieh Warshel. Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to mcpc603. *Protein Engineering, Design and Selection*, 5(3):215–228, 1992.

[22] Nadine Homeyer and Holger Gohlke. Free energy calculations by the molecular mechanics poisson- boltzmann surface area method. *Molecular Informatics*, 31(2):114–122, 2012.

[23] Robert W Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *Journal of Chemical Physics*, 22(8):1420–1426, 1954.

[24] John G. Kirkwood. Statistical mechanics of fluid mixtures. *Journal Of Chemical Physics*, 3(1935):300–313, 1935.

[25] William L. Jorgensen and C. Ravimohan. Monte carlo simulation of differences in free energies of hydration. *Journal of Chemical Physics*, 83(6):3050–3054, 1985.

[26] Matteo Aldeghi, Alexander Heifetz, Michael J Bodkin, Stefan Knapp, and Philip C Biggin. Accurate calculation of the absolute free energy of binding for drug molecules. *Chemical Science*, 7(1):207–218, 2016.

[27] Kevin Hauser, Christopher Negron, Steven K Albanese, Soumya Ray, Thomas Steinbrecher, Robert Abel, John D Chodera, and Lingle Wang. Predicting resistance of clinical abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Communications Biology*, 1(1):70, 2018.

[28] Joseph T Kim, Andrew D Hamilton, Christopher M Bailey, Robert A Domoal, Ligong Wang, Karen S Anderson, and William L Jorgensen. Fep-guided selection of bicyclic heterocycles in lead optimization for non-nucleoside inhibitors of hiv-1 reverse transcriptase. *Journal of the American Chemical Society*, 128(48):15372–15373, 2006.

[29] Ludwig Boltzmann. On certain questions of the theory of gases. *Nature*, 51(1322):413, 1895.

[30] Ludwig Boltzmann. *Lectures on gas theory*. Courier Corporation, 2012.

[31] Josiah Willard Gibbs and Henry Andrews Bumstead. *Thermodynamics*, volume 1. Longmans, Green and Company, 1906.

[32] Donald A. McQuarrie. Statistical mechanics, 1976.

[33] Mark Tuckerman. *Statistical mechanics: theory and molecular simulation*. Oxford University Press, New York, 2010.

[34] I Aleksandr and Akovlevich Khinchin. *Mathematical foundations of statistical mechanics*. Courier Corporation, 1949.

[35] George D Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

[36] Roberto Piazza. *Statistical Physics: A Prelude and Fugue for Engineers*. Springer, 2016.

[37] Born Max and Oppenheimer Robert J. *Zur quantentheorie der molekeln*, volume 389. Wiley Online Library, 1927.

[38] Wales D. *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

[39] Pettitt B. M. Brooks III CL., Karplus M. *Advances in Chemical Physics, Proteins: A Theoretical Perspective of Dynamics*, volume 71.

[40] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2001.

[41] Brock A Luty, Ilario G Tironi, and Wilfred F van Gunsteren. Lattice-sum methods for calculating electrostatic interactions in molecular simulations. *Journal of Chemical Physics*, 103(8):3014–3021, 1995.

[42] Dominik Sidler, Simon Frasch, Michael Cristòfol-Clough, and Sereina Riniker. Anisotropic reaction field correction for long-range electrostatic interactions in molecular dynamics simulations. *Journal of Chemical Physics*, 148(23):234105, 2018.

[43] Lars Onsager. Electric moments of molecules in liquids. *Journal of the American Chemical Society*, 58(8):1486–1493, 1936.

[44] John G Kirkwood. Theory of solutions of molecules containing widely separated charges with special application to zwitterions. *Journal of Chemical Physics*, 2(7):351–361, 1934.

[45] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *Journal of Chemical Physics*, 103(19):8577–8593, 1995.

[46] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

[47] Berni J Alder. Studies in molecular dynamics. i. general method. *Journal of Chemical Physics*, 31(2):459–466, 1959.

[48] Alfio Quarteroni and Alberto Valli. *Numerical Approximation of Partial Differential Equations*. Springer, 1994.

[49] JA Barker and RO Watts. Monte carlo studies of the dielectric properties of water-like models. *Molecular Physics*, 26(3):789–792, 1973.

[50] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review Letters*, 159(1):98, 1967.

[51] Herman JC Berendsen, James PM Postma, Wilfred F van Gunsteren, A RHJ DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics*, 81(8):3684–3690, 1984.

[52] Hans C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *Journal of Chemical Physics*, 72(4):2384–2393, 1980.

[53] Michael P Allen. Introduction to molecular dynamics simulation. *Computational soft matter: from synthetic polymers to proteins*, 23:1–28, 2004.

[54] Herman JC Berendsen, J. Raul Grigera, and P. Tjerk Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24):6269–6271, 1987.

[55] Kim-Hung Chow and David M Ferguson. Isothermal-isobaric molecular dynamics simulations with monte carlo volume sampling. *Computer Physics Communications*, 91(1-3):283–289, 1995.

[56] Johan Åqvist, Petra Wennerström, Martin Nervall, Sinisa Bjelic, and Bjørn O Brandsdal. Molecular dynamics simulations of water and biomolecules with a monte carlo constant pressure algorithm. *Chemical Physics Letters*, 384(4-6):288–294, 2004.

[57] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[58] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2):255–268, 1984.

[59] Vincent Kräutler, Wilfred F Van Gunsteren, and Philippe H Hünenberger. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*, 22(5):501–508, 2001.

[60] Shuichi Miyamoto and Peter A Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, 1992.

[61] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[62] Julien Michel and Jonathan W. Essex. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *Journal of Computer-Aided Molecular Design*, 24(8):639–658, August 2010.

[63] Di Wu and David A Kofke. Phase-space overlap measures. i. fail-safe bias detection in free energies calculated by molecular simulation. *Journal of Chemical Physics*, 123(5):054103, 2005.

[64] Di Wu and David A Kofke. Phase-space overlap measures. ii. design and implementation of staging methods for free-energy calculations. *Journal of Chemical Physics*, 123(8):084109, 2005.

[65] Mihaly Mezei. The finite difference thermodynamic integration, tested on calculating the hydration free energy difference between acetone and dimethylamine in water. *Journal of Chemical Physics*, 86(12):7084–7088, 1987.

[66] William L Jorgensen, J Kathleen Buckner, Stephane Boudon, and Julian Tirado Rives, J.Rives. Efficient computation of absolute free energies of binding by computer simulations. application to the methane dimer in water. *Journal of Chemical Physics*, 89(6):3742–3746, 1988.

[67] Michael K Gilson, James A Given, Bruce L Bush, and Andrew J. McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal*, 72(3):1047–1069, 1997.

[68] Christopher Woods, Antonia SJ Mey, Gaetano Calabro, and Julien Michel. *Sire molecular simulations framework*, 2016.

[69] Peter Eastman, Mark S. Friedrichs, John D. Chodera, Randall J. Radmer, Christopher M. Bruns, Joy P. Ku, Kyle A. Beauchamp, Thomas J. Lane, Lee-Ping Wang, Diwakar Shukla, Tony Tye, Mike Houston, Timo Stich, Christoph Klein, Michael R. Shirts, and Vijay S. Pande. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation*, 2013.

[70] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, 2015.

[71] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[72] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr., Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, December 2005.

[73] D A Case, J T Berryman, R M Betz, D S Cerutti, T E Cheatham III, T A Darden, R E Duke, T J Giese, H Gohlke, A W Goetz, N Homeyer, S Izadi, P Janowski, J Kaus, A Kovalenko, T S Lee, S LeGrand, P Li, T Luchko, R Luo, B Madej, K M Merz, G Monard, P Needham, H Nguyen, H T Nguyen, I Omelyan, A Onufriev, D R Roe, A Roitberg, R Salomon-Ferrer, C L Simmerling, W Smith, J Swails, R C Walker,

J Wang, R M Wolf, X Wu, D M York, and P A Kollman. *AMBER 2015*. University of California, San Francisco, 2015.

[74] Caitlin C Bannan, Kalistyn H Burley, Michael Chiu, Michael R Shirts, Michael K Gilson, and David L Mobley. Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design*, 30(11):927–944, 2016.

[75] Jian Yin, Niel M Henriksen, David R Slochower, and Michael K Gilson. The sampl5 host–guest challenge: computing binding free energies and enthalpies from explicit solvent simulations by the attach-pull-release (apr) method. *Journal of Computer-Aided Molecular Design*, 31(1):133–145, 2017.

[76] Martin Neumann and Othmar Steinhauser. The influence of boundary conditions used in machine simulations on the structure of polar systems. *Molecular Physics*, 39(2):437–454, 1980.

[77] Martin Neumann, Othmar Steinhauser, and G Stuart Pawley. Consistent calculation of the static and frequency-dependent dielectric constant in computer simulations. *Molecular Physics*, 52(1):97–113, 1984.

[78] Mika A Kastenholz and Philippe H Hünenberger. Influence of artificial periodicity and ionic strength in molecular dynamics simulations of charged biomolecules employing lattice-sum methods. *Journal of Physical Chemistry B*, 108(2):774–788, 2004.

[79] Mika A Kastenholz and Philippe H Hünenberger. Computation of methodology-independent ionic solvation free energies from molecular simulations. i. the electrostatic potential in molecular liquids. *Journal of Chemical Physics*, 124(12):124106, 2006.

[80] Mika A Kastenholz and Philippe H Hünenberger. Computation of methodology-independent ionic solvation free energies from molecular simulations. ii. the hydration free energy of the sodium cation. *Journal of Chemical Physics*, 124(22):224501, 2006.

[81] Gabriel J Rocklin, David L Mobley, Ken Dill, and Philippe H Hünenberger. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: an accurate correction scheme for electrostatic finite-size effects. *Journal of Chemical Physics*, 139(18):184103, November 2013.

[82] Maria M Reif and Chris Oostenbrink. Net charge changes in the calculation of relative ligand-binding free energies via classical atomistic molecular dynamics simulation. *Journal of Computational Chemistry*, 35(3):227–243, 2014.

[83] Niels Hansen and Wilfred F. Van Gunsteren. Practical aspects of free-energy calculations: A review. *Journal of Chemical Theory and Computation*, 10(7):2632–2647, 2014.

[84] Andrew Pohorille, Christopher Jarzynski, and Christophe Chipot. Good practices in free-energy calculations. *Journal of Physical Chemistry B*, 114(32):10235–10253, 2010.

[85] Emilio Gallicchio and Ronald M. Levy. Recent theoretical and computational advances for modeling protein and ligand binding affinities. In Christo Christov, editor, *Computational chemistry methods in structural biology*, volume 85 of *Advances in Protein Chemistry and Structural Biology*, pages 27–80. Academic Press, 2011.

[86] Xiya Lu, Dong Fang, Shingo Ito, Yuko Okamoto, Victor Ovchinnikov, and Qiang Cui. Qm/mm free energy simulations: recent progress and challenges. *Molecular Simulation*, 2016.

[87] Jeffrey M. Rickman and Richard LeSar. Free-energy calculations in materials research. *Annual Review of Materials Research*, 32(1):195–217, 2002.

[88] David R. Squire and William G. Hoover. Monte carlo simulation of vacancies in rare–gas crystals. *Journal of Chemical Physics*, 50(2):701–706, 1969.

[89] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

[90] Michael R. Mruzik, Farid F. Abraham, Donald E. Schreiber, and G. M. Pound. A monte carlo study of ion–water clusters. *Journal of Chemical Physics*, 64(2):481–491, 1976.

[91] Johan P. M. Postma, Herman J. C. Berendsen, and Jan R. Haak. Thermodynamics of cavity formation in water. a molecular dynamics study. *Faraday Symposia of the Chemical Society*, 17:55–67, 1982.

[92] Bhalachandra L. Tembe and Andrew J. McCammon. Ligand-receptor interactions. *Computers & Chemistry*, 8(4):281 – 283, 1984.

[93] David L. Beveridge and Frank M. Dicapua. Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. *Annual Review of Biophysics and Biophysical Chemistry*, 18(1):431–492, 1989.

[94] P. Tjerk Straatsma and Andrew J. MCcammon. Computational Alchemy. *Annual Review of Physical Chemistry*, 43(1):407–435, 1992.

[95] Peter. Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews*, 93(7):2395–2417, 1993.

[96] Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K. Dahlgren, Jeremy Greenwood, Donna L. Romero, Craig Masse, Jennifer L. Knight, Thomas Steinbrecher, Thijs Beuming, Wolfgang Damm, Ed Harder, Woody Sherman, Mark Brewer, Ron Wester, Mark Murcko, Leah Frye, Ramy Farid, Teng Lin, David L. Mobley, William L. Jorgensen, Bruce J. Berne, Richard A. Friesner, and Robert Abel. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015.

[97] Lingle Wang, Yuqing Deng, Yujie Wu, Byungchan Kim, David N. LeBard, Dan Wandschneider, Mike Beachy, Richard A. Friesner, and Robert Abel. Accurate modeling of scaffold hopping transformations in drug discovery. *Journal of Chemical Theory and Computation*, 13(1):42–54, 2017.

[98] David A. Pearlman. A comparison of alternative approaches to free energy calculations. *Journal of Physical Chemistry*, 98(5):1487–1493, 1994.

[99] Stefan Boresch and Martin Karplus. The role of bonded terms in free energy simulations: 1. theoretical analysis. *Journal of Physical Chemistry A*, 103(1):103–118, 1999.

[100] Sundaram Shobana, Benoît Roux, and Olaf S. Andersen. Free energy simulations: Thermodynamic reversibility and variability. *Journal of Physical Chemistry B*, 104(21):5179–5190, 2000.

[101] Michael R. Shirts, Christoph Klein, Jason M. Swails, Jian Yin, Michael K. Gilson, David L. Mobley, David A. Case, and Ellen D. Zhong. Lessons learned from comparing molecular dynamics engines on the sampl5 dataset. *Journal of Computer-Aided Molecular Design*, 31(1):147–161, 2017.

[102] Michael Schappals, Andreas Mecklenfeld, Leif Kroöger, Vitalie Botan, Andreas KoÌĹster, Simon Stephan, Edder J García, Gabor Rutkai, Gabriele Raabe, Peter Klein, et al. Round robin study: Molecular simulation of thermodynamic properties from models with internal degrees of freedom. *Journal of Chemical Theory and Computation*, 13(9):4270–4280, 2017.

[103] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics*, 117(1):1–19, 1995.

[104] Kevin J Bowers, David E Chow, Huafeng Xu, Ron O Dror, Michael P Eastwood, Brent A Gregersen, John L Klepeis, Istvan Kolossvary, Mark A Moraes, Federico D Sacerdoti, et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC 2006 conference, proceedings of the ACM/IEEE*, pages 43–43. IEEE, 2006.

[105] I Todorov and W Smith. The dl_poly user manual version 4.02. 0, 2011.

[106] Sander Pronk, Per Larsson, Iman Pouya, Gregory R. Bowman, Imran S. Haque, Kyle Beauchamp, Berk Hess, Vijay S. Pande, Peter M. Kasson, and Erik Lindahl. Copernicus: A new paradigm for parallel adaptive molecular dynamics. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, pages 60:1–60:10, New York, NY, USA, 2011. ACM.

[107] S. Kashif Sadiq, David Wright, Simon J. Watson, Stefan J. Zasada, Ileana Stoica, and Peter V. Coveney. Automated molecular simulation based binding affinity calculator for ligand-bound hiv-1 proteases. *Journal of Chemical Information and Modeling*, 48(9):1909–1919, 2008.

[108] Magnus Lundborg and Erik Lindahl. Automatic gromacs topology generation and comparisons of force fields for solvation free energy calculations. *Journal of Physical Chemistry B*, 119(3):810–823, 2015.

[109] Hannes H. Loeffler, Julien Michel, and Christopher Woods. FESetup: Automating Setup for Alchemical Free Energy Simulations. *Journal of Chemical Information and Modeling*, 55(12):2485–2490, 2015.

[110] Vivekanandan Balasubramanian, Iain Bethune, Ardita Shkurti, Elena Breitmoser, Eugen Hruska, Cecilia Clementi, Charles A. Laughton, and Shantenu Jha. Extasy: Scalable and flexible coupling of MD simulations and advanced sampling techniques. *CoRR*, abs/1606.00093, 2016.

[111] Thomas C. Beutler, Alan E. Mark, René C. van Schaik, Paul R. Gerber, and Wilfred F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539, June 1994.

[112] Martin Zacharias, P. Tjerk Straatsma, and Andrew J. McCammon. Separation–shifted scaling, a new scaling method for Lennard–Jones interactions in thermodynamic integration. *Journal of Chemical Physics*, 100(12):9025–9031, June 1994.

[113] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[114] Shuai Liu, Lingle Wang, and David L. Mobley. Is ring breaking feasible in relative binding free energy calculations? *Journal of Chemical Information and Modeling*, 2015.

[115] Stefan Boresch and Martin Karplus. The role of bonded terms in free energy simulations. 2. calculation of their influence on free energy differences of solvation. *Journal of Physical Chemistry A*, 103(1):119–136, 1999.

[116] Hannes H. Loeffler, Christoph A. Sotriffer, Rudolf H. Winger, Klaus R. Liedl, and Bernd M. Rode. Calculation of sequence–dependent free energies of hydration of dipeptides formed by alanine and glycine. *Journal of Computational Chemistry*, 22(8):846–860, 2001.

[117] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, July 2004.

[118] Araz Jakalian, Bruce L Bush, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.

[119] Araz Jakalian, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. AM1–BCC model: II. Parameterization and validation. *Journal of Computational Chemistry*, 23(16):1623–1641, 2002.

[120] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics*, 79(2):926–935, 1983.

[121] Thomas Steinbrecher, InSuk Joung, and David A. Case. Soft-core potentials in thermodynamic integration: comparing one- and two-step transformations. *Journal of Computational Chemistry*, 32(15):3253–3263, 2011.

[122] Yuqing Deng and Benoît Roux. Hydration of Amino Acid Side Chains: Non-Polar and Electrostatic Contributions Calculated from Staged Molecular Dynamics Free Energy Simulations with Explicit Water Molecules. *Journal of Physical Chemistry A*, 108:16567–16576, 2004.

[123] Levi N. Naden, Tri T. Pham, and Michael R. Shirts. Linear basis function approach to efficient alchemical free energy calculations. 1. Removal of uncharged atomic sites. *Journal of Chemical Theory and Computation*, 10(3):1128–1149, 2014.

[124] Levi N. Naden and Michael R. Shirts. Linear basis function approach to efficient alchemical free energy calculations. 2. Inserting and deleting particles with coulombic interactions. *Journal of Chemical Theory and Computation*, 11(6):2536–2549, June 2015.

[125] Jed W. Pitera and Wilfred F. van Gunsteren. A Comparison of Non-Bonded Scaling Approaches for Free Energy Calculations. *Molecular Simulation*, 28(1-2):45–65, 2002.

[126] Jamshed Anwar and David M. Heyes. Robust and Accurate Method for Free-Energy Calculation of Charged Molecular Systems. *Journal of Chemical Physics*, 122(22):224117, 2005.

[127] David L. Mobley and J. Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design*, 28(7):711–720, 2014.

[128] Guilherme Duarte Ramos Matos, Daisy Y Kyu, Hannes H Loeffler, John D Chodera, Michael R Shirts, and David L Mobley. Approaches

for calculating solvation free energies and enthalpies demonstrated with an update of the freesolv database. *Journal of Chemical & Engineering Data*, 62(5):1559–1569, 2017.

[129] Julien Michel, Marcel L. Verdonk, and Jonathan W. Essex. Protein–ligand complexes: Computation of the relative free energy of different scaffolds and binding modes. *Journal of Chemical Theory and Computation*, 3(5):1645–1655, 2007.

[130] Michael R. Shirts, David L. Mobley, John D. Chodera, and Vijay J. Pande. Accurate and efficient corrections for missing dispersion interactions in molecular simulations. *Journal of Physical Chemistry B*, 45(111):13052–13063, 2007.

[131] Stefano Bosisio, Antonia S J S Mey, and Julien Michel. Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design*, pages 1–10, August 2016.

[132] Pavel V Klimovich, Michael R Shirts, and David L Mobley. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design*, 29(5):397–411, 2015.

[133] Pavel V. Klimovich and David L. Mobley. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *Journal of Computer-Aided Molecular Design*, 24(4):307–316, 2010.

[134] Lee-Ping Wang, Todd J. Martinez, and Vijay S. Pande. Building force fields: An automatic, systematic, and reproducible approach. *Journal of Physical Chemistry Letters*, 5(11):1885–1891, 2014.

[135] Joseph W. Kaus, Levi T. Pierce, Ross C. Walker, and Andrew J. McCammon. Improving the efficiency of free energy calculations in the amber molecular dynamics package. *Journal of Chemical Theory and Computation*, 9(9):4131–4139, 2013.

[136] Michael R. Shirts, David L. Mobley, and John D. Chodera. Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? *Annual Reports in Computational Chemistry*, 3(February 2016):41–59, 2007.

[137] Piero Procacci and Chiara Cardelli. Fast switching alchemical transformations in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 10(7):2813–2823, 2014.

[138] Clara Christ and Thomas Fox. Accuracy assessment and automation of free energy calculations for drug design. *Journal of Chemical Information and Modeling*, 2013.

[139] Vytautas Gapsys, Servaas Michielssens, Daniel Seeliger, and Bert L. de Groot. pmx: Automated protein structure and topology generation for alchemical perturbations. *Journal of Computational Chemistry*, 36(5):348–354, 2015.

[140] Shuai Liu, Yujie Wu, Teng Lin, Robert Abel, Jonathan P. Redmann, Christopher M. Summa, Vivian R. Jaber, Nathan M. Lim, and David L. Mobley. Lead optimization mapper: automating free energy calculations for lead optimization. *Journal of Computer-Aided Molecular Design*, 27(9):755–770, Sep 2013.

[141] Vijayaraj Ramadoss, Francois Dehez, and Christophe Chipot. Alascan: A graphical user interface for alanine scanning free-energy calculations. *Journal of Chemical Information and Modeling*, 56(6):1122–1126, 2016.

[142] Agastya P. Bhati, Shunzhou Wan, David W. Wright, and Peter V. Coveney. Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *Journal of Chemical Theory and Computation*, 13(1):210–222, 2017.

[143] Andrea Rodil, Stefano Bosisio, Mohammed Salah Ayoup, Laura Quinn, David Bradford Cordes, Alexandra Slawin, Cormac D Murphy, Julien Michel, and David O Hagan. Metabolism and hydrophilicity of the polarised "janus face" all-cis tetrafluorocyclohexyl ring, a candidate motif for drug discovery. *Chemical Science*, 2018.

[144] Stefano Bosisio, Antonia S J S Mey, and Julien Michel. Blinded predictions of distribution coefficients in the sampl5 challenge. *Journal of Computer-Aided Molecular Design*, 30(11):1101–1114, 2016.

[145] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 64:4–17, 2012.

[146] Ewelina Rutkowska, Karolina Pajak, and K Jóźwiak. Lipophilicity–methods of determination and its role in medicinal chemistry. *Acta Poloniae Pharmaceutica*, 70(1):3–18, 2013.

[147] Albert Leo, Corwin Hansch, and David Elkins. Partition Coefficients and Their Uses. *Chemical Reviews*, 71(6):525–616, 1971.

[148] Cary T Chiou, Virgil H Freed, David W Schmedding, and Rodger L Kohnert. Partition coefficient and bioaccumulation of selected organic chemicals. *Environmental Science & Technology*, 11(5):475–478, 1977.

[149] Sujit Banerjee, Samuel H Yalkowsky, and C Valvani. Water solubility and octanol/water partition coefficients of organics. limitations of the solubility-partition coefficient correlation. *Environmental Science & Technology*, 14(10):1227–1229, 1980.

[150] Victor A Levin. Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability. *Journal of Medicinal Chemistry*, 23(6):682–684, 1980.

[151] Axel Andrés, Martí Rosés, Clara Ràfols, Elisabeth Bosch, Sonia Espinosa, Víctor Segarra, and Josep M Huerta. Setup and validation of shake-flask procedures for the determination of partition coefficients (log d) from low drug amounts. *European Journal of Pharmaceutical Sciences*, 76:181–191, 2015.

[152] Stefan H Unger, James R Cook, and John S Hollenberg. Simple procedure for determining octanol-aqueous partition, distribution, and ionization coefficients by reversed-phase high-pressure liquid chromatography. *Journal of Pharmaceutical Sciences*, 67(10):1364–1367, 1978.

[153] Ren Tomita, Nawaf Al-Maharik, Andrea Rodil, Michael Buehl, and David O Hagan. Synthesis of aryl a, a-difluoroethyl thioethers a novel structure motif in organic chemistry, and extending to aryl a, a-difluoro oxyethers. *Organic & Biomolecular Chemistry*, 2018.

[154] Chemaxon, www.chemicalize.org.

[155] Molinspiration,http://www.molinspiration.com/services/logp.html.

[156] Vcclab,http://www.vcclab.org/lab/alogps/.

[157] Peter A Kollman. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Accounts of Chemical Research*, 29(10):461–469, 1996.

[158] Scott A Best, Kenneth M Merz, and Charles H Reynolds. Free energy perturbation study of octanol/water partition coefficients: Comparison with continuum gb/sa calculations. *Journal of Physical Chemistry B*, 103(4):714–726, 1999.

[159] Bin Chen and J Ilja Siepmann. Microscopic structure and solvation in dry and wet octanol. *Journal of Physical Chemistry B*, 110(8):3555–3563, 2006.

[160] Klaus Müller. Simple vector considerations to assess the polarity of partially fluorinated alkyl and alkoxy groups. *CHIMIA International Journal for Chemistry*, 68(6):356–362, 2014.

[161] Baiwei Lin and Joseph H Pease. A novel method for high throughput lipophilicity determination by microscale shake flask and liquid chromatography tandem mass spectrometry. *Combinatorial chemistry & High Throughput Screening*, 16(10):817–825, 2013.

[162] Ariën S Rustenburg, Justin Dancer, Baiwei Lin, Jianwen A Feng, Daniel F Ortwine, David L Mobley, and John D Chodera. Measuring experimental cyclohexane-water distribution coefficients for the sampl5 challenge. *Journal of Computer-Aided Molecular Design*, 30(11):945–958, 2016.

[163] Joakim PM Jämbeck, Francesca Mocci, Alexander P Lyubartsev, and Aatto Laaksonen. Partial atomic charges and their impact on the free energy of solvation. *Journal of Computational Chemistry*, 34(3):187–197, 2013.

[164] Noel M, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011.

[165] Aleksandr V Marenich, Casey P Kelly, Jason D Thompson, Gregory D Hawkins, Candee C Chambers, David J Giesen, Paul Winget, Christopher J Cramer, and Donald G Truhlar. *University of Minnesota, Minneapolis*, 2009.

[166] Robert A Scherrer and Susan M Howard. Use of distribution coefficients in quantitative structure-activity relations. *Journal of Medicinal Chemistry*, 20(1):53–58, 1977.

[167] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[168] David Weininger, Arthur Weininger, and Joseph L Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

[169] Daniel R Roe and Thomas E Cheatham III. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9(7):3084–3095, 2013.

[170] Chad W Hopkins, Scott Le Grand, Ross C Walker, and Adrian E Roitberg. Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of Chemical Theory and Computation*, 11(4):1864–1874, 2015.

[171] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics*, 129(12), 2008.

[172] Georgios Gerogiokas, Gaetano Calabro, Richard H Henchman, Michelle W Y Southey, Richard J Law, and Julien Michel. Prediction of small molecule hydration thermodynamics with grid cell theory. *Journal of Chemical Theory and Computation*, 10(1):35–48, 2013.

[173] Julien Michel, Richard H Henchman, Georgios Gerogiokas, Michelle WY Southey, Michael P Mazanetz, and Richard J Law. Evaluation of host–guest binding thermodynamics of model cavities with grid cell theory. *Journal of Chemical Theory and Computation*, 10(9):4055–4068, 2014.

[174] George Gerogiokas, Michelle W Y Southey, Michael P Mazanetz, Alexander Hefeitz, Michael Bodkin, Richard J Law, and Julien Michel. Evaluation of water displacement energetics in protein binding sites with grid cell theory. *Physical Chemistry Chemical Physics*, 17(13):8416–8426, 2015.

[175] Georgios Gerogiokas, Michelle WY Southey, Michael P Mazanetz, Alexander Heifetz, Michael Bodkin, Richard J Law, Richard H Henchman, and J Michel. Assessment of hydration thermodynamics at protein interfaces with grid cell theory. *Journal of Physical Chemistry B*, 120(40):10442–10452, 2016.

[176] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, 2005.

[177] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[178] Andreas Klamt and Frank Eckert. COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. *Fluid Phase Equilibria*, 172(1):43–72, 2000.

[179] Andreas Klamt, Frank Eckert, Jens Reinisch, and Karin Wichmann. Prediction of cyclohexane-water distribution coefficients with COSMO-RS on the SAMPL5 data set. *Journal of Computer-Aided Molecular Design*, pages 1–9, July 2016.

[180] Nicolas Tielker, Daniel Tomazic, Jochen Heil, Thomas Kloss, Sebastian Ehrhart, Stefan Güssregen, K Friedemann Schmidt, and Stefan M Kast. The sampl5 challenge for embedded-cluster integral equation theory: solvation free energies, aqueous pk a, and cyclohexane–water log d. *Journal of Computer-Aided Molecular Design*, 30(11):1035–1044, 2016.

[181] Tyler Luchko, Nikolay Blinov, Garrett C Limon, Kevin P Joyce, and Andriy Kovalenko. Sampl5: 3d-rism partition coefficient calculations with partial molar volume corrections and solute conformational sampling. *Journal of Computer-Aided Molecular Design*, 30(11):1115–1127, 2016.

[182] Julien Michel. Current and emerging opportunities for molecular simulations in structure-based drug design. *Physical Chemistry Chemical Physics*, 16(10):4465–4477, 2014.

[183] I-Jen Chen and Nicolas Foloppe. Is conformational sampling of drug-like molecules a solved problem? *Drug Development Research*, 72(1):85–94, 2011.

[184] Thomas A Halgren and Wolfgang Damm. Polarizable force fields. *Current Opinion in Structural Biology*, 11(2):236–242, 2001.

[185] Donald J Cram and Jane M Cram. Host-guest chemistry. *Science*, 183(4127):803–809, 1974.

[186] Chia-En Chang and Michael K Gilson. Free energy, entropy, and induced fit in host- guest recognition: calculations with the second-generation mining minima algorithm. *Journal of the American Chemical Society*, 126(40):13156–13164, 2004.

[187] William L Jorgensen and Laura L Thomas. Perspective on free-energy perturbation calculations for chemical equilibria. *Journal of Chemical Theory and Computation*, 4(6):869–876, 2008.

[188] Christopher J. Woods, Maturos Malaisree, Supot Hannongbua, and Adrian J. Mulholland. A water–swap reaction coordinate for the calculation of absolute protein–ligand binding free energies. *Journal of Chemical Physics*, 134(5):054114, 2011.

[189] Matthew T Geballe, A Geoffrey Skillman, Anthony Nicholls, J Peter Guthrie, and Peter J Taylor. The SAMPL2 blind prediction challenge: introduction and overview. *Journal of Computer-Aided Molecular Design*, 24(4):259–279, April 2010.

[190] A Geoffrey Skillman. Sampl3: blinded prediction of host–guest binding affinities, hydration free energies, and trypsin inhibitors. *Journal of Computer-Aided Molecular Design*, pages 1–2, 2012.

[191] J Peter Guthrie. A blind challenge for computational solvation free energies: introduction and overview. *Journal of Physical Chemistry B*, 113(14):4501–4507, April 2009.

[192] Corinne LD Gibb and Bruce C Gibb. Binding of cyclic carboxylates to octa-acid deep-cavity cavitand. *Journal of Computer-Aided molecular design*, 28(4):319–325, 2014.

[193] Laura Gilberg, Ben Zhang, Peter Y Zavalij, Vladimir Sindelar, and Lyle Isaacs. Acyclic cucurbit [n] uril-type molecular containers: influence of glycoluril oligomer length on their function as solubilizing agents. *Organic & biomolecular chemistry*, 13(13):4041–4050, 2015.

[194] Rebecca Sure, Jens Antony, and Stefan Grimme. Blind prediction of binding affinities for charged supramolecular host–guest systems: achievements and shortcomings of dft-d3. *Journal of Physical Chemistry B*, 118(12):3431–3440, 2014.

[195] Hari S Muddana and Michael K Gilson. Prediction of sampl3 host–guest binding affinities: evaluating the accuracy of generalized force-fields. *Journal of Computer-Aided Molecular Design*, 26(5):517–525, 2012.

[196] Paulius Mikulskis, Daniela Cioloboc, Milica Andrejić, Sakshi Khare, Joakim Brorsson, Samuel Genheden, Ricardo A Mata, Pär Söderhjelm, and Ulf Ryde. Free-energy perturbation and quantum mechanical study of sampl4 octa-acid host–guest binding energies. *Journal of Computer-Aided Molecular Design*, 28(4):375–400, 2014.

[197] Gerhard König, Frank C Pickard, Ye Mei, and Bernard R Brooks. Predicting hydration free energies with a hybrid qm/mm approach: an evaluation of implicit and explicit solvation models in sampl4. *Journal of Computer-Aided Molecular Design*, 28(3):245–257, 2014.

[198] Oliver Beckstein, Anaïs Fourrier, and Bogdan I Iorga. Prediction of hydration free energies for the SAMPL4 diverse set of compounds using molecular dynamics simulations with the opls-aa force field. *Journal of Computer-Aided Molecular Design*, 28(3):265–276, 2014.

[199] Jacob I Monroe and Michael R Shirts. Converging free energies of binding in cucurbit [7] uril and octa-acid host–guest systems from sampl4 using expanded ensemble simulations. *Journal of Computer-Aided Molecular Design*, 28(4):401–415, 2014.

[200] Peter A. Kollman, Irina Massova, Carolina Reyes, Bernd Kuhn, Shuanghong Huo, Lillian Chong, Matthew Lee, Taisung Lee, Yong Duan, Wei Wang, Oreola Donini, Piotr Cieplak, Jaysharee Srinivasan, David A. Case, , and Thomas E. Cheatham. *Accounts of Chemical Research*, 33(12):889–897, 2000.

[201] Sushil K Mishra, Gaetano Calabro, Hannes H Loeffler, Julien Michel, and Jaroslav Koča. Evaluation of selected classical force fields for alchemical binding free energy calculations of protein-carbohydrate complexes. *Journal of Chemical Theory and Computation*, 11(7):3333–3345, 2015.

[202] Hari S Muddana, Andrew T Fenley, David L Mobley, and Michael K Gilson. The sampl4 host–guest blind prediction challenge: an overview. *Journal of Computer-Aided Molecular Design*, 28(4):305–317, 2014.

[203] Schrödinger release 2015-2: Maestro, version 10.2, schrödinger, 2015.

[204] David A. Kofke. Free energy methods in molecular simulation. *Fluid Phase Equilibria*, 228-229:41–48, 2005.

[205] Matthew R Sullivan, Punidha Sokkalingam, Thong Nguyen, James P Donahue, and Bruce C Gibb. Binding of carboxylate and trimethylammonium salts to octa-acid and temoa deep-cavity cavitands. *Journal of Computer-Aided Molecular Design*, 31(1):21–28, 2017.

[206] Corinne L D Gibb, Edwin D Stevens, and Bruce C Gibb. CHXR (X= Cl, Br, and I) hydrogen bonds drive the complexation properties of a nanoscale molecular basket. *Journal of the American Chemical Society*, 123(24):5849–5850, 2001.

[207] Nengfang She, Damien Moncelet, Laura Gilberg, Xiaoyong Lu, Vladimir Sindelar, Volker Briken, and Lyle Isaacs. Glycoluril-derived molecular clips are potent and selective receptors for cationic dyes in water. *Chemistry-A European Journal*, 22(43):15270–15279, 2016.

[208] Soumendranath Bhakat and Pär Söderhjelm. Resolving the problem of trapped water in binding cavities: prediction of host–guest binding free energies in the SAMPL5 challenge by funnel metadynamics. *Journal of Computer-Aided Molecular Design*, 31(1):119–132, 2017.

[209] Romelia Salomon-Ferrer, Andreas W Goötz, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald. *Journal of chemical theory and computation*, 9(9):3878–3888, 2013.

[210] Germano Heinzelmann, Niel M Henriksen, and Michael K Gilson. Attach-pull-release calculations of ligand binding and conformational changes on the first brd4 bromodomain. *Journal of Chemical Theory and Computation*, 2017.

[211] Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences*, 110(16):6358–6363, 2013.

[212] Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya, Sibsankar Kundu, Shijun Zhong, Jihyun Shim, Eva Darian, Olgun Guvench, P Lopes, Igor Vorobyov, et al. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690, 2010.

[213] Gerhard König, Stefan Bruckner, and Stefan Boresch. Unorthodox uses of bennett's acceptance ratio method. *Journal of Computational Chemistry*, 30(11):1712–1718, 2009.

[214] Nupur Bansal, Zheng Zheng, David S Cerutti, and Kenneth M Merz. On the fly estimation of host–guest binding free energies using the movable type method: participation in the SAMPL5 blind challenge. *Journal of Computer-Aided Molecular Design*, 31(1):47–60, 2017.

[215] Zheng Zheng and Kenneth M Merz Jr. Development of the knowledge-based and empirical combined scoring algorithm (kecsa) to score protein–ligand interactions. *Journal of chemical information and modeling*, 53(5):1073–1083, 2013.

[216] Octav Caldararu, Martin A Olsson, Christoph Riplinger, Frank Neese, and Ulf Ryde. Binding free energies in the sampl5 octa-acid host–guest challenge calculated with dft-d3 and ccsd (t). *Journal of Computer-Aided Molecular Design*, 31(1):87–106, 2017.

[217] Andreas Klamt. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *Journal of Physical Chemistry*, 99(7):2224–2235, 1995.

[218] Juyong Lee, Florentina Tofoleanu, Frank C Pickard, Gerhard König, Jing Huang, Ana Damjanović, Minkyung Baek, Chaok Seok, and Bernard R Brooks. Absolute binding free energy calculations of cb-clip host–guest systems in the sampl5 blind challenge. *Journal of Computer-Aided Molecular Design*, 31(1):71–85, 2017.

[219] Antonia S J S Mey, Jordi Juárez Jiménez, and Julien Michel. Impact of domain knowledge on blinded predictions of binding energies by alchemical free energy calculations. *Journal of Computer-Aided Molecular Design*, 32(1):199–210, 2018.

[220] Markus Christen and Wilfred F Van Gunsteren. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *Journal of Computational Chemistry*, 29(2):157–166, 2008.

[221] Wouter K den Otter and Willem J Briels. The calculation of free-energy differences by constrained molecular-dynamics simulations. *Journal of Chemical Physics*, 109(11):4139–4146, 1998.

[222] Maria M Reif, Vincent Krautler, Mika A Kastenholz, Xavier Daura, and Philippe HR Hünenberger. Molecular dynamics simulations of a reversibly folding β-heptapeptide in methanol: influence of the treatment of long-range electrostatic interactions. *Journal of Physical Chemistry B*, 113(10):3112–3128, 2009.

[223] Martin Neumann. Dipole moment fluctuation formulas in computer simulations of polar systems. *Molecular Physics*, 50(4):841–858, 1983.

[224] Martin Neumann and Othmar Steinhauser. On the calculation of the dielectric constant using the ewald-kornfeld tensor. *Chemical Physics Letters*, 95(4-5):417–422, 1983.

[225] Gerhard Hummer, Lawrence R Pratt, and Angel E Garcia. Free energy of ionic hydration. *Journal of Physical Chemistry*, 100(4):1206–1215, 1996.

[226] Gerhard Hummer, Lawrence R Pratt, Angel E García, Bruce J Berne, and Steven W Rick. Electrostatic potentials and free energies of solvation of polar and charged molecules. *Journal of Physical Chemistry B*, 101(16):3017–3020, 1997.

[227] Gerhard Hummer, Lawrence R Pratt, Angel E García, Shekhar Garde, Bruce J Berne, and Steven W Rick. Reply to comment on "electrostatic potentials and free energies of solvation of polar and charged molecules". *Journal of Physical Chemistry B*, 102(19):3841–3843, 1998.

[228] Thomas Simonson. Accurate calculation of the dielectric constant of water from simulations of a microscopic droplet in vacuum. *Chemical Physics Letters*, 250(5-6):450–454, 1996.

[229] Nathan Baker, Michael Holst, and Feng Wang. Adaptive multilevel finite element solution of the poisson–boltzmann equation ii. refinement at solvent-accessible surfaces in biomolecular systems. *Journal of Computational Chemistry*, 21(15):1343–1352, 2000.

[230] Federico Fogolari, Pierfrancesco Zuccato, Gennaro Esposito, and Paolo Viglino. Biomolecular electrostatics with the linearized poisson-boltzmann equation. *Biophysical Journal*, 76(1):1–16, 1999.

[231] Federico Fogolari, Alessandro Brigo, and Henriette Molinari. The poisson–boltzmann equation for biomolecular electrostatics: a tool for structural biology. *Journal of Molecular Recognition*, 15(6):377–392, 2002.

[232] Christine Peter, Wilfred F van Gunsteren, and Philippe H Hünenberger. A fast-fourier transform method to solve continuum-electrostatics problems with truncated electrostatic interactions: Algorithm and application to ionic solvation and ion–ion interaction. *Journal of Chemical Physics*, 119(23):12205–12223, 2003.

[233] Christine Peter, Wilfred F van Gunsteren, and Philippe H Hünenberger. Solving the poisson equation for solute–solvent systems using fast fourier transforms. *Journal of Chemical Physics*, 116(17):7434–7451, 2002.

[234] John D Jackson. *Classical electrodynamics*. Wiley, 1999.

[235] Thomas Simonson, Gerhard Hummer, and Benoit Roux. Equivalence of m-and p-summation in calculations of ionic solvation free energies. *Journal of Physical Chemistry A*, 121(7):1525–1530, 2017.

[236] Eugene Wigner. On the quantum correction for thermodynamic equilibrium. *Physical Review Letters*, 40(5):749, 1932.

[237] In Suk Joung and Thomas E Cheatham III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *Journal of Physical Chemistry B*, 112(30):9020–9041, 2008.

[238] Chris Oostenbrink, Alessandra Villa, Alan E. Mark, and Wilfred F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53a5 and 53a6. *Journal of Computational Chemistry*, 25(13):1656–1676, October 2004.

[239] Robert T McGibbon, Kyle A Beauchamp, Matthew P Harrigan, Christoph Klein, Jason M Swails, Carlos X Hernández, Christian R Schwantes, Lee-Ping Wang, Thomas J Lane, and Vijay S Pande. Mdtraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528–1532, 2015.

[240] Jhon E B Randles. The real hydration energies of ions. *Transactions of the Faraday Society*, 52:1573–1581, 1956.

[241] Richard M Noyes. Thermodynamics of ion hydration as a measure of effective dielectric properties of water. *Journal of the American Chemical Society*, 84(4):513–522, 1962.

[242] Yizhak Marcus. The hydration entropies of ions and their effects on the structure of water. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 82(1):233–242, 1986.

[243] Nanjie Deng, Di Cui, Bin Zhang, Junchao Xia, Jeffrey Cruz, and Ronald M Levy. Comparing alchemical and physical pathway methods for computing the absolute binding free energy of charged ligands. *Physical Chemistry Chemical Physics*, 2018.