UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF CHEMISTRY

The Automated Optimisation of a Coarse-Grained Protein Force Field Using Free Energy Data

by

Javier Caceres-Delpiano

A THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

SUPERVISOR: Prof. Jonathan W. Essex

October, 2019

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYISICAL SCIENCES
SCHOOL OF CHEMISTRY

Doctor of Philosophy

THE AUTOMATED OPTIMISATION OF A COARSE-GRAINED FORCE FIELD USING FREE ENERGY DATA

by Javier Caceres-Delpiano

Atomistic models provide a detailed representation of molecular systems, but are sometimes inadequate for simulations of large systems over long timescales. Coarse-grained models enable accelerated simulations by reducing the number of degrees of freedom, at the cost of reduced accuracy. New optimisation processes to parameterise these models could improve their quality and range of applicability. We present an automated approach for the optimisation of the SIRAH coarse-grained protein force field. A full optimisation of the SIRAH water model was performed using ForceBalance, based on experimental water properties. We implemented hydration free energy gradients as a new target for force field optimisation and applied it successfully to optimise the uncharged side-chains and the protein backbone. We managed to closely reproduce hydration free energies of atomistic models and improve agreement with experiment. An attempt was made for the optimisation of charged coarse-grained protein side-chains. Hydration free energies were improved, but at the expense of an over-fitted model, which led to an over-estimation of protein interactions. Simulations of folded proteins in water result in improved protein stabilities for the new model. We

compute the opening/closing event of a Glutamate receptor binding domain using umbrella sampling simulations, showing a clear improvement on the estimation of the PMF with previously reported studies on atomistic systems, for the ligand-free and glutamate-bound states.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Javier Caceres-Delpiano, declare that this thesis entitled "The Automated Optimisation of a Coarse-Grained Protein Force Field Using Free Energy Data" and the work presented in it are my own and has been generated by me as the result of my own original research. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as:

Signed:

Date:

Acknowledgments

Here I would like to acknowledge all the people that have been in this journey by my side, encouraging me to do better. First of all, I would like to thank my supervisor Jon Essex for all the time, effort, discussions, patience and support in the development of this thesis, as well as in my own personal development as a person. Some laughs were shared, which helped a lot in the most difficult times. I would also like to thank Lee-Ping and Maria Reif for all the time and patience, and for sharing his knowledge as an open book, thank you a lot!

All the Essex group members that have co-existed with me during this time provided me with support, and fun times! Special thanks goes to Marley, Zohra, Mabel and Ying-Chih! Thank you for all the conversations,, drinks, laughs and help to become a better me :) To all the other people that I've met in Southampton and to the Chilean group. Special thanks to Francesco for all the time shared and the music!

More than a special thanks to all my friends that I left in Chile and around the world, that unconditionally supported me. All my love goes to FV and JV for the fun talks, and constant support. Thanks to the "Bar the Benito" for just being there guys! you are awesome in your own magical way. Thanks to Scottfueibein: Rad, Serey, Caro, Tata and Cami :) And thanks to Machu, Nico and Mari for still being there sharing your time with me.

Finally, a would like to thank my family: my father Julio, my mother Amalia, and my brothers Julio and Rodrigo. And of course, all my infinite thanks to my lovely wife Paula la monita: without you I wouldn't be where I am. Your support has been the principal stone in the development of all this, my work, and my soul. Our time

in Southampton made us grown in many ways, personally, as a couple and best friends. We had worked together to finish this project, and we had fought for a better future. We will take in our hearts all the great and beeautiful moments that we spend here. Now is time to start new project together, and to see where the future is going to take us. Wherever it is, I hope that the thing that never leaves us, is happiness. Love you with all my life pepita.

Chapter

0

Introduction and Motivation

Computational tools have become very important in revealing the driving forces in biomolecular processes; in particular, molecular dynamics (MD) simulations provide a physically motivated picture based on Newton's equations of motion coupled with empirical model potentials (force fields). Classical atomistic (AT) models provide a detailed representation of the system with computational effort that scale as O(N log N) with the number of atoms, but are inadequate for simulations of very large systems over long time-scales.

Coarse-grained (CG) models currently represent one of the most important approximations for the construction and simulation of larger systems [9]. By subsuming groups of atoms into single interaction sites, much faster calculations can be realised. However, a disadvantage of CG models is the loss of accuracy associated with reducing the number of interacting particles. Moreover, coarse-graining typically smooths the energy landscape compared to classical atomistic models, diminishing the energy barriers between different states and reducing trapping in energy minima. This can greatly affect calculated thermodynamic properties such as equilibrium structures and dynamic properties such as the rates of conformational changes. Despite these drawbacks, CG models have become a widely used approximation, allowing us to extend spatial and temporal scales for the simulation of bigger and more complex systems [9]. Given this, new approaches for the optimisation of CG models are highly desirable.

Different coarse-grained protein models have been developed, which differ on the mapping and levels of resolution, secondary structure stabilisation, bead-types and MD engines that support their use [9]. One of the main reasons for the development of this type of model was to understand and perform simulations of biological processes at time scales important for their function. In comparison with atomistic force fields, most coarse-grained models smooth out the energy landscape, which accelerates the dynamics [9]. Designing these types of force field have followed, such as models derived from proteins physics and models derived from the analysis of common structural properties of proteins. The force field parameters are usually iterated "by hand", making this process tedious and prone to errors [10]. Thus the use of automated procedures is an advantage and is an important alternative to overcome these difficulties. ForceBalance [6] is a new method that let us automatically derive and optimise force field parameters in a flexible manner, combining the use of experimental and/or simulation properties as targets.

The SIRAH coarse-grained protein force field looks like a promising alternative to conventional atomistic force fields [5, 11]. Compared to MARTINI, the SIRAH force field does not use of Elastic Networks to overcome the problem of secondary structure stability. The use of a higher resolution backbone produces hydrogen bond-like interaction, which fully work to stabilise the systems. Moreover, long range electrostatic interactions are modelled with a dielectric constant of unity, compared with the MARTINI model where a dielectric constant of 15 is used [4]. Even so, the SIRAH force field has not been tested for the reproduction of conformational changes in protein, which is the main interest of our investigation.

Hydration free energies (HFEs) are an important property for aqueous systems such as proteins. They help us to directly understand biological processes such as ligand recognition and protein-protein interactions, and indirectly understand processes such as folding and conformational changes. Moreover, hydration free energies have been used for the validation of molecular force fields, and they are an integral part of the calculation and estimation of solubilities, partition coefficients and solute-solvent interactions [12]. For these reasons, use of hydration free energies as a parameterisation target for coarse-grained models may improve their performance. Moreover, it has been recently stated that there is considerable interest in methods that can automatically generate a coarse-grained model and are representative in terms of local structure and free energy changes [13].

Here, we optimise the SIRAH CG protein force field [5] using atomistic hydration free energy (HFE) data in ForceBalance [6], in which the gradient of the hydration free energy (with respect to the linear coupling parameter α, $\langle \Delta U \rangle_{\alpha}$) is optimized to match the result from an all-atom simulation, with the goal of improving the CG solvation free energies as a consequence. The approach of fitting atomistic HFE gradients has the advantage of reducing the computational cost of the parameter optimisation because it does not require full HFE calculations of the CG model at every optimisation step. Optimisation of charged side-chains beads has also been performed, with the inclusion of long-range electrostatic corrections to the calculated free energies [14, 15]. A full HFE calculation is carried out after CG model optimisation to validate the approach by comparison to atomistic and experimental HFEs. Different protein and peptide systems have been tested in terms of secondary structure stability and mean square deviations, using the newly optimised SIRAH-OBAFE (**O**ptimised **B**ased on **A**tomistic **F**ree **E**nergies) force field, to evaluate its advantages and limitations.

Chapter

General Theory

1

1.1. Molecular Dynamics

Molecular dynamics (MD) is a deterministic technique, which obeys the laws of classical mechanics and comprises the integration of Newton's equations of motion in order to follow the evolution of a set of interacting atoms [16]. As a basis, Newton's second law states:

$$\mathbf{F}_i = m_i \mathbf{a}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2} \tag{1.1}$$

where $\mathbf{F}_i$ corresponds to the force, m_i to the mass and $\mathbf{a}_i$ to the acceleration (which also corresponds to the second derivative of the position $\mathbf{r}_i$ with respect to time) for atom i in a system of N atoms.

MD is a method where the system is represented by a set of positions and momenta for each particle, where the total energy of the system corresponds to the sum of the kinetic and potential energies

$$E(\mathbf{p}, \mathbf{r}) = K(\mathbf{p}) + U(\mathbf{r}) \tag{1.2}$$

where K and U corresponds to the kinetic and potential energy, respectively, and $\mathbf{p}$ corresponds to the momenta. This method calculates a trajectory in a 6N-dimensional space (3N positions and 3N momenta). [10].

The potential energy is a function $U(\mathbf{r}_1 ... \mathbf{r}_N)$ of the atom positions $\mathbf{r}_i$ to $\mathbf{r}_N$, and is usually constructed based on the relative position of the atoms with respect to each other, rather than using their absolute configuration. The force can be derived from the gradient of the potential energy:

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r}_1 ... \mathbf{r}_N) \tag{1.3}$$

If the potential energy and the coordinates of a system are known, the forces acting on each atom can be calculated with equation 1.3, generating a new set of coordinates through the use of different integrators (see section 1.4). New forces can be calculated again, generating a trajectory that represents the time evolution of a particular system.

An important goal in MD simulations is to compare them with experimental data. In an experiment, a specific value for a property Q is just the average $\langle Q \rangle$ over space and time for a set of conformations. The property Q has to be weighted by a probability P of a conformation to occur

$$\langle Q \rangle = \int \int Q(\mathbf{p}, \mathbf{r}) P(\mathbf{p}, \mathbf{r}) d\mathbf{p} d\mathbf{r} \tag{1.4}$$

In order to get meaningful information from our MD simulations, we have to fulfil the *ergodic hypothesis*. Ergodicity is important in MD, because it is not viable to visit all possible states from finite sampled simulations, but it is possible to sample a relevant property Q for a long enough simulation. If this is true, for an ergodic system the time average of a property Q is equal to the ensemble averaged Q

$$\langle Q \rangle_{time} = \langle Q \rangle_{ensemble} \tag{1.5}$$

that is, if we manage to follow a property for a long enough period of time, we have to be able to get the same results as if we were looking to the same property from an infinitely large ensemble of systems at a single time point.

1.2. Interaction Potentials

The potential energy U is usually written as sum over pair interactions of atoms i and j, where each atom is considered only once:

$$U(\mathbf{r}_1 \dots \mathbf{r}_N) = \sum_i \sum_{j>i} \Phi(|\mathbf{r}_i - \mathbf{r}_j|) \tag{1.6}$$

where Φ corresponds to the potential, and $\mathbf{r}_i$ and $\mathbf{r}_j$ correspond to the position of atoms i and j, respectively.

In order to model the interactions of the systems of interest, a functional form and a set of parameters are necessary to compute the potential energy of the system. Each term in this functional form accounts for a specific interaction type (non-bonded and bonded terms). This set of potential functions and parameters form what is known as a force field (FF). These parameters are usually obtained based on experimental reference or high-order methods such as quantum mechanics [17]. Equation number 1.7 shows a typical expression of a force field, where molecules are defined as a group of atoms restrained by harmonic forces:

$$U = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsion} \frac{U_n}{2}[1 + \cos(n\Phi - \delta)] \tag{1.7}$$
$$+ \sum_{LJ} 4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6] + \sum_{elec} \frac{q_i q_j}{r_{ij}}$$

The first three terms describe the bonded interactions (bonds, angles, and torsions). r, r$_0$, θ and θ_0 correspond to the bond length, the reference bond length, the angle and the reference angle length, respectively. k$_b$ and k$_a$ are the force constant for bonds and angles, respectively. U$_n$ is the force constant for dihedral angles, Φ is the dihedral angle, n is the order constant, and δ is the reference dihedral angle. The last two terms in equation 1.7 describe the non-bonded interactions (Lennard-Jones and electrostatic interactions), where q$_i$ and q$_j$ represent the charge of atoms i and j, r$_{ij}$ is the distance between atoms i and j, ε_{ij} is the potential well-depth of the interaction between atoms i and j, and σ_{ij} is the distance at which the potential between atoms i and j is zero, giving a measure of how close the to atoms can get (also known as the van der Waals radius). Common potentials used in MD are described in the following sections, along with important concepts for the development of this thesis.

1.2.1. The Lennard-Jones Potential

Van der Waals interactions modelled by the Lennard-Jones (LJ) potential, describe the behaviour of the repulsion-dispersion interactions between atoms *i* and *j*:

$$\phi_{LJ}(\mathbf{r}) = 4\epsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] \tag{1.8}$$

Repulsion effects due to overlap between atoms is modelled by the r^{-12} term (collision diameter) and the attraction effect modelled by the r^{-6} term. σ_{ij} corresponds to the contact distance between atoms *i* and *j* and ϵ_{ij} the magnitude in the energy well depth. Moreover, the LJ potential has an infinite range that may be treated with a cut-off radius R_c, leaving out interactions between atoms separated by more than this distance. This simple truncation generates artefacts, where jumps in the energy are observed every time a particle crosses the cut-off. As an example, the use of shifted potentials alleviate this by making the Lennard Jones potential 0 at the cut-off radius:

$$U(\mathbf{r}) = \begin{cases} \phi_{LJ}(\mathbf{r}) - \phi_{LJ}(R_c) & \text{if } r \leq R_c \\ 0 & \text{if } r > R_c \end{cases}$$

1.2.2. Electrostatic Interactions

The electrostatic interaction energy term for a pair of atoms ij, with a specific charge $q_{i/j}$ on each of them (atomic partial charge), is given by Coulomb's law:

$$U_{coul}(r) = (4\pi\epsilon_0\epsilon_s)^{-1}\sum_i\sum_{j>i}\frac{q_iq_j}{r_{ij}} \tag{1.9}$$

where ϵ_0 and ϵ_s correspond to the vacuum and the medium relative permittivity, and r_{ij} is the distance between the pair of charges. The sum runs over all pairs of atoms i and j.

Different methods have been developed in order to reduce the computational time in order to achieve near experimental conditions [18–20], such as cut-off truncation methods and lattice summation schemes. A description for both is given below.

1.2.2.1. Cut-off Truncation Methods

In the cut-off truncation scheme (CT, or CM for molecular based cut-off truncation), similar to the LJ potential, electrostatic interactions may be limited using a cut-off radius, where the electrostatic term shown in equation 1.9 is re-formulated as a truncated sum, only including interactions between atoms i and j within a certain distance [21]. This can considerably reduce the computational cost. Even so, the cut-off noise produced by this is not negligible, given that Coulomb interactions decay as $1/r$, with r being the distance, having a significant long-range component. The following methods have been developed in order to reduce these artefacts.

Straight Cut-off Scheme: Straight cut-off schemes (SC) [22,23] rely on preserving the Coulombic form shown in equation 1.9, but truncating electrostatic interactions at a certain cut-off R_C, as:

$$U_{elec}(r) = (4\pi\epsilon_0\epsilon_s)^{-1}\sum_i\sum_{j>i}q_iq_jU_{SC} \tag{1.10}$$

$$U_{SC}(r) = r_{ij}^{-1} - R_C^{-1} \tag{1.11}$$

where R_C corresponds to the cut-off distance. The main problem in this scheme is the lack of contributions to the energy from atoms outside the cut-off, which are neglected.

Reaction Field Scheme: The reaction field (RF, or BM for Barner-Watts molecular based reaction field cut-off) scheme [18, 24–26] tries to overcome the big approximation made in SC methods, where atoms outside a certain cut-off distance are neglected in the calculation of the energy. This is done by altering the functional form as:

$$U_{elec}(r) = (4\pi\epsilon_0\epsilon_s)^{-1} \sum_{i} \sum_{j>i} q_i q_j U_{RF} \tag{1.12}$$

$$U_{RF}(r) = r_{ij}^{-1} + \frac{\epsilon_{RF} - 1}{2\epsilon_{RF} + 1} \frac{r^2}{R_C^3} - \frac{3\epsilon_{RF}}{2\epsilon_{RF} + 1} \frac{1}{R_C} \tag{1.13}$$

This method includes the omitted electrostatics beyond the cut-off distance R_C, through the assumption of a dielectric continuum surrounding the cut-off sphere of each particle, with a dielectric permittivity equal to ϵ_{RF}, rather than vacuum as in SC methods.

1.2.2.2. Lattice Summation Scheme

In this scheme, the electrostatic term from equation 1.9 is reformulated as a *lattice-sum* (LS), including pair interactions between all sites *i* and all site *j* within the computational system, and with the possible periodic copies (see below section 1.3). This pair interaction function is written, as:

$$\varphi(\mathbf{r}) = \varphi_{sr}(\mathbf{r}) + \varphi_{\ell r}(\mathbf{r}) \tag{1.14}$$

where $\varphi_{sr}(\mathbf{r})$ and $\varphi_{\ell r}(\mathbf{r})$ represent the short and long range interactions. The first term converges quickly in real-space as it decays fast and is negligible beyond a cut-off distance, similar as short-ranged electrostatic interactions (see equation 1.9) [27]. The second term converges quickly in reciprocal-space and it is written in the form of

$$\varphi_{\ell r}(\mathbf{r}) = 4\pi |\underline{\mathbf{L}}|^{-1} \sum_{\mathbf{l}, \mathbf{l} \neq 0} k^{-2} \exp\left[i\mathbf{k} \cdot \mathbf{r}_{ij}\right] \tag{1.15}$$

where $\underline{\mathbf{L}}$ corresponds to a 3x3 matrix containing the Cartesian components of the box-edge vectors in its columns, $\mathbf{l}$ is a vector with integer components (positive or negative), $\mathbf{k}$ is the associated reciprocal-lattice vector with value $\mathbf{k} = 2\pi\underline{\mathbf{L}}^{-1}\mathbf{l}$, and $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$.

An excluded-sites term U_{exc} and a self-term U_{slf} are commonly introduced. The first term is included to account for the fact that excluded sites (first and second neighbour atoms) should only be excluded in their direct Coulombic interactions, but not in terms of their intermolecular interactions across periodic boundaries, and the term U_{slf} accounts for the interaction of a site i with its own periodic copies. For more detail on the mathematical expression of this terms please refer to references [21, 27].

The Particle Mesh Ewald method [28] (or PME) is one of the most widely used methods for the calculation of long-range electrostatic interactions, which reduce the computational complexity from O(N^2) to O(N log(N)). Point charges with their continuous coordinates are replaced by a grid based charge density. Short range converged contributions are calculated in real space, and the reciprocal-space term for long-range electrostatic interactions is approximated by a discrete convolution on a grid using discrete Fast Fourier transforms.

1.3. Boundary Conditions

The type of boundary condition employed in the systems of interest are closely related with the previously mentioned electrostatic schemes. Approximations are needed to overcome the computational limits in order to simulate systems that closely represent reality, since it is not possible (at the moment) to simulate a macroscopic system (non-periodic) that is capable of calculating direct Coulombic interactions for each existing atomic pair. In the context of solvation free energies, the system has to be large enough to include all the interactions that significantly contribute to the free energy.

Different representations of boundary conditions have been proposed, and most of them are based on the fact that multiple replicas are added at the boundaries of the "physical constructed system", that is, the system that you can actually see in the visualisation software (central box). Whenever a molecule (such as water molecule, or a whole protein) crosses one of the system edges, a copy of this molecule will enter on the opposite side. Graphical representation of the most widely used boundary conditions is shown in figure 1.1.

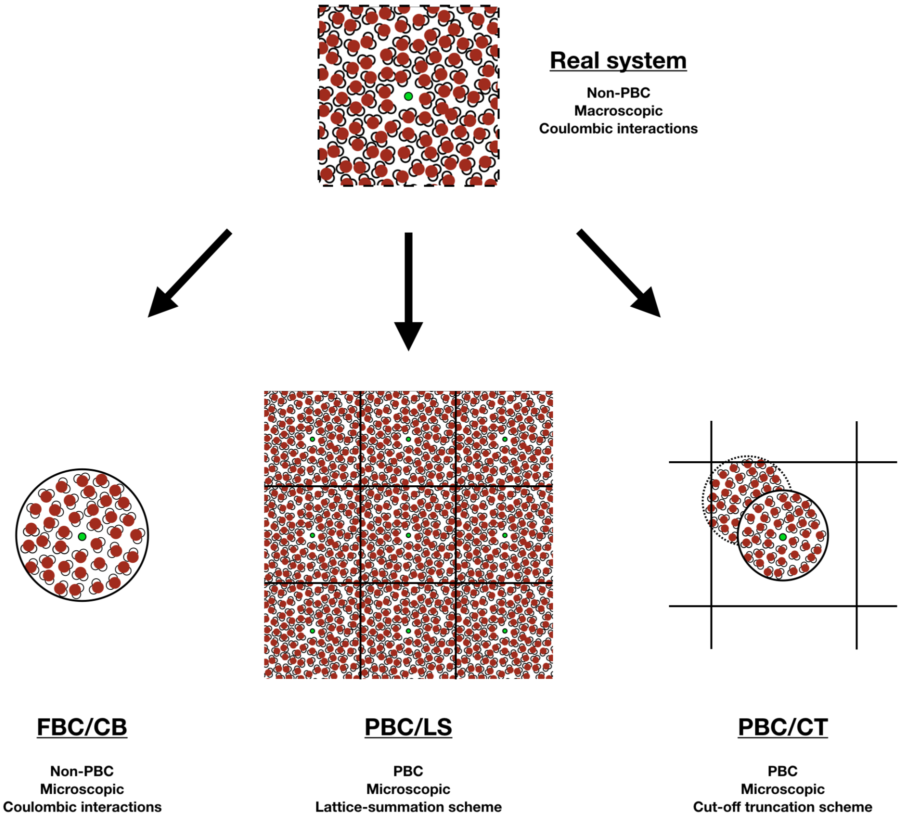


Figure 1.1

Schematic representation of periodic boundary conditions. Different types of approximations, in relation to a "real system" (top), are shown. Fixed boundary conditions with Coulomb interactions (FPB/CB, bottom left), periodic boundary condition with lattice-sum scheme (PBC/LS, bottom middle) and periodic boundary condition with straight cut-off scheme (PBC/CT, bottom right). Water molecules are represented as red and white dots and ions are represented as green dots**.**

1.4. Time Evolution

The basis of molecular dynamics is the integration of the Newton's equations of motion for the particles involved in order to follow their trajectory. The integration algorithms are based on finite difference methods where time is divided into discrete time-steps Δt. Knowing the positions of the particles and their derivatives in time, the position at time t + Δt can be known. From the potential energy term in equation 1.7, the force can be calculated (equation 1.3), and with this the time evolution of the system can be determined.

Different algorithms (known as integrators) are used to integrate the equations of motion based on finite differences, such as the Verlet [29], leapfrog [30] and Velocity Verlet [31] algorithms. As these methods are approximations, they have associated errors, such as truncation errors related to the accuracy of the finite difference method and round-off errors which are related to a particular implementation of the algorithm (e.g. finite number of digits used in the computation) [10].

The well-known Verlet and leapfrog algorithms are simple numerical schemes that are widely used in MD simulations, and they are completely equivalent algebraically. The Verlet algorithm (also known as Störmer method) is derived using a Taylor expansion over time t of the coordinate of a particle $\mathbf{r}(t)$, as:

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}(t) + \Delta t\mathbf{v}(t) + (\frac{\Delta t^2}{2})\mathbf{a}(t) + O(\Delta t^3) \tag{1.16}$$

where $\mathbf{v}(t)$ and $\mathbf{a}(t)$ are the velocity and acceleration of the coordinates at time t. If we subtract the corresponding expansion for $\mathbf{r}(t-\Delta t)$ and rearrange, we get:

$$\mathbf{r}_i(t+\Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \Delta t^2\mathbf{a}(t) + O(\Delta t^4) \tag{1.17}$$

The previously mentioned truncation error is in the order of $O(\Delta t^4)$ because of the cancellation of the Δt^3 term.

Limitations of the Verlet algorithm, is that velocities are not directly generated (see equation 1.17). Even though the velocities are not necessary to evaluate the time evolution of the system, they are usually needed to compute the Kinetic energy, which is needed for the evaluation of the conservation of the total energy [10, 32]. An alternative is the Velocity Verlet scheme, where positions are computed in a half-step manner. The Velocity Verlet goes as:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}(t) + \Delta t[\mathbf{v}(t) + (\frac{\Delta t^2}{2})\mathbf{a}(t)] \tag{1.18}$$

$$\mathbf{v}_i(t + \frac{1}{2}\Delta t) = \mathbf{v}(t) + \frac{1}{2}\Delta t\mathbf{a}(t)] \tag{1.19}$$

$$\mathbf{a}_i(t + \Delta t) = -(\frac{1}{m})\nabla U(\mathbf{r}(t + \Delta t)) \tag{1.20}$$

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}(t + \frac{1}{2}\Delta t) + \frac{1}{2}\Delta t\mathbf{a}(t + \Delta t)] \tag{1.21}$$

Another alternative, similar to Velocity Verlet, is the leapfrog method, where the calculation of positions and velocities leap between each other, and can be written as:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}(t) + \Delta t[\mathbf{v}(t) + (\frac{\Delta t}{2})\mathbf{a}(t)] + O(\Delta t^3) \tag{1.22}$$

The term multiplying Δt is $\mathbf{v}(t + \Delta t/2)$, which is then transform to:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}(t) + \Delta t\mathbf{v}(t + \frac{\Delta t}{2}) \tag{1.23}$$

Subtracting from $\mathbf{v}(t + \Delta t/2)$ the corresponding expression for $\mathbf{v}(t - \Delta t/2)$, we obtain:

$$\mathbf{v}(t + \frac{\Delta t}{2}) = \mathbf{v}(t - \frac{\Delta t}{2}) + \Delta t\mathbf{a}(t) \tag{1.24}$$

The leap-frog algorithm evolves the positions **r** and velocities **v** in parallel but shifted by a half time-step.

1.5. Free energy

Free energy can be related to the probability of finding a particular system in a specific state. Most important, the difference in free energy between two or more states is a quantity of relevance in many biological and chemistry areas, such as the analysis of receptor-ligand complex or for the in-silico analysis of drug molecules with their targets [33]. A good estimation of the free energy differences using molecular dynamics simulations has been a challenge for many decades. The possibility to access macromolecular states inaccessible by experiments (e.g. alchemical transformations of a ligand inside a protein), and the capability to greatly reduce the measurement of thermodynamic properties by experiment, has caused great interest in the area of computational free energy estimations [34, 35].

In the canonical ensemble, the Helmholtz free energy F of a system is given as:

$$F(N, V, T) = -k_B T \ln[N! h^{-3N} \int \int e^{-\beta U(\mathbf{r})} d\mathbf{r}] \tag{1.25}$$

where V is the volume, T the absolute temperature, k_B Boltzmann's constant, h is the Planck's constant, $\beta = 1/k_B T$, and U is the potential energy of the system, which is a function of the system coordinates $\mathbf{r}$.

Some free energy methods receive the name of Alchemical transformations. This is based on the old alchemists who believed that they could transform matter by changing the property of atoms. As free energy is a state function, i.e. it does not depend on the path that connects both ends, we can build a pathway with non-physical characteristics that facilitate the calculation of the difference in free energy between two states.

Thermodynamic integration and the free energy perturbation are some of the most well-known alchemical methods to compute the free energy difference between states for a given system [33–35]. These two methods are based on the fact that the free energy difference of a system can be related to the configuration integral Z as:

$$F(N, V, T) = -k_B T \ln Z \tag{1.26}$$

The configuration integral can be described as a function that represents a measure of the whole space accessible to the system of interest, and from which we can derive different thermodynamic properties such as entropy, pressure and free energy [36]. The classical configuration integral is given as:

$$Z = \int\int e^{\frac{-(U(\mathbf{r}))}{(k_B T)}} d\mathbf{r} \tag{1.27}$$

In order to get the free energy difference between a state X and Y, it is useful to include a coupling parameter to connect both ends. This coupling parameter, denominated α, changes from 0 to 1 and can be expressed as a linear function of the potential energy by:

$$U(\mathbf{r}; \alpha) = \alpha U_0(\mathbf{r}) + U_1(\mathbf{r})(1 - \alpha) \tag{1.28}$$

where $U_0(\mathbf{r})$ corresponds to the potential energy of a denominated "reference system" and $U_1(\mathbf{r})$ corresponds to the potential energy of a system of interest. α connects two or more states through a physical or non-physical pathway.

1.5.1. Free energy perturbation

Perturbation theory is one of the oldest mathematical techniques, and can be applied to different areas such as chemistry, physics, economics and engineering. We start with an initial system of interest, called the unperturbed or reference state, that is used to solve a problem of interest, called the target state, based on perturbations of the initial problem (reference state). Small perturbation steps are made by the α parameter. The initial steps and derivation of free energy perturbation theory can be found in references [37–39]. Unlike other perturbation theories, this is exact.

A reference system can be described by the Hamiltonian $H(\mathbf{p}, \mathbf{r})$, which is a function of a $3N$ coordinates of $\mathbf{r}$, and their momenta $\mathbf{p}$

$$H(\mathbf{p},\mathbf{r}) = K(\mathbf{p}) + U(\mathbf{r}) \tag{1.29}$$

$$H(\mathbf{p},\mathbf{r}) = \sum_{i=1}^{N} \mathbf{p}_i \frac{1}{2m_i} + U(\mathbf{r}_1, \ldots \mathbf{r}_N) \tag{1.30}$$

where K is the kinetic energy, U the potential energy and m_i the mass of atom i. With this, the free energy difference between two states can be expressed as the difference in the Hamiltonian

$$\Delta H(\mathbf{p},\mathbf{r}) = H_1(\mathbf{p},\mathbf{r}) - H_0(\mathbf{p},\mathbf{r}) \tag{1.31}$$

Depending on the system of study, the Hamiltonian will describe different contributions or interactions.

As stated in equation 1.22, the free energy can be related to the partition function. With this, the difference in free energy can be related to the ratio of the partition function for both states:

$$\Delta F(N,V,T) = -k_B T \ln \frac{Z_1}{Z_0} \tag{1.32}$$

Now, replacing the value for the partition function in equation 1.28:

$$\Delta F(N,V,T) = -\frac{1}{\beta} \ln \frac{\int\int e^{-\beta(H_1(\mathbf{r},\mathbf{p}))} d\mathbf{r} d\mathbf{p}}{\int\int e^{-\beta(H_0(\mathbf{r},\mathbf{p}))} d\mathbf{r} d\mathbf{p}} \tag{1.33}$$

where $\beta = (k_B T)^{-1}$. Rearranging equation 1.29

$$\Delta F(N,V,T) = -\frac{1}{\beta} \ln \frac{\int\int e^{-\beta \Delta H(\mathbf{r},\mathbf{p})} e^{-\beta(H_0(\mathbf{r},\mathbf{p}))} d\mathbf{r} d\mathbf{p}}{\int\int e^{-\beta(H_0(\mathbf{r},\mathbf{p}))} d\mathbf{r} d\mathbf{p}} \tag{1.34}$$

We can build a histogram of the different states that we are encountering during the simulation, as function of the Hamiltonian. The probability of being in a particular state across our simulation (in this case state 0) is given by the probability density function:

$$P_0(\mathbf{r},\mathbf{p}) = \frac{e^{-\beta(H_0(\mathbf{r},\mathbf{p}))}d\mathbf{r}d\mathbf{p}}{\int\int e^{-\beta(H_0(\mathbf{r},\mathbf{p}))}d\mathbf{r}d\mathbf{p}} \tag{1.35}$$

Replacing this in equation 1.28

$$\Delta F(N,V,T) = -\frac{1}{\beta}\ln\int\int e^{-\beta\Delta H(\mathbf{r},\mathbf{p})}P_0(\mathbf{r},\mathbf{p})d\mathbf{r}d\mathbf{p} \tag{1.36}$$

which is equal to

$$\Delta F(N,V,T) = -\frac{1}{\beta}\ln\langle e^{-\beta\Delta H(\mathbf{r},\mathbf{p})}\rangle_0 \tag{1.37}$$

which is the fundamental formula for the free energy perturbation method. With this, we obtain the free energy difference of two states based on the ensemble averages (represented by the angled brackets ⟨...⟩) for simulations of the reference system 0.

1.5.2. Thermodynamic integration

We can differentiate the Helmholtz free energy with respect to the coupling parameter α. First, replacing the value for the partition function in equation 1.22

$$F(N,V,T) = -k_BT\ln\int e^{\frac{-(U(\mathbf{r},\alpha))}{(k_BT)}}d\mathbf{r} \tag{1.38}$$

Now, the derivative of F with respect to α, $\frac{\partial F}{\partial\alpha}$, is given by

$$\frac{\partial F}{\partial\alpha} = -\frac{k_BT(-\frac{\frac{(\partial U(\mathbf{r},\alpha))}{\partial\alpha}}{k_BT})e^{\frac{-(U(\mathbf{r},\alpha))}{k_BT}}}{\int e^{\frac{-(U(\mathbf{r},\alpha))}{k_BT}}d\mathbf{r}} \tag{1.39}$$

$$\frac{\partial F}{\partial\alpha} = \langle\frac{\partial U(\mathbf{r},\alpha)}{\partial\alpha}\rangle_\alpha d\alpha \tag{1.40}$$

The change in the free energy, between a references state and a target state, can be computed from the integral between values of 0 (un-perturbed) and α (perturbed) of the ensemble average of the derivative of the potential energy with respect to the coupling parameter α, assuming linear coupling,

$$\Delta F_{\alpha} = F_{\alpha} - F_0 = \int_0^1 \langle \frac{\partial U(\mathbf{r}^N;\alpha)}{\partial \alpha} \rangle_{\alpha} d\alpha = \int_0^1 \langle \Delta U \rangle_{\alpha} \tag{1.41}$$

where $\Delta U = U_1 - U_0$

Equation 1.37 is known as the thermodynamic integration, and is exact for the linear form of the perturbation.

1.6. Bennett's acceptance ratio

The Bennett acceptance ratio method (BAR) [40] can also be used to calculate free energy difference for transformations from a state A to a state B. BAR requires information from the two states to estimate free energy differences (for example, two neighbouring simulations at different α values), rather than just one as is the case of FEP and TI. In the assumption that both states are in the same configuration, a pathway that connects both energy potentials can be found, and a difference in the potentials. An alternative to this method is the Multistate Bennett's Acceptance Ratio (MBAR) [41], combining data from multiple states and predicting the free energy at an un-sampled state.

1.7. Enhanced-sampling methods

Often, free energy changes can be studied as a function of some inter or intra-molecular coordinate, such as the distance between a group of atoms, a torsion angle, or the mean square deviation with respect to a reference structure. The free energy surface with respect to this chosen coordinate is known as the potential of mean force (PMF). In the case of biological systems, it is usual to see potential energy surfaces with multiple local minima, separated by high energy barriers. This is a limitation for classic MD simulations, making it difficult to sample a significant proportion of the configurational space, and leaving the system trapped in one or few minima. Important conformational changes in proteins are key to their function in most biological processes, such as membrane transport, enzyme catalysis and ligand recognition, so the correct study of these changes is of great importance. A considerable number of enhanced-sampling methods have been developed to overcome this issue of conformational sampling, allowing us to accelerate dynamics in biological systems.

These methods are roughly divided in two types:

1. Deformation and smoothing of the potential energy surface by the identification of important and slowly varying order parameters (collective variables), in order to access a broad spectrum of the configurational space. Methods such as local elevation [42], metadynamics [43] and umbrella sampling [44] fit in this class.

2. Scaling the conditions (and/or parameters) of a system, employing the use of multiple copies. Methods such a replica-exchange [45, 46] (temperature or Hamiltonian replicas [47]) can fit in this description, where copies of the system at different temperatures, or copies with different scaled force field parameters, are used.

1.7.1. Metadynamics

This method, which belongs to the first class previously mentioned, is based on the introduction of a bias potential acting on specific degrees of freedom, or collective variables (CV), discouraging the system to revisit configurations (memory-based biasing methods [42]), and directing computational resources to a broader exploration of the free energy landscape. Metadynamics defines variables which represent slow coordinates of the system of interest, mentioned before as CV. An extra potential bias is used, and it usually takes the form of a sum of Gaussians centred at a specific CV value. This bias is a function of the CV, and the CV is a function of the system coordinates **r**,

$$s(\mathbf{r}) = (s_1(\mathbf{r}), \ldots, s_n(\mathbf{r})) \tag{1.42}$$

where s corresponds to the CV, and can be a sum of a group of other CV, that in this case go from 1 to n. At time t , the extra bias potential is written as,

$$U_{\mathrm{G}}(s,t)=\int_0^t dt'\omega\exp\left(-\sum_{i=1}^{n}\frac{\left(s_i(\mathbf{r})-s_i\left(\mathbf{r}\left(t'\right)\right)\right)^2}{2\sigma_i^2}\right) \tag{1.43}$$

where ω is the energy rate, and σ_i is the Gaussian width. The energy rate is usually expressed in terms of the Gaussian height W and the frequency τ_G at which the Gaussians are added:

$$\omega=\frac{W}{\tau_{\mathrm{G}}} \tag{1.44}$$

Metadynamics lets us accelerate the sampling of rare or high energy configurations, and explore new reaction pathways. No previous knowledge of the energy landscape is required, and the low-energy configurations are first explored [48, 49]. However, limitation do exist for this method, such as the bias potential overfills the energy surface, pushing the system towards high-energy regions of the CV space, and that the choice of the CV is not trivial, and is key for the correct study of the system of interest. Solutions to this issues have been developed, such as well-tempered metadynamics [50] where the bias deposition rate decreases over time, as

$$W=W_0\exp\left(-\frac{U(s(\mathbf{r}))}{k_B\Delta T}\right) \tag{1.45}$$

where W_0 is the initial Gaussian height, and ΔT is an input parameter with the units of temperature. With this, the bias potential smoothly converges in the long time limit [50], the free energy surface can be estimated as

$$U(s,t\to\infty)=-\frac{\Delta T}{T+\Delta T}F(s)+C \tag{1.46}$$

where T is the system's temperature. Thus, the CVs will sample an ensemble at a temperature T+ΔT, which is higher than the system's temperature T. The term "bias factor" is usually encountered, which expresses the ratio between the temperature of the CVs (T+ΔT) and the system's temperature T, and its value has to be carefully chosen in order for the relevant free-energy barriers to be crossed efficiently in the time scale of the simulation.

1.7.2. Umbrella sampling

Similar to metadynamics, the umbrella sampling method attempts to overcome the difficulties in the sampling of complex free energy surface with multiple minima, but in this case, modifying the potential so un-favourable configurations are sampled sufficiently. For this, the reaction coordinate, or CV, is usually split into windows, where each is characterised by an appropriate bias potential [33, 44, 51].

Being the bias potential w_i for a window i, and that only depends on the reaction coordinate ξ,

$$U^{\mathrm{b}}(\xi) = U^{\mathrm{u}}(\xi) + w_i(\xi) \tag{1.47}$$

$U^{\mathrm{b}}(\mathbf{r})$ and $U^{\mathrm{u}}(\mathbf{r})$ correspond to the biased and unbiased potential energy, respectively. w_i usually takes the form of a quadratic form (harmonic potential):

$$w_i(\xi) = k_w(\xi - \xi_0)^2 \tag{1.48}$$

with k_w being the force constant of the harmonic potential, ξ the calculated coordinate, and ξ_0 the reference coordinate.

Configurations generated that are far from the reference position ξ_0 will have a large weighted function, and the energy function $U^{\mathrm{b}}(\xi)$ will be biased towards a relevant conformation, with a non-Boltzmann distribution. The corresponding unbiased distribution $\langle A \rangle$ can be extracted from the biased distribution by the method developed by Torrie and Valleau, 1977 (see reference for full derivation), given by,

$$\langle A \rangle = \frac{\langle A(\xi) exp[w_i(\xi)/k_B T] \rangle_w}{\langle exp[w_i(\xi)/k_B T \rangle_w} \tag{1.49}$$

where the subscript w indicated that the average is taken from a biased probability, which is determined by the modified energy $U^{\mathrm{b}}(\xi)$. It is important to notice that If the forcing potential $w_i(\xi)$ is too large for the system of study, the denominator in equation 1.43 will be dominated by contribution from $exp[w_i(\xi)]$, and the averages will take too long to converge.

Our main goal is to obtain a full distribution function, which contains all the information for the reaction coordinate of interest. The weighted histogram analysis method (WHAM) [52, 53] is usually used to combine information from multiple simulations. As mentioned before, a number of simulations are performed at different values of the reaction coordinate, with an extra umbrella biased $w_i(\xi)$ for each simulation [33]. At simulation i, the unbiased probability distribution $\mathrm{p}_i(\xi)$ can be reconstructed from the biased probability distribution $\mathrm{p}'_i(\xi)$, and presented as a normalised histogram,

$$p_i(\xi) = exp[(f_i - w_i(\xi))/k_B T] p'_i(\xi) \tag{1.50}$$

where f_i is the free energy difference between the biased state and the unbiased reference state. In order to construct an unbiased probability distribution $\mathrm{p}_0(\xi)$ from the combination of all $\mathrm{p}_i(\xi)$, and to obtain the free energy as a function of the reaction coordinate ξ, we have,

$$p_0(\xi) = \sum_{i=1}^{n} \frac{N_i exp[-(f_i - w_i)/k_B T]}{\sum_{j=1}^{n} N_j exp[(f_j - w_j)/k_B T]} p_i(\xi) \tag{1.51}$$

$$p_0(\xi) = \sum_{i=1}^{n} \frac{N_i}{\sum_{j=1}^{n} N_j exp[(f_j - w_j)/k_B T]} p'_i(\xi) \tag{1.52}$$

where n correspond to the number of simulations and N_i the number of configurations collected for simulation i. As the free energy parameters f_i are unknown in the simulations, evaluation of equations 1.46 and 1.47 is needed [33]. This is done in a self-consistent manner for all i=1 ... n,

$$exp(-\frac{f_i}{k_B T}) = \sum_{j=1}^{N_i} \frac{exp(-w_i/k_B T)}{\sum_{k=1}^{n} N_k exp[(w_k - f_k)/k_B T]} \tag{1.53}$$

1.8. Summary

Molecular dynamics is an important and useful method to study the dynamics of a wide range of research areas, such as liquids, drug design and biomolecules. The core of molecular dynamics can be divided in two: 1) the existence of a energy potential, that depends on the position of the particles, and from where we can obtain the forces involved, and 2) a time-series of the these forces, in order to follow the dynamics of the systems, which is given by an integrator. From here, we can calculate important information such as thermodynamic properties and dynamics of the system of interest.

Chapter 2

Coarse-Grain Modelling of Proteins

2.1. Introduction

Many important and crucial biological and chemical processes occur at time-scales of seconds to hours. Computational atomistic modelling of these processes is unable to achieve these kinds of timescales [9]. The computation power to perform longer simulations has been growing exponentially, according to Moore's Law [54]. Even so, the need for new solutions to bypass this problem is needed. Coarse-grained models are a widely adopted approach for improving the efficiency of computer modelling of biological systems. By subsuming groups of atoms into single interaction sites, much faster and longer calculations can be realised [9].

One of the first coarse-grained protein models developed 40 years ago was that of Levitt and Warshel [55] where the backbone is represented as pseudo-atoms centred in the Cα position, and side-chains were replaced by united pseudo-atoms at centres of their average conformations. The Lennard-Jones potential describes the interactions between beads, and the model was mainly used for protein folding studies. Brownian dynamics was used as the sampling scheme [9, 55]. In the original paper [55] they managed to fold a linear protein sequence to an almost native-like conformation. After this, Levitt proposed another model which accounted for the variable orientation of the united side chains [56]. Based on the statistical analysis of conformational properties of representative di-peptides, Levit *et al.* managed to derive the torsional potentials for the main chain degrees of freedom, making the model slightly more accurate than before. Later models based on these were derived, with statistical potentials for reside-residue interactions and used the Monte Carlo method for simulated annealing simulations of the folding process [57].

A different area was developed, related to protein-like models, using cubic lattice chains and their folding to unique three dimensional structures [58, 59]. These models were based on two types of amino-acid side-chains (hydrophobic and polar). It was shown that specific sequences are needed for the folding to unique structures, describing dynamics and thermodynamics of folding processes in these idealized systems. Lattice models of intermediate resolution (i.e. between idealistic "protein-like" and crude protein models) such as the diamond lattice [60] or “chess-knight” models [61,62] were also developed. These studies of low resolution models, independent of their crude representation, compose the foundation for the development of contemporary CG protein force fields, such as intermediate resolution models where one or two united atoms are used to approximate the geometry of the main chain and side chains, respectively (e.g. UNRES [1] and CABS [2] force fields) (see below). Other types of models use near atomistic resolution, where only small simplifications are made to speed-up simulations (e.g. PRIMO [3] and SIRAH [5] force fields) (see below), and they are designed to mainly study protein interactions and structure prediction.

The representation of coarse-grained models and force fields needs to be carefully constructed and parameterised. Different coarse-grained protein models have been developed, which differ on the mapping and levels of resolution, secondary structure stabilisation, bead-types and MD engines that support their use [1–5, 9]. One of the main reasons for the development of this type of model was to understand and perform simulations of biological processes at time scales important for their function. In comparison with atomistic force fields, most coarse-grained models smooth out the energy landscape, which accelerates the dynamics [9], but can also lead to a misinterpretation of the results. Designing these types of force field have followed two main approaches: i) based on molecular physics of proteins and ii) derived from the analysis of structural regularities found in databases of experimental protein structures. This data can come from experimental data (top-down coarse-graining) or from an underlying atomistic model (bottom-up coarse-graining) [13]. The accuracy of these force fields will depend on the empirical parameters in the model. These are usually iterated "by hand", making this process tedious and prone to errors [10]. Two main problems have been stated in the development of coarse-grained force fields: i) representability, which is related to the ability of a CG model to represent and reproduce physical properties at the state point it was parameterised, and ii) transferability, which is concerned with the ability of the same model to reproduce properties at a different state point where the parameterisation was not performed [13]. Here we briefly review some popular protein coarse-grained force fields.

2.2. UNRES Force Field

In this model [1], where the name stands for united-residues, a polypeptide chain is represented by a sequence of Cα carbons (C^{α}) linked by virtual bonds. United beads that represent a peptide group (p_i) are located between each pair of Cα carbons, and united side-chains (SC), with an ellipsoidal shape and associated rotation that mimic side-chain mobility, are directly linked to the Cα carbons (Fig. 2.1). Only the united SC and p_i beads work as interacting sites. **dX** and **dC** in figure 2.1 represent vectors that describe the geometry of the side-chain, and θ and γ represent the p-Cα-p angle and p-Cα-p-Cα dihedral, respectively [1] (see Fig. 2.1). Most of the parameters in the UNRES force field have been derived from the distribution and correlation functions from the Protein Data Bank and against free energies from atomistic polypeptide simulations in explicit water (see below) using a functional form that resembles atomistic force fields but with additional terms [1, 63–65]. A re-parameterisation was later made for most terms, based on *ab initio* and semi-empirical energy surfaces [1]. Final reparameterisations of the model have been focused on TIP3P water and AMBER force field for models for pairs of amino-acid side-chains and determining the respective PMF as functions of the distance between centres [1, 63–65].

The focus of the UNRES energy function is the potential of mean force (PMF) of the system under study in water, in which some degrees of freedom not present in the coarse-grained model have been averaged out, and it has been parameterised from PMF surfaces calculated for model systems by means of *ab initio* molecular quantum mechanics or from all-atom MD simulations [65, 66]. UNRES defines a restricted free energy function (RFE), which corresponds to averaging the energy over the degrees of freedom that are neglected in the UNRES model. The UNRES force field can be defined as a free energy of a given coarse-grain conformation given by the primary degrees of freedom X (which describe the coarse- grained degrees of freedom), with integration over the secondary degrees of freedom y (less important variables that are averaged out). The RFE is expressed as,

$$F(\mathbf{X}) = -RT \ln \left\{ \frac{1}{V_{\mathrm{Y}}} \int_{\Omega_{\mathrm{Y}}} \exp[-E(\mathbf{X};\mathbf{Y})/RT] \mathrm{d}V_{\mathrm{Y}} \right\} \tag{2.1}$$

where $V_{\mathrm{Y}} = \int_{\Omega_{\mathrm{Y}}} \mathrm{dV_Y}$, E(**X**;**Y**) is the original atomistic energy function (which is expressed as a sum of component energies, see below), R is the universal gas constant, T is the absolute temperature, Ω_{Y} is the region of **Y** of variables over which the integration is carried out, and $\mathrm{V_Y}$ is the volume of this region [1, 63–65]. Again, E(**X**;**Y**) is expressed as a sum of component energy components,

$$E(\mathbf{X};\mathbf{Y}) = \sum_{i=1}^{n} \varepsilon_i \left(\mathbf{X};\mathbf{z}_i\right) \tag{2.2}$$

where $\varepsilon_i \left(\mathbf{X};\mathbf{z}_i\right)$ is the *i*th component energy, $\mathbf{z}_i$ contains the secondary degrees of freedom on which ε_i depends, and *n* is the number of energy components. This RFE is decomposed into factors, where each corresponds to a Kubo cluster-cumulant function (i.e. a Taylor expansion with respect to an abstract element λ_i, given a function $F = \ln \left\langle e^{\lambda_1 A_1 + \lambda_2 A_2 + \dots} \right\rangle$), as,

$$
\begin{aligned}
F(\mathbf{X}) = \sum_{i} f_i^{(1)}(\mathbf{X}) + \sum_{i<j} f_{ij}^{(2)}(\mathbf{X}) + \sum_{i<j<k} f_{ijk}^{(3)}(\mathbf{X}) + \ldots \\
+ \sum_{i_1<_2\ldots<i_n} f_{i_1 i_2 \ldots i_n}^{(n)}(\mathbf{X})
\end{aligned}
\tag{2.3}
$$

Here, the factors of the first order, f(1), correspond to the PMF of isolated units (isolated amino-acid residues) or those between isolated pairs of units (pairs of interacting side chains), while factors of order 2 and higher correspond to the multibody or correlation terms. For more details on how the factors are expressed, and the explicit temperature dependence in the UNRES force field, please refer to references [1, 63–65].

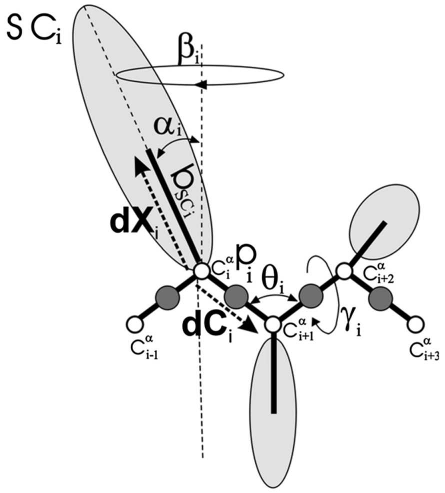


Figure 2.1

Coarse-grained mapping scheme for the UNRES protein model. The interaction sites are the Cα carbon (C_i^α), the peptide group (p_i) and the side-chain ellipsoids (SC_i). **dC** and **dX** correspond to the virtual bond vectors Cα-Cα and Cα-SC, respectively. Image taken from [1]**.**

The UNRES energy function is given by:

$$
\begin{aligned}
U =& w_{SCSC}\sum_{i<j} U_{SC_iSC_j} + w_{SCp}\sum_{i\neq j} U_{SC_ip_j} \\
&+ w_{el} f_2(T) \sum_{i<j-1} U_{p_ip_j} \\
&+ w_{tor} f_2(T) \sum_i U_{tor}(\gamma_i) \\
&+ w_{tord} f_3(T) \sum_i U_{tord}(\gamma_i, \gamma_{i+1}) \\
&+ w_b \sum_i U_b(\theta_i) + w_{rot} \sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i,\theta_i}) \\
&+ w_{bond} \sum_i U_{bond}(d_i) \\
&+ \sum_{m=2}^{N_{corr}} f_m(T) w_{corr}^{(m)} U_{corr}^{(m)} + f_3(T) w_{tum}^{(3)} U_{turn}^{(3)} \\
&+ f_4(T) w_{\mathrm{turn}}^{(4)} U_{\mathrm{turn}}^{(4)} \\
&+ w_{SS} \sum_{disulfidebonds} U_{SS;i} + n_{SS} E_{SS}
\end{aligned}
\tag{2.4}
$$

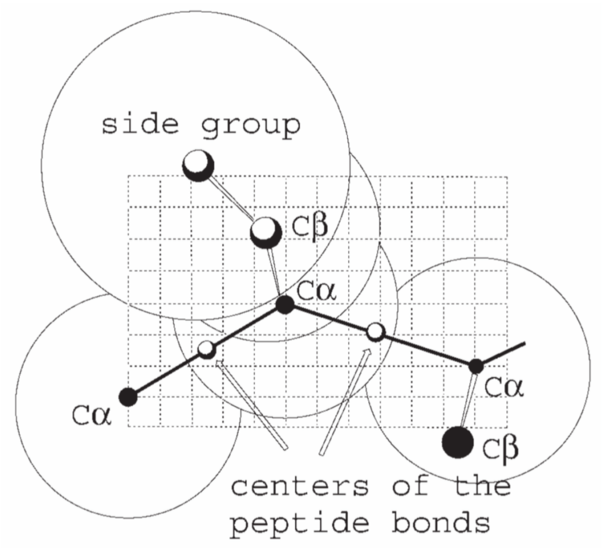
where the term $U_{SC_iSC_j}$ correspond to the mean free energy of solvent-mediated interactions between the side chains, $U_{SC_ip_j}$ correspond to the excluded-volume potential of the side chain-peptide group interactions, $U_{p_ip_j}$ correspond to the energy of mean-field electrostatic interactions between backbone peptide groups, U_{tor} and U_{tord} are the torsional and double-torsional potentials, for rotation about a given virtual bond or two consecutive virtual bonds, respectively. U_b and U_{rot} are the virtual bond angle-bending and side-chain rotamer potentials, respectively, and U_{bond} accounts for backbone and side-chain virtual-bond stretching. Correlation terms, $U_{corr}^{(m)}$ and $U_{turn}^{(m)}$ are used to reproduce secondary structures, such as α-helices and β-sheets. U_{SSi} is the energy of distortion of disulphide bonds from their equilibrium configuration, E_{SS} is the energy of formation of a disulphide bond in the chain (relative to the presence of two free cysteine residues), and n_{SS} is the number of disulphide bonds. For a more detailed derivation of the energy terms, please refer to [1]. Is important to note that the authors mention that significant improvements to the UNRES force field have to be made, where they are still missing details of local interactions [65]. As an example, torsional potentials involving the virtual Cα-SC bonds have been introduced [67]. Describing the effective SC-SC interaction potentials has also been a matter of discussion, given the actual use of Gay-Berne functional forms that imply axial symmetry, ignoring the fact that polar or charged side chains have non-polar necks and polar or charged headgroups [1].

This forcefield has been used in different protein folding studies [68–71], protein structure prediction [72], protein-protein interactions [73], protein-DNA interactions [74], replica exchange simulations [75] and the study of dynamics and interactions in larger systems [74, 76].

2.3. CABS Force Field

The name stands for Cα, β and side-chain, and has a similar resolution as UNRES, but differs in the mapping scheme, where only one bead is used for the main chain centred on the Cα position, and two beads are used for the side-chain, one centred on the Cβ position and another centred on the rest of the side-chain, when possible (Fig. 2.2) [2]. The Cα bead is restricted to a high resolution cubic lattice (0.61 Åspacing) [2, 77, 78]. The authors mention that fluctuations between the Cα-Cα virtual bond distance leads to up-to 800 different orientations [79]. The lattice representations accelerate the computation of local transitions, compared to continuous models.

The CABS force field is knowledge-based, where the parameters have been derived via Boltzmann inversion based on the statistical analysis of protein structures from the PDB database [2, 9, 77, 78]. Interactions are given by sequence-dependent short-range conformational propensities, a model of main-chain hydrogen bonds, and context-dependent potentials of pairwise interactions of side groups. Side-chain positions are based on the local main chain Cα-Cα-Cα angle and the type of amino-acid. This force field is defined as fully structural based, where the nature of the interactions will depend on the context, taking into account complex multi body effects [2, 77]. The solvent is treated implicitly. The present model does not work for simulations outside a "typical" pH, or in



Coarse-grained mapping scheme for the CABS protein model. Each large circle represents the position of each coarse-grained bead. Image taken from [2]**.**

Figure 2.2

simulation that evaluate ionic strength effects on protein structure and dynamics, making this one its main limitations [78], and a possible limitation of CG force fields.

This force field has been used in template-based or *ab initio* protein structure prediction [80], *ab initio* simulations of protein folding [78, 79, 81], modelling of protein flexibility [82] and flexible protein-peptide docking [83].

2.4. PRIMO Force Field

The name stands for **PR**otein **I**ntermediate **MO**del [3, 84, 85]. The backbone is represented with near-atomistic representation using three beads, N, Cα and CO atoms. This is a higher resolution model, compared with the UNRES and CABS models. The side-chains are represented using from one to four beads, depending on the size and shape (Fig. 2.3) [3]. The high resolution of the backbone ensures the preservation of hydrogen bond-like interaction, which are crucial for the secondary structure in proteins.

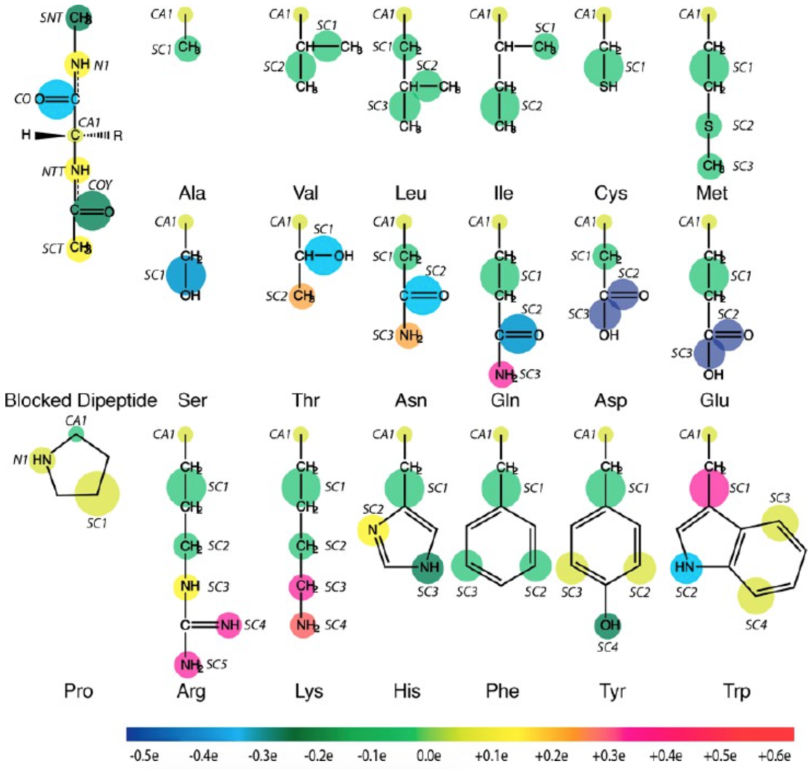


Figure 2.3

Coarse-grained mapping scheme for the PRIMO protein model side-chains. Size of each sphere corresponds to vdW radii and colours reflect the charge for each bead. Image taken from [3].

The potential energy function in PRIMO is represented with the standard force field-like form, with additional spline-based bonded terms to maintain correct bond geometries at the coarse-grained level:

$$U = U_{bond} + U_{angle} + U_{torsion} + U_{CMAP} \tag{2.5}$$

where

$$U_{bond} = \sum_{i=1}^{N_{bond}} K_i^{bond}(l_i - l_{i,0})^2 + \sum_{i=1}^{N_{bond-spline}} s_i^{1D}(i_i^{1-2}) \tag{2.6}$$

$$U_{angle} = \sum_{i=1}^{N_{angle}} K_i^{angle}(\theta_i - \theta_{i,0})^2 + \sum_{i=1}^{N_{angle-spline}} s_i^{1D}(i_i^{1-3}) \tag{2.7}$$

which correspond to covalent bonds, such as 1-2 (bonds), and 1-3 (angles) interactions, that are described by standard harmonic potentials (eq. 2.4 and 2.5, respectively. See equation 1.7 for similar term definitions). Virtual bonds are described primarily with 1D distance-based spline-interpolated potentials ($N_{bond-spline}$ and $N_{angle-spline}$ in eq. 2.4 and 2.5) to capture non-harmonic potential shapes and the presence of multiple minima. For some of the side-chains, these terms were not enough to maintain coherent geometries, where additional virtual sites are added on the fly in order to maintain correct geometries [3].

In PRIMO, 1-4 interactions are modelled by torsion terms ($U_{torsion}$), one-dimensional spline interpolated functions (U^{1-4}_{spline}) and a reduced Lennard-Jones potential (U^{1-4}_{LJ}) as follows:

$$U_{torsion} = \sum_{i=1}^{N_{torsion}} \sum_{j=1}^{N_i^{mult}} K_{ij}^{torsion}(1 + \cos(n_{ij}\phi_i - \phi_{ij,0})) \tag{2.8}$$

$$U^{1-4}_{spline} = \sum_{i=1}^{N_{torsion-spline}} s_i^{1D}(l_i^{1-4}) \tag{2.9}$$

$$U^{1-4}_{LJ} = \sum_{i=1}^{N_{atoms}-1} \sum_{j=i+1}^{N_{atom}} \epsilon_{ij}^{1-4}[(\frac{\sigma_{ij}^{1-4}}{r_{ij}})^{12} - 2(\frac{\sigma_{ij}^{1-4}}{r_{ij}})^6] \tag{2.10}$$

where ϕ correspond to the torsion angle, s_i^{1D} is the one-dimensional spline function, ϵ_{ij}^{1-4} and σ_{ij}^{1-4} are the scaled well-depth and van der Waals radius used in 1-4 interactions to avoid hard-sphere overlap (see equation 1.7 for similar term definitions). Also, a spline-interpolated two-dimensional cross-correlation term (U_{CMAP} in equation 2.3) based in the CMAP methodology is used in PRIMO to couple the sampling of CO-N-CA-CO and N-CA-CO-N torsions [3, 84, 85].

In relation to non-bonded interactions these are modelled by the classical Lennard-Jones and Coulombic functions (similar to equations 1.8 and 1.9), plus an explicit function for angles and distance dependent hydrogen bonds, that works as a complement for the reduced partial charges:

$$\begin{aligned} E_{HBOND} = & \sum_{i=1}^{N_{\text{HBOND3}}} f_3 s_i^{2\text{D}} \left(\cos\theta, l_i^{\text{N-CO}}\right) \\ & + \sum_{i=1}^{N_{\text{HBOND4}}} f_4 s_i^{2\text{D}} \left(\cos\theta, l_i^{\text{N-CO}}\right) \\ & + \sum_{i=1}^{N_{\text{HBOND}}} f_5 s_i^{2\text{D}} \left(\cos\theta, l_i^{\text{N-CO}}\right) \\ & + \sum_{i=1}^{N_{\text{HBONDN}}} f_n s_i^{2\text{D}} \left(\cos\theta, l_i^{\text{N-CO}}\right) \end{aligned} \tag{2.11}$$

where f_3, f_4, f_5 and f_N correspond to scaling factors that will adjust the strength of the interactions betweens residues i and i ±3, i ±4, i ±5, and i ±N, respectively, where N $>$ 5. These scaling factors were optimised conjointly with the partial charges by fitting total PRIMO internal energies to CHARMM atomistic energies, using a series of peptides and two protein systems, until stable trajectories were obtained [3].

Parameters in PRIMO were optimised to match energy profiles and conformational sampling from atomistic force fields (CHARMM22/CMAP [86] and CHARMM-36 [87]), using di-peptides, alanine polypeptides (AAXAA, where X corresponds to one of the 20 natural occurring amino-acids), proteins and protein complexes [3]. For the bonded terms, angle, bonds and torsion PMFs from atomistic simulations were used with an inverse-Boltzmann procedure. Non-bonded parameters were tuned from a starting guess in order to reproduce conformational energies from CHARMM [88].

This force field has been used in studies of peptide and small protein structure prediction [3], it has been extended to membrane environments [84] and has been recently used in hybrid atomistic/coarse-grained simulations of proteins [85].

2.5. MARTINI Force Field

This is one of the main force field in the area of coarse-grained simulations, mainly because of its ease of use, support on well-known MD engines (GROMACS, DESMOND, NAMD and GROMOS) and diversity in relation to the type of available macromolecules to model and the different applications, such as the support for lipids [89], water [90], proteins [4], DNA [91, 92], RNA [93], small molecules [94], nano-tubes [95], among many others [96].

The mapping in MARTINI (Fig. 2.4) is based on a four-to-one approximation, where four heavy atoms are mapped as one CG bead. Aromatic amino-acids and molecules are mapped with a higher resolution (up to two-to-one). Four main interaction sites are defined: polar (P), non-polar (N), apolar (C) and charged(Q). For each of these types, subtypes are defined according to its hydrogen bond capabilities (donors, acceptors, both or none) or by a number indicating the degree of polarity (from 1 to 5).

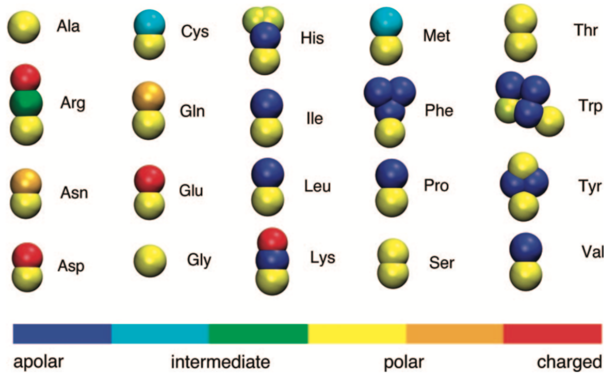


Figure 2.4

Coarse-grained mapping scheme for the MARTINI 1.0 model. Each colour represents a different particle type according to the bottom colour bar. Image taken from [4]**.**

The MARTINI model is parametrised by matching the thermodynamic partitioning free energy of amino-acid side-chains between polar and hydrophobic phases, similarly to how the 53A5 and 53A6 versions of the GROMOS force field were developed [97]. Bonded interactions are modelled by the following potentials:

$$U_b = \frac{1}{2}K_b(d_{ij} - d_b)^2 \tag{2.12}$$

$$U_a = \frac{1}{2}K_a[\cos(\phi_{ijk}) - \cos(\phi_a)]^2 \tag{2.13}$$

$$U_d = K_d[1 + \cos(n\psi_{ijkl} - \psi_d)] \tag{2.14}$$

$$U_{id} = K_{id}(\psi_{ijkl} - \psi_{id})^2 \tag{2.15}$$

where U_b, U_a, U_d and U_{id} correspond to bonds (d_{ij} distance between atoms i and j), angles (ϕ_{ijk} angle between atoms i, j and k), proper and improper dihedral potentials (ψ_{ijkl} torsion between atoms i, j, k and l), respectively (see equation 1.7 for similar term definitions). An elastic network model (based on the Cα positions) is used to improve structural stability and secondary structure [4, 98]. With this, one of the main limitations in MARTINI is the lack of dynamic secondary structure conformations, and hence its inability to be used in protein folding studies [4, 98].

Non-bonded interactions are modelled using a Lennard-Jones 12-6 potential (see eq. 1.8). Charged beads interact via Coulombic interaction using a single dielectric constant of 15 (see eq. 1.9). Parameters were optimised based on experimental hydration free energies and vaporisation free energies [4].

Different optimisation have been performed to the MARTINI protein force field [99, 100]. In the case of the MARTINI 2 version [99], reparameterisation of phenylalanine, tryptophan and proline side-chain parameters were performed, since they were too hydrophobic, as well as the reparameterisation of charged and polar interactions in a low dielectric medium, where was founds that were too weak compared to the atomistic models. The parameterisation involved the generation of new topologies for the mentioned side-chains to improve partitioning free energies, use of an off-centre charge model for the description of contact pairs of oppositely charged residues, and optimisation of polar side-chains to improve dimerisation in apolar environments. The latest version of the MARTINI force field (version 3) [100], have introduced bigger changes, especially in the fundamentals of the building blocks that characterise the force field. Briefly, changes involved the introduction of small (S) and tiny (T) beads, to fully work with the normal type beads (N). With this, aromatic rings are optimise in the way that stacking are more realistically represented (as well as better densities and partitioning free energies). More beads with hydrogen-bond capabilities are introduced, and Q-beads to model divalent ions (named Q2). The water model is improved to avoid freezing problems and to reproduce miscibility. Interactions are divided in organic, water and ions, where each block has an independent parameterisation approach to improve partitioning free energies, densities and trends in solvation free energies. With this, different interaction levels are introduced. Finally, an most importantly, backbone beads do not depend of the secondary structure, where the use of small and tiny beads in side-chains guarantees better packing and modelling of cavities [100].

This model was originally developed for lipids [89, 101] and subsequently extended to proteins [4, 99]. Some of the studies have been related to membrane self-assembly [102], slowly occurring processes like cholesterol flip-flop [103] and lipid desorption [104], bending and deformation of asymmetric bilayers [105], fusion of lipid membranes [106], organization of proteins and peptides into lipid bilayers [107, 108], protein-lipid interactions [109], protein oligomerization [110], protein self-assembly [111] and conformational changes of tertiary protein structure [9], and many more.

2.6. SIRAH Force Field

SIRAH is a top-down generic force field derived to fit structural properties. The force field contains parameters for explicit water, electrolytes, DNA and proteins. Molecular interactions are evaluated using a classical Hamiltonian.

The SIRAH force field, developed by Pantano *et al.*, is a coarse-grained force field built based on a generic Hamiltonian, similar to the most common atomistic models (equation 2.14). SIRAH is based on a top-down approach where most of its characteristics are derived to fit structural properties. At the moment, the SIRAH force field contains parameters for DNA, water, proteins and DMPC lipid.

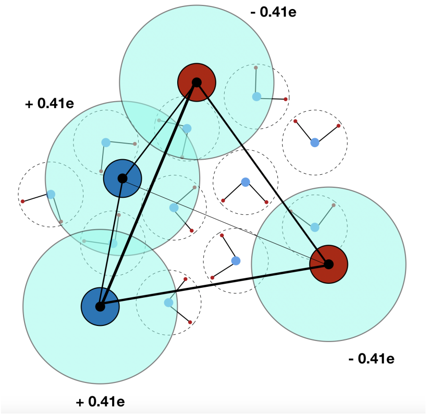
The force field has the following form,

$$
\begin{aligned}
U = \sum_{bonds} k_b(r_{ij} - r_{eq})^2 + \sum_{angles} k_\theta(\theta_{ijk} - \theta_{eq})^2 + \sum_{dihedrals} \frac{U_k}{2}[1 + \cos(n_k\Phi - \gamma_k^{eq})] \\
+ \sum_{i}^{N} \sum_{i>m}^{N} \varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] + \frac{q_l q_m}{4\pi\varepsilon_0\varepsilon_r r_{ij}}
\end{aligned}
\tag{2.16}
$$

where k_b is the force constant for bonds, r_{ij} is the bond length for atoms i and j, r_{eq} is the reference bond length, k_θ is the angle force constant, θ_{ijk} is the i-j-k angle, θ_{eq} is the reference angle, U_k is the force constant for dihedral angles, Φ is the dihedral angle, n_k is the order constant, and γ_k^{eq} is the reference dihedral angle. Terms for Lennard-Jones and electrostatic interactions have been previously defined (see eq. 1.8 and 1.9).

Water model: The water model from the SIRAH force field (known as WAT-FOUR, or WT4) is based on the transient tetrahedral clusters that are formed in pure water (Fig. 2.5). These clusters are formed of two positive and two negative beads, covalently bound (with the use of harmonic restraints) to represent 11 water molecules [7] (each bead has a total mass of 50 au. Four beads sum a total of 200 au, that if it is divided by 18 au, which is the mass of one water molecule, gives a total of 11 water molecules).

Each bead has a mass of 50 au, with a charge of +0.41e for positive beads and -0.41e for negative beads (charges applied based on the SPC water model). The CG conformation representing the tetrahedral clusters, contains an implicit water molecule in the centre. The distance between the oxygen of this central water molecule with respect to any other oxygen atom of the cluster is 0.28 nm. Based on this assumption, a distance of 0.45 nm between each bead of the CG representation is established in order to get a perfect tetrahedron. A value of 5 kcal/mol·Å^2 is used for the harmonic constant in the bonds to obtain better agreement for some water properties, such as density and diffusion coefficient. This also adds plasticity to the clusters, giving the option to adapt to temperature changes and to changes in the environment. Van der Waals and electrostatic interactions are based on the classical functional form for generic force fields



The WT4 water model. Schematic representation of the SIRAH water model WT4, composed of 4 beads arranged in a tetrahedral shape. Two beads bear a negative charge of -0.41e (red circles), and two beads with a positive charge of +0.41 (blue circles). Atomistic water molecules are represented in the background**.**

Figure 2.5

(see equations 1.8 and 1.9).

Table 2.1 summarises a comparison of bulk water properties for the WT4 and experimental data. As can be seen, most of the water properties do not agreed with the experimental data, except for the diffusion coefficient. Important properties such as the isothermal compressibility and surface tension are poorly described [7].

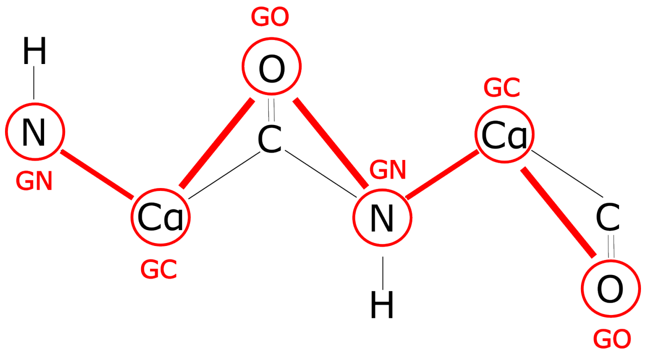
	Dielectric	Diffusion	Expansion	Mass	Isothermal
	constant	coefficient	coefficient	density	compressibility
	(ϵ)	(10^{-5} cm^{-2} s^{-1})	(10^{-4} K^{-1})	(g mL^{-1})	(GPa^{-1})
WT4	110	2.23	11.6	1.01	2.43
Exp.	78.4	2.27	2.53	0.99	0.46

Bulk water properties for the WT4 model and experimental data at 300 K. Values obtained from reference [7]**.**

Table 2.1

Advantages of the WT4 model are the existence of explicit short-range and long-range electrostatic interactions, and the possibility to perform CG/MM hybrid simulations, where regions of interest can be described in full atomistic resolution and coarser resolution to less relevant regions [112, 113]. Moreover, the authors have developed CG models for simple electrolytes such as Na^+, Cl^- and K^+. They are represented by a single CG bead, including its first solvation shell [5, 7] and they bear a single partial charge of +1e or -1e. Comparison of the hydration structure of the WT4 model around the ions, using radial distribution function plots (RDF), and profiles for their electrostatic potential, with respect to atomistic simulations were done to validate the models. Darre et al., 2010 [7] managed to reproduce the position of the second solvation shell using the WT4 mode, but not the position for the third solvation shell. Lack of first solvation shell was seen in RDF plots, hence its introduction in the ion models [5, 7].

Protein model: For the SIRAH protein model [5, 11], the backbone is represented with low granularity (i.e. higher resolution), with beads corresponding to the amino (GN), alpha carbon (GC), and carbonyl oxygen (GO) atoms (Fig 2.6). This higher resolution (similar to the PRIMO force field, see section 2.4) is discussed as an advantage over other CG force fields with lower resolution on the backbone, such as the MARTINI force field for proteins, where the backbone is represented as one CG bead for the Cα carbon [4]. This level of approximation usually needs extra constraints or restraints to maintain the secondary structure [4]. SIRAH is able to maintain the secondary structure of proteins without any extra bias. Partial charges are explicitly defined on each CG backbone bead that accounts for the formation of hydrogen-bond-like interactions. These were optimised to reproduce the electrostatic potential generated by an atomistic poly-glycine in an α-helical conformation using the GROMOS96/45a3, and AMBER99 force fields. The bond and angle values for the bonds between CG beads in the backbone where initially taken from their DNA model (with values of 41840 kJ/mol and 627.6 kJ/mol rad^2, for bond and angle constants, respectively), and consequently optimised based on atomistic simulations of glycine tri-peptides using the AMBER99 force field. The SIRAH 1.0 force field was tested for single proteins in water, protein complexes and protein aggregation, and estimations of structural stability were done based on root mean square deviations, root mean square fluctuations and secondary structure time-series [5].



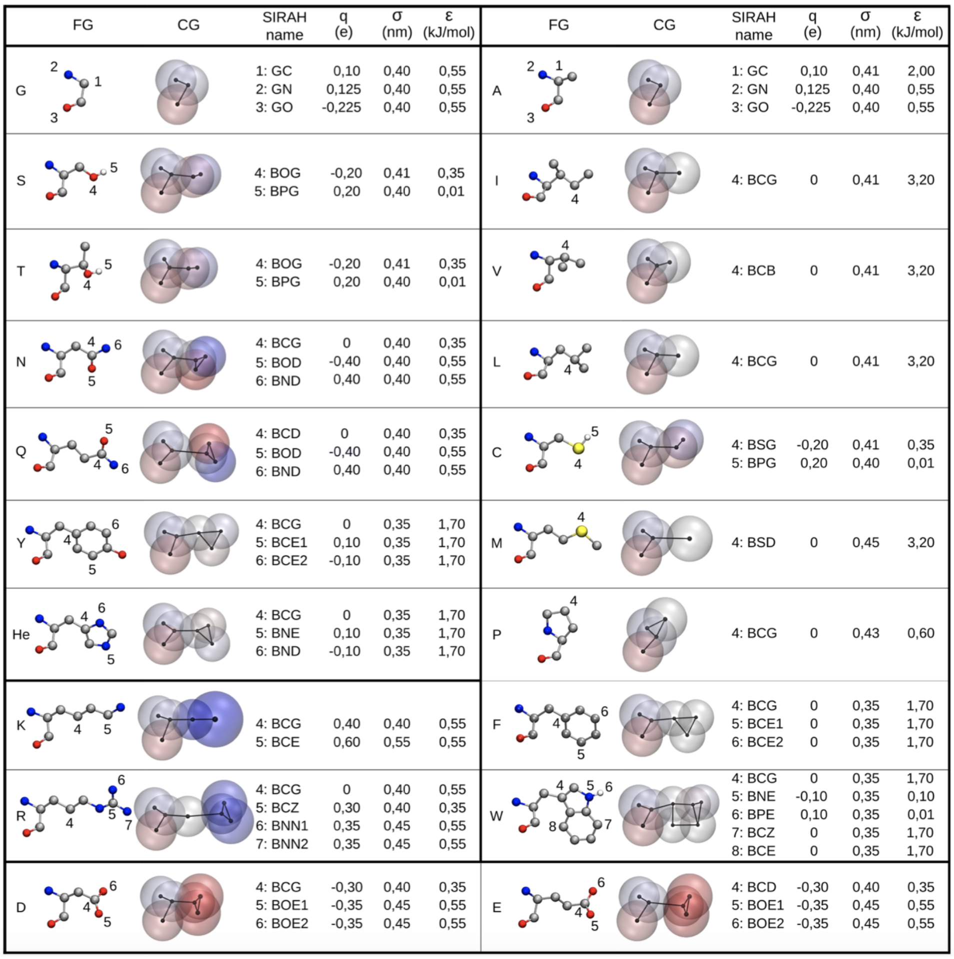
Coarse-grained mapping scheme for the SIRAH protein backbone. Red circles indicate the atoms used to defined the CG beads of the backbone (GN, GC and GO)**.**

Figure 2.6

Dihedral angles define the secondary structure for the system. Three dihedral angles exist: Ψ for the GN-GC-GO-GN plane, Φ for the GO-GN-GC-GO plane, and Ω for the GC-GO-GN-GC plane. The secondary structure is defined as helix (H) when $-180° \leq \Phi \leq 10°$ and $-120° \leq \Psi \leq 45°$, β-sheets (E) when $-180° \leq \Phi \leq 0°$ or $\Phi > 135°$ and $-180° \leq \Psi \leq 120°$ or $45° \leq \Psi \leq 180°$. Only peptides in the trans configuration are accepted for the actual version of the model, and that is why Ω is defined as a single cosine function with only one minimum at that conformation. Fourier expansion are used for dihedrals, forcing the existence of a minimum for the most stable conformations (α-helices and β-sheets).

For the side-chains, the mapping was followed based on their physicochemical characteristics (Fig. 2.7). Hydrophobic residues are represented with one bead; aromatic residues are mapped using three of five beads in a plane; and for polar and charged residues, CG beads are placed on the charged groups (Fig. 2.7). The procedure for bonds and angles was similar to the backbone, where the values were taken from the AMBER99 force field and forces were taken from the CG DNA model of SIRAH as an initial estimation.

Charges are placed in an *ad-hoc* manner, according to the amount of hydrogen bond acceptors/donors. Hydrophobic residues carry a zero charge, polar and aromatic residues



FG	CG	SIRAH name	q (e)	σ (nm)	ε (kJ/mol)	FG	CG	SIRAH name	q (e)	σ (nm)	ε (kJ/mol)
G		1: GC 2: GN 3: GO	0,10 0,125 -0,225	0,40 0,40 0,40	0,55 0,55 0,55	A		1: GC 2: GN 3: GO	0,10 0,125 -0,225	0,41 0,40 0,40	2,00 0,55 0,55
S		4: BOG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01	I		4: BCG	0	0,41	3,20
T		4: BOG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01	V		4: BCB	0	0,41	3,20
N		4: BCG 5: BOD 6: BND	0 -0,40 0,40	0,40 0,40 0,40	0,35 0,55 0,55	L		4: BCG	0	0,41	3,20
Q		4: BCD 5: BOD 6: BND	0 -0,40 0,40	0,40 0,40 0,40	0,35 0,55 0,55	C		4: BSG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01
Y		4: BCG 5: BCE1 6: BCE2	0 0,10 -0,10	0,35 0,35 0,35	1,70 1,70 1,70	M		4: BSD	0	0,45	3,20
He		4: BCG 5: BNE 6: BND	0 0,10 -0,10	0,35 0,35 0,35	1,70 1,70 1,70	P		4: BCG	0	0,43	0,60
K		4: BCG 5: BCE	0,40 0,60	0,40 0,55	0,55 0,55	F		4: BCG 5: BCE1 6: BCE2	0 0 0	0,35 0,35 0,35	1,70 1,70 1,70
R		4: BCG 5: BCZ 6: BNN1 7: BNN2	0 0,30 0,35 0,35	0,40 0,40 0,45 0,45	0,55 0,35 0,55 0,55	W		4: BCG 5: BNE 6: BPE 7: BCZ 8: BCE	0 -0,10 0,10 0 0	0,35 0,35 0,35 0,35 0,35	1,70 0,10 0,01 1,70 1,70
D		4: BCG 5: BOE1 6: BOE2	-0,30 -0,35 -0,35	0,40 0,45 0,45	0,35 0,55 0,55	E		4: BCD 5: BOE1 6: BOE2	-0,30 -0,35 -0,35	0,40 0,45 0,45	0,35 0,55 0,55

Figure 2.7

Coarse-grained mapping scheme for the SIRAH protein side-chains. Mapping of all the CG side-chain with respect to atomistic systems (FG, for fine-grained) is shown. Bead's names are presented, as well as van der Waals and charge parameters for the SIRAH 1.0 force field. Image taken from reference [5].

carry a ±0.1e, hydroxyl groups a ±0.2e charge, amide groups a ±0.4e charge and charged groups carry a unitary charge spread through the whole side-chain. In the case of histidine, epsilon and delta protonated forms are present, with similar parameters. Van der Waals parameters for the backbone are taken from the AMBER99 force field. For the side-chains, van der Waals interactions are calculated using Lorentz-Berthelot rules, except for some pairs of interaction where specific corrections to avoid over-stabilization of hydrogen bond and salt-bridge interactions are introduced *ad-hoc*. AMBER scaling factors SCNB (1–4 van der Waals interaction scaling factor) and SCEE (1–4 electrostatic interaction scaling factor) are used for 1-4 non-bonded interactions.

A recent re-parameterisation of the SIRAH protein model has been done (SIRAH 2.0) in an *ad-hoc* fashion, where changes were accepted based on the improvement of structural descriptors such as root mean square deviations (RMSD), solvent accessible surface (SAS), radius of gyration (RGYR) and native contacts [11]. Improved backbones were obtained in relation to the number of terms used in the Fourier expansions, that went from three and four terms to describe U_{Ψ}^{CG} and U_{Φ}^{CG} torsion potentials (SIRAH 1.0) [5], respectively, to two and nine terms in SIRAH 2.0. The complexity of U_{Φ}^{CG} was increased to allow parallel and anti-parallel β-sheets conformations, while the energy landscape for U_{Ψ}^{CG} remained unaffected [5, 11]. Increased side-chain flexibility is also discussed, where changes to angular force constants were made. Non-bonded interactions were updated to improve the hydrophilicity/hydrophobicity for side-chains and backbone [11].

The SIRAH force field has been used in the study of protein-DNA interactions [113], multiscale simulations and lipid simulations of DMPC membranes [114]. Machado et al. [11] have also recently presented a new supramolecular coarse-grained model for water. The WLS (WatElse) model [115] corresponds to a larger version of the actual WT4 model, where 4 beads represent 5 WT4 molecules, or 55 fine-grained water molecules. They have managed to use this in multi-scale simulations, mixing their supramolecular and coarse-grained model with finite-water molecules to study virus-like particles [115]. A speed-up of 2 orders of magnitude has been stated, based on the comparison with atomistic simulations with equivalent numbers of particles [5].

2.7. Summary

Here we have summarised a group of intermediate and higher resolution coarse-grained force fields. Advantages can be found for higher resolution models (such as PRIMO and SIRAH), where there is no need for external biases for the stabilisation of secondary structure and its flexibility (which is the case of MARTINI) (see table 2.2). Their functional forms have been shown, as well as their parameterisation and optimisation procedures. These go from an *ad-hoc* manner (SIRAH), to get stable trajectories, up to the use of functional forms similar to atomistic force field, with the purpose of reproducing specific properties, such as side-chain PMFs and partition free energies (e.g. PRIMO and MARTINI), among others. Some of the mentioned force field were parameterised for specific applications, such as folding of proteins (CABS and UNRES), or to make them more transferable in order to include a wide range of applications (MARTINI). The main limitations in the development of coarse-grained force fields are related to their representability and transferability, which correspond to their ability to represent and reproduce physical properties at the state point they were parameterised and to state point outside the parameterisation protocol. Parameterisation procedures, such as *ad-hoc* methodologies, are limited by time, and tend to be tedious and prone to errors. The use of automated procedures can be an advantage and will show itself as an important alternative to overcome these difficulties, which will be discussed in the following chapters.

Force Field	Advantage	Disadvantage
CABS	Ab initio simulations of protein folding, and considered a leading approach in CASP competitions	Limited applicabilities to protein folding studies
UNRES	Parameterised based on atomistic PMFs of protein side-chains. Diverse applications in protein folding and protein-protein interactions	Too complex functional form, with many parameters to overcome possible limitations
MARTINI	Well known force field, with vast documentations and community support. Diverse support for biomolecular models, such as lipids, DNA, proteins, sugars, etc. Classic functional form, similar as the one used in atomistic force fields	Does not support the study of systems that compromise secondary structure re-arrangements
PRIMO	Higher resolution backbone, with hydrogen bond-like interactions. Does not need external bias to maintain secondary structure	Too complex to use, lack of community support, as well as lack of documentation. Limited to peptide and small protein structure prediction
SIRAH	Supported in well-known MD engines, such as GROMACS and AMBER. Support for different types of biomolecules, such as DNA, protein and lipids. Higher resolution backbone, with hydrogen bond-like interactions. Does not need external bias to maintain secondary structure. Classic functional form, similar as the one used in atomistic force fields	Lack of studies that show applications outside plain MD simulations. Ad-hoc optimisation, and un-extensive parameter validation

Table 2.2 Protein coarse-grained force fields. Advantages and disadvantages of the force fields summarised in this chapter**.**

Chapter

3

Testing of the SIRAH Force Field

3.1. Introduction

Chapter 2 summarised intermediate and high resolution coarse-grained protein force fields. Force fields such as UNRES and PRIMO show themselves as carefully made force fields, but at expenses of complex functional forms and a large number of force field parameters. Given the advantage of near-atomistic representations of the backbone, the use of a common functional form compatible with well-known MD engines (such as GROMACS and AMBER), the availability of tools to ease its use (i.e. mapping, back-mapping and visualization) [116], the SIRAH force field looks like a promising alternative to conventional force field. Simulations using this force field have shown stable protein systems in water, either for proteins alone or in complex with other proteins or peptides [5]. Compared to MARTINI, the SIRAH force field does not need the use of elastic networks to overcome the problem of secondary structure stability. The use of higher resolution backbone and partial charges on each bead serves for the production of hydrogen bond-like interaction, which fully work to stabilise protein systems. Moreover, explicit short and long-range electrostatic interactions are modelled with a dielectric constant of 1, compared with the MARTINI model where a dielectric constant of 15 is used [4]. The SIRAH force field has not been tested for conformational changes in proteins, which is the main interest of our investigation. We believe this is a chance to evaluate the capabilities of SIRAH outside the area for what it has been optimised, and to test how the parametrisation process has been elaborated and if there are any areas of possible improvement.

In this chapter, we have tested the SIRAH 1.0 force field [5]. We have extended the simulation times from the original publication for the protein model, evaluating the overall stability of the systems. We have tested the capability of SIRAH to reproduce protein conformational changes, specifically for the S1S2 glutamate receptor binding domain and the Abl kinase.

Hydration free energies (HFEs) are an important property for aqueous systems such as proteins. They help us to understand biological processes such as ligand recognition, protein-protein interactions, folding and conformational changes. Moreover, hydration free energies have been used for the validation of molecular force fields, and they are an integral part of the calculation and estimation of solubilities, partition coefficients and solute-solvent interactions [12, 117, 118]. For these reasons, use of hydration free energies as a parameterisation target for coarse-grained models may improve their performance. Here, we have tested the prediction of hydration free energies by the SIRAH 1.0 force field, for both the side-chains and the backbone. We believe this can lead us and give us insight of possible flaws in the parameterisation process and protein dynamics using this model.

3.2. Methods

The potential energy of the system is linearly related to coupling parameter α (eq. 1.24 and 1.37). With this, the solvation free energies for the side-chain analogues, for the atomistic and coarse-grained systems, were calculated based on a decoupling approximation, where the non-bonded interactions of the solute with the solvent are gradually turned off. With this, our reference state will be our system in solution, and the final state will be the solute in vacuum. The OPLS-AA force field [119] was used for the atomistic calculations, while the SIRAH 1.0 force field [5] was used for the coarse-grained calculations. Simulations with the SIRAH 2.0 force field are later tested in chapters 4, 5 and 6.

3.2.1. Topology generation

The side-chain analogue topologies where build with the Small molecule Topology GEnerator (STaGE) [120]. We use SMILES strings as input. Topologies were generated using ACPYPE [121] and ANTECHAMBER. Assignment of the atom types was based on the OPLS-AA types. Default charges of the OPLS-AA force field were used. Figure 3.1 shows a work-flow of the topology construction. The reason for the use of the OPLS-AA force field is based on the evaluation of solvation free energies for side-chain analogues, which have shown similar results compared to experiments [8].

3.2.2. Solvation free energies

The term solvation involves a series of steps, with different consequences in terms of changes in energy. As examples, the first step involves the creation of a cavity (that would be occupied by the solute), and is enthalpically and entropically unfavourable, since the solvent-solvent interactions are lost, and the solvent ordering increases. When the solute is introduce, the interactions with the solvent increase, resulting in a enthalpically favourable process. Finally, the solute mixes in the solvent and entropy is gained. Overall, the sum of all these changed in energy, in terms of enthalpy and entropy, is called solvation energy, and solvation free energy involves the free energy change associated to transfer a molecule from the gas to the solvent phase.

The following protocol was used for both atomistic and coarse-grained simulations. Decoupling hydration free energies were calculated; that is, we started from a system with the solute-solvent interactions fully on, and we scaled them to zero. Electrostatic interactions were first turned off. If we first turn off van de Waals interactions, atoms of opposite charge will experience strong attraction, causing overlaps and consequently spurious effects. After decoupling electrostatic interactions, van de Waals interactions

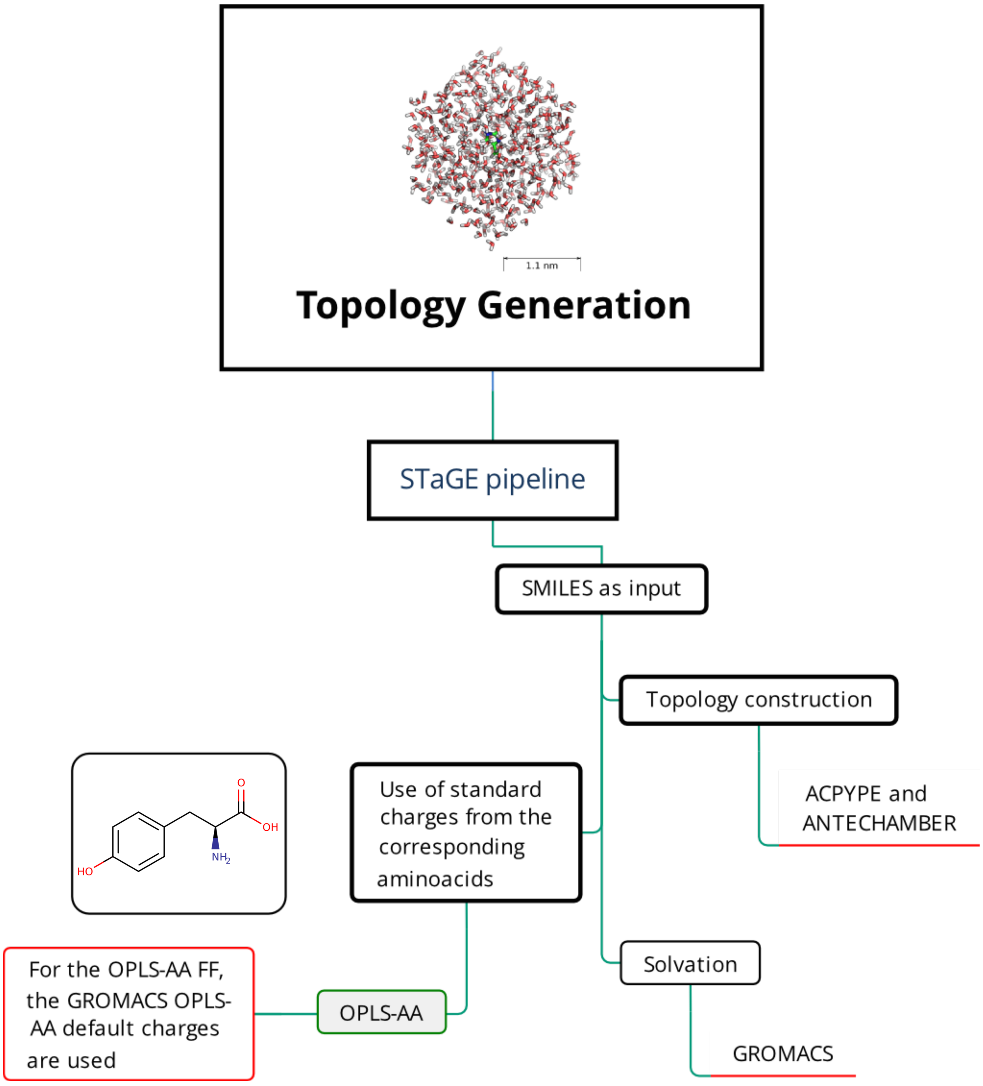


Figure 3.1

Topology generation. Workflow used for the topology construction of the atomistic side-chain analogues**.**

were turned off. Six discrete values of the coupling parameter α were used for the scaling step of both potentials: 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0 (with 0.0 being fully on and 1.0 fully off). The soft-core scaling method was used to smoothly vary the potential when atoms are nearly disappeared. A shifted version of the regular soft-core potential is implemented in GROMACS, with the form,

$$
\begin{aligned}
U_{sc}(r) &= (1-\alpha)U_0\left(r_0\right) + \alpha U_1\left(r_1\right) \\
r_0 &= \left(\alpha^{par}\sigma_0^6 P + r^6\right)^{\frac{1}{6}} \\
r_1 &= \left(\alpha^{par}\sigma_1^6(1-P+r^6\right)^{\frac{1}{6}}
\end{aligned}
\tag{3.1}
$$

where U_0 and U_1 are the normal van der Waals or electrostatic potentials, at α values of 0 and 1, respectively. α^{par} corresponds to the soft-core parameter (set as sc_{alpha} in the GROMACS input file), P is the soft-core power (set as sc_{power} in the GROMACS input file) and σ is the radius of the interaction (set as sc_{sigma} in the GROMACS input file). Parameters for these were set as α^{par}=0.5, P=1.0, and σ=0.3. Simulation were run for 1 ns, with a previous equilibration of 1 ns and the corresponding system minimisation.

The Multistate Bennett Acceptance Ratio (MBAR) [41] was used to compute the free energy difference. This method is an extension of the well- known Bennett Acceptance Ratio (BAR) [40], which needs information from two end states (in contrast to FEP or TI, which need information from only one state) to compute the free energy difference. MBAR combines data from multiple states to obtain the statistically optimum result.

3.2.3. Coarse-grained system setup and molecular dynamics

Coarse-grained (CG) models of three protein systems were constructed based on the crystal structure deposited in the Research Collaboratory for Structural Bioinformatics Data Bank (PDB). The systems with PDB code 1QYO (GFP protein, 238 amino-acids, X-ray resolution of 1.80 Å), 1RA4 (L7Ae protein, 120 amino-acids, X-ray resolution of 1.86 Å) and 1R69 (N-terminal domain of phage 434 repressor, 60 amino-acids, X-ray resolution of 2.0 Å) were selected form the original SIRAH 1.0 publication [5], in order to choose systems of different sizes and that represent different types of secondary structure such as α-helices and β-sheets.

Briefly, each structure was protonated at pH 7.0 using PDB2PQR [122]. Residue based CG representations were built using the *cgtools* code mapping procedure based on the SIRAH topology [5, 116]. Each of the three CG systems were solvated in an octahedral water box using the WT4 water model of SIRAH, where each molecule broadly represents 11 atomistic water molecules, with a distance between the protein system and the edge of the box of 1.5 nm in order to satisfy the minimum image convention. The periodic box sizes were 90x90x90 Å, 70x70x70 Åand 60x60x60 Å, for the big, medium and small system, respectively. All systems were neutralised by addition of sodium and chloride ions to a total concentration of 150 mM.

Coarse-grained molecular dynamics (CGMD) simulations were performed using GROMACS v5.0.4 [123] with the SIRAH 1.0 force field. An energy minimisation was carried out using 10000 iterations of the steepest descent algorithm. This was followed by an NPT equilibration dynamics procedure of 5 ns with positional restraints of 1000 kJ/mol nm^2 applied to all the protein beads. Production runs were performed for 3 μs for each system with an integration time-step of 20 fs. Electrostatic interactions were calculated using the Particle Mesh Ewald (PME) procedure with a direct cut-off of 1.2 nm and a grid spacing of 0.2 nm. Non-bonded interactions were modelled using the Lennard-Jones potential with a cut-off of 1.2 nm. All simulations were run at 1 bar with the Parrinello-Rahman barostat [124] and at 300 K with the v-rescale thermostat [125]. The simulation conditions were consistent with the original SIRAH publication [5].

3.2.4. Atomistic system setup and molecular dynamics

The same three systems mentioned in the CG system setup were used for simulations using an atomistic resolution. Simulations were carried out using GROMACS v.5.0.4 with the OPLS-AA force field [119]. Each system was solvated using the TIP3P water model [88] in an octahedron box with a solute-box distance of 1.5 nm in order to satisfy the minimum image convention. The periodic box sizes were 85x85x85 Å, 70x70x70 Åand 55x55x55 Å, for the systems with PDB code 1QYO (GFP protein), 1RA4 (L7Ae Archeal ribosomal protein) and 1R69 (N-terminal domain of phage 434 repressor), respectively. All systems were neutralised by addition of sodium and chloride ions to a total concentration of 150 mM.

Similar to the CG MD simulations, each system was minimised using the steepest descent algorithm with 5000 iterations. This was followed with an NPT equilibration of 10 ns with positional restraints on protein heavy atoms of 1000 kJ/mol nm^2. Production runs were performed for 200 ns for each system. Electrostatics interactions were calculated using the PME procedure with a grid spacing of 0.16 nm. Non-bonded interactions were modelled using the classical Lennard-Jones potential and a Coulombic energy function, with a cut-off of 1.2 nm each. The LINCS algorithm was applied to constrain all h-bond lengths. All simulations were run at 1 bar with the Parrinello-Rahman barostat [124] and at 300 K with the v-rescale thermostat [125].

3.2.5. Umbrella sampling simulations

As a conformational change test, the opening/closing process of the glutamate receptor ligand binding domain was used [126, 127]. A 1D reaction coordinate (ξ) was chosen as the distance between the C-α carbons of residues G451 and S652 of the CG Glutamate receptor ligand binding domain (PDB: 1FTJ, corresponding to the ligand-bound structure in its closed conformation, X-ray resolution of 1.90 Å) (Fig. 3.5). This was done for both the free-ligand (apo, by manually removing the ligand), and the glutamate-bound states (holo). The simulation were run starting from the closed state. The pull code in GROMACS was used to generate snapshots for the umbrella sampling (US) simulations from a single pulling trajectory, generating 33 umbrella windows that spanned a distance range between 0.45 to 2.0 nm, with a distance between each window of 0.05 nm. A value for the biasing harmonic potential of 2500 kJ/mol nm^{-2} was used. Each window was simulated for 300 ns. The Weighted Histogram Analysis Method (WHAM) implemented in GROMACS (g_wham) [123] was used to remove the biasing potential and get the unbiased probabilities, to finally compute the potential of mean force (PMF). The rest of the simulation settings were set exactly as in the previous CGMD section. Convergence of the PMF and overlap between the umbrella windows are shown in figures A.5 and A.6, for the apo and holo simulations, respectively.

3.2.6. Metadynamics simulations

A metadynamics simulation of the CG Abl kinase (PDB: 2G1T, X-ray resolution of 1.80 Å) was performed using the Plumed package [128] (v. 2.4.3) patched into GROMACS [123]. Simulations were run at a temperature of 310 K and for 1 μs, using the well-tempered metadynamics algorithm [50]. Two CVs were used, that compromises two dihedral angles: CV1: GN(A380)-GC(A380)-GC(D381)-BCG(D381) and CV2: GN(A380)-GC(A380)-GC(F382)-BCG(F382). Gaussians were deposited every 4 ps with a height of 2.0 kJ/mol. A bias factor of 5 was used. The width of the Gaussians was set to 0.1 rad for both dihedral CVs. Parameters for well-tempered metadynamics simulations were based on previous studies of the Abl kinase using Plumed [129], and the selection of CVs was based on a study from Roux et al., 2015 [130].

3.3. Results and discussion

3.3.1. Protein system stability and secondary structure

We have tested the SIRAH 1.0 force field for proteins in solution, with different system sizes, in order to evaluate their stability. We chose three protein systems from the original publication of SIRAH [5] to further extend the simulations, up to 3 μs (compared to 1 μs from the original publication): PDB code 1QYO (GFP protein, 238 amino-acids), 1RA4 (L7Ae protein, 120 amino-acids) and 1R69 (N-terminal domain of phage 434 repressor, 60 amino-acids).

The root mean square deviations (RMSD) histograms and time series for the three systems are shown in Figure 3.2, calculated against the crystal structure. Overall, for all the systems the average values are consistent with those on the original SIRAH publication: 0.60 nm for the big system 1QYO (0.50 nm reported in the original SIRAH publication), 0.40 nm for the medium system 1RA4 (0.36 nm reported in the original SIRAH publication) and 0.92 nm for the small system 1R69 (0.87 nm reported in the original SIRAH publication). An increase in RMSD values at around 1 μs is experienced by the small system 1R69 (Fig. 3.2 I), which can also be observed as two peaks on the histogram for this system (Fig. 3.2 G). A possible explanation for this can be observed in the time series of the secondary structure for the coarse-grained systems. It seems that the small system experiences a rupture of an α-helix to a coil structure, and the formation of an alpha helix from a beta sheet in residues 35 to 40 (Fig. 3.3, right panel). For the other two systems, no important changes in secondary structure exist (Fig. 3.3, left and middle panel). Root mean square fluctuations (RMSF) were also calculated against the crystal structure, and they are summarised in figure 3.4. In the case of 1QYO RMSF plots, similar peaks are observed compared to the atomistic RMSFs, but with a higher base line. Moreover, the peak observed for the atomistic 1QYO system at around residue 160, is not observed for the CG system. The smallest simulated systems 1RA4 and 1R69 are less stable than the 1QYO system, given the observed residue fluctuations as can be seen by the RMSF plots in figure 3.4B and 3.4C (left panel). In these cases, RMSF values go over an order of magnitude higher compared to atomistic RMSFs, with values of 1 nm around residue 50 in the case of 1R69 system. It has to be mentioned that the loss of degrees of freedom in CG models will account for fewer interactions, and can be (but not necessarily) a cause of more flexible systems.

The RMSD histograms show that the overall RMSD values maintain a stable behaviour; the RMSDs are around two times higher for 1QYO and 1RA4 systems, and one magnitude higher for the 1R69, compared to the atomistic RMSD values calculated in this work (Fig. 3.2). Usually higher RMSDs are expected for coarse-grained systems due to longer simulation times, fewer degrees of freedom, and potential energy surface smoothness, characteristics that are representative of a CG force field [9]. Calculations of RMSD values against the last frame of the trajectories can give us information if the observed RMSD values come from the intrinsic system flexibility of given some conformational changes towards other stable structures. Figure A.1 shows the RMSD time-series against the last frame of the trajectories. In the case of the 1QYO system, no sudden changes are observed (Fig. A.1 A). An important change is observed in the 1R69 system, with a possible conformational transition around 1 μs. Finally, the 1RA4 system presents a mixture with possible small conformational transitions after the first

1.5 μs.

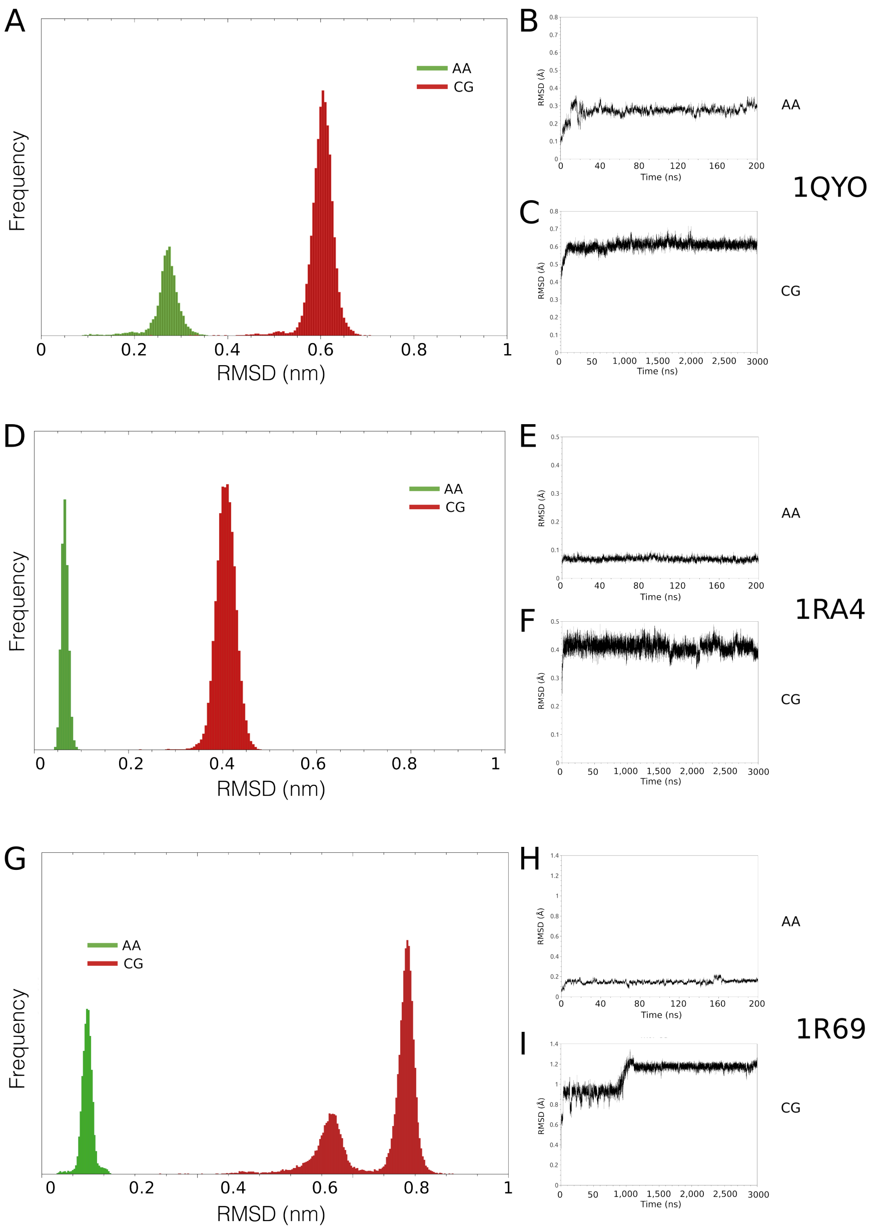


Figure 3.2

RMSD histograms and times series. Root mean square deviation frequencies (A, D and G) and time-series (B, C, E, F, H, I) at both levels of resolution, coarse-grained (red) and all-atom (green) against the crystal structures are shown, for the big system (A,B,C) (1QYO, GFP-protein, 236 residues), the medium system (D,E,F) (1RA4, L7Ae Archeal ribosomal protein, 117 residues) and the small system (G,H,I) (1R69, N-terminal domain of phage 434 repressor, 63 residues)**.**

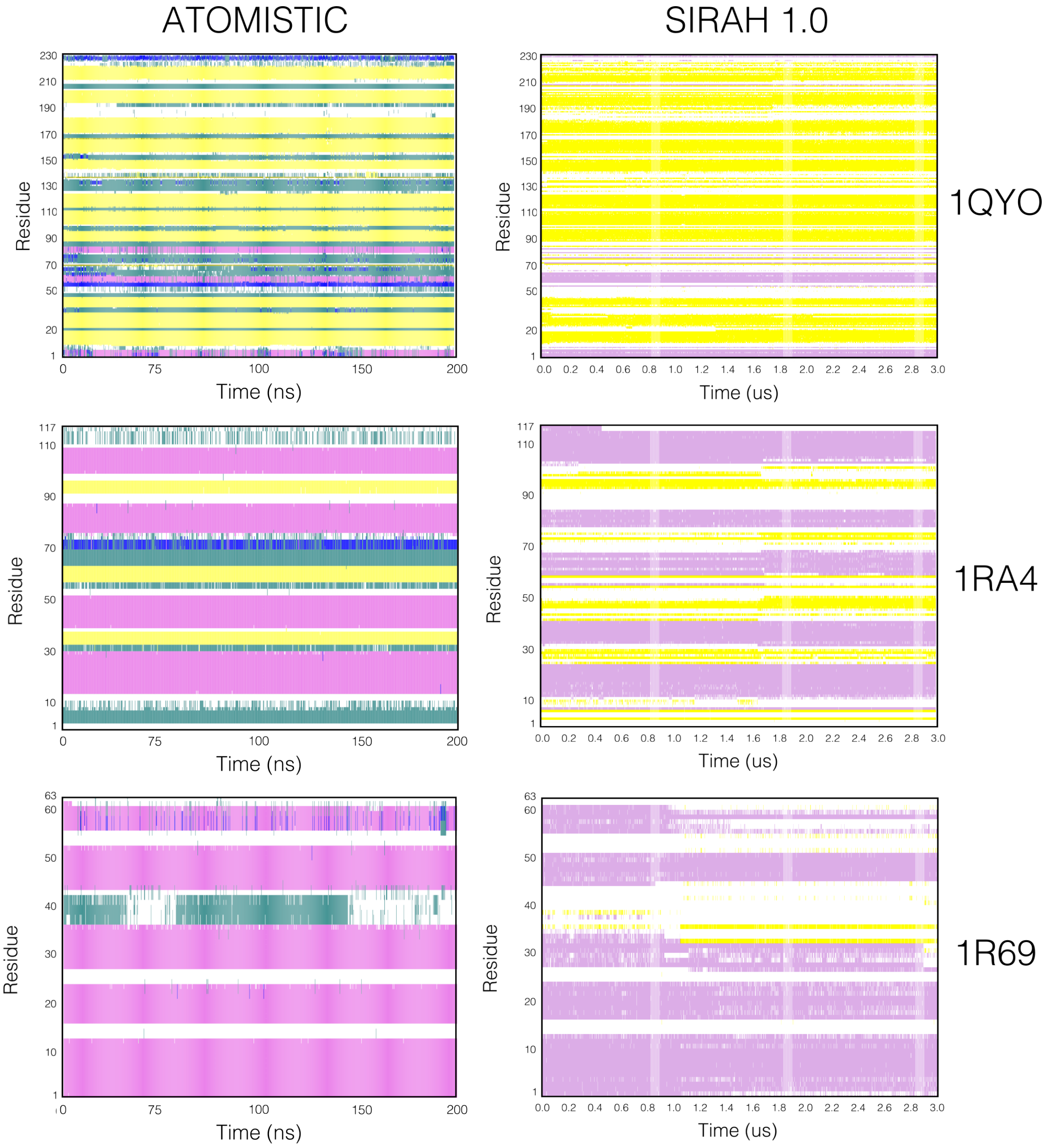


Figure 3.3

Time evolution of the secondary structure. Evolution of the secondary structure by residue for the coarse-grained models was calculated. PDB codes of the systems analysed (right side of the figure) correspond to: the big system (1QYO, GFP-protein, 236 residues), the medium system (1RA4, L7Ae Archeal ribosomal protein, 117 residues) and the small system (1R69, N-terminal domain of phage 434 repressor, 63 residues). Panels in the left hand side correspond to atomistic simulations, while panels on the right hand side correspond to coarse-grained simulations. Colours represent the evolution of α-helices (pink), β-sheets (yellow) and coil (white) secondary structures. Other colours observed in the atomistic plots represent secondary structures such as 3_{10}-helix and isolated bridge, but can not be represented by the SIRAH**.**

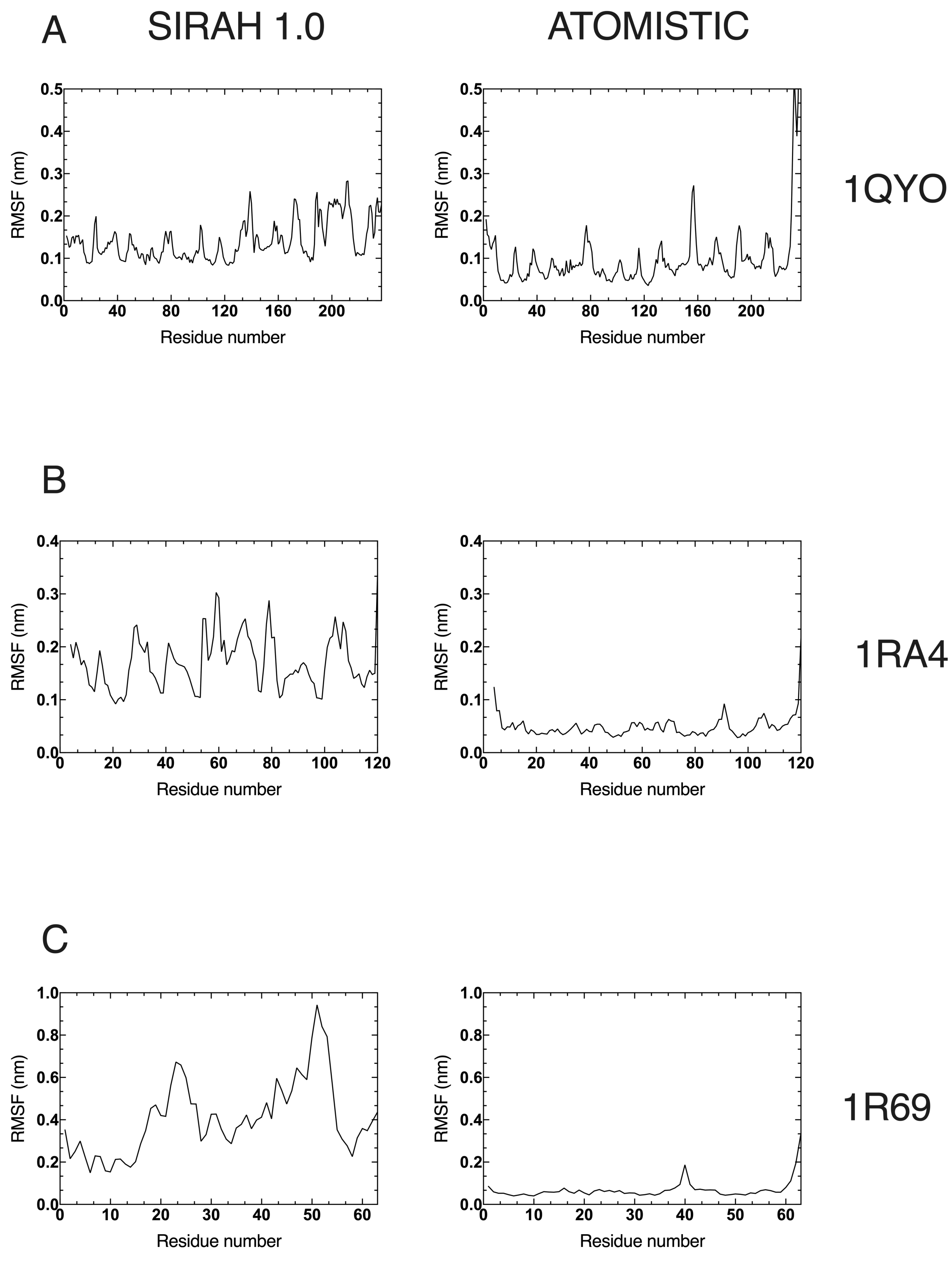


Figure 3.4

Protein RMSF. Root mean square fluctuations for each system at both levels of resolution against the crystal structure. The RMSF were calculated over the C-α carbons for the coarse-grained model, as well as in the all-atom models for (A) the big system (1QYO, 236 residues), (B) the medium system (1RA4, 117 residues) and (C) the small system (1R69, 63 residues).

3.3.2. Protein conformational change

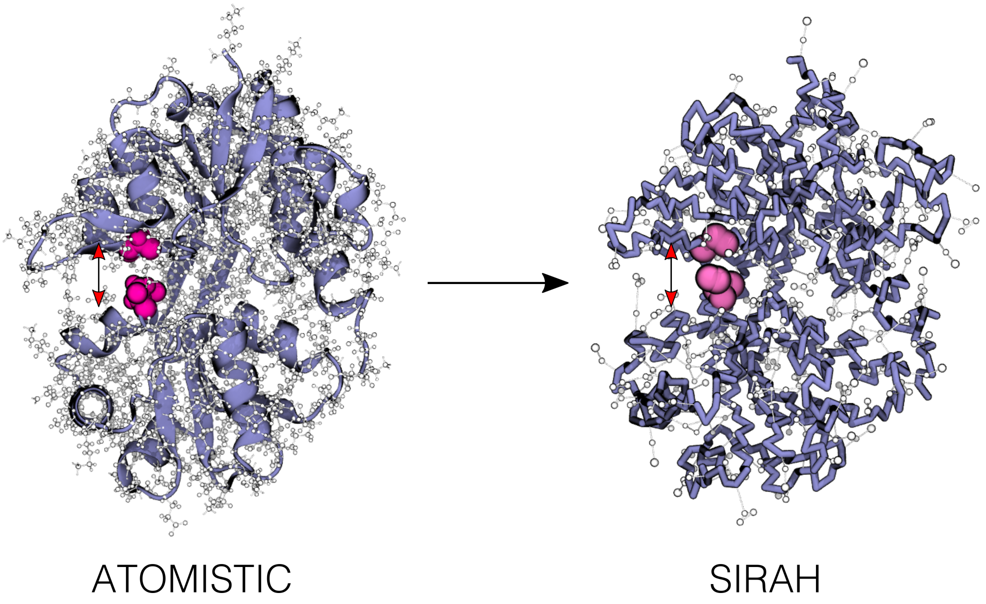
We have attempted to reproduce the free energy landscape of a protein conformational change using the original SIRAH 1.0 force field. We have studied two different processes: the opening/closing event of a Glutamate Receptor Domain, and the DFG-flip conformational transition in the Abl kinase.

3.3.2.1. Opening of the binding domain of the S1S2 glutamate receptor

We started with the opening/closing event of the ligand binding domain of the S1S2 glutamate receptor (Fig. 3.5), processes that have been extensively studied in different scenarios, such as in its unbound and bound forms with different types of ligands (e.g. glutamate, AMPA and a photo-switchable ligand) [126, 127, 131, 132]. Briefly, glutamate receptors regulate synaptic transmissions in the central nervous system. They are activated upon ligand binding, which translates to a transmembrane channel opening. The ligand binding domain (LBD) has a clamshell-like structure, that experiences a conformational change (open to closed) when an agonist binds. Even though it has been suggested that a pure structural explanation based on this cleft movement is not enough to describe the receptor mechanism [126], it has been shown that, for the case of the S1S2 glutamate receptor, a ligand-free receptor (apo) prefers more open conformations, while a glutamate-bound structure (holo) prefers a closed one [126, 127].

In order to describe the cleft movement, a one-dimensional order parameter was used, described by the distance between the centre of mass of the C-α beads of residues G451 and S652 of the ligand binding domain of the S1S2 glutamate receptor (Fig. 3.5), which can be considered as equivalent to the 1D projection of a two-dimensional coordinate used in similar studies [126, 127], that is calculated as $CV_{12} = (CV_1 + CV_2)/2$ with CV$_1$: distance between the centre-of-mass (COM) of residues 479–481 and residues 654–655, and CV$_2$: distance between the COM of residues 401–403 and residues 686–687.

As previously mentioned, it has been reported that the apo structure of the receptor



Coarse-grained mapping of glutamate receptor ligand binding domain. 3D representation of the S1S2 glutamate receptor ligand binding domain. An atomistic representation is shown, as well as the SIRAH mapping of the same structure. The residues used to represent the collective variable are shown in magenta, and the direction of the pulling is shown as red arrows**.**

Figure 3.5

prefers more open conformations. For the crystal structure, the average distance between these residues is around 1.18 nm, and the global minimum of the free energy profiles, of the apo state, from atomistic simulation lies at around the same value [126]. The original SIRAH 1.0 force field shows two shallow minima at around 1.15 nm and 1.35 nm, with a barrier between them no greater than 0.5 kcal/mol (Fig. 3.6).

Lau Y. et al., 2007 [126], showed that, for the apo structure, energies of around 4.0 kcal/mol are needed in order to achieve conformations close to the glutamate-bound state (i.e. averaged values of a distance around 0.89 nm). Energies of 1.0 kcal/mol are seen for this closed state in the case of the SIRAH 1.0 force field (Fig. 3.6). It is known that CG force fields tend to show smoother energy profiles given the fewer degrees of freedom, compared to atomistic systems [126].

In the case of the free energy landscapes for the glutamate-bound state, or holo state, a preference for a closed conformation is seen. The crystal structure for this state shows

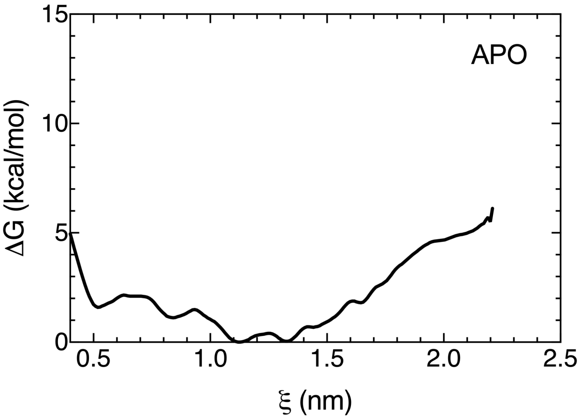


Figure 3.6 **Free energy profile of the S1S2 glutamate receptor.** PMF plot of the cleft opening/closing process of the S1S2 glutamate receptor ligand binding domain in its apo state**.**

an average distance between the designated residues of 0.86 nm, while the predicted minimum in atomistic simulations lies at around 0.89 nm [126]. In the case of the SIRAH 1.0, no preference for a closed state is seen, even showing a greater preference for a more open state compared to the apo structure (Fig. 3.7).

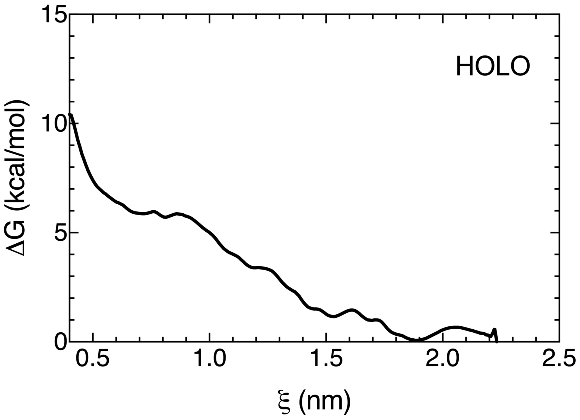


Figure 3.7 **Free energy profile of the S1S2 glutamate receptor.** PMF plot of the cleft opening/closing process of the S1S2 glutamate receptor ligand binding domain in its glutamate-bound state (holo) using the SIRAH 1.0 force field**.**

3.3.2.2. DFG-flip of the Abl kinase

Protein tyrosine kinases regulate many cellular signalling processes such as cell-growth, proliferation, metabolism, cellular differentiation and migration, hence their wide interest as targets in diseases such as cancer. It is for this reason, that the regulation of kinases has to be tightly controlled [130, 133]. Near the terminus of the "activation loop", a motif known as DFG is located, which is composed by Asp381-Phe382-Gly383 (Abl numbering, Fig. 3.8) [130]. A conformational transition in this motif exists, that can switch the enzyme from an active (DFG-in) to an inactive (DFG-out) state [130]. Phosphorylation of the activation loop is also a common mechanism for kinase activation. In the active conformation, the aspartate of the DFG motif points into an ATP-binding site coordinating the binding to Mg^{2+} ions, and the αC-helix is rotated inward towards the active site [130, 133]. Opposite to this, experimental results have shown that the inactive conformation is characterised by rotation of the aspartate to outside the binding site, loosing coordination with the Mg^{2+} ions. Morover, in the inactive form, the phenylalanine from the DFG motif occupies the binding site, disrupting ATP binding [130, 133].

Two dihedral angles have been used by Meng et al., 2015 [130] to describe the transition between the DFG-in and DFG-out conformations (in terms of the flip of aspartate and phenylalanine residues): Cβ(Ala380)-Cα(Ala380)-Cα(Asp381)-Cγ(Asp381) and Cβ(Ala380)-Cα(Ala380)-Cα(Phe382)-Cγ(Phe382). A free energy equilibrium was showed, in the case of the Abl kinase, with two clear minima located at around (60°, -100°) and (-100°, 10°), corresponding to the DFG-in and DFG-out states, respectively.

The authors used US simulation to estimate the relative free energies of the two conformational states, with a total of 5184 windows that spanned the entire range (-180°, 180°) for both dihedral angles [130]. As a straightforward and simple approach, and in order to avoid 5000 simulations, we decided to use well-tempered metadynamics. We used the same two collective variables previously mentioned, except for the case of Cβ(Ala380). Given the lack of a Cβ atom for alanine in SIRAH, we used atom GN(A380)

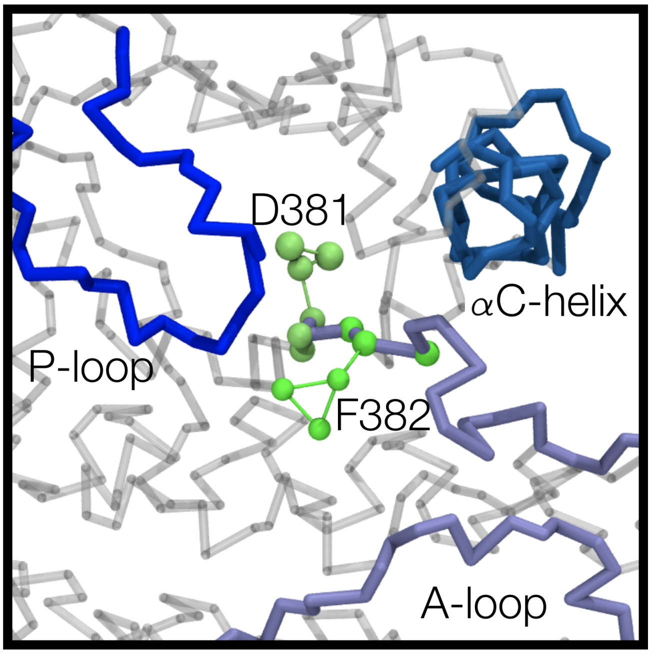


Figure 3.8

CG representation of the DFG motif. The DFG motif is shown in green, surrounded by the phosphate-binding (P-loop, light blue), the αC-helix (dark blue) and the activation loop (A-loop, purple)**.**

(which is mapped on the N atom for the peptide bond between residues 379 and 380) (Fig. 3.9). With this, the two dihedral angles were defined by the following atoms: CV1: GN(A380)-GC(A380)-GC(D381)-BCG(D381) and CV2: GN(A380)-GC(A380)-GC(F382)-BCG(F382) (see Fig. 2.6 for the side-chain mapping in SIRAH). Differences will be expected between the choices of the CVs used in atomistic and coarse-grained simulation, but we believe that the choice will still captured the rotation of aspartate and phenylalanine residues, and to be able to fully predict two conformational states (see chapter 6 for more results using this coarse-grained CV).

The 2D-PMF of the DFG transition, described using the SIRAH 1.0 force field, is shown in figure 3.10. A clear minimum exists at around (60°, 30°). Another higher basin is seen at (−150°, 60°), with relative energy values of 9-12 kcal/mol. Comparing our results with atomistic simulations [130], no equilibrium is shown for the two structural states, with an energy barrier of around 18 kcal/mol between both basins, and with a preference of just one state that resembles a DFG-in conformation, based on the dihe-

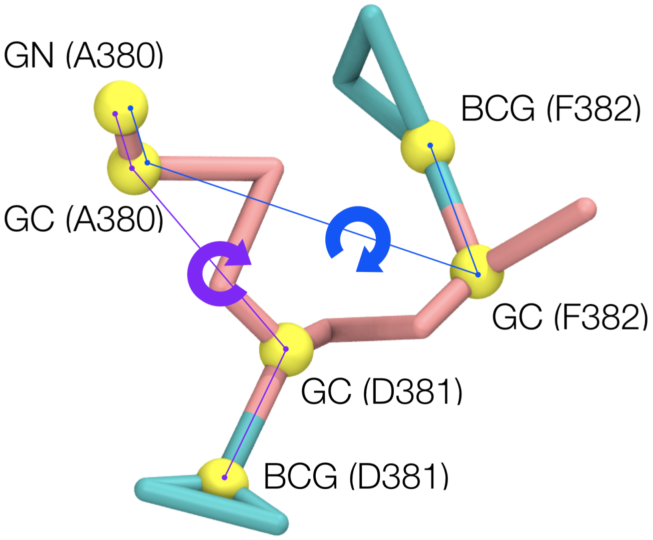


Figure 3.9

CG representation of the collective variables. Graphical representation of the two dihedral angles that were used to described the DFG transition, using the SIRAH 1.0 force field: CV1: GN(A380)-GC(A380)-GC(D381)-BCG(D381) and CV2: GN(A380)-GC(A380)-GC(F382)-BCG(F382). (see Fig. 2.6 for the side-chain mapping in SIRAH)**.**

dral angle values.

3.3.3. Hydration free energies for the SIRAH 1.0 force field

One of the main approaches used to test the quality of different force field is the evaluation of solvation free energies, for example in the case of the polarisable water model in MARTINI [134]. With this in mind, and with the intention to find potential causes for the inability of SIRAH to fully reproduce conformational changes, we decided to perform hydration free energies of protein side-chains, and the backbone beads.

As can be seen in figure 3.11, a critical discrepancy exists between the calculated values and the experimental data for the hydration free energies using the SIRAH 1.0 force field (R^2 of 0.104). OPLS-AA hydration free energies were also calculated and included for comparisson. It is clear that something important is missing in the parameterisation process of the SIRAH 1.0 force field, which can play an important role on the repre-

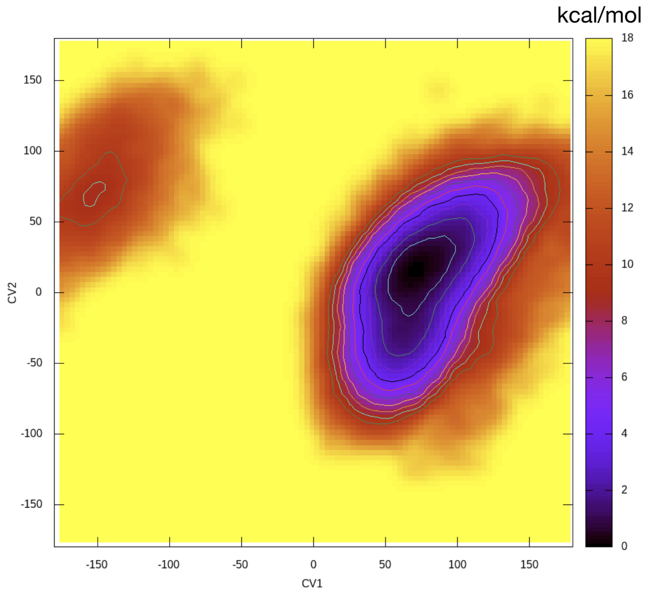


Figure 3.10

2D PMF of the DFG transition. All units are shown in kcal/mol. CV1 and CV2 correspond to two dihedral angles given by: CV1: GN(A380)-GC(A380)-GC(D381)-BCG(D381), and CV2: GN(A380)-GC(A380)-GC(F382)-BCG(F382)**.**

sentation of side-chain interactions, and a source of error on the interaction with the solvent model.

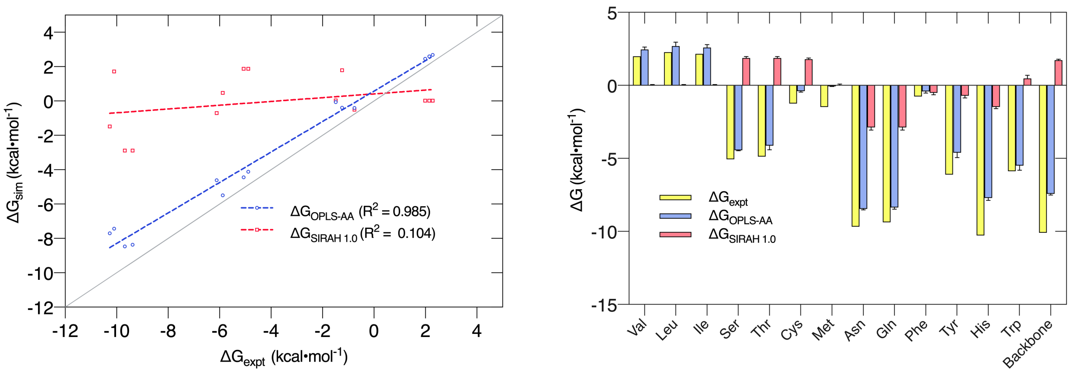


Figure 3.11

Hydration free energies for the SIRAH 1.0 side-chain and backbone beads. Comparison of the calculated hydration free energies for the OPLS-AA force field (blue) and the SIRAH 1.0 force field (red) (ΔG_{sim}) vs. experimental values (ΔG_{expt}). Error bars reported as standard errors calculated based on 3 replicas**.**

To further observe the discrepancies in the SIRAH force field, comparisons of the electrostatic component (Fig. 3.12, right panel) and the van der Waals component (Fig. 3.12, left panel) were made. As can be seen, none of the side-chains for SIRAH 1.0 has similar behaviour to that of the atomistic simulations. Some of them even gave values of different sign in the van der Waals component (Phe, Tyr, His), while others presented similarities for the same component (Ser, Thr, Asn and Gln). No similarities exist for the electrostatic component. In terms of total free energies (Fig. 3.11), Ser, Thr, Cys, Trp and backbone show energy values of different sign compared to atomistic and experimental values With this, it seems the main problem for the discrepancy between atomistic and coarse-grained solvation free energies comes from the electrostatic component. Alanine is the only amino-acid which behaves in a similar way compared to atomistic simulations, with a hydration free energy of -1.47 kcal/mol, compared to -2.13 kcal/mol for the atomistic system. Even so, its electrostatic component is one order of magnitude higher (0.81 kcal/mol vs. 0.08 kcal/mol). All the values for the solvation free energies and for each of its components are summarised in table 3.1.

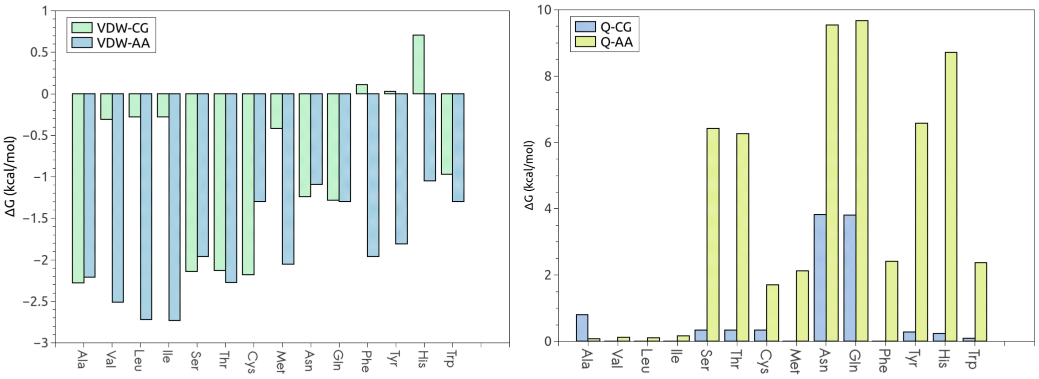


Figure 3.12

Comparison on the components of the hydration free energies. The van der Waals (left) and the electrostatic (right) contribution to the hydration free energy for the atomistic and coarse-grained systems are shown**.**

3.4. Summary

The SIRAH 1.0 force field [5] has been tested. Protein simulations have been extended to 3 μs, compared to the 1 μs time shown in the original publication of the SIRAH 1.0

	Expt.	**OPLS-AA**	**SIRAH 1.0**
Backbone (NMA)	-10.1	-7.40 ± 0.04 (AMBER-14SB)	1.73 ± 0.07
Val (propane)	1.99	2.45 ± 0.06	0.02 ± 0.01
Leu (isobutane)	2.28	2.69 ± 0.10	0.02 ± 0.01
Ile (butane)	2.15	2.59 ± 0.08	0.02 ± 0.01
Ser (methanol)	-5.06	-4.44 ± 0.01	1.87 ± 0.04
Thr (ethanol)	-4.88	-4.12 ± 0.11	1.87 ± 0.04
Cys (methanethiol)	-1.24	-0.39 ± 0.02	1.78 ± 0.03
Met (methyl-ethylsulfide)	-1.48	-0.06 ± 0.01	0.03 ± 0.02
Asn (acetamide)	-9.68	-8.46 ± 0.02	-2.87 ± 0.07
Gln (propionamide)	-9.38	-8.36 ± 0.04	-2.87 ± 0.07
Phe (toluene)	-0.76	-0.40 ± 0.04	-0.50 ± 0.05
Tyr (p-cresol)	-6.11	-4.61 ± 0.13	-0.70 ± 0.06
His (methyimidazole)	-10.27	-7.70 ± 0.06	-1.47 ± 0.04
Trp (methylindole)	-5.88	-5.55 ± 0.22	0.47 ± 0.09
MUE		1.04	5.03
MSE		-1.04	-4.13
R^2		0.98	0.10

Table 3.1 **Comparison of hydration free energies.** Hydration free energies of neutral side-chains and backbone using the OPLS-AA, AMBER-14SB and the SIRAH 1.0 force field. Mean unsigned error (MUE), mean signed error (MSE) and R^2 are shown, compared to experimental values (Expt.). All values are in units of kcal/mol. Errors reported as standard errors calculated from 3 replicas**.**

force field [5]. We have tested the capability of reproducing protein conformational changes and their corresponding free energy landscapes on two protein systems: the S1S2 glutamate receptor binding domain and the Abl kinase. We further extend our test to the estimation of hydration free energies using SIRAH 1.0, and compare them to atomistic and experimental data.

Simulation of the SIRAH 1.0 force field have shown that:

- The SIRAH 1.0 behaves well for large proteins but it shows clear instabilities for smaller protein systems.

- The capability of SIRAH to reproduce free energy profiles of conformational changes is not good, showing a lack of consistency with respect to atomistic simulations.
- Hydration free energies of the SIRAH 1.0 side-chain and backbone beads showed a clear discrepancy with atomistic and experimental values

With this, there are clear inadequacies in the parameterisation of the SIRAH 1.0 force field. Based on our results, most of the differences came from the electrostatic component of the calculated hydration free energies. Most of the parameters in the SIRAH 1.0 force field were derived with an *ad-hoc* approach. We believe that the incorporation of a better set of parameters is key to yield better agreement with experiments, and thus a better representation of the protein and solvent interactions. In the following chapters we propose the reparameterisation of the SIRAH 1.0 force field. We developed an optimisation work-flow to overcome the differences observed in terms of hydration free energies. This method has been applied for the optimisation of uncharged side-chains and the protein backbone. Important discussions are shown for the reparameterisation of charged side-chains.

Chapter

4

Optimisation of the SIRAH Force Field: uncharged side-chains and backbone

4.1. Introduction

The quality of a force field will depend on the reliability of the chosen approach to estimate the parameters that are going to be used to calculate the energies and forces of molecules and systems. Usually, the selection of the target data (i.e. the information that is used to fit our force field, which can be energies, structural information of proteins, among others) comes from supermolecule calculations such as QM calculations of representative systems of interest (e.g. simulations of an alanine di-peptide for parameterisation of a protein backbone torsion parameters). Parameterisation using experimental data can also be done, but usually such data is not available for the system of interest. Alternatives to this can be the use of implicit solvation methods, where the solvent is represented as a continuous medium [135], offering significant computational savings. Different models have been designed, such as solvent-accessible surface area models, Poisson-Boltzmann models and Generalised Born models [135]. Successful applications of such continuum models have been shown in areas of protein modelling design [136], and protein dynamics [137, 138]. Limitations have been shown for these types of models, and compared to explicit solvent models, in areas of binding free-energies [139] and solvation free energies [140].

It is clear that force field optimisation is not a trivial procedure, given that the search of paramaters is carried out based on chemical intuition to guide manual testing of parameters (*ad-hoc*), where some have classified this process as an art [141]. Other frameworks and approximations to optimise parameters have been proposed, such as methods based on genetic algorithms, hill-climbing routines and force-energy matching [142]. A basic protocol to develop a new force field includes the following four steps [17]:

1) Selection of the functional form. In order to model the system of interest, this is the first crucial step. Selecting how are we going to describe the interaction between particles is needed, where the form of the force field strongly influences this. Simple force fields (or also known as class 1) usually represent bond and angle interactions by harmonic terms, and usually Lennard-Jones and Coulomb potentials for van der Waals and electrostatic interactions. In this class, some "traditional" force field can be mentioned such as AMBER [143], CHARMM [87], OPLS-AA [119] and GROMOS [97]. Other more complex force fields (also known as class 2 and 3) use more complex functional forms that are built based on the type of applications. An example is the AMOEBA force field [144], where angles are decomposed, based on a Wilson-Decius-Cross decomposition, into in-plane and out-of-plane components, a softer 14-7 van der Waals form is used, and partial charges are replaced with polarisable atomic multipoles [144].

2) Selection of a particular data set. Choosing an interesting data set to derive the force field parameters is needed for the chosen functional form. In this context, an ideal target would be experimental data such as bond distances, angles, x-ray data, densities. In most of the cases, this data is not available for the systems that we are interested to optimise, where theoretical data such as QM calculations of energies and forces play an important role [17, 32, 145]. Other types of data include the use of osmotic coefficients for the re-parameterisation of protein force fields [146], as well as the use of

NMR data [147].

3) Parameter optimisation. Most of the parameters are mutually dependent, so changes in one can substantially affect the other [17, 145] (i.e. the parameters are coupled). Parameterisation will need an iterative process until you get stable and well behaved simulations. For example, partial charges and LJ are highly correlated: a set of van der Waals parameters determined for a specific set of charges will not work for another set of charges obtained with a different method. Moreover, bonded parameters will be correlated to non-bonded parameters, where changes in the latter will affect the behaviour of the former, for example, in the energy surface of a dihedral angle [17, 32, 145]. Historically, bond and angle parameters have been taken from small-molecule crystallography and vibrational spectroscopy [17, 32, 145]. Now, these force field parameters are usually derived from QM calculations, using small molecule vibrational frequencies [145]. In the case of dihedral terms, as experimental information on torsional barriers is often absent, quantum mechanical calculations again play an important role to determine these torsion potentials [32, 145]. The process of torsion parameterisation is as follows: a representative molecular fragment of the torsion to be modelled is chosen, where a series of configurations are generated by rotating the necessary bonds, and the energy is calculated using a QM approach. Then, the torsional potential is fitted to reproduce the energy curve [32, 145]. As mentioned before, complications in this area are the coupling that exists between torsion and 1-4 interactions, where non-bonded parameters are involved. A series of iterations have to be performed in order to reproduce the energy surface [32, 145].

Optimisation of partial charges can be done through different methods. The supramolecular approach is related to the adjustment of charges to reproduce QM interaction energies and geometries of the model compound with respect to individual water molecules, which are placed at different locations and orientations around the molecule [148]. Atomic partial charges can also be parameterised based on QM Electrostatic Potentials (ESP), which has the intention of reproducing the QM ESP mapped onto a grid surrounding the model compound [17, 141, 148]. These two methods have the advantage of being fairly quick in developing optimised parameters, but only when QM ESP are available or can be determined. Some improvements have been made for the ESP method, which now includes restraints during the fitting process (RESP) [17, 141]. This is useful for the determination of charges on buried atoms. Both of these methods are dependent on the conformation used in the optimisation process. To alleviate this, multiple conformations can be used.

Once partial charges are assigned, LJ parameters need to be optimised. Early studies of Jorgensen showed the use of condensed-phase properties (using pure liquid systems) as the target for the optimisation [149]. These parameters were adjusted to reproduce experimental heat of vaporisation, densities, heat capacities, heats of sublimation and solvation free energies [148]. Moreover, LJ parameters are optimised for atoms in a series of classes of functional groups, and once this is done, these parameter can be transferred, without any other optimisation, to other molecules that use the same class of functional groups [32, 145, 148].

4) Parameter validation. Once the parameters have been determined, a proper validation needs to be considered, which relies on comparison of simulations with experimental data. This will need to include properties for which the force field was parameterised. For example, if a force field has been parameterised based on condensed-phase properties, testing the performance and the capability to reproduce density and enthalpy of vaporisation would be a good choice. Moreover, the use of certain properties that were not included in the parameterisation process is essential. The use of hydration free energies has also been used for force field validation [12, 150], as well as partition free energies between water and different organic phases [4] and the stability of proteins and peptides [5, 11].

4.2. Force field parameter optimisation

In this section, a brief summary is provided for some actual methods and approaches related to the optimisation process. An explanation for force-matching, a well-known parameter optimisation method, is given, as well as some automated pipelines that use it. An approach related to the use of NMR data as target is also explained, mainly because it differs from previous target data used in parameter derivation (see above). At the end, the ForceBalance method is explained, which is a new method that follows a hybrid approach where simulation and experimental data are used to derive new force field parameters; this method is directly related to the development of this thesis.

4.2.1. Energy and force matching

Energy and force matching [151], is one of the most widely used methods for the optimisation of force field parameters, and it has been extensively integrated in some automated procedures [6, 152, 153] (see below). The aim of energy and force matching is to minimise the difference between a reference system potential (U_{ref}) and a simpler potential (U_{fit}) (or forces), and to get the best possible representation of the energy surface [151,154]. The fitting process is based on the minimisation of an objective function χ^2, that includes the square difference for a given property A, in relation to the calculated ($A^{LL}_{i,j}$) and reference ($A^{HL}_{i,j}$) values,

$$\chi^2 = \sum_{i=1}^{N_{\text{conf}}} w^i_{\text{conf}} \sum_{A} \sum_{j=1}^{N} w^A_{i,j} \left(A^{LL}_{i,j} - A^{HL}_{i,j} \right)^2 \tag{4.1}$$

where N_{conf} is the number of configurations, w^i_{conf} and $w^A_{i,j}$ are the weight of the configuration and that of the fitted property for the particular configuration, respectively. LL and HL stand for low-level and high-level, which represent the type of models used in the calculation of the data. For example, a low-level description model can represent an atomistic system, while a high-level description can be given by QM calculations. In this context, the property A will take the form of energies and/or forces. The use of different configurations and physical situations such as clusters, surfaces, bulk and liquid, is encouraged with the idea of achieving a potential transferability [151]. In order to get more data point contributions to the fitting of the energy surface, it has been suggested, in conjunction with forces (which include 3N data points), the use of molecular torques, dipoles and stress tensor [154].

4.2.2. ForceFit

The ForceFit package has been developed by the group of Clark [152] in order to optimise classical force fields based on the force matching approach. Here, forces from different QM configurations are used as targets in the optimisation procedure, where the difference between these forces and those calculated by classical MD are minimised.

The ForceFit workflow is as follows: a series of physically representative QM configurations need to be collected, where standard electronic structure codes can be used (Gaussian [155] and NWChem [156]). Next, the selection of the parameters to be optimised has to be performed. ForceFit support the optimisation of bonded parameters only, including pairwise, bonding, angle, dihedral and improper potentials. In order to obtain the classical force to be improved, ForceFit will run a single MD step, using DL-POLY [157], AMBER [143] or LAMMPS [158]. The only implemented method in ForceFit for the parameter optimisation is the Powell algorithm [159], which corresponds to a least square based method. The squared difference between the classical and QM forces is given as:

$$U = \sum_i [f_i(\lambda) - F_i]^2 w_i \tag{4.2}$$

where the i subscript will run over all the QM geometries in the so-called database, which is just the collection of all given available geometries and forces. f_i and F_i stands for the classical and QM forces, respectively. λ corresponds to the parameters to be optimised in the classical force field. The weight w_i determines where one particular geometry is more physically represented in the fitting process.

4.2.3. ParamFit

ParamFit corresponds to another automated optimisation protocol for classical force fields [153]. It is based on the use of forces and energies from *ab initio* simulations or experimental data as target for the optimisation of only bonded parameters in Amber.

In order to fit energies, ParamFit will try to minimise the following expression:

$$f(N, E_{QM}, K) = \sum_{i=1}^{N} [(E_{MM}(i) - E_{QM}(i) + K)^2] \tag{4.3}$$

where N$_{\text{atoms}}$ and N are the number of atoms in the system and the number of configurations, respectively. E_{QM} and E_{MM} are the quantum energy and the calculated energy from classical simulations. K is a constant that accounts for different origins in the quantum and classical energies, and allows minimisation to zero to be conducted. When fitting to forces, a similar approach is followed, given as:

$$f(N, N_{atoms}, F_{QM}) = \sum_{i=1}^{N} \sum_{atom=1}^{N_{atoms}} |\mathbf{F}(i, atom)_{MM} - \mathbf{F}(i, atom)_{QM}|^2 \tag{4.4}$$

where F_{QM} and F_{MM} are the quantum force and the calculated force from classical simulations. All the other terms are the same as previously defined.

In order to diminish the noise, forces errors are summed for all atoms in the molecules or just for the atoms directly involved in the parameters to be optimised. A combination of a hybrid genetic algorithm approach is implemented, with a later refinement through the use of a simplex algorithm, as the authors claim that this increased the convergence speed compared to the use of a genetic algorithm alone. ParamFit has been used in the development of the GAFF14 [160] and Lipid14 [161] force fields in AMBER, optimising dihedral parameters in lipid tails.

4.2.4. Optimisation based on NMR data

Force matching is one of the most common approaches used in the optimisation of classical force fields, but other groups have focused attention on a new approach for the type of target data. Li and Brüschweiler have developed a strategy for the optimisation of protein force fields based on NMR chemical shifts and residual dipolar couplings (RDCs) [147].

The calculations of these properties is performed for each snapshot on the MD trajectory, based on a parent force field defined as U_{old}. Time and ensemble averaged chemical shifts are calculated for each snapshot, using equal weights, $p_{old}(i) = 1/N$. These are compared with the experimental values by means of RMSD. They reweight a parent trajectory performed with the old force field U_{old} for a new test force field defined as U_{new}, using the Boltzmann relationship:

$$p_{new}(i) = p_{old}(i)e^{-U_{new}(i)/k_B T}/e^{-U_{old}(i)/k_B T} \tag{4.5}$$

where U^{old} and U^{new} are the energies for the i^{th} snapshot, for the old and new force field, respectively. p_{old} and p_{new} are the relative weights, k_B is the Boltzmann constant and T is the simulation temperature. The force field is iteratively optimised: for each new trial force field U^{new}, the new weights p_{new} are used to calculate new NMR parameters. The optimised force field is the one that minimise the differences between the calculated and reference properties [147].

This approximation has been used for the optimisation of the ϕ and ψ backbone dihedral angles in the AMBER99 force field, validating it in around 20 proteins using chemical shifts. The newly produced force field was better in the reproduction of RMSD values against crystallographic structures and NMR data, and it was also tested in the study of unfolded peptides and intrinsically disordered proteins using molecular dynamics simulations [147].

4.2.5. Machine learning force fields

In recent years, a data-driven approach has been in constant development, where new force-fields have been optimised based on the use of machine learning (ML) methods. This has lead to the creation of flexible and adaptive force fields, where energies or forces may be learned by induction [162–164]. Different advantages of these ML force fields have been stated, such as their accuracy and versatility compared to conventional inter-atomic potentials, the lesser computational cost and comparable accuracy when compared to QM methods [162]. Deep learning neural networks [165, 166] and non-linear regression processes [167] have been the methods of choice for models describing atomic interactions

A basic work-flow for the development or optimisation of a ML force field will involve: i) a reference data set appropriate for the desired prediction (e.g forces computed using a QM method or atom configurations), ii) using a subset of the reference data set as training set, so the learning process for algorithm of choice can be optimised, and ensuring that this training set is representative of the diversity observed in the reference data, and iii) evaluation of the force field with a test data set, that includes data outside the training set, to see the predictive power and applicability of the new optimised force field.

Recent advances have also shown that forces experienced by a particular atom can be learned and predicted given a set of configurations of atoms. In contrast to the conventional calculation of the potential energy, the use of the force is determined by its immediate environment, and this is the only input that some material simulations need, such as geometry optimisation and MD. AGNI force fields are an example of this type of force field [162]. The authors have used a non-linear kernel ridge regression (KRR) method as the learning algorithm, with a diverse reference data set of equilibrium configurations (and their associated forces) that mimic the environments an atom could exist in, such as defect free bulk, surfaces, point defects, and isolated clusters. The force field has been tested in predicting structural, transport or vibrational properties of materials, as well as more complex phenomena, such as surface melting and stress-strain behaviour.

Another example is the symmetric gradient domain machine learning (sGDML) model [163, 164]. A reference data set of geometries with their corresponding total energy and atomic-forces has been used. It has been stated that the sGDML model can achieve spectroscopic accuracy in the calculation of energies for small molecules like toluene, it can calculate forces and energies 4 to 8 time faster than QM methods, is around 3 orders of magnitude slower compared to conventional atomistic force fields [164], and it has been used to reproduce the potential energy surface and free energy surfaces (based on probability distributions) of small molecules such as aspirin, ethanol and paracetamol [163]. A python package to reconstruct sGDML methods has been recently published [164].

Most recent developments in this area have included the use of hybrid approaches, were ML and classical potentials are mixed to predict energies and forces, and are the first steps for the study of materials science problems such as phase nucleation and forest hardening [168].

4.2.6. ForceBalance

In ForceBalance [6, 169, 170] (Fig. 4.1), accurate force field parameters (λ) are derived based on the use of experimental and/or ab initio calculations (i.e QM data). An objective function, which is just a weighted sum of squared difference of simulated values and the target data, is constructed. Multiple residuals X can be added to one objective function X^2, which will be integrated over the whole configuration space R using MD, for a suitable measure P($\mathbf{r}$; λ) that reflects the thermodynamic ensemble of interest. In the early publications of the ForceBalance method [169], a forcefield optimisation was automated based on QM energies and forces (force and energy matching), with an objective function of the form,

$$X^2 = \int_R P(\mathbf{r};\lambda)|X(\mathbf{r};\lambda)|^2 d\mathbf{r} \tag{4.6}$$

$$|X(r;\lambda)|^2 = w[\frac{(\Delta E(\mathbf{r};\lambda) - \langle \Delta E \rangle)^2}{\langle E_{QM}^2 \rangle - \langle E_{QM} \rangle^2}] + \frac{1-w}{3N_{atoms}}[\Delta F(\mathbf{r};\lambda)^T Cov(\mathbf{F}_{QM})^{-1} \Delta F(\mathbf{r};\lambda)] \tag{4.7}$$

where

$$\Delta E = E_{MM} - E_{QM} \tag{4.8}$$

$$\Delta F = F_{MM} - F_{QM} \tag{4.9}$$

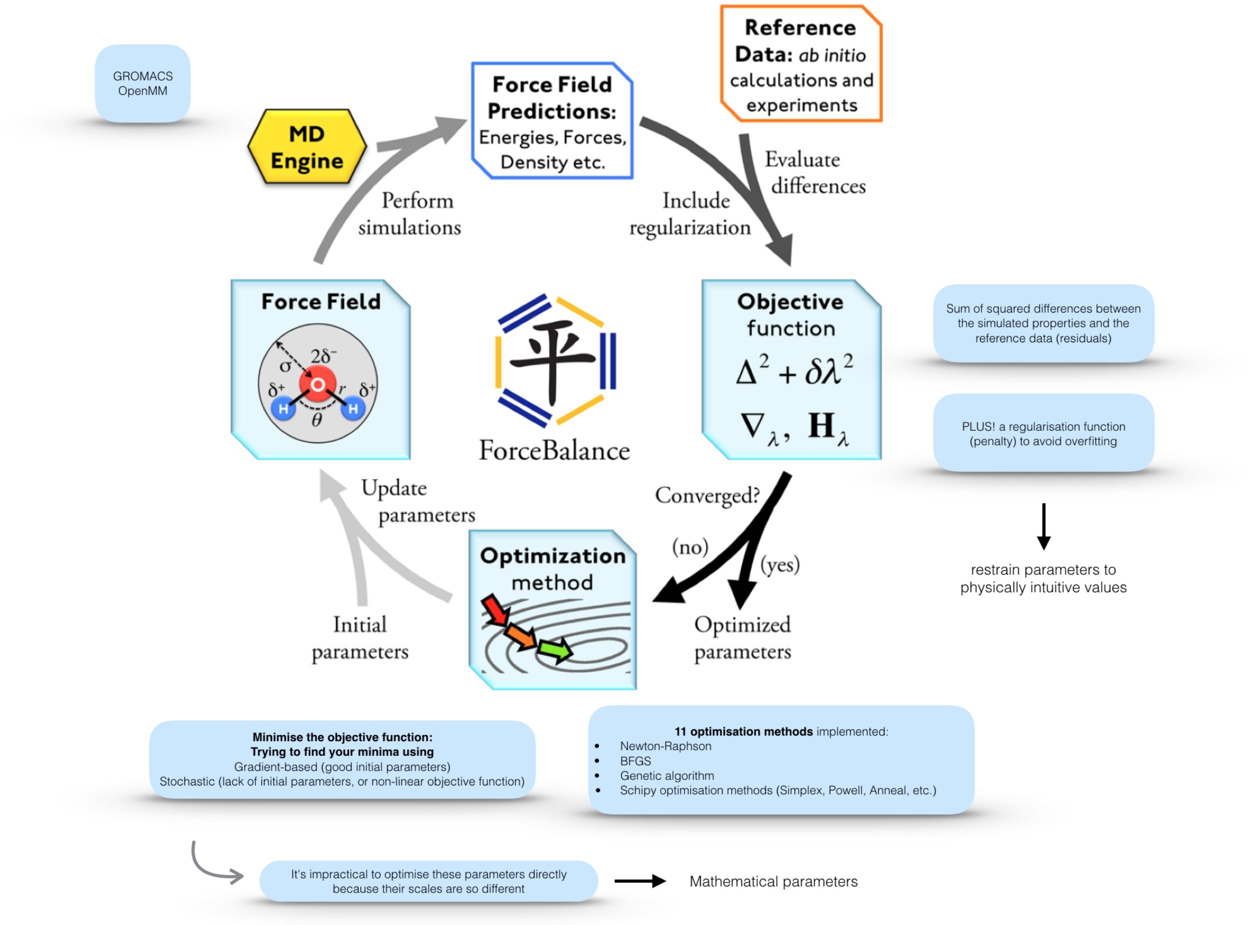
ΔE and ΔF are the energy and force residuals, w weights the importance of the residuals and Cov($\mathbf{F}_{QM}$) is the covariance of the reference QM forces. This objective function will be minimised through the used of different optimisation methods implemented in ForceBalance, such as Newton-Rhapson, BFGS, genetic algorithm, simplex, powell and conjugate gradient, among many others [6, 169, 170].

Physical values, which are the parameter values printed in the force field file, and the ones that are going to be optimised, have to be rescaled. It is impractical to optimise parameter values with different units at the same time (e.g. bonds and angles). The physical parameters are going to be related to variables of order 1; the mathematical parameters:

$$O^{phys} = O_0^{phys} + S(O^{math}) \tag{4.10}$$

Where O^{phys} represents the physical parameter that is being printed to the force field during the optimisation, O_0^{phys} is the original values that is read from the force field, S is the scaling factor that transforms between physical and mathematical parameter, and O^{math} is the mathematical parameter (mval) that is stored internally.

Then, simulations are performed, where calculations for each target property are made, and calculation of the objective function is performed. Changes to the parameters are made and new simulations are performed with an updated version of the parameters, until convergence for the objective function is achieved.



GROMACS
OpenMM

Sum of squared differences between the simulated properties and the reference data (residuals)

PLUS! a regularisation function (penalty) to avoid overfitting

restrain parameters to physically intuitive values

Minimise the objective function:
Trying to find your minima using
Gradient-based (good initial parameters)
Stochastic (lack of initial parameters, or non-linear objective function)

11 optimisation methods implemented:

- Newton-Raphson
- BFGS
- Genetic algorithm
- Schipy optimisation methods (Simplex, Powell, Anneal, etc.)

It's impractical to optimise these parameters directly because their scales are so different

Mathematical parameters

Figure 4.1

The ForceBalance workflow. The calculation starts with a set of parameters to optimise, based on experimental and/or theoretical data. Simulations are made, and an objective function is built based on the sum of the square difference of the simulation results and the targets. The optimisation method updates the parameters in order to minimise the objective function, until a convergence criteria is achieved. Image taken from [6].

In order to relate condensed phase properties to the force field parameters, ForceBalance uses a fluctuation formula similar to the Hamiltonian Gibbs-Duhem integration [171, 172]. We can add a value λ as a new thermodynamic value, where the change in Gibbs free energy is given by

$$dG = -SdT + VdP + X_G d\lambda \tag{4.11}$$

$$X_G = (\frac{\partial G}{\partial \lambda})_{P,T} \tag{4.12}$$

In that sense, the expression for a property derivative is similar to Gibbs-Duhem integration because both give the estimate of a property at some other value of the force field parameter. One can derive the ensemble average of a generic thermodynamic property A with respect to the force field parameters λ based on a previously mentioned study [169]. Assuming a generic thermodynamic property A, its ensemble average is given by

$$\langle A \rangle_\lambda = \frac{1}{Z} \int A(\mathbf{r}, V; \lambda) e^{-\beta(U(\mathbf{r};\lambda))} d\mathbf{r} \tag{4.13}$$

where Z is the partition function, $\beta = (k_B T)^{-1}$ is the thermodynamic beta, U the potential energy, and $\langle A \rangle_\lambda$ is the ensemble average of the property A in the thermodynamic ensemble of the force field, parameterised by λ.

In more detail (and omitting parameter dependence for clarity), to obtain the analytical derivative of the property with respect to the parameters (λ), one can simply follow the product rule for three terms given as:

$$\begin{aligned} \frac{d\langle A \rangle_\lambda}{d\lambda} = \frac{\partial Z^{-1}}{\partial \lambda} &\int A e^{-\beta E} d\mathbf{r} dV \\ +Z^{-1} &\int \frac{\partial A}{\partial \lambda} e^{-\beta E} d\mathbf{r} dV \\ +Z^{-1} &\int A(\frac{\partial e^{-\beta E}}{\partial \lambda}) d\mathbf{r} dV \end{aligned} \tag{4.14}$$

$$\begin{aligned}\frac{d\langle A\rangle_\lambda}{d\lambda} = &-\frac{1}{Z^2}\frac{dZ}{d\lambda}\int Ae^{-\beta E}d\mathbf{r}dV \\ &+Z^{-1}\int\frac{\partial A}{\partial\lambda}e^{-\beta E}d\mathbf{r}dV \\ &+Z^{-1}\int A(-\beta\frac{\partial E}{\partial\lambda})e^{-\beta E}d\mathbf{r}dV\end{aligned} \tag{4.15}$$

$$\frac{d\langle A\rangle_\lambda}{d\lambda} = -\frac{1}{Z^2}\frac{dZ}{d\lambda}\int Ae^{-\beta E}d\mathbf{r}dV + \langle\frac{\partial A}{\partial\lambda}\rangle_\lambda - \beta\langle A\frac{\partial E}{\partial\lambda}\rangle_\lambda \tag{4.16}$$

From there,

$$\begin{aligned}-\frac{1}{Z^2}\frac{dZ}{d\lambda}\int Ae^{-\beta E}d\mathbf{r}dV = &-\frac{1}{Z}\frac{dZ}{d\lambda}\frac{1}{Z} \\ &\int Ae^{-\beta E}d\mathbf{r}dV \\ &= -\frac{1}{Z}\frac{dZ}{d\lambda}\langle A\rangle_\lambda\end{aligned} \tag{4.17}$$

and now,

$$\frac{dZ}{d\lambda} = \frac{d[\int e^{-\beta E}]}{d\lambda}d\mathbf{r}dV = -\beta\int\frac{\partial E}{\partial\lambda}e^{-\beta(E(\mathbf{r},V;\lambda))}d\mathbf{r}dV \tag{4.18}$$

Replacing 4.18 in 4.17

$$\begin{aligned}-\frac{1}{Z^2}\frac{dZ}{d\lambda}\int Ae^{-\beta E}d\mathbf{r}dV &= \beta\langle A\rangle_\lambda\frac{1}{Z}\int\frac{\partial E}{\partial\lambda}e^{-\beta E}d\mathbf{r}dV \\ &= \beta\langle A\rangle_\lambda\langle\frac{\partial E}{\partial\lambda}\rangle_\lambda\end{aligned} \tag{4.19}$$

Finally,

$$\frac{d\langle A\rangle_\lambda}{d\lambda} = \langle\frac{\partial A}{\partial\lambda}\rangle_\lambda - \beta(\langle A\frac{\partial E}{\partial\lambda}\rangle_\lambda - \langle A\rangle_\lambda\langle\frac{\partial E}{\partial\lambda}\rangle_\lambda) \tag{4.20}$$

This final expression can be followed for any thermodynamic property that has an explicit dependence on the force field parameters. As an example, the heat of vaporisation depends directly on the energy (which depends on the parameters) in the liquid and gas phase (as $(\partial E_{liquid}/\partial\lambda - \partial E_{gas}/\partial\lambda)$), but in the case of density the term $\langle\partial A/\partial\lambda\rangle$ disappears, given that the parameters do not directly change the volume of the system.

The numerical derivatives of the potential energy with respect to the parameters are calculated using finite differences, but in principle they could also be done by an analytical approach. The main reason for choosing a numerical method is that it is a less time-consuming process and does not need to be directly included in the MD code, as would be the case for an analytical approach. The process for numerical derivatives is as follows,

1. ForceBalance starts an initial simulation with a specified prior for each parameter and a finite difference step size in mathematical parameter space (of usually 0.001).
2. The energy for the initial simulations is stored, where no change has been made to any parameter. The energy for each snapshot is stored as a variable called F0.
3. A first change is attempted, for just one parameter, changing it by [h*prior], where h is the finite difference step, and prior corresponds to the prior width (or rescaling factor) for an specific mval.
4. The energy is calculated again, after the previous change, and is stored as a variable called F1.
5. A second change is attempted, now by [-h*prior].
6. The energy is calculated again, after the previous change, and is stored as a variable called Fm1.
7. Now, the change of potential energy with respect to the parameter is done by:

$$\frac{\partial E}{\partial \lambda} = \frac{F1 - Fm1}{F0} \tag{4.21}$$

8. Steps 1 to 7 are performed for each parameter that is to be optimised.

Another important concept to be aware of in the optimisation process is the possibility of overfitting. Overfitting can occur when we try to fit a complex model (i.e. too many parameters) with insufficient or poor data. With this, an over-fitted model will have poor predictive performance. In ForceBalance, overfitting is treated by regularisation, where parameters will be penalised if they go too far away from the original parameters. Mathematically speaking, we will add a "regularisation term" to our objective function in order to prevent a perfect fitting. The ridge regression is used in ForceBalance, and involves adding a quadratic penalty to the objective function that restrains parameters to their initial values. The quadratic penalty function arises from imposing a Gaussian prior distribution on the force field parameters. Gaussian widths for each parameter reflects the intrinsic scale of that parameter and provides a form of dimensional rescaling that is required to treat parameters with different physical units in the same context. For each parameter, the centre of the prior is given by its initial value, and the prior width is the rescaling factor specified at the start of the optimisation. The prior width needs to be specified based on physical knowledge of the parameters, which affect the scaling of parameters in the parameter space, and consequently, how much they are allowed to vary.

Clear advantages are observed in ForceBalance with respect to the other mentioned methods, such as the use of multiple properties in a single objective function. These properties can be obtained from experiments and/or simulations, and they are used to fully optimise different parameter types, such as non-bonded and bonded parameters together. This can overcome problems such as the mutual dependence of force field parameters.

As mentioned in the previous chapters, hydration free energies are an important property for aqueous systems such as proteins, because they help us to understand important biological processes such as protein-protein interactions and conformational changes. Hydration free energies have been used for the validation of molecular force fields in the past [4, 12, 98], indicating that hydration free energies as a parameterisation target for coarse-grained models might improve their performance. It has been recently stated that there is considerable interest in methods that can automatically generate a coarse-grained model and are representative in terms of local structure and free energy changes [13]. In this chapter, we show the optimisation of the SIRAH 1.0 force field with the use of the ForceBalance software. We have used experimental water properties to optimise the WT4 model in SIRAH. Moreover, we present a new optimisation method using free energy gradients to improve hydration free energies of the un-charged SIRAH protein side-chains, as well as the backbone beads. We compare our results with atomistic simulations, SIRAH 1.0, the updated version of SIRAH 1.0 (SIRAH 2.0) and experimental data.

4.3. Methods

4.3.1. Optimisation of the WT4 water model.

For the WT4 model optimisation, three condensed-phase properties were optimized: density, enthalpy of vaporization and dielectric constant. Experimental values (taken from ref. [6]) for these properties were used as targets, at 298.15 K and 1 atm. The trust-radius Newton-Raphson algorithm was used to minimize the objective function. In trust region methods, there is a region of search space in which it is assumed the local derivative information is a good approximation of the objective function being minimized. After each optimization step, the trust radius may be increased or decreased based on the quality Q of the steps taken, i.e. the ratio of the objective function change between steps i and i+1 and the expected change from the local derivative information at step i. The following formula is used to adjust the adaptive trust radius after the step is taken:

$$R_{i+1} = \max\left(R_{\min}, \frac{R_i}{1+a}\right) \quad Q < 0.25 \tag{4.22}$$

$$R_{i+1} = R_i\left[1 + a\exp\left[-b\left(\frac{R_i}{R_0} - 1\right)\right]\right] \quad Q > 0.75 \tag{4.23}$$

Here R_i is the current trust radius; $R_{(i+1)}$ the trust radius at the next iteration; R_0 the default trust radius, set to 0.1; and R_{min} the minimum trust radius, set to 0.05. The parameter a, called "adapt-fac” in ForceBalance, which is related to how much the step size is increased, is set to 1.0; b, called “adapt-damp”, parameter that ties down the trust radius, is set to 0.2. The exponential term biases the current trust radius toward the default value, i.e. the trust radius increases by larger factors if the current value is smaller than the default, and vice versa if larger.

To prevent the optimisation from changing the parameters too much and to avoid overfitting, an additional penalty is applied to the objective function (regularization). For this work, the optimisation was regularized using a Gaussian prior (L2 type, or ridge regression) that is centred on the original SIRAH 1.0 parameters. Only non-bonded parameters were optimized, including van der Waals sigma (σ) and epsilon (ϵ) values, and partial charges.

100 optimisation cycle iterations were run, with the following simulation protocol: the system was minimized for 5000 steps using a steepest descent algorithm followed by an NPT equilibration time of 5 ns. Production runs were performed for 15 ns. A leap-frog algorithm was used to integrate Newton’s equations of motion with a time-step of 20 fs. Electrostatic interactions are calculated using the Particle mesh Ewald method [28] with a direct cut-off of 1.2 nm and a grid spacing of 0.2 nm. A 1.2 nm cut-off was used for van der Waals interactions. The V-rescale thermostat [125] and the Parrinello-Rahman barostat [124] were used to maintain the temperature at 298.15 K and the pressure at 1 atm. The simulation protocol was based on the original publication of the SIRAH protein force field [5]. All simulations were run with GROMACS v. 2018.2 [123]. Sta-

tistical fluctuations in the thermodynamic properties dominated the objective function after 30 iterations, and the sets of parameters with the lowest objective function were chosen as the best solution.

The above protocol was also tested for the optimisation of density, enthalpy of vaporization and dielectric constant with a data set that included experimental liquid phase measurements of water at different temperatures. In order to test the best combination for optimisation, simulation times of 1 ns, 5 ns and 15 ns were used. Table 4.1 shows all the combination of targets, simulation times and temperatures used in the optimisation procedure. The new optimised model will be referred to as WT4-FB from now on.

	Simulation time (ns)	**Temperature range (K)**	**Targets**
Combination 1	1	261, 269, 281, 289, 298, 309, 321, 329, 341, 349, 360	Exp. water density
Combination 2	1	281, 289, 298, 309, 321	Exp. water density and enthalpy of vaporisation
Combination 3	5	281, 289, 298, 309, 321	Exp. water density, enthalpy of vaporisation and dielectric constant
Combination 4	15	298	Exp. water density, enthalpy of vaporisation and dielectric constant

Table 4.1

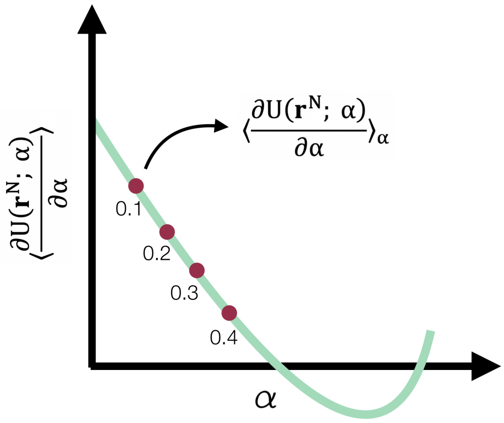
Combination of simulation times and targets for the optimisation of the WT4 parameters. Tested protocols that were used in the attempt to optimise the CG water model in SIRAH. They mainly vary in the simulation time used in the optimisation, as well as the temperature range**.**

4.3.2. Optimisation based on free energy gradients: overview

We have implemented a new mathematical expression for the optimisation of coarse-grained force field parameters based on free energy gradients from atomistic simulations and the thermodynamic integration theory (see section 1.5.2 and equation 1.37). Starting with a set of simulations that evaluate $\langle \Delta U \rangle_\alpha$ for AT systems at selected values of α, we fit these values in our CG simulations by optimising the CG parameters, which indirectly improves the hydration free energies (Fig. 4.2). The derivatives of the free energy gradients with respect to the force field parameters are calculated and used to build an objective function, which is the squared sum of the differences between the AT and CG gradients. The analytical derivative of $\langle \Delta \text{U} \rangle_\alpha$ with respect to the force field parameters can be obtained based on equations 4.14 to 4.20 as:

$$\frac{\partial \langle \Delta U \rangle_\alpha}{\partial \lambda} = \langle \frac{\partial \Delta U}{\partial \lambda} \rangle_\alpha - \beta(\langle \Delta U \frac{\partial E}{\partial \lambda} \rangle_\alpha - \langle \Delta U \rangle_\alpha \langle \frac{\partial E}{\partial \lambda} \rangle_\alpha) \tag{4.24}$$

where λ corresponds to the force field parameter, $\langle \Delta \text{U} \rangle_\alpha$ is the ensemble average of the energy difference between α = 0.0 and α = 1.0, simulated at a defined α value, ΔU corresponds to the instantaneous energy difference for each snapshot between α = 0.0 and α = 1.0, and E is the potential energy of the system at the corresponding α value. Rather than optimising the free energies directly, we optimise against the ensemble average of the free energy gradients at specific α values, $\langle \Delta \text{U} \rangle_\alpha$. This choice was made because it presents itself as a less time-consuming and complex approach, given that we only use a small group of α values rather than the whole range, based on the assumption that there is a linear relationship, if one of the free energy gradients is correct, we expect a better performance across the whole range of α values, and that coarse-grained and atomistic systems should have similar free energy gradient profiles. The derivative of the free energy gradients $\langle \Delta \text{U} \rangle_\alpha$ with respect to the force field parameters λ is composed of ensemble averages of instantaneous ΔU values, and derivatives of ΔU and the potential energy with respect to the FF parameters, at each α point used, where both derivatives are obtained numerically by finite difference using snapshots from the corresponding trajectories.



Optimisation based on free energy gradients. Free energy gradients are collected from specific α points (red dots) using atomistic simulations and used later as fitting for the same point using coarse-grained simulations**.**

Figure 4.2

4.3.3. Optimisation of the SIRAH protein force field: uncharged side-chains and backbone

A work-flow showing the steps followed in this work, and separated into four main stages, is presented in figure 4.3. Briefly, hydration free energies for atomistic systems are calculated by decoupling both van der Waals and charge parameters. Then, atomistic free energy gradients are collected as an average of ΔU values, at simulations with different α values, $\langle \Delta U \rangle_\alpha$. These data are used to optimise the parameter of each specific CG side-chain (or the backbone) with its corresponding $\langle \Delta U \rangle_\alpha$ value. Then, parameters corresponding to the smallest objective function are collected. These parameters are then used to re-calculate new hydration free energies of the CG side-chains.

Stage A: The interaction energy terms between the solute and solvent are linearly related to the coupling parameter α. With this, the solvation free energies for the side-chain analogues, for the atomistic and coarse-grained systems, were calculated based on a decoupling approximation. That is, interactions between the solute and the solvent are gradually turned off. Our reference state will be our system in solution, and the final state will be the solute in vacuum. The OPLS-AA [119] and the AMBER-14SB [173] force fields were used for the atomistic side-chain analogues and the backbone, respec-

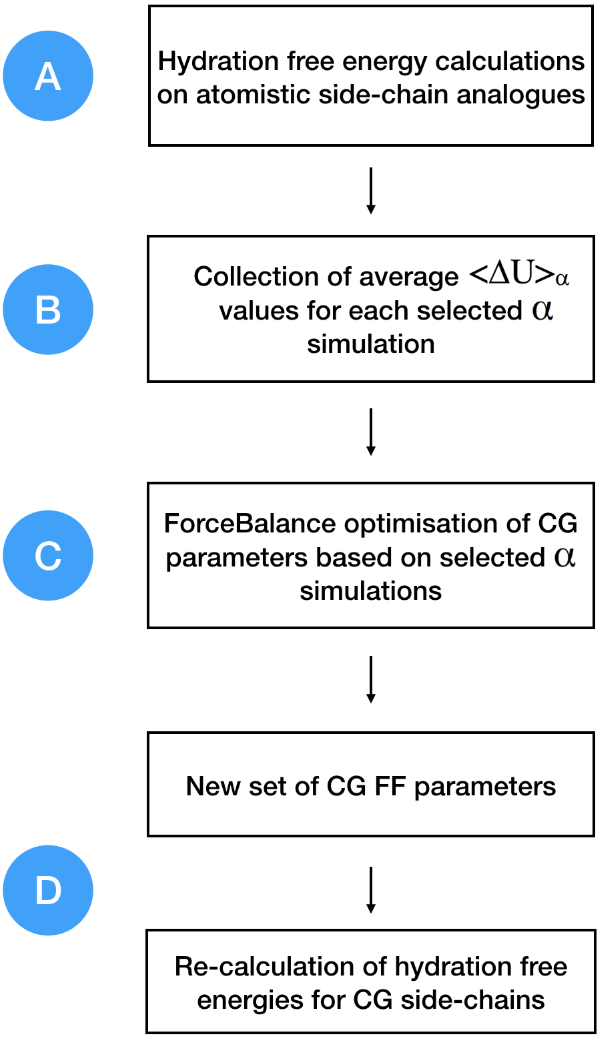

A
Hydration free energy calculations on atomistic side-chain analogues
B
Collection of average <ΔU>α values for each selected α simulation
C
ForceBalance optimisation of CG parameters based on selected α simulations
New set of CG FF parameters
D
Re-calculation of hydration free energies for CG side-chains

Figure 4.3

General workflow for the SIRAH force field optimisation. Free energy gradients are collected from atomistic simulations and used as optimisation targets in ForceBalance. New parameters are obtained and later used in the re-calculation of hydration free energies for CG beads (side-chains and backbone). Letters from A to D correspond to each of the main stages in the optimisation and validation process**.**

tively, where in all cases systems were solvated in a TIP3P water box. Since we are comparing our calculations with previous studies, especially those that give closer results to experiments, we have tried to be consistent with the force fields and simulation protocols used in those studies, hence the choice of different force fields and α values in the optimisation process for side-chains and backbone. The SIRAH 1.0 protein force field [5] was used for the CG side-chains and backbone beads, solvated in a WT4 [7] water box. Electrostatic and van der Waals interactions were turned off together. Eleven discrete values of the coupling parameter α were used for the scaling of both CG and AT potentials: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0, where 0.0 and 1.0 represent the fully on and fully off systems. In the case of N-methylacetamide (NMA), which was used as a representation of the backbone beads, twenty-five values were used: 0.0, 0.5, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.87, 0.89, 0.91, 0.93, 0.95, 0.97 and 1.0. The soft-core scaling potential for Lennard-Jones (with α-LJ = 0.5) and Coulombic interactions were used to smoothly vary the potentials [174, 175]. A shifted version of the regular soft-core potential is implemented in GROMACS (see equation 3.1 for details).

Simulations were run for 5 ns per window, with a previous equilibration of 1 ns and the corresponding system minimization. All the simulations were run using the NPT ensemble. The Multistate Bennett Acceptance Ratio (MBAR) [41] was used to compute the free energy difference.

For the AT simulations, a leap-frog stochastic dynamics integrator was used for integration of Newton's equations of motion with a time-step of 2 fs. Electrostatics interactions were calculated using the PME procedure with a real-space cut-off of 1.2 nm and a Fourier grid spacing of 0.12 nm. Van der Waals interactions were modelled using the classical Lennard-Jones potential with a cut-off of 1.2 nm. The LINCS algorithm [176] was applied to constrain all bond lengths. AT simulations were run at 1 atm with the Parrinello-Rahman barostat [124] and at 298.15 K with the Berendsen thermostat [177]. The choice of temperature was based on the temperature at which the WT4 model was optimised, and as a matter of consistency with the studies that we are using to compare our results. This applies as well for the choice of thermostat.

For the CG simulations, a leap-frog stochastic dynamics integrator was used for integration of Newton's equations of motion with a time-step of 20 fs. Electrostatic interactions were calculated using the PME procedure with a grid spacing of 0.2 nm. Nonbonded interactions were modelled using the classical Lennard-Jones potential and a

Coulombic energy function, with a cut-off of 1.2 nm each. The LINCS algorithm was applied to constraint all bond lengths. All simulations were run at 1 atm with the Parrinello-Rahman barostat [124] and at 298.15 K with the v-rescale thermostat [125]. All simulations were run with GROMACS v. 2018.2 [123].

Stage B: Collection of AT $\langle \Delta \mathrm{U} \rangle_\alpha$ values. $\langle \Delta \mathrm{U} \rangle_\alpha$ were collected from the AT simulations in stage A, at different α values. For most of the side-chains, α simulations at 0.0 were not used due to the large magnitudes of $\langle \Delta \mathrm{U} \rangle_\alpha$ values and differences between AA and CG that could not be closely fitted. Val, Cys and Trp were the only exceptions where these higher values were not found. Values were collected with an in-house Python code created for this purpose, averaging ΔU values for each frame in the trajectories. Table 4.2 summarise the α values used for each of the simulated side-chains in their ForceBalance optimisation.

Stage C: Optimisation of SIRAH CG side-chains and backbone. The derivatives of the free energy gradients ($\langle \Delta U \rangle_\alpha$) with respect to the force field parameters are calculated. These are used to build an objective function, which is a squared sum of the differences between the AA and CG $\langle \Delta \mathrm{U} \rangle_\alpha$ values. The optimisation was carried out using ForceBalance using the same settings described in the WT4 model development, except the adapt-fac and adapt-damp parameters were set to 0.2 and 0.5, respectively, for a more variable step-size. Only 10 sets of parameters were optimised, 9 of them corresponding to 13 uncharged amino acid side-chains, as some of the side-chains are described by identical parameters, and 1 set corresponding to the backbone beads. In this case, the targets were atomistic free energy gradients at 2 or 3 different α simulation values (table 4.2). Proline is the only side-chain that has not been optimised given the lack of side-chain analogues, keeping its previous parameter values. Only non-bonded parameters were optimized, including van der Waals epsilon (ϵ) values, and charges, mainly given the parameter sensitivity observed (see results section for more details). All the optimisation simulations for the SIRAH beads were run with the optimized WT4-FB model (this work). 100 optimization cycles were carried out, and the optimal parameters were taken from the lowest value of the objective function. Systems were minimized for 5000 steps using a steepest descent algorithm followed by a NPT equilibration time of 5 ns. Production runs were performed for 10 ns. A leap-frog algorithm was used for integration of Newton's equations of motion with a time-step of 20 fs. Electrostatic interactions are calculated using the Particle Mesh Ewald method with a direct cut-off of 1.2 nm and a grid spacing of 0.2 nm. A 1.2 nm cut-off was used for van der Waals interactions. The V-rescale thermostat [125] and the Parrinello-Rahman barostat [124] were used to maintain the temperature at 298.15 K and the pressure at 1 atm, respectively. The simulation conditions were consistent with the original SIRAH publication [5]. All simulations were run with GROMACS v. 2018.2. All specific non-bonded pairs, previously set to the original SIRAH 1.0 force field, between the backbone beads (GC, GN and GO) and water beads (WT) have been removed, and we have set those interactions using Lorentz-Berthelot combing rules.

Stage D: Re-calculation of CG hydration free energies. The new optimised SIRAH force field was used for the re-calculation of the coarse-grained hydration free energies. The same protocol in stage A was used, based on the original publication of the SIRAH 1.0 protein force field [5]. For all simulations, a leap-frog stochastic dynamics integrator was used for integration of Newton's equations of motion with a time-step of 20 fs. Electrostatic interactions were calculated using the PME procedure with a grid spacing of 0.2 nm. Non-bonded interactions were modelled using the classical Lennard-Jones potential and a Coulombic energy function, with a cut-off of 1.2 nm each. All simulations were run at 1 atm with the Parrinello-Rahman barostat and at 298.15 K with the v-rescale thermostat. All simulations were run with GROMACS v. 2018.2. The new hydration free energies with the set of optimized parameters are shown in table 4.4, and they are compared with hydration free energies calculated from atomistic systems, with the original SIRAH 1.0 protein force field and the updated SIRAH 2.0 version.

Side-chain	α values
Asn (acetamide)	0.1, 0.2, 0.5
Cys (methanethiol)	0.0, 0.2, 0.4
His (methylimidazole)	0.1, 0.2, 0.4
Met (methyl-ethylsulfide)	0.1, 0.2, 0.4
Phe (toluene)	0.5, 0.6
Ser (methanol)	0.1, 0.2, 0.5
Trp (methylindole)	0.0, 0.4, 0.5
Tyr (p-cresol)	0.1, 0.4, 0.5
Val (propane)	0.0, 0.5
Backbone (N-methylacetamide)	0.1, 0.3

Table 4.2

Gradient sets in the optimisation process. α simulation values used for the collection of $\langle \Delta U \rangle_\alpha$ values, that correspond to the targets in the optimisation of the CG beads in ForceBalance. Atomistic analogues used are shown in parenthesis**.**

4.4. Results and Discussion

4.4.1. Optimisation of the WT4 water model

We start our ForceBalance calculation with the optimisation of the WT4 water model, where only non-bonded parameters were optimized (charges, sigma and epsilon values). Three condensed-phase properties for liquid water were used as reference data: density, enthalpy of vaporization and dielectric constant. All these target properties were used in different combination in relation to temperature range, and simulation time, as stated in the methodology section. The original WT4 model is able to reproduce experimental thermodynamic properties such as the water density at 298 K, but it is less satisfactory in the prediction of other properties (i.e. dielectric constant, expansion coefficient, surface tension, etc.) [7].

In order to test the strength of ForceBalance, we started the optimisation using only experimental data of densities for 11 temperature values between 260 K and 360 K (see table 4.1), and running the simulations for 1 ns. As shown in figure 4.4, ForceBalance is able to greatly improve the WT4 densities. Even so, and as expected, values for other thermodynamic parameters not included in the optimisation, such as the dielectric constant, move far away from the experimental values (see Fig. A.3).

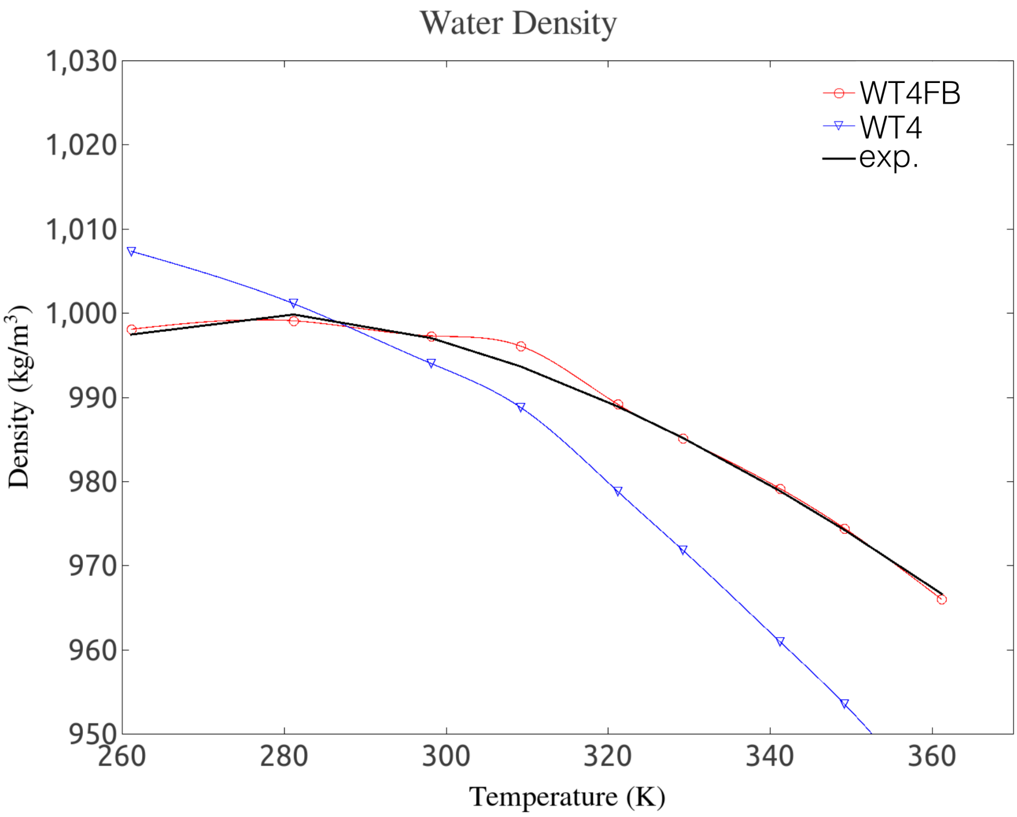
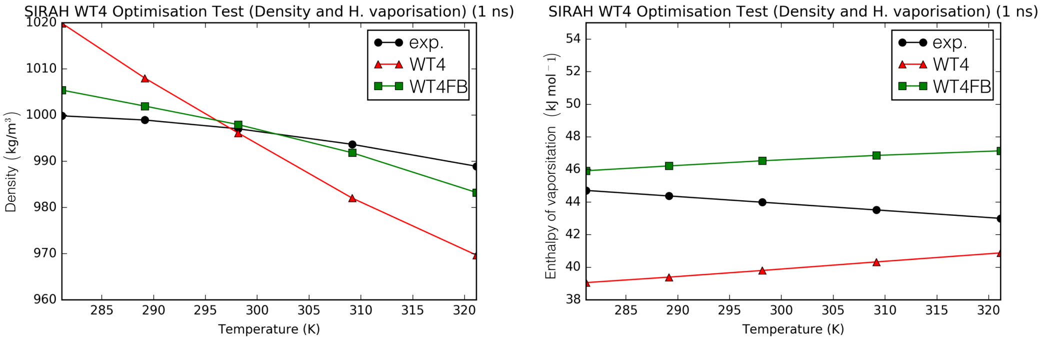


Figure 4.4

WT4 optimisation using density as target. Comparison on the performance of the optimised WT4 model (WT4FB) with respect to the unoptimised force field (WT4) for simulations of 1 ns and using only experimental data of water densities (exp.) at 11 temperatures [6].

Using fewer temperatures to fit multiple thermodynamic properties for the WT4 force field does not change the outcome. Simulations of 1 ns at 5 different temperatures between 280 K and 320K (see table 4.1) are shown in figure 4.5. As can be seen, the density and the enthalpy of vaporisation get closer to the experimental values compared to the unoptimised force field. Even so, the performance is not good for the whole range of tested temperatures, particularly for the enthalpy of vaporisation where an increase of the simulated values is seen, rather as the decreased observed in the experimental data.

In ForceBalance, the objective function is a sum of squared residuals between simula-



WT4 optimisation using density and enthalpy of vaporisation as target. Comparison on the performance of the optimised WT4 model (WT4FB) with respect to the unoptimised force field (WT4) for simulations of 1 ns and using experimental data of water densities and enthalpy of vaporisation (exp.) at 5 temperatures**.**

Figure 4.5

tion results and/or experimental data. The simulation results are going to be ensemble averages, which come from finite length simulations. With this, an amount of random noise is going to be added to the estimation of these averages. With this said, the longer the simulations, the smaller the noise is going to be and the more accurate the objective function should be.

Given this observation, in order to continue the optimisation of multiple thermodynamic properties, longer simulations were performed. The same temperature range between 280 K and 320 K was used (see table 4.1), but now extending the simulation time to 5 ns. Moreover, the dielectric constant was also added to the optimisation procedure. Figure 4.6 summarises the performance results. As the simulation length increases in this case, the agreement of the simulated properties with experiments improves for the optimised WT4 model, showing that the values of density and enthalpy of vaporisation are closer to the experimental values. The behaviour of the dielectric constant is better for the optimised force field, but still with important deviations from experimental values and far away from being perfect across the whole tested temperature range. A possible solution to this problem is to extend further the simulation times in order to reduce the noise. Either way, there is a chance that no further improvement can be achieved given the number of force field parameters that are available, which is an important limitation.

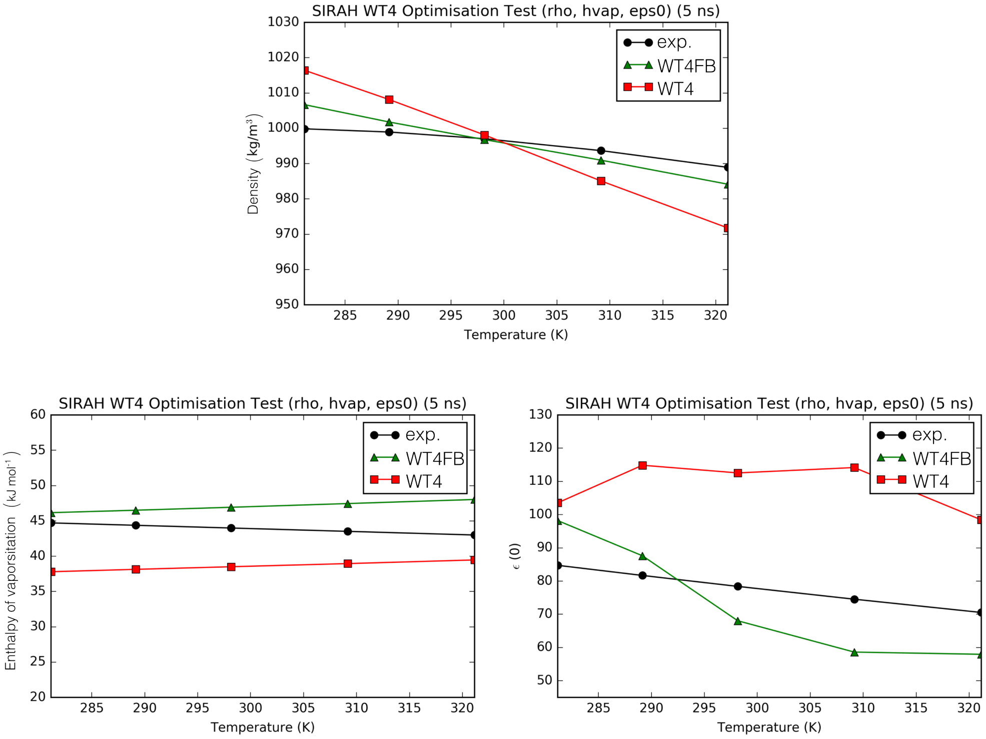


Figure 4.6

WT4 optimisation using density, enthalpy of vaporisation and dielectric constant as target. Comparison of the performance of the optimised WT4 model (WT4FB) with respect to the un-optimised force field (WT4) for simulations of 5 ns and using experimental data of water densities (top), enthalpy of vaporisation (bottom left) and dielectric constant (bottom right) at 5 temperatures**.**

We therefore decided to stay with the last combination, i.e. a simulation time of 15 ns for the optimisation and just one temperature at 298 K. This decision was made on the basis that this is the temperature in which we are interested for future application on protein systems. We have optimised the force field parameter of the WT4 model using water density, enthalpy of vaporisation and dielectric constant. The new WT4 model (now called WT4-FB) overcomes the previous issue with the dielectric constant in the original model by accurately reproducing experimental values for the three properties together (table 4.3). Calculations of the expansion coefficient yield similar results to those of the original model (11.8*x*10^{-4} K^{-1} vs. 11.6*x*10^{-4} K^{-1}) [7]. Thus, optimizing WT4 with ForceBalance does not necessarily improve all properties; the level of accuracy obtainable will depend on the granularity of the CG representation, the choice of force field functional form the number of parameters available to be optimised.

Property	Expt.	WT4	WT4-FB (this work)
ρ (kg/m^3)	997.045	996.6 ± 0.3	995.4 ± 1.5
ΔH_{vap} (kJ/mol)	43.989	39.8 ± 0.2	43.7 ± 0.2
ϵ_r	78.409	123.7 ± 14.2	74.2 ± 12.3
α (10^{-4} K^{-1})	2.572	11.6 ± 2.4	11.8 ± 2.7

Table 4.3

Comparison of WT4 and WT4-FB models against experimental water properties at 298 K and 1 atm. The calculated properties correspond to density (ρ), enthalpy of vaporization (ΔH_{vap}), dielectric constant (ϵ_r) and expansion coefficient (α). Experimental properties were obtained from reference [6]. Error bars are reported as standard errors**.**

4.4.2. Optimisation of the SIRAH protein force field: uncharged side-chains and backbone.

Initially, a screening test was performed to evaluate the parameter dependence of $\langle \Delta U \rangle_\alpha$ with respect to the force field parameters, i.e. to evaluate the changes in $\langle \Delta U \rangle_\alpha$ based on changes in the force field parameters. Only van der Waals σ (VDWσ) and van der Waals ϵ (VDWϵ) parameters were tested. As an example, figure 4.7 shows simulations at a value of α=0 (fully charged), where different scaled parameters were tested to evaluate changes in the $\langle \Delta U \rangle_0$. Based on the parameter dependence observed in figure 4.7 for the case of Val, the $\langle \Delta U \rangle_0$ values do not significantly change within a sensible range of VDWσ values (Fig. 4.7, left panel). On the contrary, an important parameter dependence is shown with respect to VDWϵ values, with clear changes in the values of $\langle \Delta U \rangle_0$ (Fig. 4.7, right panel). In figure 4.7, the van der Waals parameters are plotted in the form of internal optimisation variables in ForceBalance ("mathematical parameters"), which are related to the physical parameters (i.e. the parameters that are actually printed in the force field file) as a shifted displacement form the original value:

$$K_{\text{phys}} = K_{\text{phys}0} + SF * K_{\text{math}} \tag{4.25}$$

where K_{phys} corresponds to the parameter that is used in the simulation after the optimisation process, $K_{\text{phys}0}$ is the initial parameter before the optimisation, SF is the scaling factor and K_{math} the mathematical value used in the optimisation process. The following scaling factors were used: VDWσ = 0.0529 and VDWϵ = 2.4789, which correspond to the default values in ForceBalance. This can be translated to a physical value range of 0.3770 nm to 0.4829 nm for VDWσ, and a value range of -0.6850 kJ/mol to 1.7894 kJ/mol. Finally, and based on these observations, only VDWϵ values and charges were optimised given the parameter sensitivity observed (see Fig. 4.7) and that the optimisation yields improvements in the side-chain hydration free energies using this protocol.

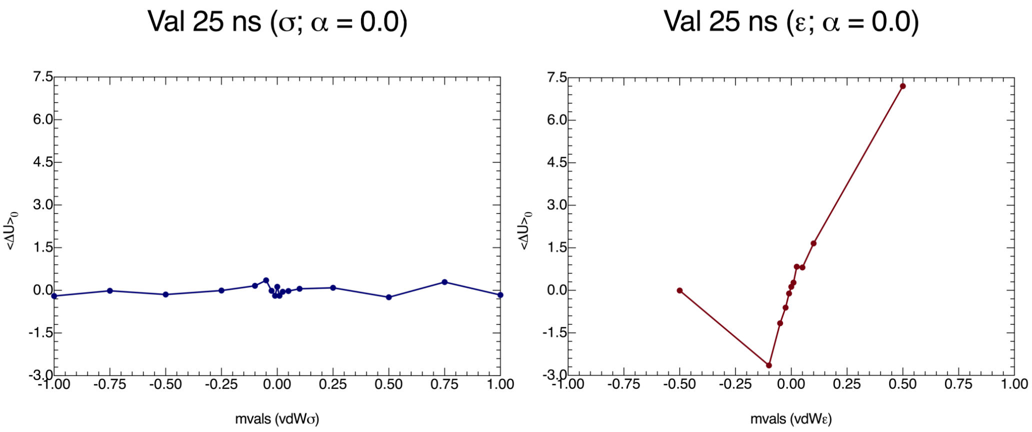


Figure 4.7

Parameter dependence for Val. Changes of $\langle \Delta U \rangle_0$ with respect to the vdW sigma and epsilon values are shown (left and right panel, respectively). Simulations were run at α = 0.0 (fully charged) for 25 ns. The simulation conditions were the same as the ones used for the side-chain optimisations (see stage A). Van der Waals parameters are plotted as mathematical values (mvals).

In relation to the atomistic gradient choice, it is important to note that for most cases, AT gradients at α=0 were too high to be fitted by ForceBalance due to the large magnitude of the gradient and the large difference between AT and CG observed. Inclusion of the α=0.0 point would have introduced a very large contribution to the objective function and worsened the quality of fit of all the other α points. We are assuming that the gradients should behave in a similar manner between the all-atom and coarse-grained systems, but this might not be the case. Using the free energy gradients as a proxy for the free energies, instead of the free energy itself, relies on the assumption that 1) if one of the free energy gradients is correct, we expect a better performance across the whole range of α values, and 2) coarse-grained and atomistic systems should have similar free energy gradients. Neither of these is necessarily true.

One of the main points that encouraged the development and improvement of these CG models, and also an important limitation of the SIRAH 1.0 force field (see chapter 3, section 3.3), is the lack of accuracy for hydration free energies of amino acid side-chains (section 3.3.3, Fig. 3.11), which could limit the force fields predictive power in protein simulations. Calculations of the SIRAH 1.0 hydration free energies yield completely different results compared to all-atom OPLS-AA simulations, with mean unsigned errors (MUE) of 5.03 kcal/mol vs. 1.04 kcal/mol, for SIRAH 1.0 and all-atom systems, respectively, calculated against experimental values (Table 4.4).

10 sets of CG parameters were optimized representing 13 uncharged side-chains because of the shared mapping scheme and bead types for some groups of side chains; e.g. Asn/Gln share the same mapping, as do Ser/Thr and Val/Leu/Ile, and the backbone. Figure 4.8 and table 4.4 summarise the performance of our new set of parameters for uncharged side-chains and the backbone, now called SIRAH-OBAFE (**O**ptimised **B**ased on **A**tomistic **F**ree **E**nergies), together with the new WT4-FB force field, against HFEs from the atomistic force field OPLS-AA [119] (for side-chains), the AMBER14SB [143] (for the backbone) the original SIRAH 1.0 force field [5], the updated SIRAH 2.0 force field [11], and experimental data [6].

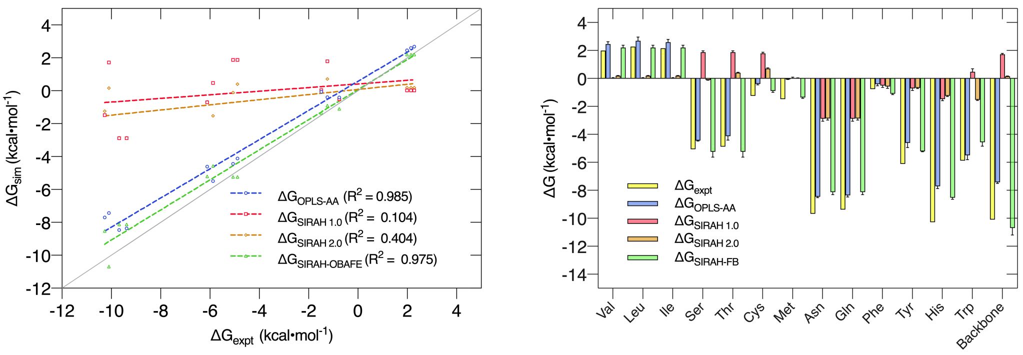


Figure 4.8

Comparison of HFEs from the new set of optimised parameters (SIRAH-OBAFE) against atomistic simulations, the original SIRAH 1.0 force field, the latest version SIRAH 2.0 and experimental data. (A) Linear regression of predicted ΔG values for OPLS-AA (blue), SIRAH 1.0 (red), SIRAH 2.0 (orange) and SIRAH-OBAFE (green) against experimental data. Each point represents a specific side-chain. The grey line shows a perfect fit (x=y), and R^2 values are given in the inset legends. (B) Bar plot comparison of predicted ΔG values for OPLS-AA (blue), SIRAH 1.0 (red), SIRAH 2.0 (orange) and SIRAH-OBAFE (green) against experimental data (yellow; y axis) for all the neutral side-chains. Error estimates were calculated as standard errors based on three repeat simulations. For some cases, red bars appear to be missing given they are too small to be seen on the scale of the plot**.**

As can be seen, the original set of parameters in SIRAH 1.0 do not perform well for the prediction of HFEs, with an R^2 of 0.104 against experimental values (Fig. 4.8A). A similar case is observed for the latest version SIRAH 2.0, with an R^2 of 0.404 (Fig. 4.8A). SIRAH-OBAFE is able to greatly improve the agreement with experimental HFEs to be as good as OPLS-AA, with an R^2 of 0.982 and 0.975, respectively. SIRAH-OBAFE reproduces the correct sign of several neutral side-chains where the previous SIRAH model predicted the wrong sign, such as Ser, Thr, Cys and Trp (Fig. 4.8B). Significant improvements have been made to the HFEs of hydrophobic residues such as Val, Leu, and Ile; these share the same representation in SIRAH using just one bead. The original SIRAH 1.0 and the updated SIRAH 2.0 models predict -0.02 ± 0.01 and -0.18 ± 0.01 kcal/mol for the HFE, respectively (table 2), whereas SIRAH-OBAFE achieves a value of -2.26 ± 0.03 kcal/mol (table 2); the latter value is much closer to OPLS-AA simulations and experiment which predict HFEs of (-2.45, -2.69, -2.59) and (-1.99, -2.28, -2.15) kcal/mol, for Val, Leu, and Ile, respectively (table 4.4). In the case of methionine, SIRAH-OBAFE produced even more accurate HFE values than the OPLS-AA model that provided the

HFE gradients to which the CG model was fitted; we think this result is fortuitous and the differences are within the residual errors of the CG model vs. the AT reference (see below).

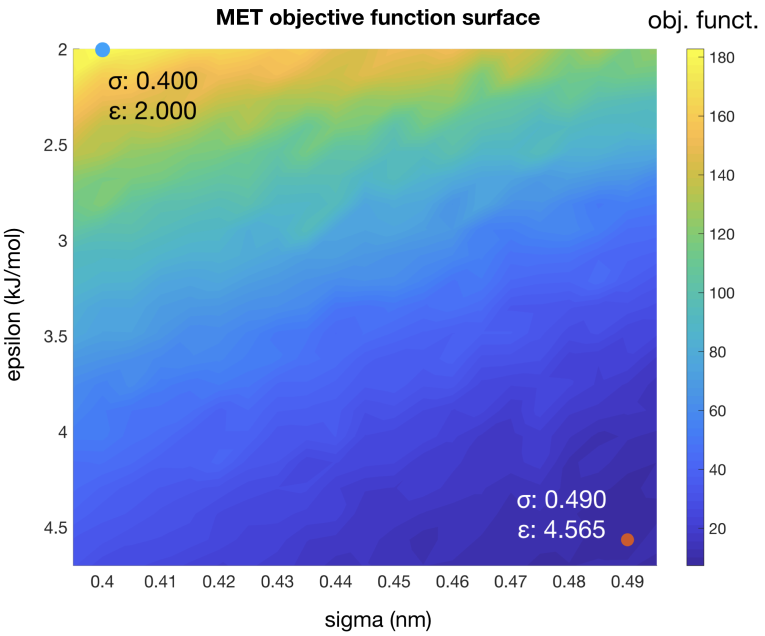


Figure 4.9

Methionine objective function surface. 441 combinations (21 x 21) of vdWσ and vdWϵ simulations were performed, and single calculations of the objective function were extracted and plotted. The maximum and minimum values for the objective function are shown as blue and red dots, respectively, and for each of these the parameter combination is shown. Simulations were done at α=0 (fully charged)**.**

The optimisation of methionine is an example where our method has worked finding a minimum, i.e. the optimal set of parameters to minimise the objective function. Figure. 4.9 shows a manual search of 441 parameter combinations that led to similar results to those obtained for the full optimisation of methionine in ForceBalance, with values of vdWσ = 0.49 nm, vdWϵ = 4.56 kJ/mol, and vdWσ = 0.48 nm and vdWϵ = 4.22 kJ/mol, for the manual search and the automated optimisation in ForceBalance, respectively. However, the assumption that similar free energy gradient profiles between AT and CG systems will lead to the correct hydration free energies is not met here, with the optimised parameters giving a worse agreement with the calculated AT hydration free energy (see table 4.4). Figure 4.10 shows the free energy gradients for the atomistic and coarse-grained methionine side-chain. The overall shape of the profile is maintained, but differences exists in the magnitude of the gradients. This may account for the differences observed for the calculated HFEs. Fortuitously, the optimised parameters led to better agreement with experimental hydration free energies.

In the case of phenylalanine, the optimised parameters simulated with the WT4-FB model performed worse compared to the original SIRAH 1.0 force field parameters calculated with the WT4-FB water model, with values of 1.96 ± 0.04 kcal/mol vs. 1.12 ± 0.06 kcal/mol, respectively. We believe this is mainly due the complexity on the free energy gradient profile for this residue. Moreover, later optimisation runs of side-chains that share the A2C bead-types with phenylalanine (such as His, Tyr, and Trp) were performed using this parameter fixed.

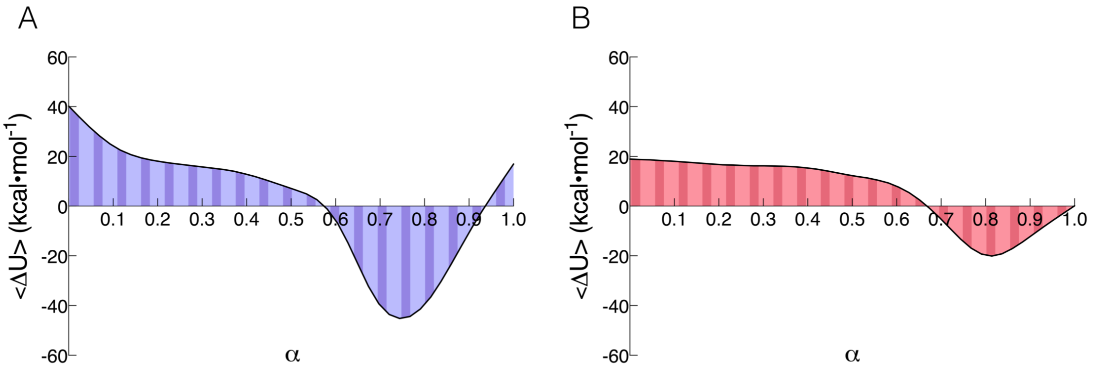


Figure 4.10

Free energy gradients for methionine. (A) atomistic free energy gradients and (B) coarse-grained free energy gradients. The results represent 11 α simulations with average $\langle \Delta U \rangle_\alpha$ values for each of those simulations shown, at α=0 (fully charged)**.**

4.5. Summary

In this chapter we have shown the optimisation of a coarse-grained water model based on experimental data. Moreover, the optimisation of uncharged CG protein side-chains and backbone, by incorporating atomistic free energy data into the target function, has been successful, representing a new and promising approach in force field development.

	Expt.	OPLS-AA	SIRAH 1.0	SIRAH 2.0	SIRAH-OBAFE
Backbone	-10.1	-7.40 ± 0.01 (AMBER-14SB)	1.73 ± 0.01	-0.16 ± 0.01	-10.91 ± 0.05
Val	1.99	2.45 ± 0.04	0.02 ± 0.01	0.18 ± 0.07	2.26 ± 0.04
Leu	2.28	2.69 ± 0.06	0.02 ± 0.01	0.18 ± 0.01	2.26 ± 0.04
Ile	2.15	2.59 ± 0.08	0.02 ± 0.01	0.18 ± 0.01	2.26 ± 0.03
Ser	-5.06	-4.44 ± 0.01	1.87 ± 0.04	-0.10 ± 0.09	-5.26 ± 0.10
Thr	-4.88	-4.12 ± 0.11	1.87 ± 0.04	0.40 ± 0.01	-5.26 ± 0.10
Cys	-1.24	-0.39 ± 0.02	1.78 ± 0.03	0.71 ± 0.01	-0.92 ± 0.07
Met	-1.48	-0.06 ± 0.01	0.03 ± 0.02	0.01 ± 0.03	-1.36 ± 0.02
Asn	-9.68	-8.46 ± 0.02	-2.87 ± 0.07	-2.85 ± 0.04	-8.12 ± 0.05
Gln	-9.38	-8.36 ± 0.04	-2.87 ± 0.07	-2.86 ± 0.04	-8.12 ± 0.05
Phe	-0.76	-0.40 ± 0.04	-0.50 ± 0.05	-0.57 ± 0.05	-1.12 ± 0.06
Tyr	-6.11	-4.61 ± 0.13	-0.70 ± 0.06	-0.67 ± 0.02	-5.10 ± 0.04
His	-10.27	-7.70 ± 0.06	-1.47 ± 0.04	-1.24 ± 0.02	-8.46 ± 0.08
Trp	-5.88	-5.55 ± 0.22	0.47 ± 0.09	-1.52 ± 0.02	-4.51 ± 0.04
MUE		1.04	5.03	4.45	0.68
MSE		-1.04	-4.13	-3.61	-0.43
R^2		0.98	0.10	0.40	0.97

Table 4.4

Comparison of hydration free energies. Hydration free energies of neutral side-chains and backbone using the OPLS-AA, AMBER-14SB, SIRAH 1.0, SIRAH 2.0 and the newly optimised SIRAH-OBAFE force fields. Mean unsigned error (MUE), mean signed error (MSE) and R^2 are shown, compared to experimental values (Expt.) obtained from reference [8]. All values are in units of kcal/mol. Errors are shown as standard errors**.**

Our new approach for CG FF optimisation is based on using derivatives of the free energy gradients (i.e. $\langle \Delta U \rangle_\alpha$ at different values of the coupling parameter α) with respect to the force field parameters. We choose to work with free energy gradients due to their linear relationship with the easily computed "vertical energy gap", $\langle \Delta U \rangle_\alpha$. In practice, the thermally averaged CG $\langle \Delta U \rangle_\alpha$, is fitted to atomistic $\langle \Delta U \rangle_\alpha$, where one or more selected values of the coupling parameter α are used to carry out the simulations. Our implementation of this method into ForceBalance enables full automation of the complex optimisation procedure and the incorporation of flexible choices of target data. The results in this chapter show that:

- The new WT4-FB model presents a clear improvement in the target data, especially in the estimation of the dielectric constant at 298 K and 1 atm.

- The new set of parameters clearly improve the agreement between experimental and simulated hydration free energies compared to the original SIRAH 1.0 and the updated SIRAH 2.0

Non-bonded interaction parameters of uncharged side-chains and the backbone were optimised against higher resolution models and experimental hydration free energies, yielding a new parameter set called SIRAH-OBAFE. Comparative analyses of the newly optimised parameters for the uncharged protein side-chains and the backbone reproduce well hydration free energies compared to the previous version of the SIRAH force field, with increased R^2 values of 0.97 for the new SIRAH-OBAFE parameter set, compared to values of 0.1 and 0.4 for the SIRAH 1.0 and SIRAH 2.0 sets, versus experimental values. The next step involves the evaluation of the new optimised SIRAH-OBAFE force field to protein systems, and the calculation of free energy surfaces. These calculations will be compared with the results found in chapter 3, related to the testing of the SIRAH 1.0 force field, as well as with previously published atomistic studies. The final set of parameters are summarised in chapter 6, table 6.1 for the side-chain and backbone parameters, and table 6.2 for the WT4-FB model.

Chapter

5

Optimisation of the SIRAH Force Field: charged side-chains

5.1. Hydration free energies of charged systems.

The calculation of hydration free energies for charged systems is a more complex process compared to uncharged systems. The standard raw hydration free energy ($\Delta G^{\ominus}_{hyd}$) for an ion is calculated as the sum of three processes: charging (ΔG_{chg}), cavitation (ΔG_{cav}) and a standard state correction term ($\Delta G^{\ominus}_{std}$ [14, 178, 179] (see below at the end of section 5.2)), as:

$$\Delta G^{\ominus}_{hyd} = \Delta G_{chg} + \Delta G_{cav} + \Delta G^{\ominus}_{std} \tag{5.1}$$

These calculations are especially sensitive to the chosen simulation methodology, and different corrections have been introduced to alleviate this effect [14, 178, 179]. These ion solvation free energies present different sources of errors, that come from approximations in the charging component, and that are considered with respect to an ideal system (i.e. an ion in a macroscopic and non-periodic solution). With this, and until recently, accurate calculations of ion solvation free energies were not possible due to the approximated treatment of electrostatics and boundary conditions, which differ from an ideal macroscopic and non-periodic system with explicit treatment of electrostatic interactions using Coulomb's Law (see below points A to D, and more details in section 5.2) [14, 178]. Accurate *ex post* correction terms can be included to solve the problems mentioned, and have been shown by different publications from Kastenholz, Reif and Hunenberger [14, 15, 178–180], to lead to methodological independence in the calculation of ion solvation free energies. The following corrections are applied (and is recommended that the reader refers to section 1.2.2 for abbreviations of the electrostatic schemes mentioned in this chapter):

A Approximate representation of the electrostatic interactions (non-Coulombic) which lead to a deviation of the solvent polarization around the ion relative to an idealized Coulombic system, with also incomplete interactions of the ion with the solvent beyond the cut-off. This type A correction is specific for the electrostatic scheme used. This type of correction does not apply for lattice-summation schemes (PME), which are Coulombic in the limit of infinite system sizes, but it does apply for cut-off truncation (CT) or reaction field schemes (BM) (please refer to section 1.2.2 for the abbreviations of the corresponding electrostatic schemes). The type A correction is specific for the electrostatic potential used, and is evaluated using the same potential, but in the idealized context of a macroscopic and non-periodic system. Moreover, it can be sub-divided into corrections A_1 and A_2 for CT schemes, which apply beyond the cut-off sphere of the ion and within it, respectively.

B Approximation of the size of the systems (finite), which do not follow a macroscopic regime. This leads to deviations on the solvent polarization, relative to the polarisation of an ideal system (macroscopic). A clear example is the use of a computational box simulated under periodic boundary conditions. This type B correction is applied for the specific electrostatic scheme in the simulation (e.g., LS, CT or BM scheme). Although this correction sounds similar to type-A, the main difference is the source of the error (approximation of electrostatic schemes vs. approximations of the size of the system).

C Deviation of the solvent generated electrostatic potential at the ion site relative to a "correct" electrostatic potential, which is a consequence of the use of an inappropriate summation scheme for the calculation of electrostatic interactions (i.e. P scheme, which stands for summing over individual charges, and a M scheme, which stands for summing over whole solvent molecules). This type C correction is applied for a specific electrostatic scheme and choice of boundary conditions, and can be subdivided in type C_1 and C_2 corrections (see below for details).

D Approximate force-field representations, especially related to the wrong dielectric constant for the solvent model used.

Based on the above description, these correction are written as ΔG_A^{cor}, ΔG_B^{cor}, ΔG_C^{cor} and ΔG_D^{cor}. A graphic representation of the corrections applied are shown in figure 5.1. As a brief explanation, and since most of the corrections cannot be calculated in the context of atomistic simulations, continuum-electrostatic (CE) calculations are used for their estimation (blue boxes in figure 5.1), which account for the estimation of the error using simulation conditions in the CE calculations (right blue boxes in figure 5.1) and the estimation of the proper quantity under full macroscopic and Coulombic conditions (left blue boxes in figure 5.1). The difference between these two calculations represents the deviation in the potential, and the value of the correction to be used.

5.2. Correction terms

5.2.1. Type A corrections

This type of correction accounts for the misinterpretation of the solvent polarisation around the ion, and the lack of interactions between the ion with the polarised solvent [21]. The main reason for this is the use of a non-Coulombic electrostatic scheme for the whole range of interactions between the atoms in the simulation. This type of correction can be evaluated in CE calculations, which are numerical and represent an idealised non-periodic and infinite system, where the electrostatic interaction are represented by Coulomb's Law. Then, these calculations can be compared to the calculations made using a cut-off truncation scheme [181, 182]. The corresponding perturba-

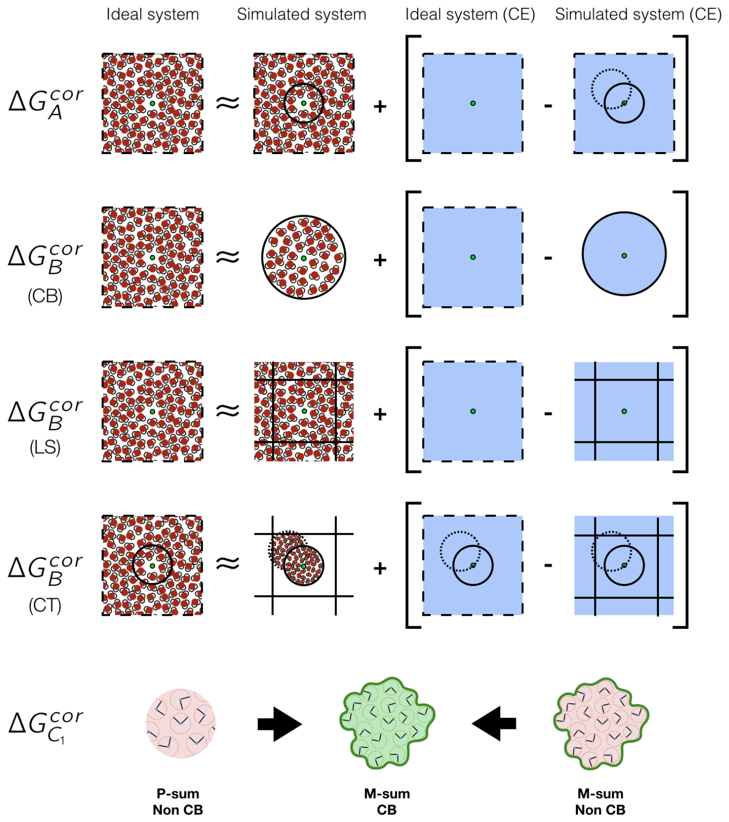


Figure 5.1

Correction terms for raw ionic solvation free energies. Schematic representation of the nature of each correction applied to the calculation of ionic solvation free energies. The purpose is to approximate the perturbation induced by each error by the corresponding perturbation within an idealized model using continuum-electrostatic calculations (CE) (presented here as an ion surrounded by a homogeneous dielectric medium in blue). The illustrated corrections are Type A, Type B (for each specific electrostatic scheme: Coulombic (CB), Lattice-summation (LS) and cut-off truncation (CT), and type C_1)**.**

tion can be added to the calculated raw hydration free energies as a correction term ΔG_A^{cor}, where "cor" corresponds to the implemented electrostatic potential (LS, BM, CM, CT, etc., please refer to section 1.2.2).

For a CB electrostatic scheme (Coulombic in the whole range of interactions in the simulation, see Fig. 1.1), there is no type-A correction,

$$\Delta G_A^{CB} = 0 \tag{5.2}$$

In the case of a LS electrostatic scheme, that is formally non-Coulombic for a finite box size, the LS interaction between two charges separated by a finite box size, becomes Coulombic in the limit of an infinite box size [15, 21, 178–180],

$$\Delta G_{A}^{LS} = 0 \tag{5.3}$$

In the case of a CT type scheme, a type-A1 correction will include all the neglected interactions that exist beyond the cut-off sphere, that are calculated on an idealised systems (CE),

$$\Delta G_{A_1}^{CT} = -(8\pi\epsilon_0)^{-1} N_A q_i^2 (1 - \epsilon_s'^{-1}) R_C^{-1} \tag{5.4}$$

where ϵ_0 and ϵ_s' correspond to the vacuum permittivity and the model solvent permittivity, respectively, N_A is Avogadro's number, q_i is the ion charge and R_C is the cut-off radius.

A second type-A correction (A_2) has to be applied for the error in the polarisation within the cut-off sphere. This correction can be obtained by getting the contribution to the charging free energy of the solvent within the cut-off sphere (interactions within the cut-off distance), and comparing it to the idealised non-periodic and infinite system with electrostatics interaction computed by Coulomb's Law (Born model) [21]. An empirical (parameterised) equation [178] can be use to describe the type-A_2 correction, which is only applied to CT electrostatic schemes (i.e. SC, BM. See section 1.2.2 for more details), and is zero for all the others, and are given as,

$$\begin{aligned}\Delta G_{A_2}^{\mathrm{SC}} = {} & (8\pi\epsilon_o)^{-1}\, N_A q_I^2 R_C^{-1} 10^{-1} a_1 \\ & \times \left[1 - a_2 \epsilon_S'^{-1} + 10 a_3 \left(1 + a_4 \epsilon_S' + 10 a_5 \epsilon_S'^2\right)^{-1}\right] \\ & \times \left[1 - a_6 \left(R_C^{-1} R_I\right)^2\right]\end{aligned} \tag{5.5}$$

and,

$$
\begin{aligned}
\Delta G^{\mathrm{BW}}_{A_2} = (8\pi\epsilon_o)^{-1} N_A q_I^2 R_C^{-1} \Big\{ & -10^{-1} b_1 \Big[1 - b_2 \epsilon'^{-1}_S - b_3 \epsilon'^{-2}_S \\
& + b_4 \left(\epsilon'^{2}_S \left(1 + 10^{-1} b_5 \epsilon'_S + 10^{-2} b_6 \epsilon'^{2}_S \right) \right)^{-1} \Big] \\
& + b_7 \left(R_C^{-1} R_I \right)^3 \left(1 - b_8 \epsilon'^{-1}_S + b_9 \epsilon'^{-2}_S + 10^{-1} b_{10} \epsilon'_S \right) \\
& \left(\epsilon'_S + b_{11} \right)^{-1} \Big\}
\end{aligned}
\tag{5.6}
$$

The dimension-less optimised fitting coefficients, a_i, with i = 1, ..., 6, and b_j, with j = 1, ..., 11, are listed on the work of Reif et al., 2011, table 1 [178], and were fitted, by trial and error, with a data set of 59800 ΔG_{A_2} values, with an ion of charge 1e, cut-off between 0.8 and 2 nm, using an ion radius ranging from 0.1 to 0.8 nm, and ϵ'_s ranging from 2 to 100. The process was evaluated by successively fitting the dependence of ΔG_{A_2} on R_I (for a given R_C and ϵ'_s), fitting the dependence of the resulting coefficients on R_C (for a given ϵ'_s), and fitting the dependence of the resulting coefficients on ϵ'_s [178].

Type-A corrections depend on: 1) the type of CT electrostatic scheme used, 2) the cut-off distance, 3) the ion radius, 4) the permittivity of the solvent model used, and 5) the ion charge.

5.2.2. Type B corrections

Type-B corrections account for the error that arises from the use of a finite size system in the simulations, and the use of periodic boundary conditions for the specific electrostatic scheme used. As for type-A corrections, this correction is evaluated in CE calculations, and represents an idealised non-periodic and infinite system, where the electrostatic interactions are represented by Coulomb's Law.

In the case of a cut-off truncation scheme, the electrostatic interactions are non-Coulombic because they are range limited. One has to compare the outcome of a charging free energy calculation with CT electrostatics within an infinite non-periodic system to an analogous calculation involving CT electrostatics in the context of a finite box size [21] (see Fig. 5.1). The difference between these two charging free energies provides an estimate for the perturbation in the atomistic system,

$$\Delta_s G_B^{CT} = \Delta_s G_B^{(CT,NPBC)} - \Delta_s G_B^{(CT,PBC)} \tag{5.7}$$

where $\Delta_s G_B^{(CT,NPBC)}$ and $\Delta_s G_B^{(CT,PBC)}$ stand for the free energy in a non-periodic and periodic system with a CT scheme, respectively.

A type-B correction, in the case of a system with fixed boundary conditions and Coulomb interactions (i.e. a spherical droplet), is calculated as:

$$\Delta_s G_B^{CB} = -\left(8\pi\epsilon_o\right)^{-1} N_A q_I^2 \left(1 - \epsilon_S'^{-1}\right) S^{-1} \tag{5.8}$$

with S being the droplet radius. For a system with PBC and non-Coulombic interactions such as a LS scheme, a type-B correction is calculated as:

$$\begin{aligned}\Delta_s G_B^{LS} = &\left(8\pi\epsilon_o\right)^{-1} N_A q_I^2 \left(1 - \epsilon_S'^{-1}\right) L^{-1} \\ &\times \left[\alpha_{LS} + \frac{4\pi}{3}\left(\frac{R_I}{L}\right)^2 - \frac{16\pi^2}{45}\left(\frac{R_I}{L}\right)^5\right]\end{aligned} \tag{5.9}$$

where $\alpha_{LS} = -2.837297$, and corresponds to the LS self-term constant [14,21,178]. This correction has to be applied to the raw charging free energy excluding the contribution of the LS self energy term, which accounts for interaction of a site i within the system box with its own periodic copies (but not with itself), and with the homogeneous neutralising background charge density [21].

Type-B corrections depend on: 1) the type of electrostatic scheme and boundary conditions, 2) the cut-off distance, 3) the ion radius, 4) the droplet radius, 5) the box edge length (for LS and CT schemes), 6) the permittivity of the solvent model used, and 7) the ion charge.

5.2.3. Type C corrections

Type-C corrections are a consequence of the use of an inappropriate summation scheme for the calculation of electrostatic interactions (i.e. P scheme, which stands for summing over individual charges, and a M scheme, which stands for summing over whole solvent molecules) (Fig. 5.1 and 5.2). Both electrostatic potentials will converge to specific values for large cut-off radii, but a difference between both is maintained, known as the exclusion potential ($\xi_{AT or CG}$, which accounts for a AT or CG model) (see section 5.3.1 for details on the calculation of this term). The proper summation scheme to use would involve the use of the potential generated by entire solvent molecules (M-scheme), as has been stated previously in references [14, 15, 21, 178, 179].

Type-C corrections can be further divided in type C$_1$, which stand for the error in the evaluation of the electrostatic potential at the ion site, and C$_2$, which accounts for the possible presence of a constant offset in the potential at the cavity centre. Type-C$_2$ corrections are small in the case of LS schemes, and they are usually neglected [15, 179]. Type-C$_2$ have also been neglected in this work, since most of the calculations have been done using a LS scheme (PME), and will not be covered in detail.

Type-C$_1$ corrections can be calculated as follow. For a system with fixed boundary conditions with Coulombic interactions, whole solvent molecules are used in the calculation, then:

$$\Delta_s G_{C_1}^{CB} = 0 \tag{5.10}$$

In the case of a system with PBC and a LS-scheme, type-C$_1$ are calculated as:

$$\Delta_s G_{C_1}^{LS} = -N_A q_I \left(1 - \frac{4\pi R_I^3}{3L^3}\right) \xi'_S \tag{5.11}$$

Here, the exclusion potential (ξ'_S, which can stand for an atomistic ξ_{AT}, or coarse-grained model ξ_{CG}) comes in play. In the case of monatomic ions and solvent models with a single Lennard-Jones site (e.g. TIP3P model) [15], this value can be calculated based on the quadrupole-moment trace of the solvent, as:

$$\xi'_S = (6\epsilon_o)^{-1} N_A M_S^{-1} \rho'_S \overline{\mathcal{Q}}'_S \tag{5.12}$$

where M_S is the molar mass of the solvent model used in the simulations. $\overline{\mathcal{Q}}'_S = \mathcal{Q}'_S$, which is the quadrupole-moment trace of the solvent model relative to its molecular centre M,

$$\overline{\mathcal{Q}}'_S = \sum_{i}^{N_s} q_i r_i^2 \tag{5.13}$$

where q_i and r_i are the partial charges and distances from M associated to the N_s point charges within the solvent molecule [21].

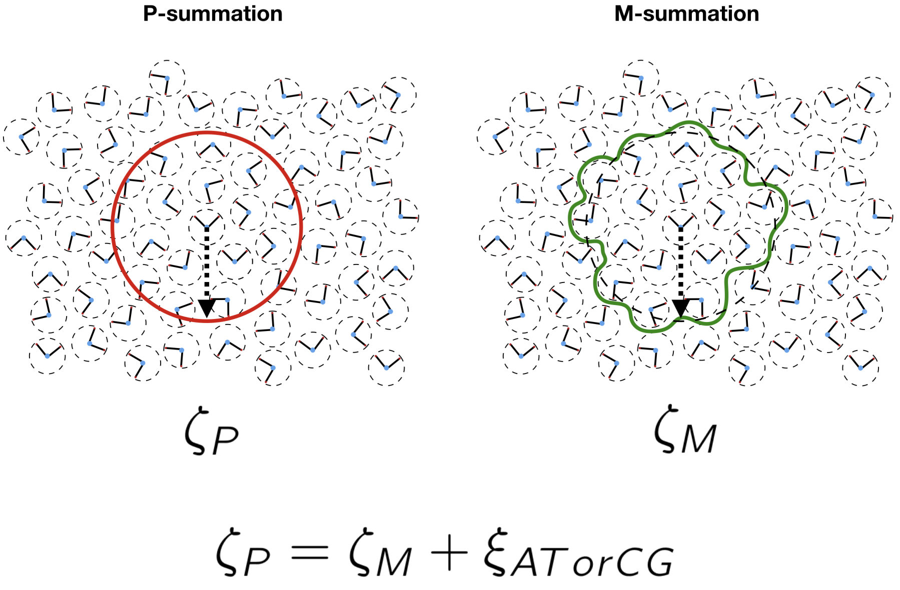


Figure 5.2

Summation schemes. Schematic representation of the types of summation implemented in MD simulations. P-summation stands for the sum of the Coulombic potential from the solvent based on atoms within a cut-off, up-to a distance R (left panel). M-summation stands for the sum of the Coulombic potential from the solvent based on whole molecules within a cut-off (usually based on a molecular centre, such as the oxygen in water), up-to a distance R (right panel). The electrostatic potential of both summations converges for large radii R, but they differ by a specific amount known as the exclusion potential ($\xi_{AT or CG}$).

Type-C corrections depend on: 1) the type of electrostatic scheme and boundary condition, 2) the cut-off distance, 3) the ion radius, 4) the droplet radius, 5) the box edge length, 6) the solvent model used (via the exclusion potential and the quadrupole moment trace), and 7) the ion charge.

5.2.4. Type-D corrections

Last but not least, type-D corrections are needed for the possible inaccuracy of the relative dielectric permittivity of the solvent model used in the simulations (ϵ'), versus the corresponding experimental dielectric permittivity (ϵ_s). This correction does not depend on the electrostatic scheme employed, being similar for all of them. It is easily calculated by comparing the results of the Born equation applied with the two permittivities, as,

$$\Delta_s G_D = (8\pi\epsilon_o)^{-1} N_A q_I^2 \left(\epsilon_S^{-1} - \epsilon_S'^{-1}\right) R_I^{-1} \tag{5.14}$$

This correction is usually small, and it tends to be calculated at the same time as corrections A and B (see below). Type-D corrections depend on the 1) the solvent model used, 2) the ion radius and 3) the ion charge.

5.2.5. Corrections for polyatomic ions

Most of the previously mentioned corrections and expressions, have been derived for monoatomic ions [14, 178]. In the case of polyatomic ions [179], numerical solutions of the Poisson equation are needed to obtain an estimation of the charging free energy in an idealised system that obeys a macroscopic regime (non-periodic with Coulombic electrostatic interactions) and based on the experimental solvent permittivity (ΔG_{chg}^{NPBC}). Simulations of periodic systems with a specific electrostatic scheme and based on the model solvent permittivity are also needed ($\Delta G_{chg}^{PBC,ES*}$ for a periodic boundary condition system using a specific electrostatic scheme (ES*)). The sum of corrections A, B and D can be obtained based on these two continuum-electrostatic simulations, as:

$$\Delta G_{A+B+D}^{ES*} = \Delta G_{chg}^{NPBC} - \Delta G_{chg}^{PBC,ES*} \tag{5.15}$$

where $ES*$ corresponds to the corresponding electrostatic scheme (i.e. LS or BM). The two terms on the right side of equation 5.1 are charging free energies obtained with the Poisson equation solver from references [183–185], for non-periodic and periodic systems with Coulombic electrostatic interactions.

Finally, and summarising all the necessary methodology-dependent corrections, standard hydration free energies can be calculated as:

$$\Delta G^{\ominus}_{hyd} = (\Delta G^{raw}_{chg} + \Delta G_{cav}) + \Delta G_{A+B+D} + \Delta G_{C_1} + \Delta G^{\ominus}_{std} \tag{5.16}$$

Results obtained from free energy simulations correspond to the transfer of an ion from a fixed point in the gas phase to a fixed point in bulk water (or vice-versa), which is not standard. Conversion of this non-standard hydration free energy ($\Delta G^{\odot}_{\mathrm{hyd}}$) to a standard quantity that characterises transfer from a random location within the molar volume accessible to the ion in the ideal gas at pressure P and temperature T, to a random location within the molar volume accessible to the aqueous ion in an ideal solution at pressure P, molality b°, and temperature T ($\Delta G^{\ominus}_{\mathrm{hyd}}$) [21, 178, 179], is given as,

$$\Delta G^{\ominus}_{\mathrm{hyd}} = \Delta G^{\odot}_{\mathrm{hyd}} + \Delta G^{\ominus}_{std} \tag{5.17}$$

where $\Delta G^{\ominus}_{std} = RT \ln \left[(P)^{-1} RT b^{\circ} \rho^{\ominus}_{\mathrm{w}} \right]$, with $\rho^{\ominus}_{\mathrm{w}}$ being the density of water and R the ideal-gas constant. At a density $\rho^{\ominus}_{\mathrm{w}} = 997 kg/m^3$ and P of 1 atm, the standard state term is equal to 7.95 kJ/mol [21, 178, 179].

In this chapter we have attempted to mix our proposed optimisation method (section 4.3.2) with the necessary corrections for charged ions. We show again that our method works, but with the consequence of parameters with unphysical values. PMFs were computed for charged side-chain pairs and a manual optimisation of the parameters was made to fit to atomistic results.

5.3. Methods

5.3.1. Hydration free energies of charged side-chains

The calculation of hydration free energies for charged systems is a more complex process compared to the classical use for uncharged systems. The standard hydration free energy ($\Delta G^{\ominus}_{\mathrm{hyd}}$) for an ion is calculated as the sum of three processes: charging (ΔG_{chg}), cavitation (ΔG_{cav}) and a standard convention term ($\Delta G^{\ominus}_{\mathrm{std}}$, which is equal to 7.95 kJ/mol, considering a water density of 997 kg/m^3 at a pressure of 1 atm) [14, 178, 179], as:

$$\Delta G^{\ominus}_{\mathrm{hyd}} = \Delta G_{\mathrm{chg}} + \Delta G_{\mathrm{cav}} + \Delta G^{\ominus}_{\mathrm{std}} \tag{5.18}$$

The calculation of raw charging free energies is especially sensitive to the chosen simulation methodology [15, 21, 178–180] where different corrections have been introduces to alleviate these effects (see section 5.2).

Following these corrections [15, 21, 178–180], raw hydration free energies and the energy components of these corrections (see section 5.2) were used to calculate the corrected values for the different charged side-chains, as:

$$\Delta G_{\mathrm{chg}} = \Delta G^{\mathrm{raw}}_{\mathrm{chg}} + \Delta G_{\mathrm{cor}} \tag{5.19}$$

The raw charging free energies (equation 5.1) have been calculated using a lattice-summation scheme (PME) by decoupling the interactions (electrostatic and van der Waals together) of the ion (side-chains) with the solvent (excluding intramolecular interactions). Eleven lambda values have been used (0.0, 0.1, . . . , 0.9, 1.0) for all the charged side-chains, using the GROMOS 54A8 [179, 186] for atomistic systems, and the original SIRAH 1.0 [5], the updated SIRAH 2.0 force field [11], and SIRAH-OBAFE force fields in GROMACS v.2018.2. The choice of the atomistic force field was based on the study of Reif et al. [179], where a parameter optimisation of the GROMOS force field was performed to closely reproduce hydration free energies of charged side-chain analogues, including the necessary corrections mentioned in section 5.2. This is similar to the choices made in chapter 4, section 4.3.3, where we choose force fields that closely reproduce hydration free energies for un-charged side-chains and the protein backbone. The simulation conditions and soft-core potential settings were identical to the ones used in the calculation of hydration free energies for uncharged side-chains (Stage A from section 4.3.3). All the reported raw free energies exclude the self-interaction energy contribution, which is related to the interactions of the ion with its periodic copies.

The sum of corrections A, B and D can be obtained based on two continuum-electrostatic simulations that correspond to idealised systems given by Coulomb's Law, and non-periodic of infinite size, and the system with the approximated electrostatic scheme, with periodic boundary conditions and finite size, as stated in equation 5.15. In this work, the sum of corrections A, B and D was obtained using an LS scheme, and can be calculated (based on equation 5.15) as,

$$\Delta \mathrm{G}^{\mathrm{LS}}_{\mathrm{A+B+D}} = \Delta G^{\mathrm{NPBC}}_{\mathrm{chg}} - \Delta G^{\mathrm{PBC,LS}}_{\mathrm{chg}} \tag{5.20}$$

In this work, a relative permittivity of 78.4 for water has been used in the calculation of ΔG^{NPBC}_{chg}. A relative permittivity of 63.84 for the optimised WT4-FB water model was used, as calculated based on reference [15], in the calculations of $\Delta G^{PBC,LS}_{chg}$. Briefly, simulations of pure water were performed in a NVT ensemble at 298.15 K, and to a density of approximately 997 kg/m^3. The simulations were run with a BM scheme (reaction-field, molecule-based cut-off), with a reaction-field permittivity ϵ_{RF} = 60. The permittivity ϵ'_S was calculated as,

$$(\epsilon'_S - 1)\left(\frac{2\epsilon_{RF} + 1}{2\epsilon_{RF} + \epsilon'_S}\right) = \frac{\langle \mathbf{M}^2 \rangle - \langle \mathbf{M} \rangle^2}{3\epsilon_o L^3 k_B T} \tag{5.21}$$

where **M** corresponds to the total dipole moment, T is the temperature, k_B is the Boltzman's constant, ϵ_0 is the dielectric permittivity of vacuum and angle brackets correspond to ensemble averages. The dielectric permittivity calculated here differs with the value previously reported in chapter 4 (with a value of 74.2 in table 4.3 using a LS scheme), but is within the error. Moreover, it has been reported that dielectric permittivities calculated using a reaction-field scheme are more sensitive to the choice of simulation parameters such as the non-bonded cut-off [187]. Given this, the lack of agreement is not unexpected, but as a matter of consistency with previous studies [15], we decided to use the dielectric permittivity calculated in this section for the evaluation of the exclusion potential. Moreover, the dielectric permittivity for the WT4 water model calculated in this section is similar to the one calculated by Reif et.al [15] with a reported value of 66.7 using the SPC model.

Continuum-electrostatic calculations were performed with the GROMOS++ pre-MD and analysis software v.1.4.1 [188] and were based on a single structure taken as the final configuration of the hydration free energy simulations of the charged side-chains. The appropriate boundary conditions and electrostatic scheme were used for each case, with a grid spacing of 0.02 nm and a threshold of 10^{-6} kJ/mol for the convergence of the electrostatic free energy, based on reference [179]. The box lengths used in the simulations were 3.22 nm for Glu, 3.25 nm for Lys and 3.45 nm for Arg, and the number of WT4 water molecules involved in the simulations were 99 for Glu, 111 for Lys and 125 for Arg.

A type C1 correction is required for LS and BM (reaction field) schemes, and corrects the P-summation (atom-based cut-off) implied by these schemes to a proper M-summation (molecule-based cut-off). For a LS scheme, this is given by equation 5.11. In there, ξ'_S corresponds to the exclusion potential of the solvent model used (in this case, the WT4-FB model). For fully rigid models with a single van der Waals interaction site, this last term has been usually calculated based on the quadrupole moment trace of the solvent model [14, 21, 178]. For more complex solvent models, different methods have been derived for the calculations of their exclusion potentials [15]. In this work, we have employed method IV from the work of Reif and Hunenberger, 2016 [15], which relies on the comparison of the raw potentials within a cavity using two different electrostatic schemes, assuming that the corrected potentials are equal. For this, we have used a cut-off truncation (CM) and reaction field schemes (BM). The difference in the raw potentials are related to ξ'_S as:

$$\xi'_S = -\left[\frac{2\left(\epsilon'_S - 1\right)}{2\epsilon'_S + 1}\left(1 - \frac{R_I^3}{R_C^3}\right)\right]^{-1}\left(\Phi^{*,raw,CM} - \Phi^{*,raw,BM}\right) \tag{5.22}$$

where R_I is the effective ionic radius, R_C is the cut-off, $\phi^{*,raw,CM}$ and $\phi^{*,raw,BM}$ are the raw electrostatic potentials within an uncharged cavity of the size of a CG sodium ion, and ϵ'_S corresponds to the dielectric permittivity of the solvent model, which has been calculated as previously mentioned. Simulations of an uncharged sodium ion (similar as the work of Reif and Hunenberger, 2016 [15]) solvated in the optimised WT4-FB model were run for 1 ns using a BM scheme, with a reaction field permittivity ϵ_{RF} equal to 80. Electrostatic potentials at the cavity were obtained for both CM and BM schemes based on the electrostatic interaction of the cavity with the solvent within the cut-off R_C of 1.4 nm, using a Python script created for this purpose. Simulation settings were identical to the previous ones used in this work.

Type C2 corrections correct for the presence of an interfacial potential at the ion surface. This term is proportional to the ratio of the ionic volume to the box volume. With this, its magnitude is very small for the systems used in this work, and has been neglected in the calculation of the corrected hydration free energies.

5.3.2. Optimisation of charged side-chains.

Optimisation for the charged side-chains were performed in a similar fashion as the case for uncharged side-chain and the protein backbone (section 4.3.2), where free energy gradients from atomistic simulation were used as optimisation target for the CG parameters (see equation 4.27). As this type of free energy calculations are methodology dependent, the inclusion of corrections must be performed. Assuming that the final free energies between the atomistic and coarse-grained models must be equal, the sum between their free energy gradients and the necessary correction gradients must be equal as well. Since the corrections are added *ex post*, the fitting data used is given as,

$$\frac{\partial \left\langle \Delta G^{\mathrm{raw}}_{\mathrm{chg}} \right\rangle_{\alpha,\mathrm{AT}}}{\partial \lambda} + \frac{\partial \left\langle \Delta G_{\mathrm{cor}} \right\rangle_{\alpha,\mathrm{AT}}}{\partial \lambda} = \frac{\partial \left\langle \Delta G^{\mathrm{raw}}_{\mathrm{chg}} \right\rangle_{\alpha,\mathrm{CG}}}{\partial \lambda} + \frac{\partial \left\langle \Delta G_{\mathrm{cor}} \right\rangle_{\alpha,\mathrm{CG}}}{\partial \lambda} \tag{5.23}$$

and moving the property that we want to optimise to one side,

$$\frac{\partial \left\langle \Delta G^{\mathrm{raw}}_{\mathrm{chg}} \right\rangle_{\alpha,\mathrm{CG}}}{\partial \lambda} = \frac{\partial \left\langle \Delta G^{\mathrm{raw}}_{\mathrm{chg}} \right\rangle_{\alpha,\mathrm{AT}}}{\partial \lambda} + \frac{\partial \left\langle \Delta G_{\mathrm{cor}} \right\rangle_{\alpha,\mathrm{AT}}}{\partial \lambda} - \frac{\partial \left\langle \Delta G_{\mathrm{cor}} \right\rangle_{\alpha,\mathrm{CG}}}{\partial \lambda} \tag{5.24}$$

where $\frac{\partial \left\langle \Delta G^{\mathrm{raw}}_{\mathrm{chg}} \right\rangle_{\alpha,\mathrm{AT}}}{\partial \lambda}$ and $\frac{\partial \left\langle \Delta G^{\mathrm{raw}}_{\mathrm{chg}} \right\rangle_{\alpha,\mathrm{CG}}}{\partial \lambda}$ correspond to the derivative of the raw hydration free energy gradients of the charged side-chains with respect to the force field parameters (at a specific α value), for an atomistic and coarse-grained system, respectively. $\frac{\partial \left\langle \Delta G_{\mathrm{cor}} \right\rangle_{\alpha,\mathrm{AT}}}{\partial \lambda}$ and $\frac{\partial \left\langle \Delta G_{\mathrm{cor}} \right\rangle_{\alpha,\mathrm{CG}}}{\partial \lambda}$ are the derivatives of the free energy corrections with respect to the force field parameters (at a specific α value), for an atomistic and coarse-grained system, respectively. The derivatives of the corrections were calculated using central finite differences, using a set of α values between 0.4 and 1.0, where the parameters were scaled accordingly (i.e. for α=0.9, parameters were scaled to 90% of their original value).

All the optimisation simulations for the SIRAH beads were run with the optimized WT4-FB model (this work). 100 optimisation cycle iterations were run for only charge and van der Waals ϵ values. Systems were minimised for 5000 steps using a steepest descent algorithm followed by an NPT equilibration time of 5 ns. Production runs were performed for 10 ns. A leap-frog algorithm was used for integration of Newton's equations of motion with a time-step of 20 fs. Electrostatic interactions are calculated using the Particle Mesh Ewald method [28] with a direct cut-off of 1.2 nm and a grid spacing of 0.2 nm. A 1.2 nm cutoff was used for van der Waals interactions. The V-rescale thermostat [125] and the Parrinello-Rahman barostat [124] were used to maintain the temperature at 298.15 K and the pressure at 1 atm, respectively. The simulation conditions were consistent with the original SIRAH publication [5]. All simulations were run with GROMACS v. 2018.2 [123]. All specific heteronuclear non-bonded Lennard Jones parameters, previously set in the original SIRAH force field between charged beads have been removed, and we have set those interactions using Lorentz-Berthelot combining rules, which provide the non-bonded interaction energy between two particles i and j, given by $\sigma_{ij} = \frac{\sigma_{ii}+\sigma_{jj}}{2}$ and $\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}$

In order to continue with the optimisation of charged side-chains, and to possibly improve the optimised parameters with unphysical values, PMFs for charged side-chain pairs were calculated, and parameters were manually changed to fit to atomistic PMFs of the same side-chain pairs. For the PMF calculations, the distance between the centre of mass of the BCG bead of Glu$^-$, and the BCE and BCZ bead of Lys$^+$ and Arg$^+$, respectively, were used as collective variables (see Fig. 2.6 for bead names). A total of 78 windows have been used, with distance spanning from 0.38 nm to 1.8 nm, using a spring force constant of 5000 kJ/mol nm^2. Simulations settings were identical to the previous ones used: a leap-frog stochastic dynamics integrator was used for integration of Newton's equations of motion with a time-step of 20 fs. Electrostatic interactions were calculated using the PME procedure with a grid spacing of 0.2 nm. Non-bonded interactions were modelled using the classical Lennard-Jones potential and a Coulombic energy function, with a cut-off of 1.2 nm each. All simulations were run with GROMACS v. 2018.2 at 1 atm with the Parrinello-Rahman barostat and at 298.15 K with the v-rescale thermostat, and were preceded by the corresponding minimisation and NPT equilibration.

5.4. Results

5.4.1. Optimisation of the SIRAH protein force field: charged side-chains

A different approach, compared to the optimisation of uncharged side-chains and backbone (chapter 4), was followed for the charged side-chains. We started with Force-

Balance optimisation procedures as explained in the methodology section 5.3.2, where the gradients of the raw charged free energies plus the gradient of the methodology-dependent corrections were used. Most of the ForceBalance optimisation results yield good agreement with experimental and AT hydration free energies (table 5.2, denoted as HFE-fitted), but the parameters were driven to unphysical values (see Fig. 5.3 and 5.4, for charge and Lennard Jones parameters, respectively), with a clear over-estimation of the side-chain interactions (Fig. 5.5 and 5.6, as HFE-fitted). Given this, we conclude that our optimisation procedure works, but given the existence of few parameters to represent charged side-chain in SIRAH, over-fitting might be an unavoidable consequence in this case. Moreover, coarse-graining is an important simplification of the physics, where the option to fully reproduce complex properties, such as the free energy or charged entities, might not be possible.

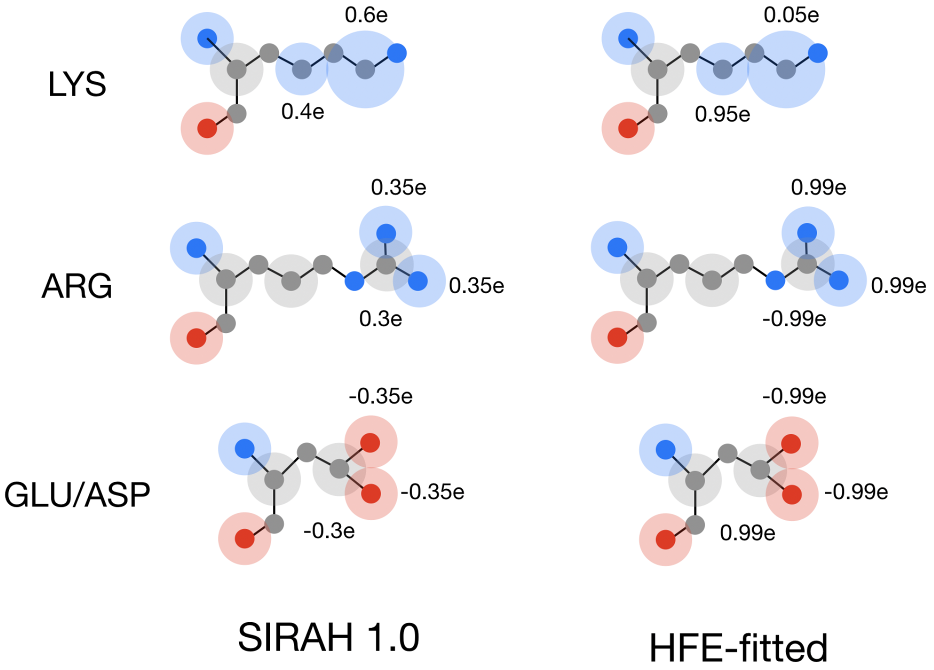


Figure 5.3

HFE-fitted charge values for the optimised charged side-chain. Schematic representation of the three optimised charged side-chains (Lys, Arg and Glu/Asp). Parameter values for the charges are shown for the corresponding beads (see Fig. 2.6 for details), for the original SIRAH 1.0 and the HFE-fitted parameter set (after the initial optimisation)**.**

Given the size of the charges found in the initial optimisation process (Fig. 5.3), higher

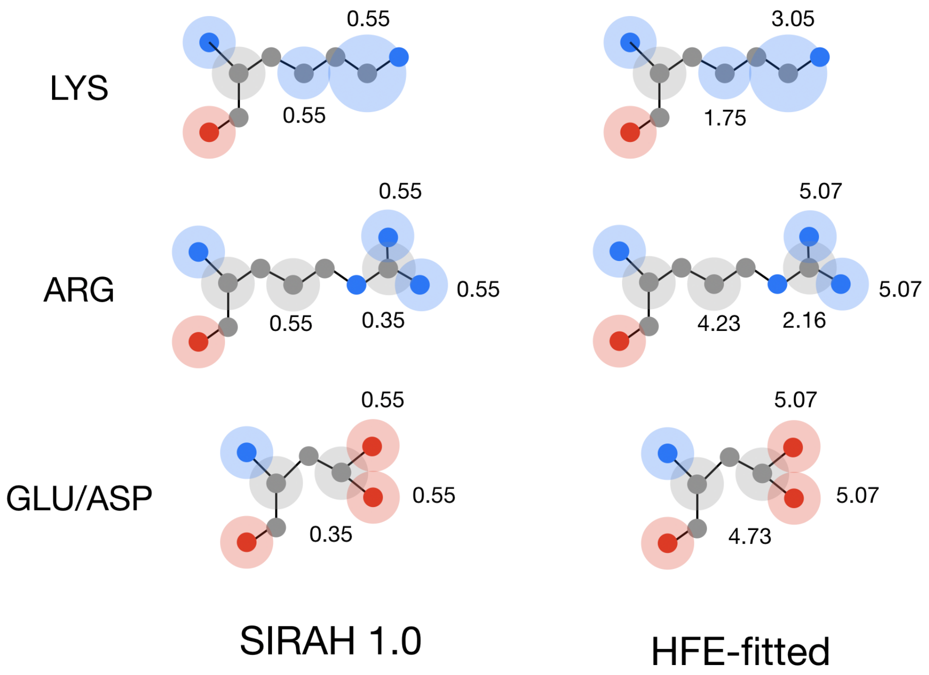


Figure 5.4

HFE-fitted VDWϵ values for the optimised charged side-chain. Schematic representation of the three optimised charged side-chains (Lys, Arg and Glu/Asp). VDWϵ parameters are shown for the corresponding beads (see Fig. 2.6 for details), for the original SIRAH 1.0 and the HFE-fitted parameter set (after the initial optimisation). All values are in units of kJ/mol**.**

values for Lennard-Jones ϵ parameters were found as well (Fig. 5.4). We decided to follow the optimisation process with a manual refinement of the parameters to reproduce atomistic side-chain PMFs rather than solely hydration free energies, which could lead to a reasonable representation of protein-protein interactions within the model, and the balance of charge-charge and charge-solvent interactions [150]. For this, different sets of parameters were manually generated, where only four of these sets are shown to demonstrate the gradual improvement in the PMF plots (Fig. 5.5 and 5.6, as set 1 to 4). As a general observation, it seems that the scaling of Lennard Jones ϵ parameters improves the well-depth prediction on the PMFs for different side-chain pairs (see below, and Fig. 5.5 and 5.6). Detailed explanations for this follow. Details related to the parameters corresponding to each set are shown in table 5.1.

In the case of Glu^{-}/Arg^{+} a well-depth between -4.0 kcal/mol and -5.3 kcal/mol has been reported for atomistic simulation with explicit water using the TIP3P model [189]

	SIRAH bead name	VDWσ (nm)	VDWϵ (kJ/mol)	Charge (e)
		Set 1		
Lys	C1Ck	SaO	1.75	0.95
	C7Nk	SaO	3.05	0.05
Arg	C2Cr	SaO	4.23	-0.99
	C3Cr	SaO	2.16	0.99
	C5N	SaO	5.07	0.99
Glu/Asp	C4Ce/d	SaO	2.16	0.99
	C6O	SaO	5.07	-0.99
		Set 2		
Lys	C1Ck	SaO	1.75	0.95
	C7Nk	SaO	2.05	0.05
Arg	C2Cr	SaO	4.23	-0.99
	C3Cr	SaO	2.16	0.99
	C5N	SaO	4.07	0.99
Glu/Asp	C4Ce/d	SaO	2.16	0.99
	C6O	SaO	4.07	-0.99
		Set 3		
Lys	C1Ck	SaO	1.75	0.35
	C7Nk	SaO	1.75	0.65
Arg	C2Cr	SaO	2.23	-0.99
	C3Cr	SaO	2.16	0.99
	C5N	SaO	3.07	0.99
Glu/Asp	C4Ce/d	SaO	2.16	0.99
	C6O	SaO	2.07	-0.99
		Set 4		
Lys	C1Ck	SaO	1.75	0.35
	C7Nk	SaO	1.75	0.65
Arg	C2Cr	SaO	2.23	-0.99
	C3Cr	SaO	2.16	0.99
	C5N	SaO	2.07	0.99
Glu/Asp	C4Ce/d	SaO	2.16	0.99
	C6O	SaO	2.07	-0.99

Table 5.1

Parameter sets used in the PMF fitting procedure. Parameters are shown for all the charged side-chains used in the optimisation of the SIRAH force field. Bead types were taken from the original SIRAH 1.0 publication. Parameters directly taken from the original SIRAH 1.0 force field are mentioned as SaO (Same as Original, see section 4.4.2 and Fig. 4.7 for the rationalisation of this)**.**

and for implicit solvent simulations [150]. The SIRAH force field implements the use of specific non-bonded pair interaction between the charged beads, which are unable to represent a correct well-depth in the PMF plots (Fig. 5.5 as SIRAH 1.0). Just by removing these specific pair interactions, and rather using Lorentz-Berthelot combining-rules, an improvement is observed with a well-depth of -2.3 kcal/mol (Fig. 5.5, shown as w/ Comb. rules, which stands for "with Combining rules"). The HFE-fitted results obtained in the initial optimisation, with the parameters shown in figures 5.3 and 5.4, yield highly attractive behaviour, with a well-depth of -16.94 kcal/mol (Fig 5.5). Scaling set 4 gives the optimum result (table 5.1 for details), with a well-depth of -5.3 kcal/mol at around 0.59 nm (Fig. 5.5).

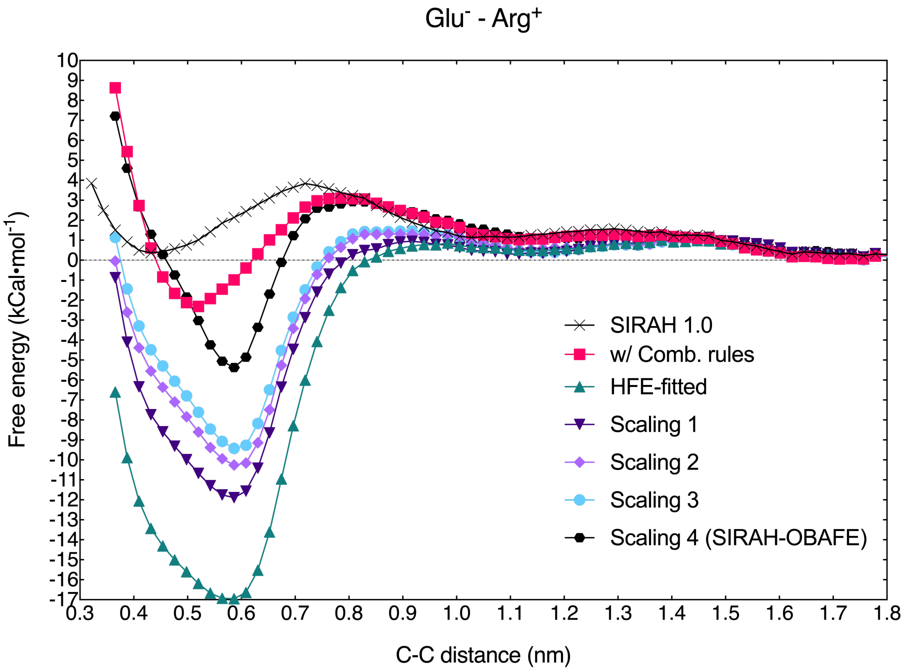
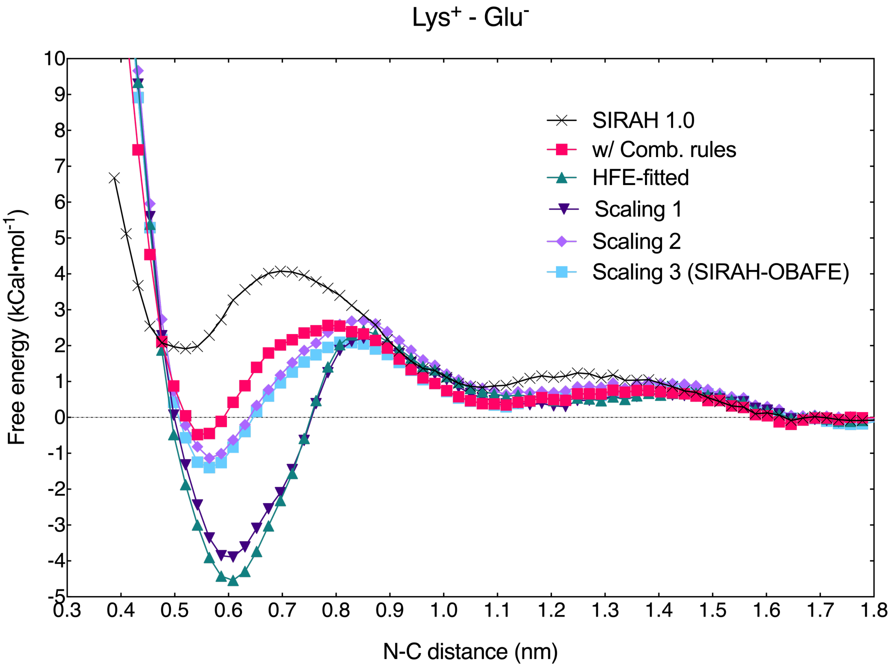


Figure 5.5

Potentials of mean force for charged side-chain pairs. PMF plot for the Glu-/Arg+ pair is shown, using the original SIRAH 1.0 force field and for the set of optimised parameters, as well as the SIRAH-OBAFE force field, and the HFE-fitted (i.e. the parameters set optimised with ForceBalance)**.**

For the case of Lys$^+$/Glu$^-$, a well-depth between -2.2 kcal/mol and -2.4 kcal/mol has been reported in the context of atomistic simulation with explicit water using the TIP3P model [189] and for implicit solvent simulations [150]. Again, the use of specific non-bonded pairs for the charged side-chains in SIRAH 1.0 fails to reproduce this (Fig. 5.6). The use of Lorentz-Berthelot combining rules again improves the well-depth (Fig. 5.6). The HFE-fitted result is highly attractive with a well-depth of -4.5 kcal/mol, but the scaling of non-bonded interactions rapidly solves this, with a well-depth now of -1.5 kcal/mol for set 3 (table 5.1 for details), at a distance of 0.56 nm (Fig. 5.6). Future force field optimisation protocols would benefit from the inclusion of both PMF and free energy data together, to avoid the over-fitting behaviour observed in these cases.



Potentials of mean force for charged side-chain pairs. PMF plot for the Lys+/Glu- pair is shown, using the original SIRAH 1.0 force field and for the set of optimised parameters, as well as the SIRAH-OBAFE force field, and the HFE-fitted (i.e. the parameters set optimised with ForceBalance)**.**

Figure 5.6

The re-calculation of standard state hydration free energies for the charged side-chains using the new set of parameters, refined using the PMF calculations, was performed. Table 5.2 summarises the results, where not much improvement is seen with respect to the previously calculated hydration free energies using the SIRAH 1.0 force field. In terms of the predicted hydration free energies, by the PMF-optimised parameters, Arg^+ and Glu^- shows a decrease of around 30 kJ/mol in both cases, compared to the original SIRAH 1.0 force field, and the updated SIRAH 2.0 force field (table 5.2).

Even though the PMF fitting shows itself as a promising route for the optimisation of the charged side-chains, some preliminary tests of protein simulations showed an overestimation of protein interactions, probably from higher charge values on the charge side-chain beads (see Fig. 5.3, which are of the order of 1e), which translates to a higher level of stiffness in the simulated systems. Figure 5.7 shows simulations of some proteins in water (with simulation protocols similar to the ones mentioned in section 3.2.3 and section 3.2.5) using the SIRAH-OBAFE force field after PMF scaling. Even though the average RMSD values are lower than the ones observed using the SIRAH 1.0 force field (Fig. 5.7A and 5.7B), and that can be considered as an improvement on the stability of the systems, the overall variation and flexibility of the proteins systems must play an important role in their dynamics, such as in conformational change processes. Adding to this, the calculation of the free energy landscape for the ligand binding domain of the S1S2 gluatamate receptor (identical protocol to that mentioned in section 3.2.5), showed quantitative overestimation of the interactions (Fig. 5.7C), seen as a single minimum focused on the closed conformation of this domain, where a preference for open conformations has been previously established [126, 127].

As a matter of discussion, we have obtained PMFs for 3 uncharged side-chain pairs, and 2 charged-polar side-chain pairs, using the parameters derived on chapter 4 (SIRAH-OBAFE) for the uncharged side-chains (Fig. 5.8), the original parameters for the charged side-chains (which are the ones that we have decided to keep, see below) and the WT4-FB water model. These PMFs have also been compared with PMF plots of the same side-chain pairs using the original SIRAH 1.0 force field and the original WT4 water model.

In the case of hydrophobic side-chain models, such as Tyr/Val and Leu/Val, a PMF minima with a depth of around -1 kcal/mol has been observed in atomistic simulations [66]. In the case of the Tyr/Val pair combination, SIRAH-OBAFE predicts a minimum with a depth of around 2.5 kcal/mol, while the SIRAH 1.0 force field predicts a minimum with

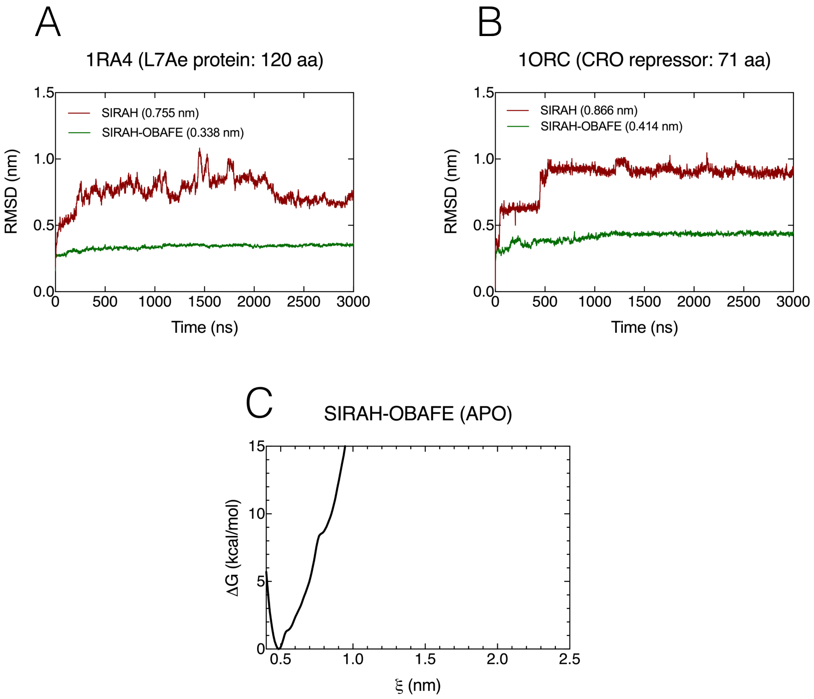


Figure 5.7 **Overestimation of the interactions in the SIRAH-OBAFE force field after PMF scaling.** RMSD timeseries for two protein systems: A) L7Ae Archeal ribosomal protein and B) the CRO repressor. C) PMF plot of the cleft opening/closing process of the S1S2 glutamate receptor ligand binding domain in its ligand free form (apo). Protocols as previously discussed in sections 3.2.3 and 3.2.5**.**

a depth of 1.8 kcal/mol (Fig. 5.8A). For the Leu/Val combination, SIRAH-OBAFE predicts a minimum of depth 1.5 kcal/mol, while SIRAH 1.0 predicts a minimum of around 1 kcal/mol (Fig. 5.8B).

In the case of polar-polar side-chain combinations, such as Ser/Asn, a minimum with depth between -0.5 kcal/mol and -1.0 kcal/mol has been predicted from atomistic simulations [66]. In the case of SIRAH-OBAFE a minimum is found with a depth around 2.8 kcal/mol, while SIRAH 1.0 predicts a minimum of around 2 kcal/mol (Fig. 5.8C).

In the case of atomistic PMF plots for charged side-chains and polar amino-acids, for positively charged side-chains such as the Arg$^+$/Asn, a minimum of depth between -1.0 kcal/mol to -2.0 kcal/mol has been predicted, while a minimum of depth between -0.5 kcal/mol to -1.0 kcal/mol has been observed for negative pairs such as Glu$^-$/Asn. In our case, the pair Glu^{-}/Asn presents a minima of depth around 1.5 kcal/mol using the SIRAH-OBAFE force field, while a minimum of depth of 2.0 kcal/mol is predicted using the SIRAH 1.0 force field (Fig. 5.8D). In the case of the Arg^{+}/Asn combination, a minimum with depth of 1.8 kcal/mol is observed using the SIRAH-OBAFE force field. while a minimum of the same value is found using the SIRAH 1.0 force field (Fig. 5.8E).

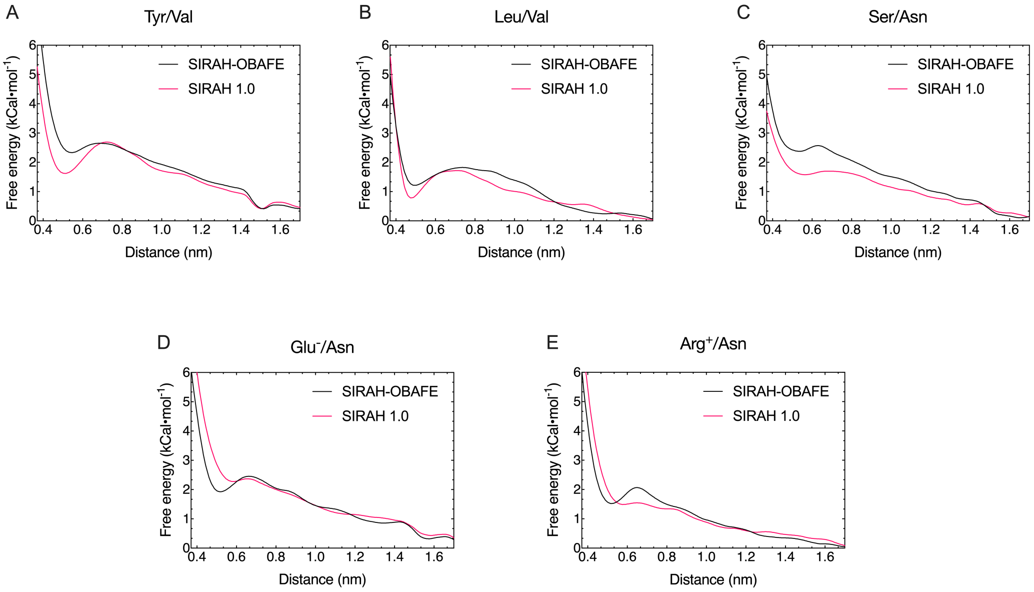
It is clear that important discrepancies are presented, compared with atomistic simulations, for the PMF of uncharged side-chains, as well as the combination of charged-polar side-chains (Fig. 5.8), where most of the PMFs show a repulsion effect for the tested pairs. Even though the parameters for the uncharged side-chains were sufficient to predict correct hydration free energies using the SIRAH-OBAFE (see Fig. 4.8 and table 4.4), they are not capable of reproducing PMF plots, as was the case of the charged-charged combinations tested in this chapter (see Fig. 5.5 and 5.6). We believe that the simplification of the physics observed in coarse-grained force fields, such as the SIRAH model, present a challenge for the reproduction of multiple properties. The few parameters available in SIRAH are insufficient to fit our proposed target for the case of charged side-chains (i.e. hydration free energy gradients from a higher resolution model), where they become unphysical in terms of the size of the charge and Lennard-Jones values, as well as estimate the interactions between the charged side-chains.

We have decided to continue with the original SIRAH 1.0 parameters for charged side-chains, in combination with the optimised parameters for backbone and uncharged side-chains. The scaled parameters presented an improvement in the estimation of the PMFs for charged-charged side-chain pairs. Although, the interactions in the proteins systems were still over-estimated, as was observed in figure 5.7, with really stiff RMSD values and the incorrect prediction of a protein PMF.

Force field	**Expt.**	$\Delta G_{chg}^{raw} + \Delta G_{cav}$	ΔG_{A+B+D}	ΔG_{C_1}	$\Delta G^{\ominus}_{std}$	$\Delta G^{\ominus}_{hyd}$
			ARG			
54A8	-276.50	-137.89 ± 0.4	-58.63	-67.88	7.95	-256.46
SIRAH 1.0		-149.01 ± 0.6	-57.50	-7.25	7.95	-205.82
SIRAH 2.0		-149.85 ± 0.6	-57.50	-7.25	7.95	-206.65
SIRAH-OBAFE		-186.64 ± 0.5	-57.50	-7.25	7.95	-243.45
HFE-fitted		-223.70 ± 0.5	-54.81	-7.25	7.95	-277.81
			LYS			
54A8	-289.50	-180.08 ± 0.6	-58.83	-67.88	7.95	-298.85
SIRAH 1.0		-134.49 ± 0.6	-54.70	-7.49	7.95	-188.74
SIRAH 2.0		-130.90 ± 0.5	-57.50	-7.25	7.95	-185.15
SIRAH-OBAFE		-132.34 ± 0.7	-57.50	-7.49	7.95	-186.58
HFE-fitted		-178.60 ± 0.6	-57.50	-7.49	7.95	-235.84
			GLU			
54A8	-315.4	-349.84 ± 0.5	-58.98	67.88	7.95	-332.99
SIRAH 1.0		-156.38 ± 0.4	-58.83	7.52	7.95	-199.25
SIRAH 2.0		-153.62 ± 0.6	-57.50	-7.25	7.95	-196.49
SIRAH-OBAFE		-181.77 ± 0.6	-58.83	7.52	7.95	-224.64
HFE-fitted		-252.75 ± 0.6	-58.34	7.52	7.95	-295.62
			ASP			
54A8	-321.2	-349.38 ± 0.5	-58.98	67.88	7.95	-332.54
SIRAH 1.0		-156.38 ± 0.4	-58.83	7.52	7.95	-199.25
SIRAH 2.0		-153.62 ± 0.6	-57.50	-7.25	7.95	-196.49
SIRAH-OBAFE		-181.77 ± 0.6	-58.83	7.52	7.95	-224.64
HFE-fitted		-252.75 ± 0.6	-58.34	7.52	7.95	-295.62

Table 5.2

Hydration free energies for charged side-chains. Hydration free energy values of all the charged side-chains optimised in this work, using the GROMOS 54A8, SIRAH and SIRAH-OBAFE force fields. HFE-fitted values are also included with the sole intention of comparison and discussion. All values are in the units of kJ/mol. Error bars were modelled as standard errors across 3 repeat simulations**.**



PMF for uncharged and charged side-chain pair combinations. PMF plots for a set of 3 uncharged side-chains pair combinations, and 2 charged-polar side-chains combinations, are shown. (A) Tyr/Val, (B) Leu/Val, (C) Ser/Asn, (D) Glu$^-$/Asn, and (E) Arg$^+$/Asn**.**

Figure 5.8

5.5. Summary

In this chapter we have summarised the process to obtain hydration free energies of charged molecules, and we have outlined the necessary corrections for the calculation of these energies in charged protein side-chains. An optimisation procedure has been followed, using gradients for the corresponding hydration free energies, as well as the gradient of their corrections. The results of this chapter have shown that:

- Our optimisation method works for the improvement of charged hydration free energies using a combination of free energy gradients, and the gradient of the corrections with respect to the α values.

- Even though an important improvement has been made in the overall charged hydration free energies for the SIRAH CG force field, compared to both experimental and simulation data, an overestimation of the protein interaction has been observed. This was represented by higher values in charge and Lennard-Jones parameters, which directly influenced in the over-stabilisation of protein simulations in water and in the calculation of a PMF for a protein conformational change.

- As a second route, the fitting of the PMFs for charged-charged side-chain pairs was achieved. The overall size of the charge parameters in the protein beads was not changed, but scaled parameters for VDWϵ were enough to improve the estimation of the side-chain PMFs. However, the overestimation of the interactions was still present as it was shown in the calculation of protein RMSDs and a protein PMF in figure 5.7.
- The result of this overestimations must come from the use of a property as complicated as the hydration free energy gradients, where the use of just a few parameters can lead to over-fitted models. Moreover, the WT4 model is too coarse, where subtleties of the PMFs cannot be screened properly. The simplifications in coarse-grained models take out details that are important for the physics of the tested systems, and the full estimation of multiple properties such as hydration free energies and PMFs for side-chain pairs, might not be possible, as was shown by the results in this chapter related to charged molecules, and for the outlined uncharged PMFs.

The next chapter summarises the optimised parameter set for the SIRAH-OBAFE force field. Different tests have been performed: evaluation of protein stability, simulations of a protien complex, and the stability of a peptide. We finalise with the recalculation for the PMFs shown in chapter 3.

Chapter

6

Testing of the optimised force field: SIRAH-OBAFE

6.1. Introduction

In chapters 4 and 5, we have shown the application of a new optimisation method where the hydration free energy of a lower resolution model can be improved by fitting of the free energy gradients of higher resolution models, as in this case for a CG model based on atomistic data. Our implementation of this method into ForceBalance enables full automation of the complex optimisation procedure and the incorporation of flexible choices of target data, which can be mixed with other experimental or modelled data available, such as QM energies.

The full list of optimised parameters for the new SIRAH-OBAFE force field (which stands for **O**ptimised **B**ased on **A**tomistic **F**ree **E**nergies) and the optimised WT4-FB water model are shown in tables 6.1 and 6.2. All the changes were made for the un-charged side-chain and the backbone. As described in chapter 5, charged side-chains remained as in the original SIRAH 1.0 force field. Bead names can be seen in figure 2.6.

	Bead type	**VDWσ (nm)**	**VDWϵ (kJ mol^{-1})**	**Charge (e)**
Asn/Gln	P3Cn/q	SaO	3.5217E-01	0.00
	P5N	SaO	5.5453E-01	5.9527E-01
	P4O	SaO	5.547E-01	-5.9527E-01
Cys	P1S	SaO	1.0547	-6.0817E-01
	P2P	SaO	2.2622E-01	6.0817E-01
His (epsilon protonated)	A2C	SaO	SaO	0.00
	A5E	SaO	1.7084	5.0449E-01
	A5D	SaO	1.7023	-5.0449E-01
His (delta protonated)	A2C	SaO	SaO	0.00
	A5E	SaO	1.7084	-5.0449E-01
	A5D	SaO	1.7023	5.0449E-01
Met	Y3Sm	SaO	4.7181	0.00
Phe	A2C	SaO	SaO	0.00
	A1C	SaO	SaO	0.00
Ser/Thr	P1O	SaO	4.4658E-01	-9.1874E-01
	P2P	SaO	2.2622E-01	9.1874E-01
Trp	A2C	SaO	SaO	0.00
	A7N	SaO	6.9916E-01	-3.5323E-01
	A8P	SaO	1.5469	3.5323E-01
	A1Cw	SaO	3.1449	0.00
Tyr	A2C	SaO	SaO	0.00
	A4O	SaO	2.0418	-3.5107E-01
	A3P	SaO	2.0491	3.5107E-01
Val/Leu/Ile	Y4Cv/Y1C	SaO	5.0887E-01	0.00
Backbone	GC	SaO	5.5058E-01	4.2176E-01
	GO	SaO	5.2511E-01	-6.7336E-01
	GN	SaO	5.5058E-01	2.5161E-01
Arg	C2Cr	SaO	SaO	SaO
	C3Cr	SaO	SaO	SaO
	C5N	SaO	SaO	SaO
Lys	C1Ck	SaO	SaO	SaO
	C7Nk	SaO	SaO	SaO
Asp/Glu	C4Ce/d	SaO	SaO	SaO
	C6O	SaO	SaO	SaO

Table 6.1

The optimised SIRAH-OBAFE protein force field. Optimised parameters for the SIRAH-OBAFE protein force field, obtained in chapter 4. Bead types were taken from the original SIRAH 1.0 publication [5] (Fig. 2.6). Parameters taken from the SIRAH 1.0 protein force field publication are mentioned as SaO (Same as Original) [5]**.**

Bead type	VDWσ (nm)	VDWϵ (kJ mol^{-1})	Charge (e)
WN1	4.2474E-01	7.6717E-01	-2.6730E-01
WN2	4.2474E-01	7.6717E-01	-5.6223E-01
WP1	4.2474E-01	7.6717E-01	2.6730E-01
WP2	4.2474E-01	7.6717E-01	5.6223E-01

Table 6.2

The optimised WT4-FB force field. Optimised parameters for the WT4-FB water model, obtained in chapter 4, are shown. Bead types were taken from the original SIRAH 1.0 publication [5]..

In this chapter we summarise a series of test for the new optimised SIRAH-OBAFE force field, ranging from protein stability in water, secondary structure and conformational changes.

6.2. Methods

6.2.1. Validation set of the new optimised SIRAH-OBAFE force field

A set of 6 protein systems chosen from the original SIRAH 1.0 publication [5], of sizes ranging from 585 to 63 residues, have been tested in terms of protein stabilty: (A) Serum albumin (PDB: 1E7I, 585 residues, X-ray resolution of 2.70 Å), (B) GFP protein (PDB: 1QYO, 238 residues, X-ray resolution of 1.80 Å), (C) Gamma-adaptin domain (PDB: 1GYV, 120 residues, X-ray resolution of 1.71 Å), (D) L7Ae Archeal ribosomal protein (PDB: 1RA4, 120 residues, X-ray resolution of 1.86 Å), (E) CRO repressor (PDB: 1ORC, 71 residues, X-ray resolution of 1.54 Å) and (F) the N-terminal domain of phage 434 repressor (PDB: 1R69, 63 residues, X-ray resolution of 2.0 Å). RMSD values were calculated based on the C-α carbons and compared against the crystal structures, using GROMACS v.2018.2. Analysis of secondary structure stability has been performed using the Calmodulin system, with (PDB: 3CLN) and without (PDB: 1LKJ) calcium. Proteins from the group previously analysed for protein stability were also analysed in terms of secondary structure stability, comparing the updated version SIRAH 2.0 and the optimised SIRAH-OBAFE force fields. Simulations of the small Trp-

Cage peptide (20 residues) in its folded form (PDB: 1L2Y) have been performed to further push the limits of both the original SIRAH 1.0, SIRAH 2.0 and the optimised SIRAH-OBAFE force fields, comparing the overall protein stability and secondary structure variations. Finally, the same systems presented in chapter 3, section 3.3.2.1 and 3.3.2.2, related to the calculation of the PMF for the S1S2 glutamate receptor binding domain, and the Abl kinase, were evaluated again using the optimised SIRAH-OBAFE force field (see below section 6.2.2 and 6.2.3 for simulation details).

6.2.2. Coarse-grained molecular dynamics simulations

For all the previously mentioned systems (except for the PMF calculations, see below), simulations were performed as follows. Coarse-grained molecular dynamics simulations were performed using the SIRAH1.0/WT4, SIRAH2.0/WT4 (when stated) and SIRAH-OBAFE/WT4-FB force field combinations. An energy minimisation was carried out with 10000 iterations of the steepest descent algorithm. This was followed by an NPT equilibration dynamics procedure of 20 ns with positional restraints of 1000 $kJmol^{-1}nm^{-2}$ applied to all the protein beads. Production runs were performed for 3 μs for each system with an integration time-step of 20 fs. Electrostatic interactions were calculated using the Particle Mesh Ewald procedure [28] with a direct cut-off of 1.2 nm and a grid spacing of 0.2 nm. Non-bonded interactions were modelled using the Lennard-Jones potential with a cut-off of 1.2 nm. All simulations were run at 1 bar with the Parrinello-Rahman barostat [190] and at 298.15 K with the v-rescale thermostat [191]. Simulations, RMSF and RMSD time series were calculated with GROMACS v.2018.2 [123]. Evolution of the secondary structure of our CG systems was also tested, using the *sirah-vmdtk.tcl* script supplied in SIRAH tools with the original SIRAH force field [116]. In SIRAH, the secondary structure is calculated as a function of dihedral angles along the backbone beads (see chapter 2, Fig. 2.5), which span between $-180°$ and $+180°$. From this definition, each residue is assigned to three of the corresponding categories: Helix (H), Extended (E) and Coil (C) (see reference [5], and the description of the secondary structure in section 2.6).

6.2.3. Atomistic molecular dynamics simulations

In the case of the Calmodulin systems, atomistic simulations were run for systems, with (PDB: 3CLN, X-ray resolution of 2.2 Å) and without (PDB: 1LKJ, obtained through NMR) the presence of calcium. Simulations were carried out using GROMACS v.2018.2 with the AMBER14SB force field [143]. Each system was solvated using the TIP3P water model [88] in an octahedral box with a solute-box distance of 1.5 nm in order to satisfy the minimum image convention. All systems were neutralised by addition of sodium and chloride ions to a total concentration of 150 mM. Simulation were run with calcium when mentioned, using the SIRAH model for that ion (see chapter 2 for details).

Each system was minimised using the steepest descent algorithm with 10000 iterations. This was followed with an NPT equilibration of 10 ns with positional restraints on protein heavy atoms of 1000 kJ/mol nm^2. Production runs were performed for 400 ns for each system. Electrostatics interactions were calculated using the PME procedure with a grid spacing of 0.16 nm. Non-bonded interactions were modelled using the classical Lennard-Jones potential and a Coulombic energy function, with a cut-off of 1.0 nm each. The LINCS algorithm was applied to constrain all h-bond lengths. All simulations were run at 1 bar with the Parrinello-Rahman barostat [124] and at 300 K with the v-rescale thermostat [125].

6.2.4. Umbrella sampling

As a new application of the optimised force-field, the reproduction of the opening/closing event of a glutamate receptor binding domain was attempted, as in section 3.2.4. A 1D order reaction coordinate (ξ) was chosen as the distance between the C-α carbons of residues G451 and S652 of the Glutamate receptor ligand binding domain (PDB: 1FTJ, bound form, X-ray resolution of 1.90 Å). This was done for both the free-ligand (apo, manually removing the ligand from the binding site from the structure with PDB code 1FTJ, based on the protocol of reference [126]), and the glutamate-bound states (holo). The SIRAH representation for the glutamic acid side-chain was used as ligand. The pull code in GROMACS was used to generate snapshots for the umbrella sampling simulations from a single pulling trajectory, generating 23 umbrella windows that spanned a distance range between 0.45 to 2.0 nm, with a distance between each window of 0.05 nm. A value for the biasing harmonic potential of 2500 kJ/mol nm^2 was used. Each window was simulated for 300 ns. The Weighted Histogram Analysis Method (WHAM) implemented in GROMACS (g_wham) [123] was used to remove the biasing potential and get the unbiased probabilities, to finally compute the potential of mean force (PMF). The rest of the simulation settings were set exactly as in the validation process. The results were compared with our previous results using the SIRAH 1.0 force field (chapter 3, section 3.3.2.1), and with the SIRAH 2.0 force field. Convergence of the PMF and overlap between the umbrella windows are shown in figures A.5 and A.6, for the apo and holo conformations, respectively.

6.2.5. Metadynamics simulations

A metadynamics simulation of the Abl kinase (PDB: 2G1T, X-ray resolution of 1.80 Å) was performed using the Plumed package [128] (v. 2.4.3) patched into GROMACS [123]. Simulations were run at a temperature of 310 K and for 1 μs, using the well-tempered metadynamics algorithm [50]. Two CVs were used, that comprise two dihedral angles: CV1: GN(A380)-GC(A380)-GC(D381)-BCG(D381) and CV2: GN(A380)-GC(A380)-GC(F382)-BCG(F382). Gaussians were deposited every 4 ps with a height of 2.0 kJ/mol. A bias factor of 5 was used. The width of the Gaussians was set to 0.1 rad for both dihedral CVs. Parameters for well-tempered metadynamics simulations were based on previous studies of the Abl kinase using Plumed [129], and the selection of CVs was based on a study from Roux et al., 2015 [130].

6.3. Results and Discussion

6.3.1. Protein stability

To test the performance of SIRAH-OBAFE in protein simulations, an RMSD analysis was performed on 6 protein systems of different sizes. Simulations using the optimised SIRAH-OBAFE with the optimised WT4-FB were run for 3 μs. While the computed

RMSDs are generally larger compared to atomistic simulations, all the simulations that used the optimised SIRAH-OBAFE model show improvement in protein stability with lower RMSD values throughout the whole trajectory with respect to the original SIRAH 1.0 and the updated SIRAH 2.0 (Fig. 6.1). Even though the overall behaviour of the optimised SIRAH-OBAFE FF does not yield exact results compared to atomistic RMSDs, it shows an important improvement compared to the original SIRAH FF. As a simple comparison, the three protein atomistic systems previously run for 200 ns in chapter 3, section 3.3.1, with PDB codes 1QYO (GFP-protein), 1RA4 (L7Ae Archeal ribosomal protein) and 1R69 (N-terminal domain of phage 434 repressor) (Fig. 3.2, section 3.3.1), showed average RMSD values (with respect to the crystal structure) of 0.270 nm, 0.06 nm and 0.148 nm, respectively. In this section, the original SIRAH 1.0 force field shows averaged RMSD values (with respect to the crystal structure) of 0.723 nm, 0.755 nm and 0.804 nm, for the same three systems, while our optimised SIRAH-OBAFE force field shows averaged RMSD values (with respect to the crystal structure) of 0.453 nm, 0.491 nm and 0.635 nm, for the same three cases, 1QYO, 1RA4 and 1R69, respectively. In the case of the updated SIRAH 2.0 force field, the overall behaviour of the RMSD timeseries is similar to the optimised SIRAH-OBAFE, except for two biggest systems with PDB codes 1E7I and 1QYO (Fig. 6.1A and 6.1B), with average RMSD values of 0.543 nm and 0.601 nm, respectively. Even though the new RMSD values are not close to the atomistic RMSD (presented in section 3.3.1) and we would not necessarily expect them to be, there is as an improvement in the stability of protein systems based on our new optimisation approach.

Calculations of RMSDs against the last frame of the trajectories were also performed using the updated SIRAH 2.0 and the optimised SIRAH-OBAFE force fields, allowing us to tell whether the large RMSDs are due to a big number of fluctuations or a change in conformation to a rigid conformer. Figure A.2 summarise the results. As can be seen in the case of the 1E7I system, a big change of conformation is seen at around 1μs, which stabilise afterwards. In the other systems, it seems that the higher RMSD values observed in figure 6.1 are due to different fluctuations across the simulation, with simi-

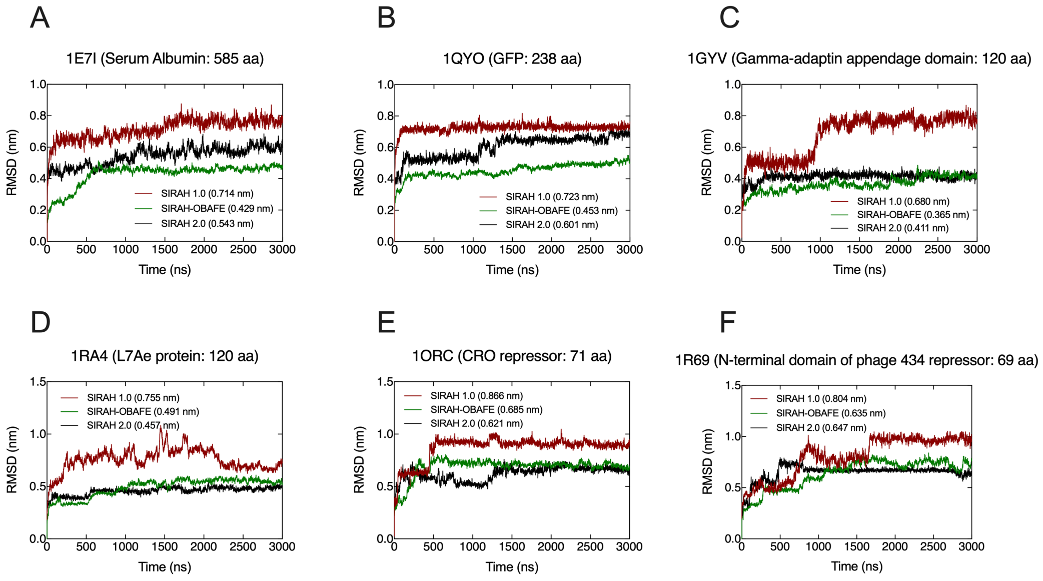


Figure 6.1

RMSD time series comparison. RMSD trajectory analysis is shown as a time series comparison with respect to the C-α carbons of the CG representation to the crystal structure for (A) Serum albumin, (B) GFP protein, (C) Gamma-adaptin domain, (D) L7Ae Archeal ribosomal protein, (E) CRO repressor and (F) the N-terminal domain of phage 434 repressor. PDB codes are shown in the figure titles and running averaged RMSD values are shown in the figure legends. Simulations were run using the SIRAH 1.0 (red), SIRAH 2.0 (black) and SIRAH-OBAFE (green) force fields**.**

lar behaviours for both the updated SIRAH 2.0 and the optimised SIRAH-OBAFE force fields.

6.3.2. Secondary structure stability

Analysis of secondary structure stability was tested in CG simulations of Calmodulin, a protein that has shown high levels of flexibility and which poses a challenge in CG simulations. Calmodulin is a calcium dependent protein composed of two EF-hand domains (Fig. 6.2), and acts as an important regulator in different biological processes, such as the cell cycle and intracellular signalling. Moreover, different studies have described Calmodulin in its apo (Ca^{2+} free) and holo (Ca^{2+} bound) forms, and the importance in the process for the exposure of different hydrophobic residues in both EF-hands for the recognition of different targets, such as target peptides derived from protein kinases or enzymes involved in metabolic processes [192, 193].

We started with the basic calculation of RMSD time series for both the apo and holo forms of Calmodulin, using SIRAH 1.0 and the optimised SIRAH-OBAFE. Previous simulations of the same Calmodulin systems have been performed in the publication of the SIRAH 2.0 force field [11]. In this section, SIRAH 1.0 shows averaged RMSD values of 1.479 nm and 1.05 nm, for the apo (PDB: 1LKJ) and holo (PDB: 3CLN) forms (Fig. 6.3A and 6.3B). The presence of calcium ions might be a possible explanation of the reduction of movement in the holo form, even though, in the case of SIRAH-OBAFE no bigger differences are observed for the apo and holo forms, with averaged RMSDs values of 1.277 nm and 1.339 nm, respectively (Fig. 6.3A and 6.3B). In the publication for the updated SIRAH 2.0 force field, the authors showed RMSD values for the Calmodulin system in its holo form (Ca^{2+}-bound) of up-to 2 nm, for simulations of 1.5 μs. Moreover, they mention structural distortions on the EF-hands, but without further details. This

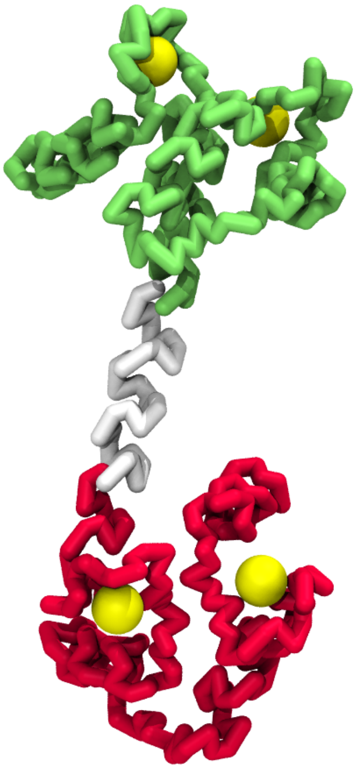


Figure 6.2

Calmodulin 3D structure. Calmodulin structure in its Ca^{2+} bound-state (Holo). The two domains are coloured as N-term (green) and C-term (red), while the central linker is coloured in white. Calcium is presented as a yellow VDW sphere bound to the EF-hand motifs, two for each domain. Structure taken from PDB code 3CLN. The N-terminal domain is coloured in green, C-terminal in red, and the linker in white**.**

can be compared with structural distortions observed in the protein structures from figure 6.4 and 6.5. In the case of Apo structure, important disruptions are observed at 1 μs using the SIRAH 1.0 force field, followed by a collapse of the EF-hands towards the central linker at 3 μs (Fig. 6.4). This important disruptions are not observed in using the SIRAH-OBAFE force fields, where only torsions of the EF-hands are observed (Fig. 6.4). On the other hand, the Holo structure are more stable using the SIRAH 1.0, with small movements towards the central linker (Fig. 6.5). In the case of the SIRAH-OBAFE simulations, rotations of the EF-hands are again observed (Fig. 6.5). Atomistic systems are less flexible, compared to the CG models, with averaged RMSDs values of 0.984 nm and 0.948 nm for the apo and holo forms, respectively (Fig. 6.3C, calculated in this work). Overall, there are no significant differences in the averaged RMSDs values between the SIRAH 1.0, SIRAH-OBAFE and atomistic systems, for both apo and holo forms. High RMSD values were expected, since different studies have shown that the linker between the two domains posses certain flexibility, probably related to the peptide recognition process [193–196], and it has been shown further destabilisation when Ca^{2+} is removed [193, 194].

We also evaluate secondary structure stability, for all cases, throughout the whole trajectory. Important things to notice are the bigger secondary structure changes observed in the apo form. In the case of the SIRAH 1.0 system, bigger changes are observed, for example after 1.4 μs where the first α-helix (residues 6-18) experiences a total disruption (Fig. 6.3D). This is also observed as an increase in the RMSD values in figure 6.3A, at around the first microsecond. In the case of the SIRAH-OBAFE system, this α-helix is partially observed, where the disruption possibly happened on the equilibration stages. The holo form is presented as a more stable system in relation to secondary structure. Still, the SIRAH 1.0 force field fails to maintain again the first α-helix, with a disruption after the first 400 ns, which is then transformed to two separate β-sheets, results that are inconsistent with any previous data. Qualitatively speaking, the SIRAH-OBAFE and the atomistic system share a “frayed” behaviour, which can be observed, for example, for the α-helices that span between residues 65-75 and 78-92

(Fig 6.3D, holo form). Moreover, two of the four small β-sheets observed in the crystal structure and noted in UNIPROT (Fig. 6.3E, β-sheets of residues 99-100 and 136-137) are much more stable in the SIRAH-OBAFE runs, similar to the atomistic systems (Fig 6.3D). Either way, both CG models do not present a perfect match compared to the atomistic system results, but SIRAH-OBAFE shows more consistencies both quantitively and qualitatively.

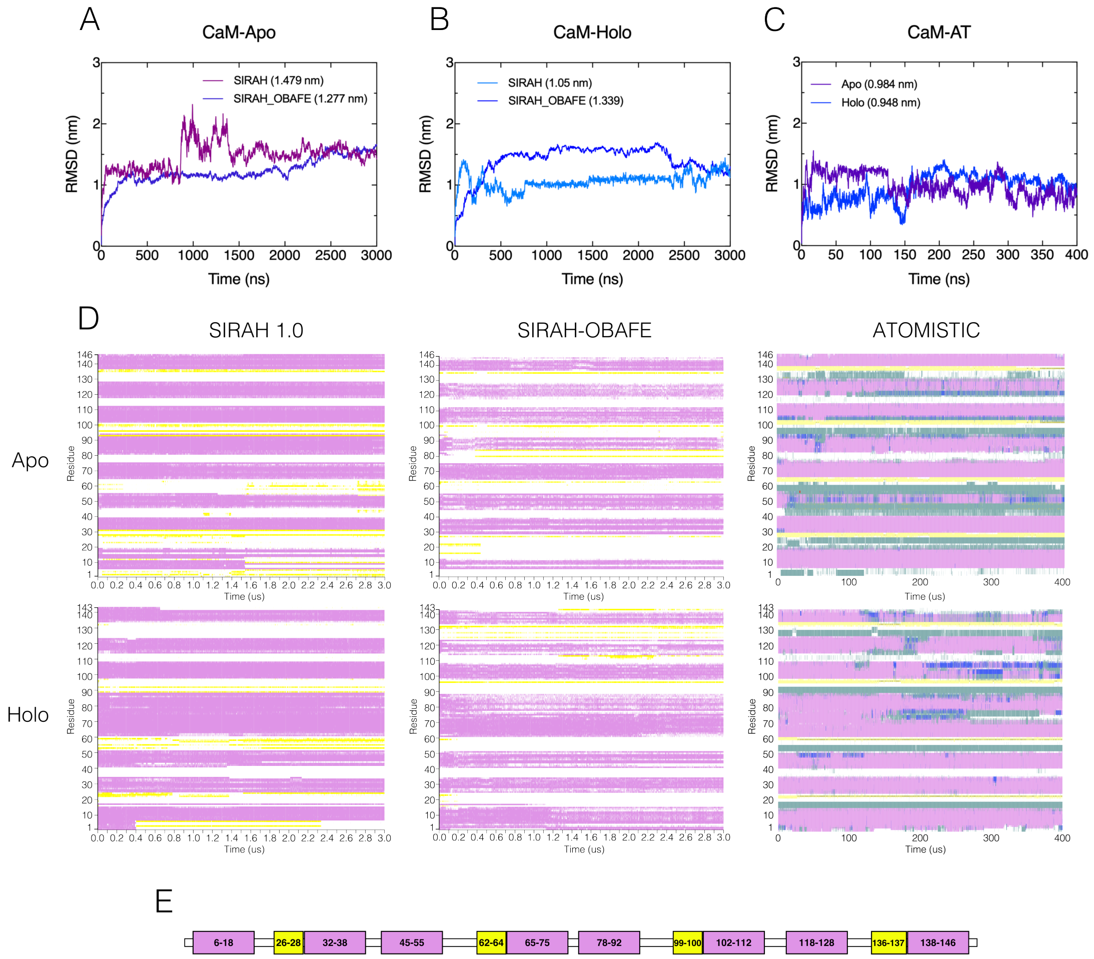
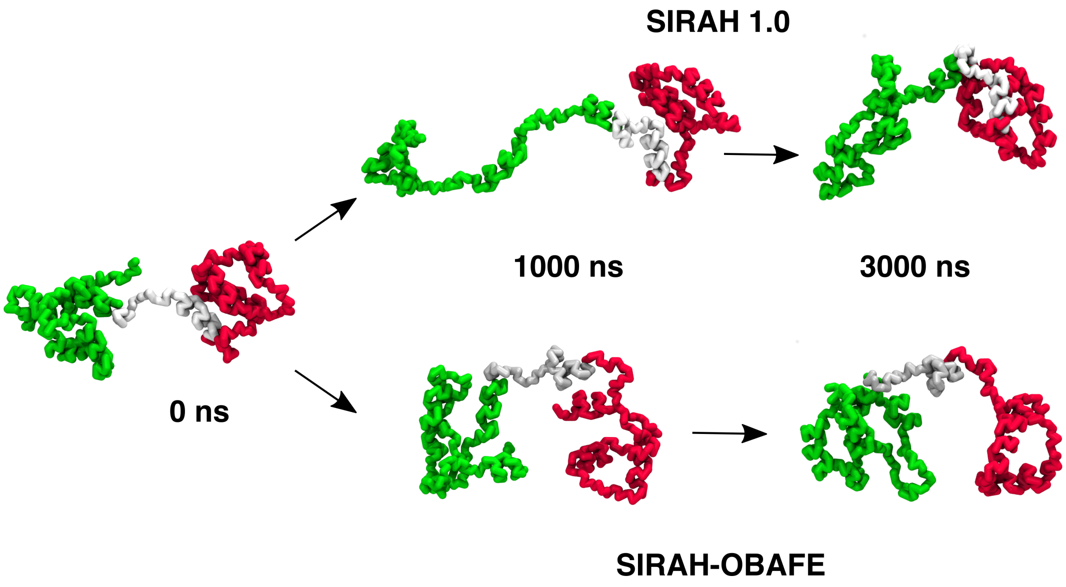


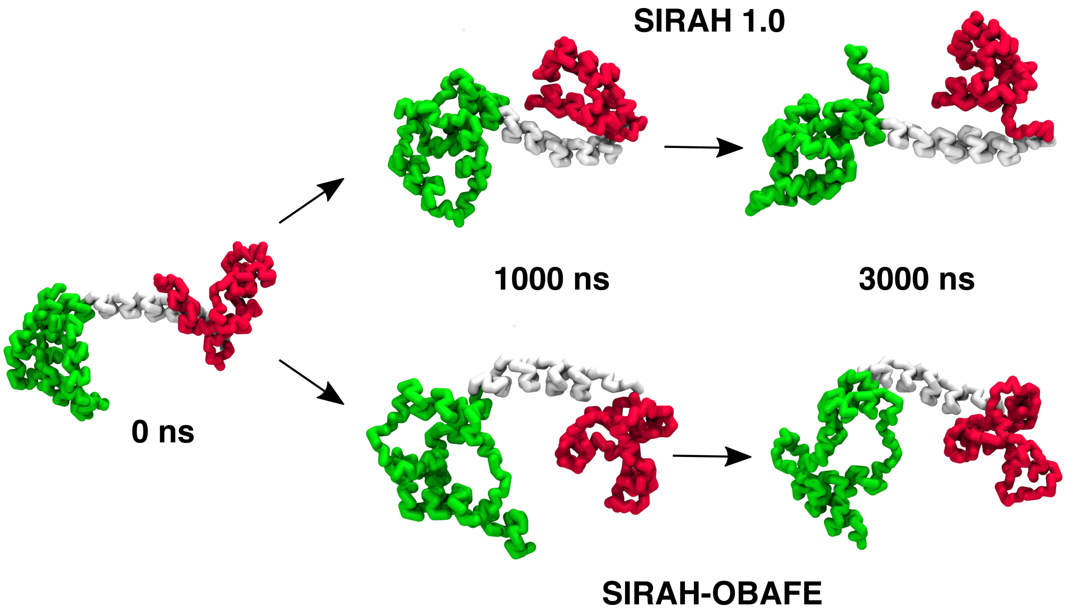
Figure 6.3

Calmodulin structural analysis. RMSD trajectory timeseries of Cα carbons for Calmodulin, in its apo and holo forms, compared between (A) the original SIRAH force field, (B) the optimised SIRAH-OBAFE force field and (C) atomistic systems. (D) Secondary structure timeseries for the apo and holo forms for the same three cases. Each colour represents a specific secondary structure: α-helix (pink), β-sheets (yellow) and coil (white). All other colours observed in the atomistic system are not relevant for this case, since they cannot be described by SIRAH. (E) Secondary structure based on the UNIPROT P0DP29 notation**.**



Calmodulin Apo conformations. Three-dimensional structures of Calmodulin in its apo conformation (PDB 1LKJ), using the SIRAH 1.0 (upper structures) and SIRAH-OBAFE (bottom structures) force fields. Structures were taken at 1 μs and 3 μs for simulations using both force fields. The N-terminal domain is coloured in green, C-terminal in red, and the linker in white.

Figure 6.4



Calmodulin Holo conformations. Three-dimensional structures of Calmodulin in its holo conformation (PDB 3CLN), using the SIRAH 1.0 (upper structures) and SIRAH-OBAFE (bottom structures) force fields. Structures were taken at 1 μs and 3 μs for simulations using both force fields. The N-terminal domain is coloured in green, C-terminal in red, and the linker in white.

Figure 6.5

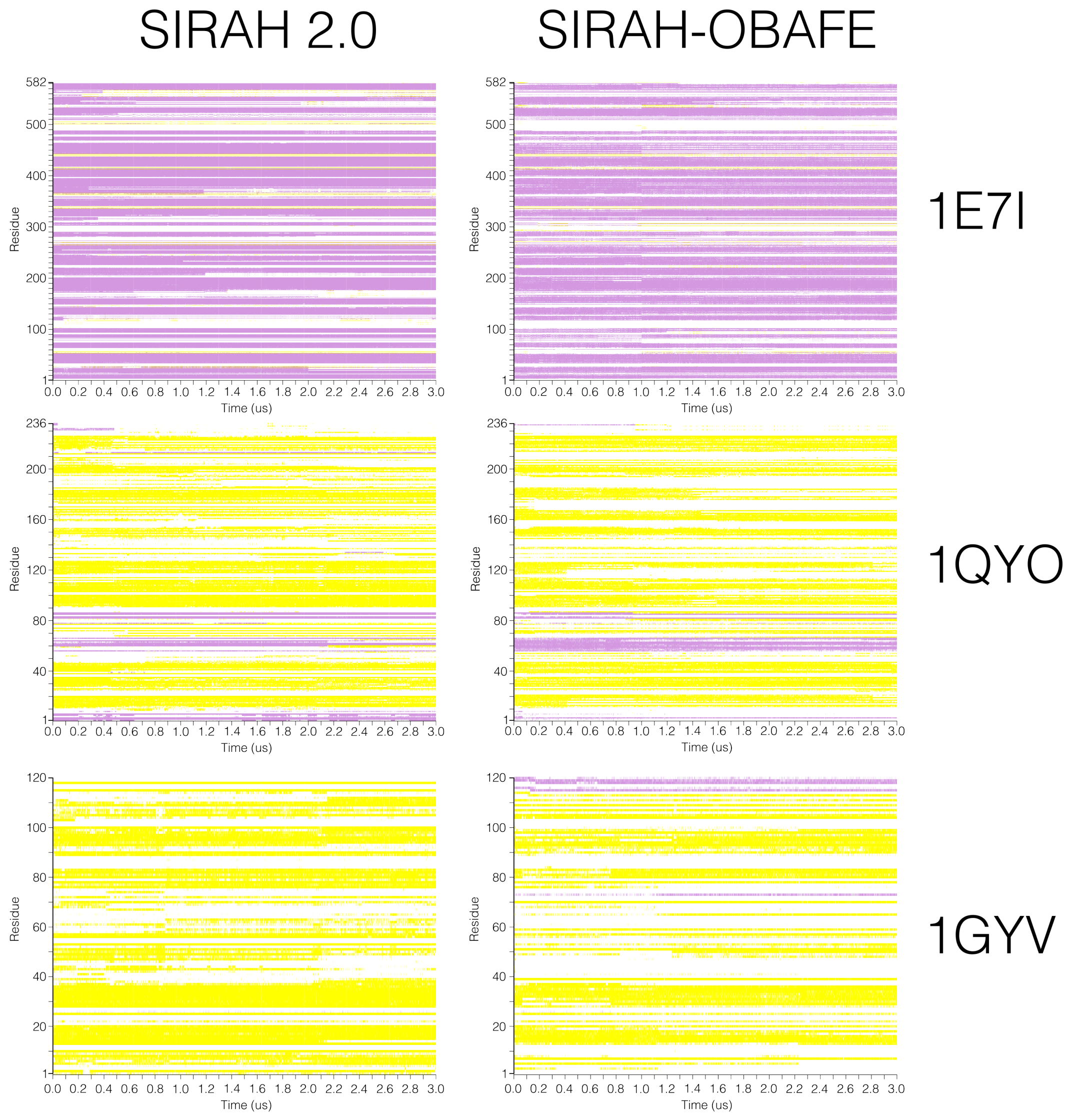
Early studies of the Calmodulin system have shown a more flexible N-terminal domain (N-CaM), compared to the C-terminal domain (C-CaM), for the holo form (Ca^{2+}-bound) [192, 195–197], where an opposite trend in the apo form is mentioned (Ca^{2+}-free) [194]. Moreover, it has been reported that the N-CaM binding may need a more flexible mechanism, with intermediate states to facilitate deep binding, compared to the C-CaM binding which may be dominated by conformational selection [193]. To study this, RMSD calculations were done for both domains (N-CaM and C-CaM), for both forms, apo and holo. Table 6.3 summarises averaged RMSD values with respect to the crystal structures. In the case of SIRAH 1.0, simulation showed a more flexible N-CaM domain, in both states (apo and holo). Differences are observed for the N-CaM domain, with much lower values in its holo state (apo: 1.273 nm, holo: 0.658 nm). In the case of the SIRAH-OBAFE simulations, a more flexible C-CaM domain is observed for the apo form, with values of 0.706 nm for the N-CaM domain, vs. 0.928 nm for the C-CaM domain. An opposite behaviour is seen in the holo form, with a more flexible N-CaM domain compared to the C-CaM domain, with values of 0.932 nm and 0.585 nm respectively. This correlates with the trend observed in the previously mentioned atomistic studies [192, 195–197]. Interestingly, our atomistic simulations did not show this trend, with no actual differences for any of the states (see table 6.3). Still, the SIRAH-OBAFE force field correlates with the finding that a more flexible N-CaM domain may be needed for ligand recognition in its active form (Ca^{2+}-bound) [193], expressed as higher RMSDs in the holo form for the N-CaM domain.

		SIRAH 1.0	**SIRAH-OBAFE**	**AMBER14SB**
Apo	N-term	1.273 nm	0.706 nm	0.286 nm
	C-term	0.537 nm	0.928 nm	0.294 nm
Holo	N-term	0.658 nm	0.932 nm	0.238 nm
	C-term	0.535 nm	0.585 nm	0.298 nm

Table 6.3

RMSD values for the Calmodulin domains. Averaged RMSD values, calculated with respect to the crystal structure, are shown for the N-CaM and C-CaM domains, in its apo and holo forms. Simulations were done using the SIRAH 1.0, SIRAH-OBAFE and the AMBER14SB force field..

As a final test for the behaviour and stability of the secondary structure with the optimised SIRAH-OBAFE force field, secondary structure time-series were calculated for the group of proteins previously used in section 6.3.1 (see section 6.2.1 for details on the proteins used). These results were compared with the most updated version of the SIRAH force field (i.e. SIRAH 2.0) [11]. Figures 6.6 and 6.7 show these results, and were divided in two figures for clarity. As an overall picture, both force field behave similarly, and some minor differences are observed in the smaller systems (i.e. 1RA4, 1GYV, 1ORC and 1R69). In the case of 1GYV (corresponding to the Gamma-adaptin domain), some α-helices are observed around residue 70 and towards the end of the structure, using the SIRAH-OBAFE force field, while these α-helices are absent with SIRAH 2.0 (Fig. 6.6). Comparing these results with the secondary structure information in UniProtKB, code P16117, these small α-helices are reported, which match with our finding using our optimised force field. In the case of 1R4A (corresponding to the L7Ae Archeal ribosomal protein), the α-helix between around residues 30 to 40 is presented in SIRAH-OBAFE, while an important disruption is present in SIRAH 2.0 (Fig. 6.7). On the other hand, the small α-helix around residues 80 to 85 is disrupted in SIRAH-OBAFE, but present in SIRAH 2.0 (Fig. 6.7). All these α-helices are reported in UniProtKB, code P54066. The same "frayed" behaviour, previously observed in figure 6.3, is also present here.



Protein secondary structure time-series. Evolution of the secondary structure thorughout the simulation for Serum albumin (PDB: 1E7I, 585 residues), GFP protein (PDB: 1QYO, 238 residues), and Gamma-adaptin domain (PDB: 1GYV, 120 residues), using the updated SIRAH 2.0 and the optimised SIRAH-OBAFE force fields. Each colour represents a specific secondary structure: α-helix (pink), β-sheets (yellow) and coil (white).

Figure 6.6

6.3.3. SNARE complex

As in the original SIRAH 1.0 publication, the SNARE complex was used as a measure of errors in the force field, where the authors mention that small errors in single protein could be amplified in bigger and more complex systems [5]. 4 α-helical proteins form the SNARE complex including the plasma-membrane-associated proteins syntaxin and SNAP-25 (two copies), and the vesicular protein synaptobrevin, and they are an impor-

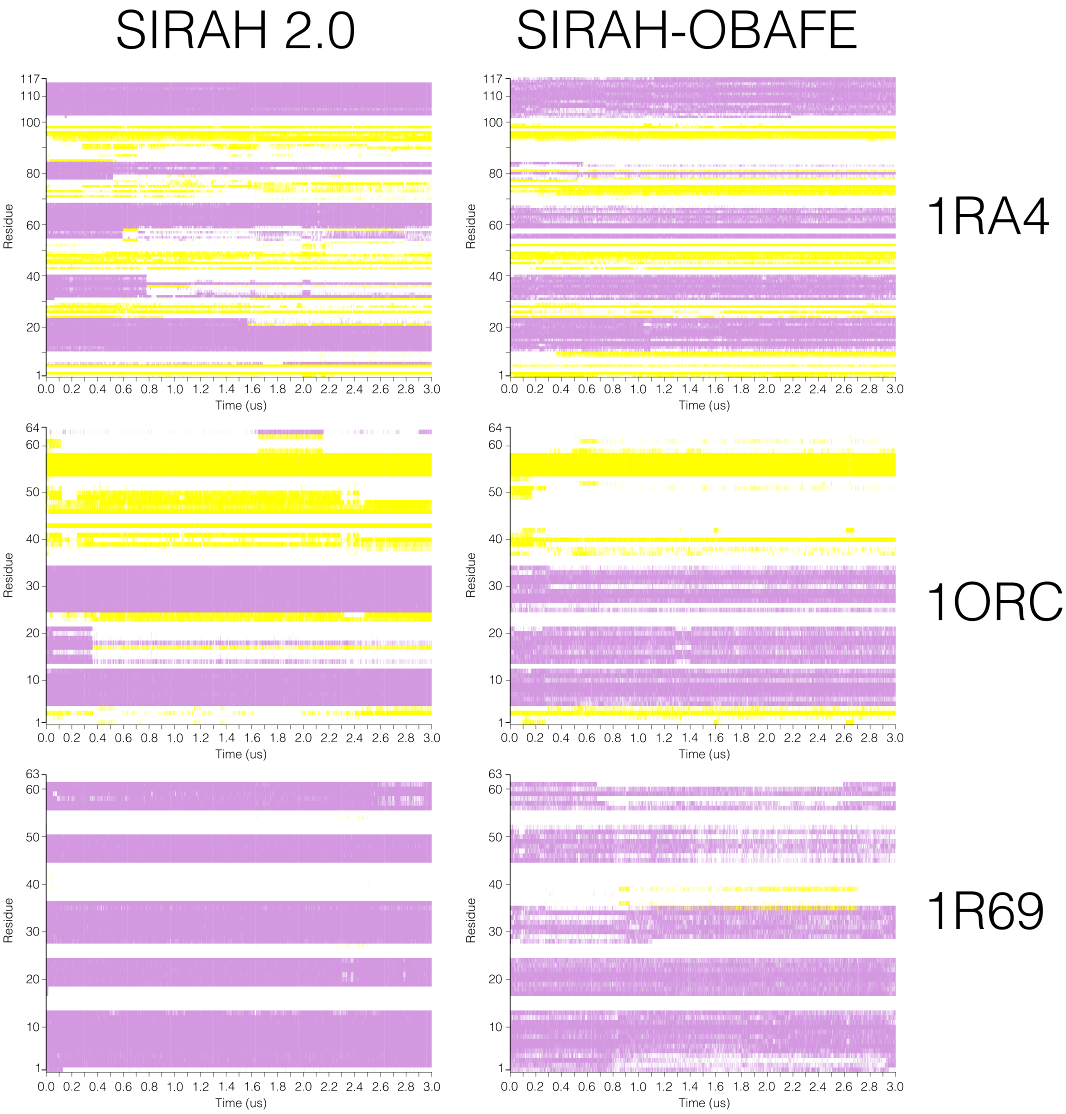


Figure 6.7

Protein secondary structure time-series. Evolution of the secondary structure thorughout the simulation for L7Ae Archeal ribosomal protein (PDB: 1RA4, 120 residues), CRO repressor (PDB: 1ORC, 71 residues) and the N-terminal domain of phage 434 repressor (PDB: 1R69, 63 residues), using the updated SIRAH 2.0 and the optimised SIRAH-OBAFE force fields. Each colour represents a specific secondary structure: α-helix (pink), β-sheets (yellow) and coil (white).

tant component in the synaptic vesicle fusion process [198,199]. In the protein complex core, these 4 helices form hydrophobic layers, numbered between -7 to 8, with respect to a central 0 layer, known as the zero ionic layer, that is composed of a central arginine surrounded by 3 glutamine residues (Fig. 6.8F). This central layer is of great interest given that is the only hydrophilic region in the SNARE complex, and that it is usually conserved in the SNARE superfamily [200].

C-α carbon RMSD time series were evaluated for each of the components of the SNARE complex with respect to the crystal structure, where averaged values between 0.15 to 0.3 nm are observed using the SIRAH-OBAFE force field (Fig. 6.8A). Previous atomistic studies of the SNARE complex in explicit and implicit solvents have reported RMSD values between 0.06 to 0.3 nm [199], while values of 0.3 nm have been reported in the original SIRAH 1.0 publication [5]. The secondary structure stability is consistent with the expected 4 α-helices (Fig. 6.8C). RMSF analyses show a great similarity with previously reported atomistic RMSFs, with values of around 0.1 nm (Fig 6.8D) [199, 201]. Moreover, these values are much lower compared to the ones reported on the original SIRAH 1.0 publication [5]. The expected rigidity of the central part of the SNARE complex is reproduced using the optimised force field, which is also consistent with the B-factor values from the crystal structure; this was not reproduced with the original SIRAH 1.0 [5].

6.3.4. Trp-Cage peptide

We sought to evaluate the limits of these CG force fields in peptide simulations. We used the Trp-Cage system, a 20-residue mini-protein (with sequence NLYIQWLKDGGPSSGRPPPS) that has been extensively used in folding studies [202–204]. A central hydrophobic core exists around TRP6, that is surrounded by three prolines located towards the C-terminal end.

We tested the stability of this mini-protein through the same analysis as the previous protein cases in sections 6.3.1 and 6.3.2. Simulation were run starting from its folded form. As can be seen in figure 6.9, secondary structure stability is improved using the SIRAH-OBAFE force field, compared to the original SIRAH 1.0 (Fig 6.9A and 6.9B). For the latter, the N-terminal α-helix is disrupted at around 7 μs, which is then transformed to two β-sheets, and finally to coil (Fig. 6.9A). The overall secondary structure is main-

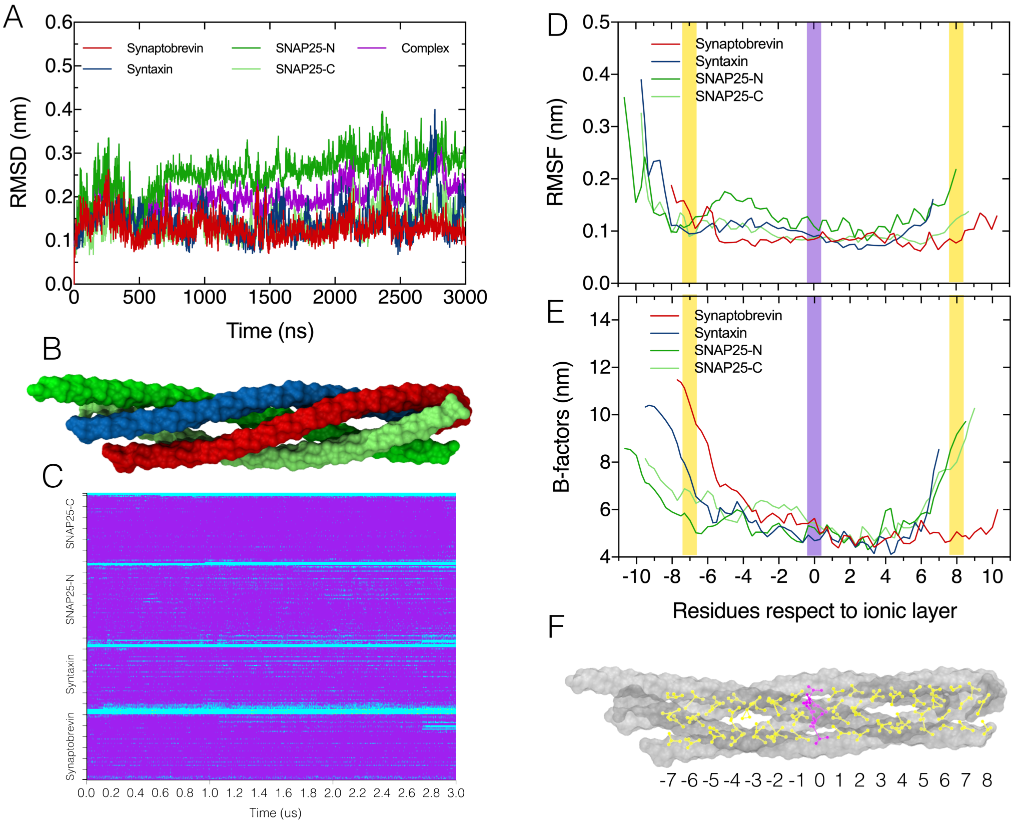


Figure 6.8

Structural analysis of the SNARE complex. (A) RMSDs time series, based on C-α carbons, for the SNARE complex and each component of the complex: Synaptobrevin (red), Syntaxin (blue), SNAP25 N-terminal (dark green) and SNAP25 C-terminal (light green). (B) 3D structure of the SNARE complex with the same colour code as figure A (PDB: 1KIL). (C) Secondary structure stability of each component of the SNARE complex. (D) RMSF analysis, based on C-α carbons, for each component of the SNARE complex. Residues are noted based on the ionic layer. (E) B-factors for the crystal structure (PDB: 1KIL). Residues are noted based on the ionic layer. (F) 3D structure of the SNARE complex coloured based on the 15 hydrophobic layers (yellow) and the central ionic layer (purple).

tained across the whole simulation for the SIRAH-OBAFE force field (Fig. 6.9B). This stability can also be observed based on a contact map of the Trp-Cage protein, showing the percentage of interaction time of certain protein regions, throughout the whole simulation (Fig. 6.9C and 6.9D). The interactions around the N-terminal α-helix are maintained for 40-60% of the simulation time in the case of the SIRAH 1.0 force field (Fig. 6.9C), while these are maintained for most of the simulation time using the SIRAH-OBAFE force field (Fig. 6.9D), compared to the insets shown in both figures, that describes the existing interactions in the crystal structure. It is also clear that important interactions are lost in the 3_{10}-helix domain for SIRAH 1.0, which are maintained for less than 2 μs (Fig. 6.9A and 6.9C). The last frame of the CG simulations using the SIRAH 1.0 and SIRAH-OBAFE force fields are shown in figure 6.9E and 6.9F, respectively, to graphically represent how the observed disruptions that are affecting the overall structure and the well-known U-shape of Trp-Cage. An improvement can also be seen for the secondary structure time series using the SIRAH 2.0 (Fig. 6.10A). Even so, a disruption of the first portion of the α-helix happens before the first microsecond (Fig. 6.10A), and some of the interactions between the 3_{10}-helix and the coil are lost (Fig. 6.10B). Moreover, the orientation of the coil with respect to the α-helix is shifted, losing its U-shape (Fig. 6.10C and 6.10D).

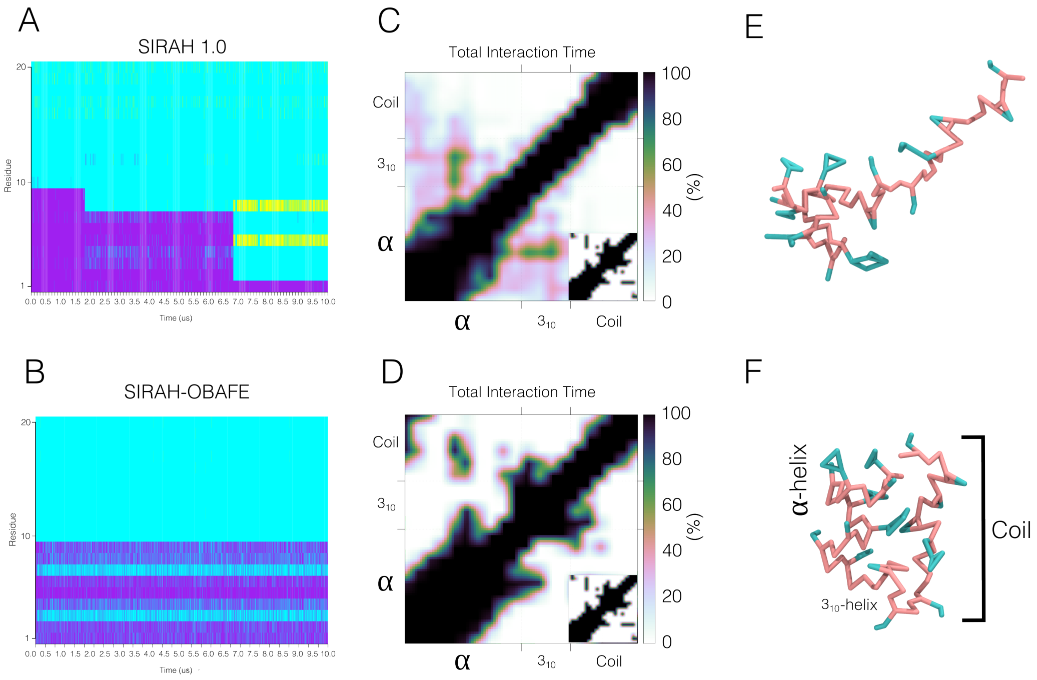


Figure 6.9

Structural analysis of the Trp-Cage miniprotein. Secondary structure time series for the Trp-Cage mini-protein using (A) the SIRAH 1.0 force field, and (B) the optimised SIRAH-OBAFE force field. Total interaction time using (C) the SIRAH 1.0 force field, and (D) the optimised SIRAH-OBAFE force field. The same inset is shown in both figures, which represents the interactions observed using the crystal structure (PDB: 1L2Y). Last frames in the CG simulations of Trp-Cage using (E) the original SIRAH 1.0 force field and (F) the optimised SIRAH-OBAFE force field are shown**.**

As a side note, we have attempted to reproduce folding studies of the TrpCage mini-

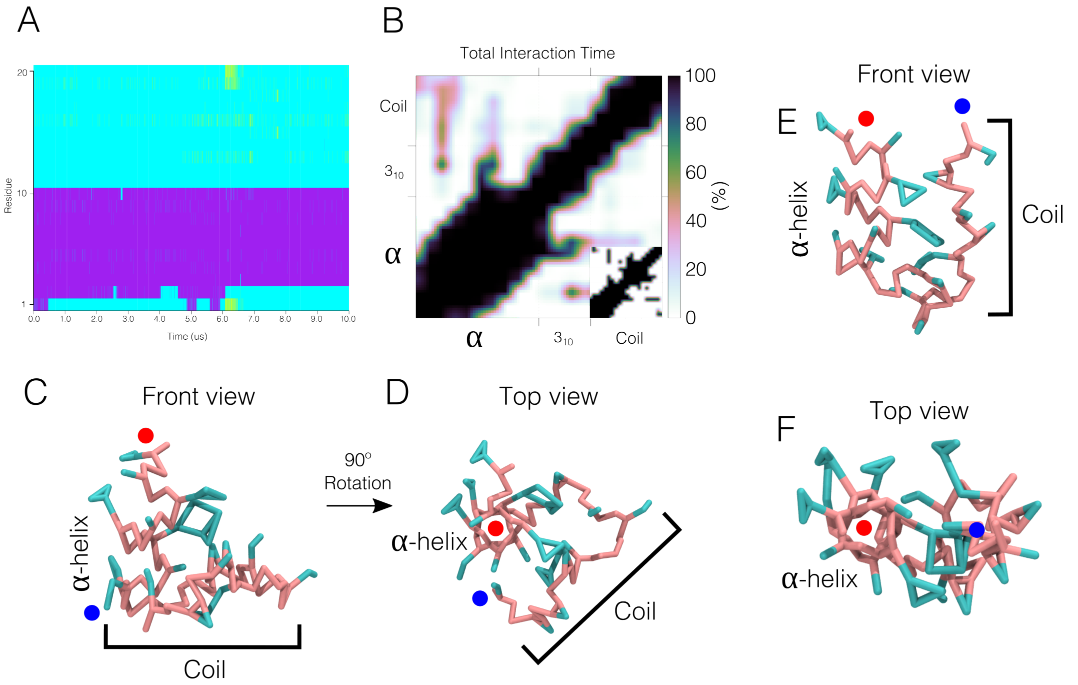


Figure 6.10 **Structural analysis of the Trp-Cage miniprotein using SIRAH 2.0.** (A) Secondary structure time series for the Trp-Cage mini-protein. (B) Contact map of the percentage of interaction time throughout the whole simulation. The inset represents the interactions observed using the crystal structure (PDB: 1L2Y). (C and D) Two views of the TrpCage structure, with a 90° rotation. (E and F) Same rotations as previously shown, but for the crystal structure. The red and blue dots are located on top of the α-helix and at the C-terminal end, respectively, and can be used as a guide for the rotation of the structure**.**

protein in terms of plain MD, and replica-exchange simulations. For both cases, simulations were started from a fully un-folded TrpCage structure, manually created in PyMol. A folding time of 4 μs has been measured, and reported as being the fastest protein folding known [205]. Plain MD simulations were run for up to 10 μs, with simulation protocols identical to the ones shown in section 6.2.3. Unfortunately, we did not observe any folding within the simulated time-scale. With the intention to improve sampling, replica-exchange simulations were run. Briefly, 36 replicas were run that span a temperature range from 273 K to 460 K, similarly as was performed in reference [206]. Simulations were run for 3 μs for each replica. Again, the folding process of TrpCage has not been observed. One reason could be the lack of agreement for the side-chain PMFs seen in chapter 5, where the interactions are underestimated. We believe this is a demonstration of a limitation of the newly optimised SIRAH-OBAFE force field, and an important point of which other workers should be aware if they desire to use this optimised force field in their personal studies.

6.3.5. Free energy landscapes

As a final test of the optimised SIRAH-OBAFE force field, we have attempted to reproduce the free energy landscape of a protein conformational change, as performed in sections 3.3.2.1 and 3.3.2.2, for the S1S2 glutamate receptor [126, 127] and for the DFG transition in the Abl kinase [130, 133].

To describe the cleft movement, a one-dimensional order parameter was used, defined by the distance between the centre of mass of the Cα beads of residues G451 and S652 of the ligand binding domain of the S1S2 glutamate receptor (Fig. 6.9C), which can be considered as an equivalent to the one dimensional projection of a two-dimensional coordinates used in similar studies [126, 127]. This calculation was performed using the original SIRAH 1.0, SIRAH 2.0, and the optimised SIRAH-OBAFE force fields.

As previously mentioned, it has been reported that the apo structure of the receptor prefers more open conformations. For the apo crystal structure, the averaged distance between these residues is around 1.18 nm, and the global minimum of the free energy profiles from atomistic simulation lies at around the same value [126]. The original SIRAH 1.0 force field shows two shallow minima at around 1.15 nm and 1.35 nm, with a barrier between them no greater than 0.5 kcal/mol (Fig. 6.11A, top panel). No improvement is seen using the updated SIRAH 2.0 force field, where the minimum is now located at 0.87 nm, resembling a closed conformation (Fig. 6.12A). In the case of the SIRAH-OBAFE force field, a clear minimum is seen at 1.17 nm (Fig. 6.11B, top panel). Even though both the SIRAH 1.0 and SIRAH-OBAFE force fields predict the right minima for the apo structure, the SIRAH 1.0 force field presents a broader profile compared to the optimised SIRAH-OBAFE. Lau Y. et al., 2007 [126], showed that, for the apo structure, energies of around 4.0 kcal/mol are needed in order to achieve conformations close to the glutamate-bound state (i.e. averaged values of around 0.89 nm).

Energies of 1.0 kcal/mol are seen towards this closed state in the case of the SIRAH 1.0 force field (Fig. 6.11A, top panel), while energies of around 3.0 kcal/mol are needed to achieve this state using the SIRAH-OBAFE force field (Fig. 6.11B, top panel). It is known that CG force fields tend to show smoother energy profiles given the fewer degrees of freedom, compared to atomistic system [9]. However, the apo PMFs derived using the SIRAH-OBAFE force field show great similarity with previous atomistic simulations of opening/closing process in the S1S2 glutamate receptor in its apo state [126, 127].

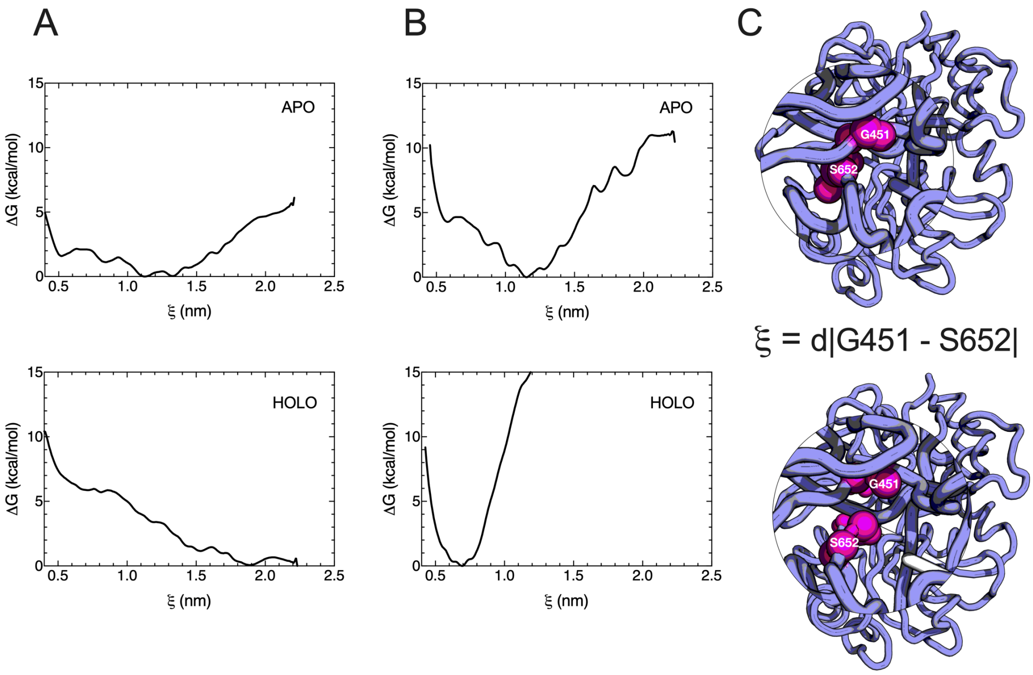


Figure 6.11

Free energy landscapes for the ligand binding domain of the S1S2 glutamate receptor. PMF plots for the apo and glutamate-bound structures, using (A) the original SIRAH 1.0 force field and (B) the optimised SIRAH-OBAFE force field. (C) Structural visualization of the S1S2 glutamate receptor CG model (purple) showing both residues used in the description of the collective variable ξ (magenta). Both conformations for the closed (upper) and open (lower) structures are shown, obtained from the closed structure with PDB code 1FTJ**.**

In the case of the free energy landscapes for the glutamate-bound state, or holo state, a preference for a closed conformation has been seen. The crystal structure for this state shows an average distance between the identified residues of 0.86 nm, while the predicted minimum on atomistic simulations lies at around 0.89 nm [126]. In the case of the SIRAH 1.0, no preference for a closed state is seen, with a greater preference for a more open state compared to the apo structure being shown (Fig. 6.11A, lower panel). This behaviour is repeated using the SIRAH 2.0 force field (Fig. 6.12B). In the case of the optimised SIRAH-OBAFE force field, a clear minimum is seen at around 0.7 nm (Fig. 6.11B, lower panel). Even though the energy minimum is not located where the predicted minimum has been shown in atomistic simulations, the new optimised force field is able to fully predict a preference for a closed state in the glutamate-bound structure.

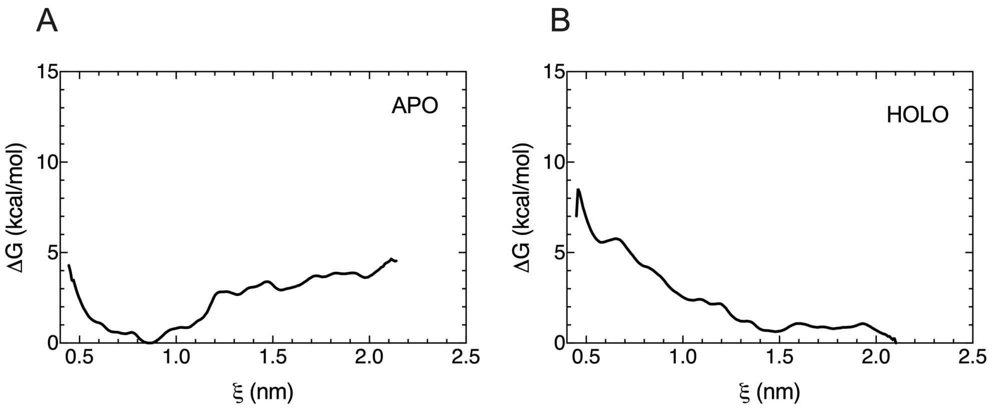


Figure 6.12

Free energy landscapes for the ligand binding domain of the S1S2 glutamate receptor using the SIRAH 2.0 FF. (A) the apo and (B) holo structures**.**

As a second case, the DFG-flip transition was simulated. Please refer to section 3.3.2.2 for an explanation on the DFG transition, and the activation of this type of protein kinase.

In the case of the Abl kinase metadynamics simulations, similarly to the simulation performed in section 3.3.2.2, the DFG-flip transition was re-computed, now using the optimised SIRAH-OBAFE force field. Two dihedral angles were used by Meng Y. et. al, 2015 [130] to describe the transition between the DFG-in and DFG-out conformations: Cβ(Ala380)-Cα(Ala380)-Cα(Asp381)-Cγ(Asp381) and Cβ(Ala380)-Cα(Ala380)-Cα(Phe382)-Cγ(Phe382). A free energy equilibrium was shown, in the case of the Abl kinase, with two clear minima located at around (60°, $-100°$) and ($-100°$, 10°), corresponding to the DFG-in and DFG-out states, respectively [130]. In our case, similar

CVs were used, given the lack of the C-β atom of Ala380, that we believe will capture the DFG transition: CV1: GN(A380)-GC(A380)-GC(D381)-BCG(D381), and CV2: GN(A380)-GC(A380)-GC(F382)-BCG(F382).

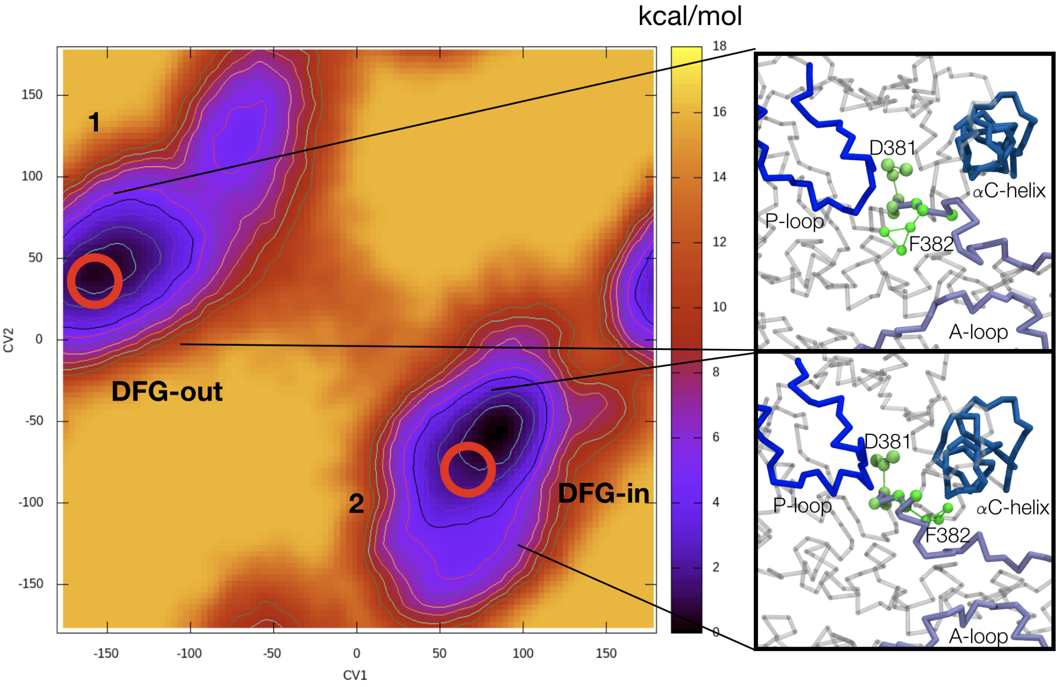


Figure 6.13

2D PMF of the DFG transition. All units are shown in kcal/mol. CV1 and CV2 correspond to two dihedral angles given by: CV1: GN(A380)-GC(A380)-GC(D381)-BCG(D381), and CV2: GN(A380)-GC(A380)-GC(F382)-BCG(F382). Simulation snapshot are shown for both basins, showing the DFG-out (top) and DFG-in (bottom) conformations. The original CVs were calculated using representative backmapped structures of the two minima found (labelled as 1 and 2), and are shown as red circles for comparison with previous atomistic results**.**

Figure 6.13 shows the new 2D PMF of the DFG transition. As can be seen, two large minima exist which are located at around (80°, −70°) and (−150°, 40°) degrees (labelled as 1 and 2). As an approximation, and with the intention to make a closer comparison with previous atomistic studies [130], representative structures located at the centre of both CG minima were back-mapped to their atomistic representation using SIRAH tools [116] (Fig. 6.14 and 6.16). Briefly, atom positions are generated on a per-residue basis from the location of CG beads, while bond distances and angles are derived from rough organic chemistry considerations stored in backmapping

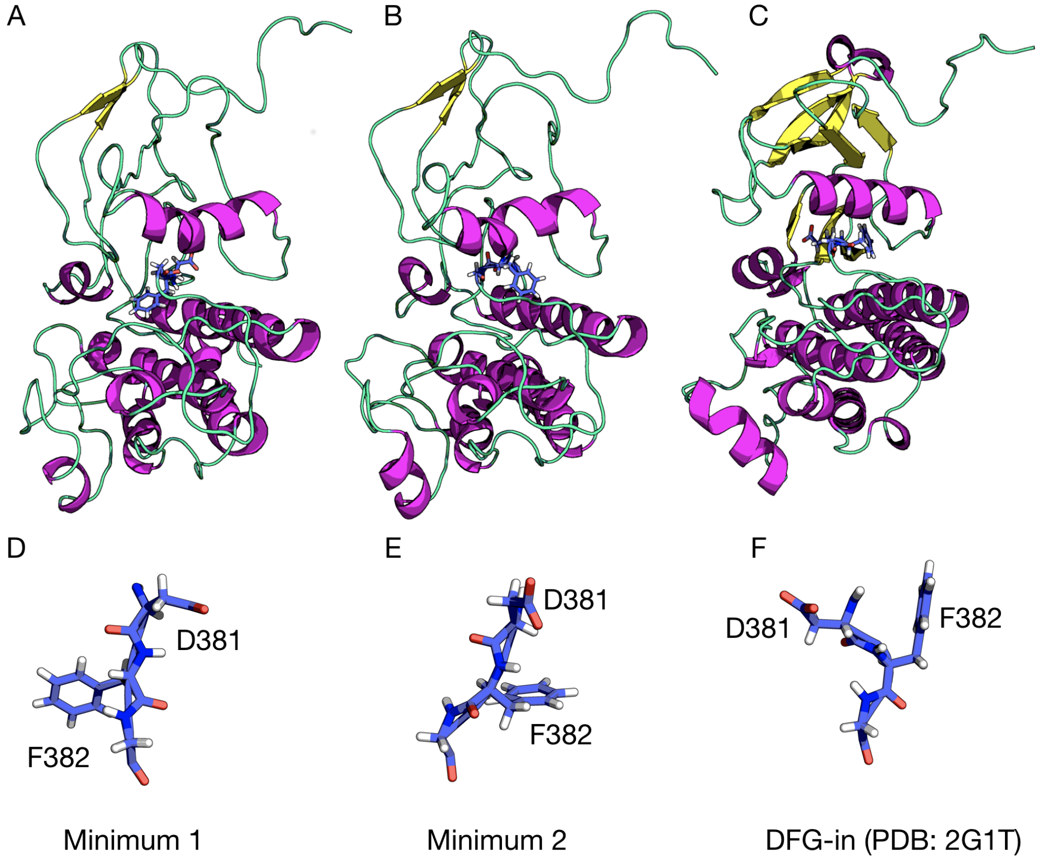


Figure 6.14

SIRAH-OBAFE backmapped structures. Atomistic backmapped structures calculated from representative structures from the two main minima shown in figure 6.11: (A) minimum 1 and (B) minimum 2. (C) Crystal structure shown for comparison. Helices are shown in magenta, β-sheets in yellow and coils in green-cyan. Close-ups to the DFG motif are shown (with the active site on their left-hand side) for (D) minimum 1, (E) minimum 2 and (F) the crystal structure**.**

libraries [116]. The current implementation of SIRAH tools runs 100 steps of steepest-descent followed by 50 steps of conjugated gradient minimisation in vacuum using the sander module of AmberTools [143] to correct for poor stereochemistry conformations [116]. With this, dihedral angle calculations were performed using the 2 collective variables previously mentioned for atomistic systems [130]: Cβ(Ala380)-Cα(Ala380)-Cα(Asp381)-Cγ(Asp381) and Cβ(Ala380)-Cα(Ala380)-Cα(Phe382)-Cγ(Phe382). With this, the back-mapped angles translate to (68°, $-64°$) degrees for the DFG-in conformation, and ($-168°$, 45°) degrees for the DFG-out conformation. This corresponds to differences, with respect to the study from Meng et al., 2015 [130], of (8°, $-36°$) and ($-68°$, 35°) degrees for the DFG-in and DFG-out conformations, respectively (see Fig. 6.13 for the location of the back-mapped angles). Even though the two minima are not exactly in the place predicted in atomistic simulations [130], they resemble structures similar to the DFG-in and DFG-out conformations (Fig. 6.13). This is partially confirmed by 3D structures shown in figures 6.14, 6.15D and 6.15E, where residue F382 is seen outside and inside the active site, movement that has been previously described and that is involved in the disruption of ATP binding [130]. Also, a transition barrier is observed between the two basins, that goes to values of around 12-14 kcal/mol (Fig. 6.13). A transition was also described in atomistic simulations, with values of around 7 kcal/mol [130]. Given the lack of Cβ carbon on alanine in SIRAH, we used atom GN(A380) for alanine, which could also play and important role in the difference observed in the position of the minima and the barrier height. Moreover, the movement of residue D381 is not totally captured with the use of our CVs, which can also be a factor in the observed differences. Either way, we believe this is a notable and important improvement from the previous attempt using the SIRAH 1.0 (Fig. 3.10), where only one big minimum was found and that was considerably shifted from the predicted minima in atomistic studies [130].

Simulations were also performed using the SIRAH 2.0 force field based on the same simulation protocols. Figure 6.15 shows the PMF plots, based on the same collective variables previously used for the CG metadynamics simulations. Compared to the results using the SIRAH 1.0 force field (section 3.3.2.2), two main minima (labelled as 1 and 2) are now observed located at ($-50°$, $-130°$) and ($-150°$, $40°$) degrees. The latter, is located in the same position as predicted by SIRAH-OBAFE, suggesting the approximate location of a DFG-out conformation (Fig. 6.12 and 6.14). Although, backmapped structures confirm something different, showing a wrongly located DFG-in like conformation at minima 1 (Fig. 6.14D). Minima 2 at ($-50°$, $-130°$) degrees is completely shifted (Fig. 6.15) compared to atomistic studies and the prediction by SIRAH-OBAFE (Fig. 6.13), with no clear orientation for residue F382 (i.e. neither inside or outside the active site) (Fig. 6.14E). Moreover, a third minima is observed at around ($100°$, $50°$) degrees (Fig. 6.15), where no comparison can be made since this has not been previously predicted.

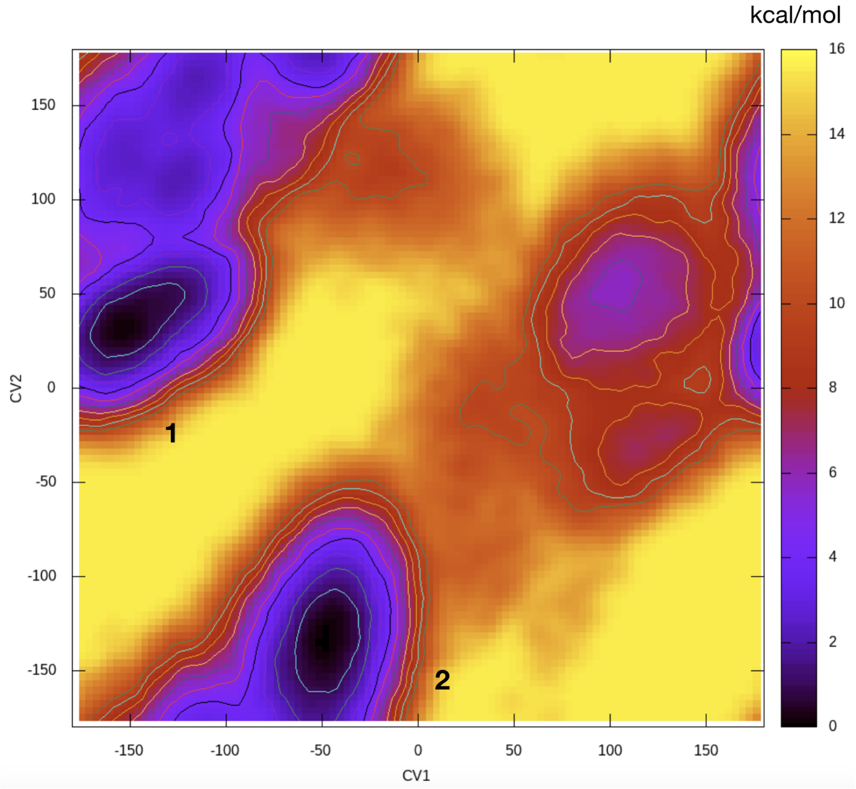


Figure 6.15

2D PMF of the DFG transition using SIRAH 2.0. All units are shown in kcal/mol. CV1 and CV2 correspond to two dihedral angles given by: CV1: GN(A380)-GC(A380)-GC(D381)-BCG(D381), and CV2: GN(A380)-GC(A380)-GC(F382)-BCG(F382). The two main minima are labelled as 1 and 2**.**

6.4. Summary

In this chapter we have evaluated the capabilities of the newly optimised SIRAH-OBAFE force field. We have compared our results with the previous SIRAH 1.0 and SIRAH 2.0 force fields, and with previously published atomistic studies and experimental findings. The main conclusions from this chapter are:

- The SIRAH coarse-grained force field is an alternative to conventional atomistic force fields, especially in the study of protein systems. An updated version of the original 1.0 version of the SIRAH force field [5], called SIRAH 2.0, has been developed by Machado et al., 2019 [11], in an ad-hoc manner, as was the case for SIRAH 1.0 where most of the parameters were based on previously published atomistic force fields (such as AMBER99) and physical intuition (e.g. partial charges based on the amount of hydrogen acceptors/donors) [5]. In this chapter, our newly op-

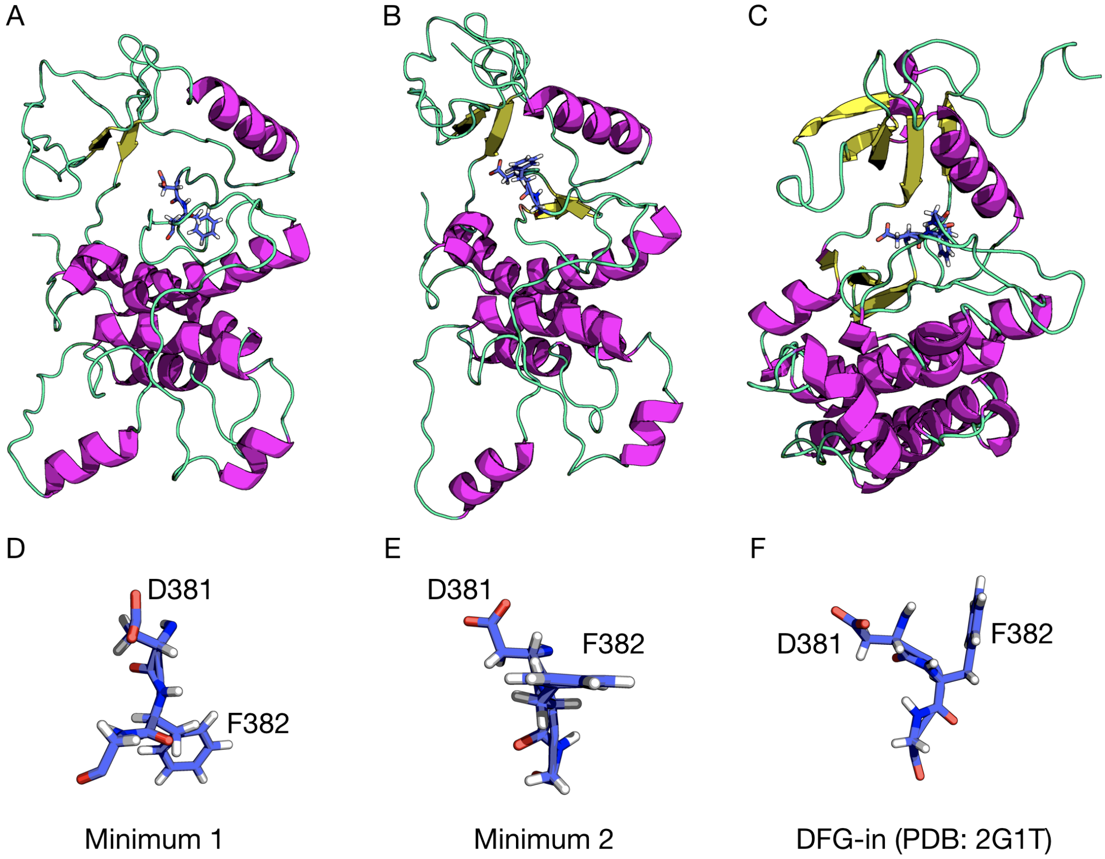


Figure 6.16

SIRAH 2.0 backmapped structures. Atomistic backmapped structures calculated from representative structures from the two main minima shown in figure 6.13: (A) minimum 1 and (B) minimum 2. (C) Crystal structure shown for comparison. Helices are shown in magenta, β-sheets in yellow and coils in green-cyan. Close-ups to the DFG motif are shown (with the active site on their left-hand side) for (D) minimum 1, (E) minimum 2 and (F) the crystal structure**.**

timised SIRAH-OBAFE force field has shown improvement in protein stabilities. RMSD values were reduced by an average of 0.25 nm across the protein systems tested, compared to the original SIRAH 1.0 force field. No important differences were observed between the optimised SIRAH-OBAFE force field and the updated SIRAH 2.0 force field in terms of RMSD, except for the big systems (i.e. 1E7I: Serum albumin, and 1QYO: GFP protein). In these cases, reduced values of 0.15-0.20 nm were found in the SIRAH-OBAFE simulations.

- Secondary structure stability has been evaluated. The stability of secondary structure has not been affected and similarities with atomistic studies have been found. In the case of Calmodulin, the SIRAH-OBAFE and the atomistic system share a “frayed” behaviour in the secondary structure time-series plots, while SIRAH 1.0 shows a "smooth" behaviour (i.e. much less variability in secondary structure). This frayed behaviour was seen again for simulation using the SIRAH 2.0 and SIRAH-OBAFE, for the same group of 6 proteins tested in terms of RMSD values (section 6.3.1). Small differences were found, especially in the stability of α-helices in smaller systems (i.e. 1GYV: Gamma-adaptin domain and 1RA4: L7Ae Archeal ribosomal protein).

- Simulations of the SNARE protein complex system were stable, and showed similarities with previous atomistic studies. RMSD values were observed with values between 0.15-0.30 nm, which is comparable to RMSD values previously reported in the original SIRAH 1.0 publication of 0.3 nm [5], while atomistic studies have reported values of 0.06 to 0.3 nm. The RMSF values were consistent with the expected rigidity of the central section of the SNARE protein complex and the reported RMSF values were in the range of 0.1 to 0.4 nm, consistent with previously reported values of 0.1 nm for atomistic studies.

- The structural stability of the small Trp-Cage peptide was evaluated using the optimised SIRAH-OBAFE force field, and compared with simulations using the original SIRAH 1.0, and the updated SIRAH 2.0 force fields. After 10 μs using SIRAH-OBAFE, Trp-Cage maintains its characteristic U-shape and its corresponding secondary structure, which was not achieved by SIRAH 1.0 and 2.0, where important disruptions were observed, such as unfolding and disruption in the coil section, with respect to the N-terminal α-helix. Attempts were made to study the folding process of TrpCage, using plain MD and replica exchange simulations. Unfortunately, this could not be achieved and we believe that an important factor in this is the underestimation of side-chain interactions, observed in chapter 5 in the PMF of side-chain pairs.

- The calculation of free energy landscapes has been considerably improved using the SIRAH-OBAFE force field, versus the SIRAH 1.0 and 2.0. Test cases for a Glutamate Receptor and the Abl Kinase showed important improvements on finding the correct position of predicted minima compared to atomistic simulations, as well as the size of the barriers. However, our results do not show a perfect fit, but this is a significant step in the optimisation of CG force fields of this kind.

Chapter

7

Conclusions and outlook

Coarse-grained (CG) models currently represent one of the most important approximations for the construction and simulation of larger systems, where much faster calculations can be realised, and allowing us to extend spatial and temporal scales for the simulation of bigger and more complex systems, such as proteins. Different coarse-grained protein models have been developed throughout the years, with differences in terms of mapping, type of functional forms employed, and MD engines where they are supported (see chapter 2, sections 2.2 to 2.6). Intermediate resolution models have been developed, such as UNRES and CABS force field, and have been created for specific tasks such as protein folding. More complex force fields, with a near atomistic mapping representation have also been developed, such as PRIMO and SIRAH. These type of force fields extend the application in areas such as membrane protein simulations and hybrid atomistic/coarse-grained simulations of protein systems.

The SIRAH force field (section 2.6) is a promising alternative compared to other CG protein force fields: no external bias is needed to maintain secondary structures given its higher resolution backbone, long range electrostatic interactions are modelled with a dielectric constant of unity, and it is supported by well-known MD engines such as GROMACS and AMBER. A recent publication of an updated version of SIRAH (named

SIRAH 2.0) has been released, where an ad-hoc optimisation was performed for bonded and non-bonded interactions, improving protein stability. With this, in chapter 3 we have tested the original SIRAH 1.0, and the updated SIRAH 2.0 force fields, extending protein simulations in water, similar to the ones seen in the original publications. Based on our results, we observed that simulations of larger proteins (over 100 residues) were more stable than smaller systems, where clear instabilities were seen, such as sudden changes in the the RMSD values given by the rupture of secondary structure segments and unfolding. The capability of SIRAH to reproduce conformational changes was also tested for the case of the opening/closing process in the S1S2 glutamate receptor ligand binding domain and the DFG transition in the Abl kinase. In the case of the S1S2 glutamate receptor, some of the minima were found but not all in the right place. The overall shape of the PMF showed what is usually expected for CG force field: a smoother energy landscape. Simulation of the ligand binding domain with glutamate did not show similarities with previous atomistic results. In the case of the DFG transition in the Abl kinase, only one minimum was found, compared to two minima observed in previous atomistic studies. Hydration free energies of the CG side-chain were calculated, where important discrepancies were found compared to atomistic and experimental data. We believe this is the main source of error for the interpretation of protein energy landscapes, given the possible misinterpretation of protein interactions between side-chains and side-chains with the solvent.

Generally, force fields are built to reproduce experimental thermodynamic properties and/or *ab initio* calculations, where the parameters are usually iterated "by hand", making this process tedious. Different automated approaches have been developed, such as force matching where parameters are optimised to fit QM energies, machine learning force fields, a data-driven approach where algorithms can be trained to optimise parameters based on a specific training set, and ForceBalance, a hybrid approach where experimental and theoretical data can be combined (see chapter 4, section 4.2). It has been recently stated that there is considerable interest in methods that can automatically generate a coarse-grained model and are representative in terms of local structure and free energy changes [13]. Given this, a new optimisation method was proposed, where hydration free energies of the CG model can be optimised based on hydration free energy gradients from higher resolution models (in this case, atomistic models). This optimisation method was implemented in ForceBalance, enabling a full automation in the optimisation process. Hydration free energy gradients of the CG model were fitted to the atomistic gradients, and parameters were optimised to minimise an objective function that includes square differences between the CG and atomistic gradients.

In chapter 4, we started with the optimisation of the WT4 model, which corresponds to the CG water model implemented in SIRAH. The optimisation was performed using experimental data for water, with the use of properties such as density, enthalpy of vaporisation and dielectric constant, and it was focused on optimising non-bonded parameters only. Initial attempts were made for the optimisation of density for a range of 11 temperatures, between 261 K and 360 K. An improvement was observed, but other properties not included in the optimisation process behaved worse than before. A smaller temperature range was used, now including the three previously mentioned thermodynamic properties. Different ranges of simulation times were used, in order to minimise the noise in the optimisation. However, the optimised parameters did not behave accordingly for the whole temperature range. As 298 K is the temperature where our protein simulations will be run, we decided to proceed with the optimisation at just this temperature. All these properties were successfully improved at 298 K.

Following this, in chapter 4 we continue with the optimisation of the protein side-chains. We started with evaluation of the parameter dependence of the hydration free energies, in order to evaluate how much this property varies upon changes in the force field parameters. Only non-bonded parameters were tested, where we observed no parameter dependence for the free energy gradients with respect to van der Waals σ parameters. Based on this, we proceed with the optimisation of van der Waals ϵ and charge parameters. Comparative analyses of the newly optimised parameters for the uncharged protein side-chains and the backbone excel in the prediction of hydration free energies compared to the previous versions of the SIRAH force field, with increased R^2 values (against experimental data) of 0.97 for the new SIRAH-OBAFE parameter set, compared with values of 0.1 and 0.4 for the SIRAH 1.0 and SIRAH 2.0 sets. To validate our method, a manual parameter search (using 441 parameter combinations) was performed. We achieve a similar minimum in our manual search, compared with the parameters found in ForceBalance.

In chapter 5, an extensive discussion was given of the complications in the calculation of hydration free energies for charged entities, and how this property varies with respect to the simulation methodology. The necessary corrections that need to be introduced to obtain methodological-independent hydration free energies were explained. Attempts were made for the optimisation of charged side-chains, using free energy gradients (similar to chapter 4), but now adding the necessary correction (in the form of gradients of the corrections themselves). Ion-ion PMFs were calculated to evaluate and possibly improve the force field parameters. An improvement on the CG hydration free energies were obtained using this method, but at the expense of exaggerated (and possibly overfitted) parameters. Partial charges, of the order of 1e, for the charged side-chain beads were obtained. To alleviate possible artefacts with the charge sizes, parameters were manually scaled to fit atomistic PMFs of charged side-chain pairs. Even though the PMF fitting was achieved, an over-estimation of the protein interactions was obtained, expressed as stiffer protein systems and over-estimated PMFs of protein conformational changes. One of the main explanations for this kind of behaviour can be that the fewer parameters presented in the SIRAH CG protein force field are not enough to improve the hydration free energies by themselves, and possibly the improvement of both PMFs and free energies is not possible because there is not enough granularity to capture the physics. Given the complexity presented in this process, and to avoid the use of unphysical parameters, the original charged parameters from SIRAH 1.0 were maintained. Either way, a new perspective on the performance of these charged parameters in the estimation of important properties, such as the interaction between them and the solvent, has been shown.

In chapter 6 we tested our newly optimised SIRAH-OBAFE force field. Validations were made on protein systems, of different size, in water. The stability of the secondary structure of flexible and small systems was evaluated. Improvements using the new parameter set have been found, showing better agreement with higher resolution studies in relation to RMSD and RMSF values, as well as the secondary structure stability throughout the simulations. RMSD values were reduced by an average of 0.25 nm across the protein systems tested. The secondary structure stability was improved in small peptide simulations, compared to the original SIRAH 1.0 force field. The stability of protein complexes was also improved, where we also achieved results that correlate with previous atomistic studies on the same systems. The approximate reproduction of free energy landscapes for the conformational change of a soluble protein was achieved, showing great similarities with previous atomistic studies, such as finding the correct position of predicted minima compared to atomistic simulations, as well as the size of the barriers.

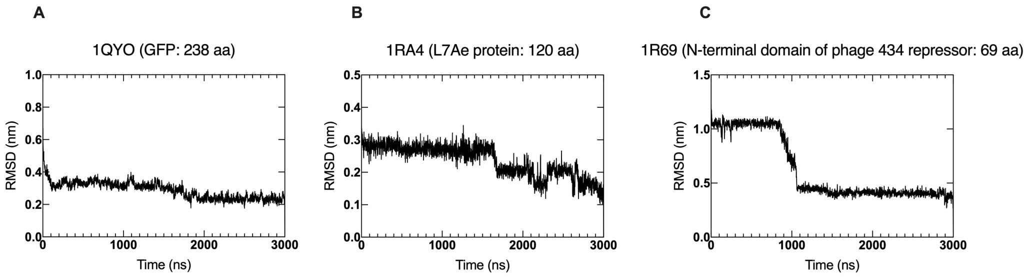
As a final summary, the results in this work show that the SIRAH 1.0 force field is highly promising for protein coarse-grained simulations, but further improvements in parameterisation and more extensive validation studies were needed to study the structural and dynamical behaviour of these systems. It has been stated that there is considerable interest in methods that can automatically generate a coarse-grained model, and that are representative in terms of local structure and free energy changes. Our method paves the way to new optimisation procedures that rely on the use of free energy data as a target. The structural stability of proteins has been improved with the use of the new SIRAH-OBAFE force field, as well as the agreement with higher resolutions studies of systems such as single proteins in water, protein complexes and small peptides. The new optimised SIRAH-OBAFE force field is able to approximately reproduce atomistic PMFs for different protein systems, and by using different methods to estimate the free energy profiles such as umbrella sampling and metadynamics. This improves the possible application of the force field in the study of conformational changes in proteins using a much cheaper approach, which is one of the main advantages of a CG force field. The lack of agreement in the side-chain PMFs observed in chapter 5 is worrying. We believe that limitations in the optimisation methodology are the main cause of this, and is mainly given by the size of the parameter set that is available to optimise the property of interest, where there is not enough granularity to capture the physics involved in the calculation of hydration free energies and side-chain PMFs. The few parameters available in CG models will likely limit the applicability of our proposed optimisation method. To better understand the implications, future studies could be related to the use of more complex CG protein force fields (near an atomistic resolution) in the optimisation process, testing if a more complex set of parameters, or a different functional form, might bring improvements to this limitation. Significant and more validation is needed. The study of the new optimised force field in the area of membrane proteins is of great interest. For this, testing of the lipid force field supported by SIRAH could be performed for simple cases, in order to check protein stabilities and the balance between hydrophobic/hydrophilic interactions. This will greatly increase the applicabilities and areas of research for the SIRAH-OBAFE force field. Furthermore, the parameterisation approach opens a new path to developing CG force fields for other classes of biomolecules such as carbohydrates, nucleic acids, lipids and metabolites, where experimental data is not as readily available, and we look forward to the development and application of such models in the near future.

Appendix A

A.1. CG RMSDs: conformation and flexibility

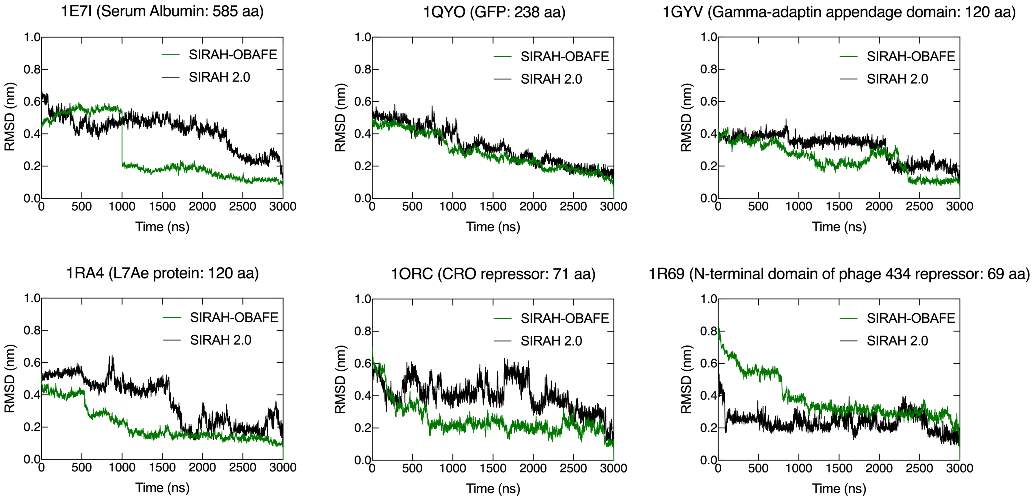
In order to evaluate if the observed high CG RMSD values throughout this thesis correspond to system flexibility or change of conformations to other stable structures, RMSD values were calculated against the last frame of the trajectories. Sudden changes in the RMSD time-series followed by a *plateau* will mean changes in the protein conformation, while constant higher values will reflect that the observed values are related to system flexibility.

Figure A.1 shows RMSD values corresponding to the structures used in chapter 3, section 3.3.1 and figure A.2 shows RMSD values corresponding to the structures used in chapter 6, section 6.3.1.



RMSD times series against the last frame. Root mean square deviation time-series against the last frame of the trajectories are shown for the CG system, using the SIRAH 1.0 force field. (A) (1QYO, GFP-protein, 236 residues), (B) (1RA4, L7Ae Archeal ribosomal protein, 117 residues) and (C) (1R69, N-terminal domain of phage 434 repressor, 63 residues)**.**

Figure A.1



RMSD times series against the last frame. RMSD trajectory analysis is shown as a time series comparison with respect to the C-α carbons of the CG representation to the last frame of the trajectory for (A) Serum albumin, (B) GFP protein, (C) Gamma-adaptin domain, (D) L7Ae Archeal ribosomal protein, (E) CRO repressor and (F) the N-terminal domain of phage 434 repressor. PDB codes are shown in the figure titles. Simulations were run using the SIRAH 2.0 (black) and SIRAH-OBAFE (green) force fields**.**

Figure A.2

A.2. WT4: Estimation of properties outside the optimisation set

In chapter 4, section 4.4.1, the optimisation of the WT4 model was started with the use of density only, in a temperature range between 260 K and 360 K. As mentioned, values for other thermodynamic parameters not included in the optimisation move far away from the experimental values. Figure A.3 shows the behaviour of the dielectric constant based on the parameters derived using combination 1 shown in table 4.1.

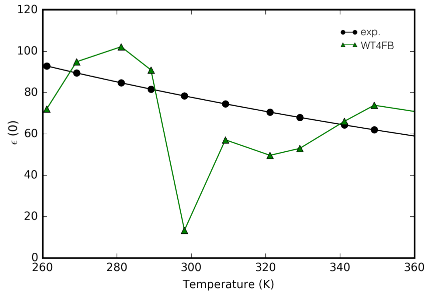


Figure A.3

Behaviour of the dielectric constant. Performance of the dielectric constant using combination 1 (table 4.1). The optimisation was performed using density only as target, in a temperature range between 260 K and 360 K**.**

A.3. 3D protein structures

In order to give the reader a 3D point of view of some the protein structures used in the analysis of protein stability, fluctuations and secondary structure time-series, figure A.4 shows protein structures corresponding to chapter 3, section 3.3.1, and chapter 6, section 6.3.1.

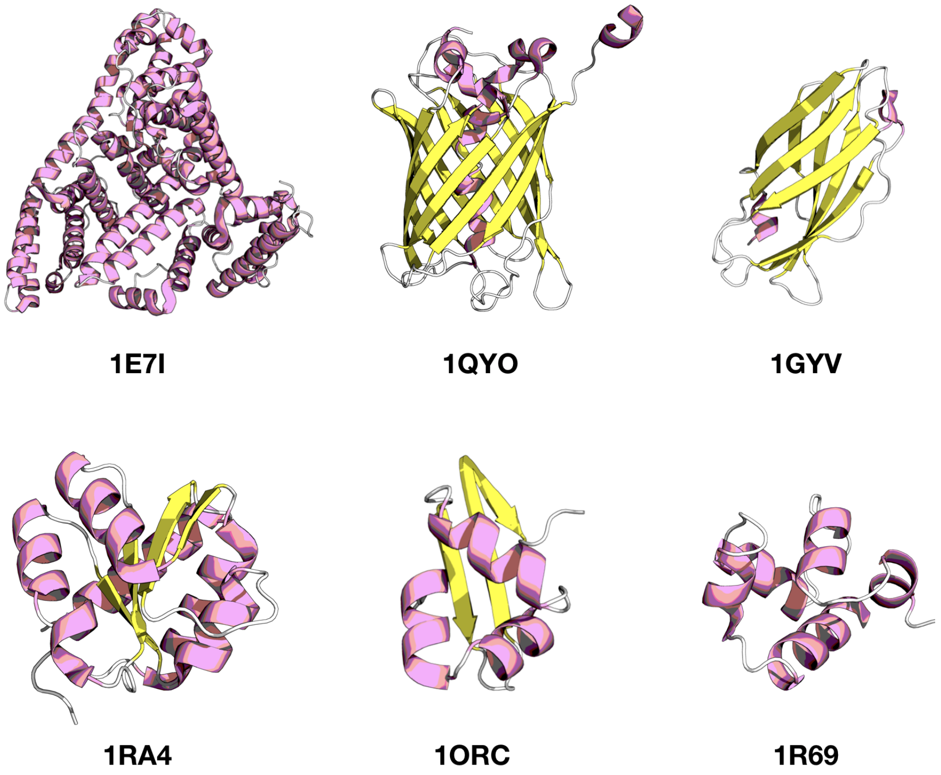


Figure A.4

3D protein structures. Protein structures are shown, coloured by secondary structure (Helix: pink, Sheets: yellow, and Coil: white). Proteins are shown with their corresponding PDB code, which corresponds to: 1E7I: Serum albumin, 1QYO: GFP protein, 1GYV: Gamma-adaptin domain, 1RA4: L7Ae Archeal ribosomal protein, 1ORC: CRO repressor, and 1R69: the N-terminal domain of phage 434 repressor**.**

A.4. Umbrella sampling: convergence

The correct interpretation of the PMFs shown in figures 3.6, 3.7, 6.9 and 6.10 was measured in terms of US window overlap, and convergence in the PMF in terms of calculation at different simulation times. Figure A.5 summarises these results using the SIRAH 1.0, SIRAH 2.0 and SIRAH-OBAFE force fields.

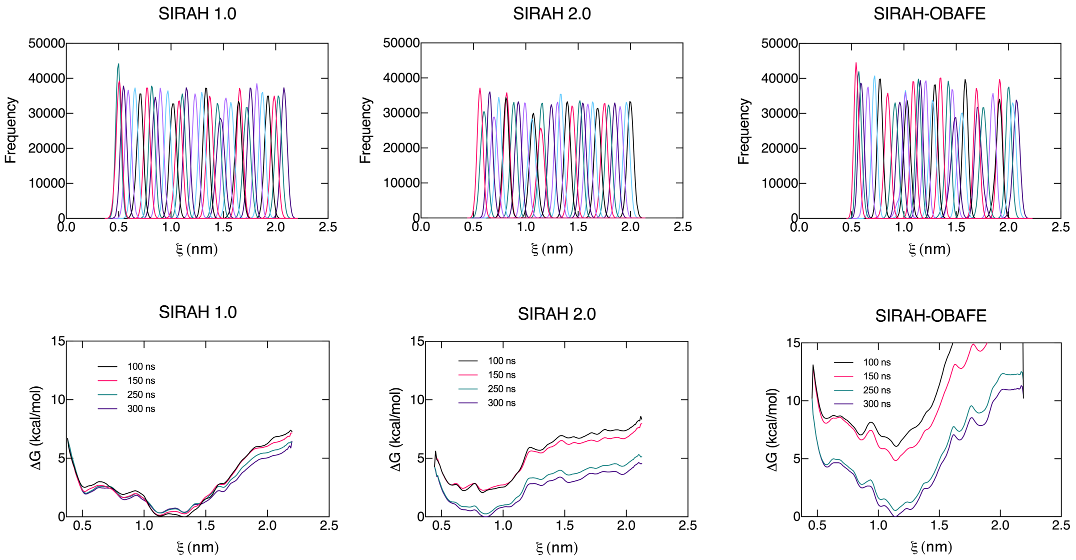
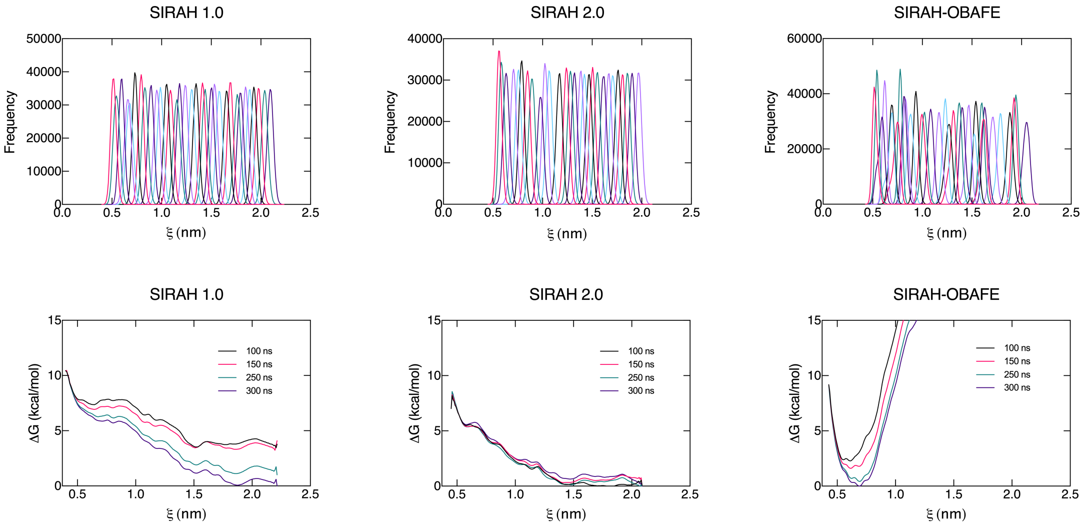


Figure A.5

Umbrella sampling convergence. Analyses of the overlap between the simulated US windows are shown in the upper plots, while convergence of the PMFs, calculated at different simulation times, are shown in the bottom plots. Simulations were performed for the apo (ligand-free) conformation, using the SIRAH 1.0, SIRAH 2.0 and SIRAH-OBAFE force fields**.**



SIRAH 1.0

SIRAH 2.0

SIRAH-OBAFE

Figure A.6

Umbrella sampling convergence. Analyses of the overlap between the simulated US windows are shown in the upper plots, while convergence of the PMFs, calculated at different simulation times, are shown in the bottom plots. Simulations were performed for the holo (glutamate-bound) conformation, using the SIRAH 1.0, SIRAH 2.0 and SIRAH-OBAFE force fields**.**

Bibliography

[1] Adam Liwo, Yi He, and Harold A Scheraga. Coarse-grained force field: general folding theory. *Physical Chemistry Chemical Physics*, 13(38):16890–16901, 2011.

[2] Andrzej Koliński et al. Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, 51, 2004.

[3] Parimal Kar, Srinivasa Murthy Gopal, Yi-Ming Cheng, Alexander Predeus, and Michael Feig. Primo: a transferable coarse-grained force field for proteins. *Journal of Chemical Theory and Computation*, 9(8):3769–3788, 2013.

[4] Luca Monticelli, Senthil K Kandasamy, Xavier Periole, Ronald G Larson, D Peter Tieleman, and Siewert-Jan Marrink. The martini coarse-grained force field: extension to proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, 2008.

[5] Leonardo Darre, Matias Rodrigo Machado, Astrid Febe Brandner, Humberto Carlos Gonzalez, Sebastian Ferreira, and Sergio Pantano. Sirah: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *Journal of Chemical Theory and Computation*, 11(2):723–739, 2015.

[6] Lee-Ping Wang, Todd J Martinez, and Vijay S Pande. Building force fields: an automatic, systematic, and reproducible approach. *The Journal of Physical Chemistry Letters*, 5(11):1885–1891, 2014.

[7] Leonardo Darre, Matias R Machado, Pablo D Dans, Fernando E Herrera, and Sergio Pantano. Another coarse grain model for aqueous solvation: Wat four? *Journal of Chemical Theory and Computation*, 6(12):3793–3807, 2010.

[8] Jaeeon Chang, Abraham M Lenhoff, and Stanley I Sandler. Solvation free energy of amino acids and side-chain analogues. *The Journal of Physical Chemistry B*, 111(8):2098–2106, 2007.

[9] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016.

[10] F. Ercolessi. A molecular dynamics primer. 1997.

[11] Matías R Machado, Exequiel E Barrera, Florencia Klein, Martín Sóñora, Steffano Silva, and Sergio Pantano. The sirah force field 2.0: Altius, fortius, citius. *Journal of Chemical Theory and Computation*, 2019.

[12] Richard T Bradshaw and Jonathan W Essex. Evaluating Parametrization Protocols for Hydration Free Energy Calculations with the AMOEBA Polarizable Force Field. *Journal of Chemical Theory and Computation*, 12(8):3871–3883, August 2016.

[13] Thomas D Potter, Jos Tasche, and Mark R Wilson. Assessing the transferability of common top-down and bottom-up coarse-grained molecular models for molecular mixtures. *Physical Chemistry Chemical Physics*, 21(4):1912–1927, 2019.

[14] Maria Reif and Philippe Hunenberger. Computation of methodology-independent single-ion solvation properties from molecular simulations. iv. optimized lennard-jones interaction parameter sets for the alkali and halide ions in water. *The Journal of Chemical Physics*, 134(14):144104, 2011.

[15] Maria M Reif and Philippe H Huunenberger. Origin of asymmetric solvation effects for ions in water and organic solvents investigated using molecular dynamics simulations: The swain acity–basity scale revisited. *The Journal of Physical Chemistry B*, 120(33):8485–8517, 2016.

[16] Dennis C Rapaport, Robin L Blumberg, Susan R McKay, Wolfgang Christian, et al. The art of simulation. *Computers in Physics*, 10(5):456–456, 1996.

[17] MA Gonzalez. Force fields and molecular dynamics simulations. *Ecole Thematique de la Societe Franccaise de la Neutronique*, 12:169–200, 2011.

[18] Ilario G Tironi, René Sperb, Paul E Smith, and Wilfred F van Gunsteren. A generalized reaction field method for molecular dynamics simulations. *The Journal of Chemical Physics*, 102(13):5451–5459, 1995.

[19] JW Eastwood, RW Hockney, and DN Lawrence. P3m3dp-the three-dimensional periodic particle-particle/particle-mesh program. *Computer Physics Communications*, 35, 1984.

[20] Tom Darden, Lalith Perera, Leping Li, and Lee Pedersen. New tricks for modelers from the crystallography toolkit: the particle mesh ewald algorithm and its use in nucleic acid simulations. *Structure*, 7(3):R55–R60, 1999.

[21] Philippe Hunenberger and Maria Reif. *Single-ion solvation: experimental and theoretical approaches to elusive thermodynamic quantities*, volume 3. Royal Society of Chemistry, 2011.

[22] Michael P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 2017.

[23] Wilfred F van Gunsteren and Herman JC Berendsen. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angewandte Chemie International Edition in English*, 29(9):992–1023, 1990.

[24] Howard Alper and Ronald M Levy. Dielectric and thermodynamic response of a generalized reaction field model for liquid state simulations. *The Journal of Chemical Physics*, 99(12):9847–9852, 1993.

[25] Philippe H Hünenberger and Wilfred F van Gunsteren. Alternative schemes for the inclusion of a reaction-field correction into molecular dynamics simulations: Influence on the simulated energetic, structural, and dielectric properties of liquid water. *The Journal of Chemical Physics*, 108(15):6117–6134, 1998.

[26] Gerald Mathias, Bernhard Egwolf, Marco Nonella, and Paul Tavan. A fast multipole method combined with a reaction field for long-range electrostatics in molecular dynamics simulations: The effects of truncation on the properties of water. *The Journal of Chemical Physics*, 118(24):10847–10860, 2003.

[27] Thierry Matthey. Plain ewald and pme. *Protomol. Sourceforge. net/ewald. pdf*, 2005.

[28] Darden, Tom, York, Darrin, and Pedersen, Lee. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, June 1993.

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

[30] Roger W Hockney. The potential calculation and some applications. *Methods Comput. Phys.*, 9:136, 1970.

[31] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982.

[32] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[33] Christophe Chipot and Andrew Pohorille. *Free energy calculations*. Springer, 2007.

[34] Clara D Christ, Alan E Mark, and Wilfred F Van Gunsteren. Basic ingredients of free energy calculations: a review. *Journal of Computational Chemistry*, 31(8):1569–1582, 2010.

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

[36] Thijs JH Vlugt, JPJM Van der Eerden, Marjolein Dijkstra, Berend Smit, and Daan Frenkel. Introduction to molecular simulation and statistical thermodynamics. *Delft, The Netherlands*, 2008.

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

[38] E Lifeshitz and L Landau. *Statistical physics*. The Charendon Press, Oxford, 1938.

[39] John G Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

[40] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

[41] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008.

[42] Thomas Huber, Andrew E Torda, and Wilfred F Van Gunsteren. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *Journal of Computer-Aided Molecular Design*, 8(6):695–708, 1994.

[43] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[44] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[45] Robert H Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607, 1986.

[46] Koji Hukushima and Koji Nemoto. Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

[47] Pu Liu, Byungchan Kim, Richard A Friesner, and BJ Berne. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proceedings of the National Academy of Sciences*, 102(39):13749–13754, 2005.

[48] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843, 2011.

[49] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

[50] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical Review Letters*, 100(2):020603, 2008.

[51] Johannes Kästner. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942, 2011.

[52] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[53] Michael Andrec. The weighted histogram analysis method (wham), 2010.

[54] Michele Vendruscolo and Christopher M Dobson. Protein dynamics: Moores law in molecular biology. *Current Biology*, 21(2):R68–R70, 2011.

[55] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, 1975.

[56] Michael Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, 104(1):59–107, 1976.

[57] Charles Wilson and Sebastian Doniach. A computer model to dynamically simulate protein folding: studies with crambin. *Proteins: Structure, Function, and Bioinformatics*, 6(2):193–209, 1989.

[58] EI Shakhnovich and AM Gutin. Formation of unique structure in polypeptide chains: theoretical investigation with the aid of a replica approach. *Biophysical Chemistry*, 34(3):187–199, 1989.

[59] Ken A Dill, Sarina Bromberg, Kaizhi Yue, Hue Sun Chan, Klaus M Ftebig, David P Yee, and Paul D Thomas. Principles of protein folding—a perspective from simple exact models. *Protein Science*, 4(4):561–602, 1995.

[60] Jeffrey Skolnick, Andrzej Kolinski, and Robert Yaris. Monte carlo simulations of the folding of beta-barrel globular proteins. *Proceedings of the National Academy of Sciences*, 85(14):5057–5061, 1988.

[61] Jeffrey Skolnick and Andrzej Kolinski. Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *Journal of Molecular Biology*, 221(2):499–531, 1991.

[62] Ming-Hong Hao and Harold A Scheraga. Monte carlo simulation of a first-order transition for protein folding. *The Journal of Physical Chemistry*, 98(18):4940–4948, 1994.

[63] Adam Liwo, Mey Khalili, Cezary Czaplewski, Sebastian Kalinowski, Stanisław Ołdziej, Katarzyna Wachucik, and Harold A Scheraga. Modification and optimization of the united-residue (unres) potential energy function for canonical simulations. i. temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *The Journal of Physical Chemistry B*, 111(1):260–285, 2007.

[64] Adam Liwo, Cezary Czaplewski, Stanisław Ołdziej, Ana V Rojas, Rajmund Kaz, Mariusz Makowski, Rajesh K Murarka, Harold A Scheraga, et al. Simulation of

protein structure and dynamics with the coarse-grained unres force field. In *Coarse-graining of Condensed Phase and Biomolecular Systems*, pages 118–133. CRC Press, 2008.

[65] Adam Liwo, Maciej Baranowski, Cezary Czaplewski, Ewa Gołaś, Yi He, Dawid Jagieła, Paweł Krupa, Maciej Maciejczyk, Mariusz Makowski, Magdalena A Mozolewska, et al. A unified coarse-grained model of biological macromolecules based on mean-field multipole–multipole interactions. *Journal of Molecular Modeling*, 20(8):2306, 2014.

[66] Mariusz Makowski, Adam Liwo, and Harold A Scheraga. Simple physics-based analytical formulas for the potentials of mean force of the interaction of amino acid side chains in water. vii. charged–hydrophobic/polar and polar–hydrophobic/polar side chains. *The Journal of Physical Chemistry B*, 121(2):379–390, 2017.

[67] Paweł Krupa, Adam K Sieradzan, S Rackovsky, Maciej Baranowski, Stanisław Ołdziej, Harold A Scheraga, Adam Liwo, and Cezary Czaplewski. Improvement of the treatment of loop structures in the unres force field by inclusion of coupling between backbone-and side-chain-local conformational states. *Journal of Chemical Theory and Computation*, 9(10):4620–4632, 2013.

[68] Adam Liwo, Mey Khalili, and Harold A Scheraga. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2362–2367, 2005.

[69] Gia G Maisuradze, Patrick Senet, Cezary Czaplewski, Adam Liwo, and Harold A Scheraga. Investigation of protein folding by coarse-grained molecular dynamics with the unres force field. *The Journal of Physical Chemistry A*, 114(13):4471–4485, 2010.

[70] Khatuna Kachlishvili, Gia G Maisuradze, Osvaldo A Martin, Adam Liwo, Jorge A Vila, and Harold A Scheraga. Accounting for a mirror-image conformation as a

subtle effect in protein folding. *Proceedings of the National Academy of Sciences*, 111(23):8458–8463, 2014.

[71] Rui Zhou, Gia G Maisuradze, David Suñol, Toni Todorovski, Maria J Macias, Yi Xiao, Harold A Scheraga, Cezary Czaplewski, and Adam Liwo. Folding kinetics of ww domains with the united residue force field for bridging microscopic motions and experimental measurements. *Proceedings of the National Academy of Sciences*, 111(51):18243–18248, 2014.

[72] Yi He, Magdalena A Mozolewska, Paweł Krupa, Adam K Sieradzan, Tomasz K Wirecki, Adam Liwo, Khatuna Kachlishvili, Shalom Rackovsky, Dawid Jagieła, Rafał Ślusarz, et al. Lessons from application of the unres force field to predictions of structures of casp10 targets. *Proceedings of the National Academy of Sciences*, 110(37):14936–14941, 2013.

[73] Adam K Sieradzan, Adam Liwo, and Ulrich HE Hansmann. Folding and self-assembly of a small protein complex. *Journal of Chemical Theory and Computation*, 8(9):3416–3422, 2012.

[74] Yanping Yin, Adam K Sieradzan, Adam Liwo, Yi He, and Harold A Scheraga. Physics-based potentials for coarse-grained modeling of protein–dna interactions. *Journal of Chemical Theory and Computation*, 11(4):1792–1808, 2015.

[75] Cezary Czaplewski, Sebastian Kalinowski, Adam Liwo, and Harold A Scheraga. Application of multiplexed replica exchange molecular dynamics to the unres force field: tests with α and α+ β proteins. *Journal of Chemical Theory and Computation*, 5(3):627–640, 2009.

[76] Ewa Gołas, Gia G Maisuradze, Patrick Senet, Stanisław Ołdziej, Cezary Czaplewski, Harold A Scheraga, and Adam Liwo. Simulation of the opening and closing of hsp70 chaperones by coarse-grained molecular dynamics. *Journal of Chemical Theory and Computation*, 8(5):1750–1764, 2012.

[77] Sebastian Kmiecik and Andrzej Kolinski. Folding pathway of the b1 domain of

protein g explored by multiscale modeling. *Biophysical journal*, 94(3):726–736, 2008.

[78] Sebastian Kmiecik and Andrzej Kolinski. Simulation of chaperonin effect on protein folding: a shift from nucleation–condensation to framework mechanism. *Journal of the American Chemical Society*, 133(26):10283–10289, 2011.

[79] Sebastian Kmiecik, Dominik Gront, Maksim Kouza, and Andrzej Kolinski. From coarse-grained to atomic-level characterization of protein dynamics: transition state for the folding of b domain of protein a. *The Journal of Physical Chemistry B*, 116(23):7026–7032, 2012.

[80] Joseph D Bryngelson, Jose Nelson Onuchic, Nicholas D Socci, and Peter G Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.

[81] Maciej Blaszczyk, Michal Jamroz, Sebastian Kmiecik, and Andrzej Kolinski. Cabsfold: server for the de novo and consensus-based prediction of protein structure. *Nucleic Acids Research*, 41(W1):W406–W411, 2013.

[82] Michal Jamroz, Andrzej Kolinski, and Sebastian Kmiecik. Cabs-flex: server for fast simulation of protein structure fluctuations. *Nucleic Acids Research*, 41(W1):W427–W431, 2013.

[83] Mateusz Kurcinski, Michal Jamroz, Maciej Blaszczyk, Andrzej Kolinski, and Sebastian Kmiecik. Cabs-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Research*, 43(W1):W419–W424, 2015.

[84] Parimal Kar, Srinivasa Murthy Gopal, Yi-Ming Cheng, Afra Panahi, and Michael Feig. Transferring the primo coarse-grained force field to the membrane environment: simulations of membrane proteins and helix–helix association. *Journal of Chemical Theory and Computation*, 10(8):3459–3472, 2014.

[85] Parimal Kar and Michael Feig. Hybrid all-atom/coarse-grained simulations of

proteins by direct coupling of charmm and primo force fields. *Journal of Chemical Theory and Computation*, 13(11):5753–5765, 2017.

[86] Matthias Buck, Sabine Bouguet-Bonnet, Richard W Pastor, and Alexander D MacKerell Jr. Importance of the cmap correction to the charmm22 protein force field: dynamics of hen lysozyme. *Biophysical Journal*, 90(4):L36–L38, 2006.

[87] Jing Huang and Alexander D MacKerell Jr. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of Computational Chemistry*, 34(25):2135–2145, 2013.

[88] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[89] Siewert J Marrink, Alex H De Vries, and Alan E Mark. Coarse grained model for semiquantitative lipid simulations. *The Journal of Physical Chemistry B*, 108(2):750–760, 2004.

[90] Julian Michalowsky, Lars V Schäfer, Christian Holm, and Jens Smiatek. A refined polarizable water model for the coarse-grained martini force field with long-range electrostatic interactions. *The Journal of Chemical Physics*, 146(5):054501, 2017.

[91] Jaakko J Uusitalo, Helgi I Ingolfsson, Parisa Akhshi, D Peter Tieleman, and Siewert J Marrink. Martini coarse-grained force field: extension to dna. *Journal of Chemical Theory and Computation*, 11(8):3932–3945, 2015.

[92] Vishal Maingi, Jonathan R Burns, Jaakko J Uusitalo, Stefan Howorka, Siewert J Marrink, and Mark SP Sansom. Stability and dynamics of membrane-spanning dna nanopores. *Nature Communications*, 8:14784, 2017.

[93] Jaakko J Uusitalo, Helgi I Ingólfsson, Siewert J Marrink, and Ignacio Faustino. Martini coarse-grained force field: extension to rna. *Biophysical Journal*, 113(2):246–256, 2017.

[94] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H De Vries. The martini force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[95] Jirasak Wong-Ekkabut, Svetlana Baoukina, Wannapong Triampo, I-Ming Tang, D Peter Tieleman, and Luca Monticelli. Computer simulation study of fullerene translocation through lipid membranes. *Nature Nanotechnology*, 3(6):363–368, 2008.

[96] Amy Y Shih, Anton Arkhipov, Peter L Freddolino, and Klaus Schulten. Coarse grained protein- lipid model with application to lipoprotein particles. *The Journal of Physical Chemistry B*, 110(8):3674–3684, 2006.

[97] 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, 2004.

[98] Helgi I Ingólfsson, Cesar A Lopez, Jaakko J Uusitalo, Djurre H de Jong, Srinivasa M Gopal, Xavier Periole, and Siewert J Marrink. The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(3):225–248, 2014.

[99] Djurre H de Jong, Gurpreet Singh, WF Drew Bennett, Clement Arnarez, Tsjerk A Wassenaar, Lars V Schafer, Xavier Periole, D Peter Tieleman, and Siewert J Marrink. Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1):687–697, 2012.

[100] The martini 3.0 cg force field: Open beta (version 3.0.b.3.2). 2018.

[101] Riccardo Baron, Daniel Trzesniak, Alex H de Vries, Andreas Elsener, Siewert J Marrink, and Wilfred F van Gunsteren. Comparison of thermodynamic properties of coarse-grained and atomic-level simulation models. *ChemPhysChem*, 8(3):452–461, 2007.

[102] Siewert J Marrink and Alan E Mark. Molecular dynamics simulation of the formation, structure, and dynamics of small phospholipid vesicles. *Journal of the American Chemical Society*, 125(49):15233–15242, 2003.

[103] WF Drew Bennett, Justin L MacCallum, Marlon J Hinner, Siewert J Marrink, and D Peter Tieleman. Molecular view of cholesterol flip-flop and chemical potential in different membrane environments. *Journal of the American Chemical Society*, 131(35):12714–12720, 2009.

[104] Fumiko Ogushi, Reiko Ishitsuka, Toshihide Kobayashi, and Yuji Sugita. Rapid flip-flop motions of diacylglycerol and ceramide in phospholipid bilayers. *Chemical Physics Letters*, 522:96–102, 2012.

[105] Santi Esteban-Martín, H Jelger Risselada, Jesús Salgado, and Siewert J Marrink. Stability of asymmetric lipid bilayers assessed by molecular dynamics simulations. *Journal of the American Chemical Society*, 131(42):15194–15202, 2009.

[106] Peter M Kasson, Erik Lindahl, and Vijay S Pande. Atomic-resolution simulations predict a transition state for vesicle fusion defined by contact of a few lipid tails. *PLoS Computational Biology*, 6(6):e1000829, 2010.

[107] Peter J Bond and Mark SP Sansom. Insertion and assembly of membrane proteins via simulation. *Journal of the American Chemical Society*, 128(8):2697–2704, 2006.

[108] Kia Balali-Mood, Peter J Bond, and Mark SP Sansom. Interaction of monotopic membrane enzymes with a lipid bilayer: a coarse-grained md simulation study. *Biochemistry*, 48(10):2135–2145, 2009.

[109] Artturi Koivuniemi, Timo Vuorela, Petri T Kovanen, Ilpo Vattulainen, and Marja T Hyvönen. Lipid exchange mechanism of the cholesteryl ester transfer protein clarified by atomistic and coarse-grained simulations. *PLoS Computational Biology*, 8(1):e1002299, 2012.

[110] Satyan Sharma and Andre H Juffer. An atomistic model for assembly of trans-

membrane domain of t cell receptor complex. *Journal of the American Chemical Society*, 135(6):2188–2197, 2013.

[111] Xavier Periole, Adam M Knepp, Thomas P Sakmar, Siewert J Marrink, and Thomas Huber. Structural determinants of the supramolecular organization of g protein-coupled receptors in bilayers. *Journal of the American Chemical Society*, 134(26):10959–10965, 2012.

[112] Humberto C Gonzalez, Leonardo Darre, and Sergio Pantano. Transferable mixing of atomistic and coarse-grained water models. *The Journal of Physical Chemistry B*, 117(46):14438–14448, 2013.

[113] Matias R Machado and Sergio Pantano. Exploring laci–dna dynamics by multiscale simulations using the sirah force field. *Journal of Chemical Theory and Computation*, 11(10):5012–5023, 2015.

[114] Exequiel E Barrera, Ezequiel N Frigini, Rodolfo D Porasso, and Sergio Pantano. Modeling dmpc lipid membranes with sirah force-field. *Journal of Molecular Modeling*, 23(9):259, 2017.

[115] Matias R Machado, Humberto C González, and Sergio Pantano. Md simulations of virus-like particles with supra cg solvation affordable to desktop computers. *Journal of Chemical Theory and Computation*, 2017.

[116] Matías R Machado and Sergio Pantano. Sirah tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics*, 32(10):1568–1570, 2016.

[117] Michael R Shirts, Jed W Pitera, William C Swope, and Vijay S Pande. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *The Journal of Chemical Physics*, 119(11):5740–5761, September 2003.

[118] Silvia A Martins, Sergio F Sousa, Maria João Ramos, and Pedro A Fernandes. Prediction of Solvation Free Energies with Thermodynamic Integration Using

the General Amber Force Field. *Journal of Chemical Theory and Computation*, 10(8):3570–3577, July 2014.

[119] W L Jorgensen and J Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, March 1988.

[120] Magnus Lundborg and Erik Lindahl. Automatic gromacs topology generation and comparisons of force fields for solvation free energy calculations. *The Journal of Physical Chemistry B*, 119(3):810–823, 2014.

[121] Alan W Sousa da Silva and Wim F Vranken. Acpype-antechamber python parser interface. *BMC Research Notes*, 5(1):367, 2012.

[122] Todd J Dolinsky, Jens E Nielsen, J Andrew McCammon, and Nathan A Baker. Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic Acids Research*, 32(suppl 2):W665–W667, 2004.

[123] 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*, 1:19–25, 2015.

[124] M Parrinello and A Rahman. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Physical Review Letters*, 45(14):1196–1199, October 1980.

[125] Bussi, Giovanni, Donadio, Davide, and Parrinello, Michele. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, January 2007.

[126] Albert Y Lau and Benoît Roux. The Free Energy Landscapes Governing Conformational Changes in a Glutamate Receptor Ligand-Binding Domain. *Structure*, 15(10):1203–1214, October 2007.

[127] Albert Y Lau and Benoît Roux. The hidden energetics of ligand binding and activation in a glutamate receptor. *Nature Structural & Molecular Biology*, 18(3):283–287, March 2011.

[128] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.

[129] Silvia Lovera, Ludovico Sutto, Ralitza Boubeva, Leonardo Scapozza, Nicole Dolker, and Francesco L Gervasio. The different flexibility of c-src and c-abl kinases regulates the accessibility of a druggable inactive conformation. *Journal of the American Chemical Society*, 134(5):2496–2499, 2012.

[130] Yilin Meng, Yen-lin Lin, and Benoît Roux. Computational study of the "dfg-flip" conformational transition in c-abl and c-src tyrosine kinases. *The Journal of Physical Chemistry B*, 119(4):1443–1456, 2015.

[131] Yanan Guo, Tino Wolter, Tomáš Kubař, Martin Sumser, Dirk Trauner, and Marcus Elstner. Molecular Dynamics Investigation of gluazo, a Photo-Switchable Ligand for the Glutamate Receptor GluK2. *PloS One*, 10(8):e0135399, 2015.

[132] Huan-Xiang Zhou. Gating Motions and Stationary Gating Properties of Ionotropic Glutamate Receptors: Computation Meets Electrophysiology. *Accounts of Chemical Research*, 50(4):814–822, April 2017.

[133] RSK Vijayan, Peng He, Vivek Modi, Krisna C Duong-Ly, Haiching Ma, Jeffrey R Peterson, Roland L Dunbrack Jr, and Ronald M Levy. Conformational analysis of the dfg-out kinase motif and biochemical profiling of structurally validated type ii inhibitors. *Journal of Medicinal Chemistry*, 58(1):466–479, 2014.

[134] Semen O Yesylevskyy, Lars V Schäfer, Durba Sengupta, and Siewert J Marrink. Polarizable water model for the coarse-grained martini force field. *PLoS Computational Biology*, 6(6):e1000810, 2010.

[135] Jens Kleinjung and Franca Fraternali. Design and application of implicit sol-

vent models in biomolecular simulations. *Current Opinion in Structural Biology*, 25:126–134, 2014.

[136] Richard A Friesner, Robert Abel, Dahlia A Goldfeld, Edward B Miller, and Colleen S Murrett. Computational methods for high resolution prediction and refinement of protein structures. *Current Opinion in Structural Biology*, 23(2):177–184, 2013.

[137] Jens Kleinjung, Peter Bayley, and Franca Fraternali. Leap-dynamics: efficient sampling of conformational space of proteins and peptides in solution. *FEBS Letters*, 470(3):257–262, 2000.

[138] Bojan Zagrovic, Eric J Sorin, and Vijay Pande. β-hairpin folding simulations in atomistic detail using an implicit solvent model. *Journal of Molecular Biology*, 313(1):151–169, 2001.

[139] Julien Michel and Jonathan W Essex. Hit identification and binding mode predictions by rigorous free energy simulations. *Journal of Medicinal Chemistry*, 51(21):6654–6664, 2008.

[140] Jin Zhang, Haiyang Zhang, Tao Wu, Qi Wang, and David van der Spoel. Comparison of implicit and explicit solvent models for the calculation of solvation free energy in organic solvents. *Journal of Chemical Theory and Computation*, 13(3):1034–1043, 2017.

[141] Thomas A Halgren. Potential energy functions. *Current Opinion in Structural Biology*, 5(2):205–210, 1995.

[142] Andres Jaramillo-Botero, Saber Naserifar, and William A Goddard III. General multiobjective force field optimization framework, with application to reactive force fields for silicon carbide. *Journal of Chemical Theory and Computation*, 10(4):1426–1439, 2014.

[143] DA Case, DS Cerutti, TE Cheateham, TA Darden, RE Duke, TJ Giese, H Gohlke, AW Goetz, D Greene, N Homeyer, et al. Amber16 package. 2016.

[144] Jay W Ponder, Chuanjie Wu, Pengyu Ren, Vijay S Pande, John D Chodera, Michael J Schnieders, Imran Haque, David L Mobley, Daniel S Lambrecht, Robert A DiStasio Jr, et al. Current status of the amoeba polarizable force field. *The Journal of Physical Chemistry B*, 114(8):2549–2564, 2010.

[145] Luca Monticelli and D Peter Tieleman. Force fields for classical molecular dynamics. In *Biomolecular Simulations*, pages 197–213. Springer, 2013.

[146] Mark S Miller, Wesley Kayser Lay, Shuxiang Li, William Charles Hacker, Jiadi An, Jianlan Ren, and Adrian Hamilton Elcock. Reparameterization of protein force field nonbonded interactions guided by osmotic coefficient measurements from molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 2017.

[147] Da-Wei Li and Rafael Bruschweiler. Iterative optimization of molecular mechanics force fields from nmr data of full-length proteins. *Journal of Chemical Theory and Computation*, 7(6):1773–1782, 2011.

[148] Alexander D Mackerell. Empirical force fields for biological macromolecules: overview and issues. *Journal of Computational Chemistry*, 25(13):1584–1604, 2004.

[149] William L Jorgensen. Optimized intermolecular potential functions for liquid alcohols. *The Journal of Physical Chemistry*, 90(7):1276–1284, 1986.

[150] Julien Michel, Richard D Taylor, and Jonathan W Essex. The parameterization and validation of generalized born models using the pairwise descreening approximation. *Journal of Computational Chemistry*, 25(14):1760–1770, November 2004.

[151] Furio Ercolessi and James B Adams. Interatomic potentials from first-principles calculations: the force-matching method. *EPL (Europhysics Letters)*, 26(8):583, 1994.

[152] Benjamin Waldher, Jadwiga Kuta, Samuel Chen, Neil Henson, and Aurora E

Clark. Forcefit: a code to fit classical force fields to quantum mechanical potential energy surfaces. *Journal of Computational Chemistry*, 31(12):2307–2316, 2010.

[153] Robin M Betz and Ross C Walker. Paramfit: automated optimization of force field parameters for molecular dynamics simulations. *Journal of Computational Chemistry*, 36(2):79–87, 2015.

[154] Marco Masia, Elvira Guàrdia, and Paolo Nicolini. The force matching approach to multiscale simulations: merits, shortcomings, and future perspectives. *International Journal of Quantum Chemistry*, 114(16):1036–1040, 2014.

[155] MJ Frisch, GW Trucks, Hs B Schlegel, GE Scuseria, MA Robb, JR Cheeseman, JA Montgomery Jr, TKKN Vreven, KN Kudin, JC Burant, et al. Gaussian 03, revision c. 02; gaussian, inc. *Wallingford, CT*, 26, 2004.

[156] EJ Bylaska, WA De Jong, N Govind, K Kowalski, TP Straatsma, M Valiev, D Wang, E Apra, TL Windus, J Hammond, et al. Nwchem, a computational chemistry package for parallel computers, version 5.1. *Pacific Northwest National Laboratory, Richland, Washington*, 99352:0999, 2007.

[157] W Smith, CW Yong, and PM Rodger. Dl_poly: Application to molecular simulation. *Molecular Simulation*, 28(5):385–471, 2002.

[158] Steve Plimpton, Paul Crozier, and Aidan Thompson. Lammps-large-scale atomic/molecular massively parallel simulator. *Sandia National Laboratories*, 18, 2007.

[159] MJD Powell. A method for minimizing a sum of squares of non-linear functions without calculating derivatives. *The Computer Journal*, 7(4):303–307, 1965.

[160] Callum J Dickson, Lula Rosso, Robin M Betz, Ross C Walker, and Ian R Gould. Gafflipid: a general amber force field for the accurate molecular dynamics simulation of phospholipid. *Soft Matter*, 8(37):9617–9627, 2012.

[161] Callum J Dickson, Benjamin D Madej, Åge A Skjevik, Robin M Betz, Knut Teigen, Ian R Gould, and Ross C Walker. Lipid14: the amber lipid force field. *Journal of Chemical Theory and Computation*, 10(2):865–879, 2014.

[162] Venkatesh Botu, Rohit Batra, James Chapman, and Rampi Ramprasad. Machine learning force fields: Construction, validation, and outlook. *The Journal of Physical Chemistry C*, 121(1):511–522, 2016.

[163] Huziel E Sauceda, Stefan Chmiela, Igor Poltavsky, Klaus-Robert Müller, and Alexandre Tkatchenko. Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *The Journal of Chemical Physics*, 150(11):114102, 2019.

[164] Stefan Chmiela, Huziel E Sauceda, Igor Poltavsky, Klaus-Robert Müller, and Alexandre Tkatchenko. sgdml: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications*, 2019.

[165] Jörg Behler. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Physical Chemistry Chemical Physics*, 13(40):17930–17955, 2011.

[166] Mário RG Marques, Jakob Wolff, Conrad Steigemann, and Miguel AL Marques. Neural network force fields for simple metals and semiconductors: construction and application to the calculation of phonons and melting temperatures. *Physical Chemistry Chemical Physics*, 2019.

[167] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13):136403, 2010.

[168] Peng Wang, Yecheng Shao, Hongtao Wang, and Wei Yang. Accurate interatomic force field for molecular dynamics simulation by hybridizing classical and machine learning potentials. *Extreme Mechanics Letters*, 24:1–5, 2018.

[169] Lee-Ping Wang, Jiahao Chen, and Troy Van Voorhis. Systematic parametrization of polarizable force fields from quantum chemistry data. *Journal of Chemical Theory and Computation*, 9(1):452–460, 2012.

[170] Lee-Ping Wang, Teresa Head-Gordon, Jay W Ponder, Pengyu Ren, John D Chodera, Peter K Eastman, Todd J Martinez, and Vijay S Pande. Systematic im-

provement of a classical molecular model of water. *The Journal of Physical Chemistry. B*, 117(34):9956, 2013.

[171] Jose LF Abascal and Carlos Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *The Journal of Chemical Physics*, 123(23):234505, 2005.

[172] C Vega, E Sanz, and JLF Abascal. The melting temperature of the most common models of water. *The Journal of Chemical Physics*, 122(11):114507, 2005.

[173] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.

[174] Jamshed Anwar and David M Heyes. Robust and accurate method for free-energy calculation of charged molecular systems. *The Journal of Chemical Physics*, 122(22):224117, 2005.

[175] Yue Shi, Chuanjie Wu, Jay W Ponder, and Pengyu Ren. Multipole electrostatics in hydration free energy calculations. *Journal of Computational Chemistry*, 32(5):967–977, 2011.

[176] 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.

[177] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

[178] Maria Reif and Philippe Hunenberger. Computation of methodology-independent single-ion solvation properties from molecular simulations. iii. correction terms for the solvation free energies, enthalpies, entropies, heat capacities, volumes, compressibilities, and expansivities of solvated ions. *The Journal of Chemical Physics*, 134(14):144103, 2011.

[179] Maria Reif, Philippe Hunenberger, and Chris Oostenbrink. New interaction parameters for charged amino acid side chains in the gromos force field. *Journal of Chemical Theory and Computation*, 8(10):3705–3723, 2012.

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

[181] Andrew W Hakin and Colin L Beswick. Single-ion enthalpies and entropies of transfer from water to aqueous urea solutions at 298.15 k. *Canadian Journal of Chemistry*, 70(6):1666–1670, 1992.

[182] Stefania Taniewska-Osińska, Bożenna Nowicka, Anetta Pietrzak, Stanisław Romanowski, and Tomasz M Pietrzak. Enthalpies of transfer of selected ions from water to water—propan-2-ol mixtures. some quantum chemical aspects of the ionic solvation. *Thermochimica Acta*, 225(1):9–16, 1993.

[183] Malcolm E Davis, Jeffry D Madura, Brock A Luty, and J Andrew McCammon. Electrostatics and diffusion of molecules in solution: simulations with the university of houston brownian dynamics program. *Computer Physics Communications*, 62(2-3):187–197, 1991.

[184] Jeffry D Madura, James M Briggs, Rebecca C Wade, Malcolm E Davis, Brock A Luty, Andrew Ilin, Jan Antosiewicz, Michael K Gilson, Babak Bagheri, L Ridgway Scott, et al. Electrostatics and diffusion of molecules in solution: simulations with the university of houston brownian dynamics program. *Computer Physics Communications*, 91(1-3):57–95, 1995.

[185] Philippe H Hünenberger and J Andrew McCammon. Effect of artificial periodicity in simulations of biomolecules under ewald boundary conditions: a continuum electrostatics study. *Biophysical Chemistry*, 78(1-2):69–88, 1999.

[186] Maria M Reif, Moritz Winger, and Chris Oostenbrink. Testing of the gromos force-field parameter set 54a8: structural properties of electrolyte solu-

tions, lipid bilayers, and proteins. *Journal of Chemical Theory and Computation*, 9(2):1247–1264, 2013.

[187] Jonathan W Essex. The application of the reaction-field method to the calculation of dielectric constants. *Molecular Simulation*, 20(3):159–178, 1998.

[188] Andreas P Eichenberger, Jane R Allison, Jozica Dolenc, Daan P Geerke, Bruno AC Horta, Katharina Meier, Chris Oostenbrink, Nathan Schmid, Denise Steiner, Dongqi Wang, et al. Gromos++ software for the analysis of biomolecular simulation trajectories. *Journal of Chemical Theory and Computation*, 7(10):3379–3390, 2011.

[189] Artëm Masunov and Themis Lazaridis. Potentials of Mean Force between Ionizable Amino Acid Side Chains in Water. *Journal of the American Chemical Society*, 125(7):1722–1730, February 2003.

[190] M. Parrinello and A Rahman. Crystal structure and pair potentials: A molecular-dynamics study. *Physical Review Letters*, 45(14):1196, 1980.

[191] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007.

[192] Willy Wriggers, Ernest Mehler, Felicia Pitici, Harel Weinstein, and Klaus Schulten. Structure and dynamics of calmodulin in solution. *Biophysical Journal*, 74(4):1622–1639, 1998.

[193] Lucie Delemotte and Annie Westerlund. Effect of ca2+ on the promiscuous target-protein binding mechanism of calmodulin. *BioRxiv*, page 277327, 2018.

[194] Craig M Shepherd and Hans J Vogel. A molecular dynamics study of ca2+-calmodulin: evidence of interdomain coupling and structural collapse on the nanosecond timescale. *Biophysical Journal*, 87(2):780–791, 2004.

[195] Yuto Komeiji, Yutaka Ueno, and Masami Uebayasi. Molecular dynamics simulations revealed ca2+-dependent conformational change of calmodulin. *FEBS Letters*, 521(1-3):133–139, 2002.

[196] NP Barton, CS Verma, and LSD Caves. Inherent flexibility of calmodulin domains: A normal-mode analysis study. *The Journal of Physical Chemistry B*, 106(42):11036–11040, 2002.

[197] Cheng Yang, Gouri S Jas, and Krzysztof Kuczera. Structure and dynamics of calcium-activated calmodulin in solution. *Journal of Biomolecular Structure and Dynamics*, 19(2):247–271, 2001.

[198] R Bryan Sutton, Dirk Fasshauer, Reinhard Jahn, and Axel T Brunger. Crystal structure of a snare complex involved in synaptic exocytosis at 2.4 åresolution. *Nature*, 395(6700):347, 1998.

[199] Marie-Pierre Durrieu, Richard Lavery, and Marc Baaden. Interactions between neuronal fusion proteins explored by molecular dynamics. *Biophysical Journal*, 94(9):3436–3446, 2008.

[200] Dirk Fasshauer, R Bryan Sutton, Axel T Brunger, and Reinhard Jahn. Conserved structural features of the synaptic fusion complex: Snare proteins reclassified as q-and r-snares. *Proceedings of the National Academy of Sciences*, 95(26):15781–15786, 1998.

[201] Mohammad Mehdi Ghahremanpour, Faramarz Mehrnejad, and Majid Erfani Moghaddam. Structural studies of snare complex and its interaction with complexin by molecular dynamics simulation. *Biopolymers: Original Research on Biomolecules*, 93(6):560–570, 2010.

[202] J Juraszek and PG Bolhuis. Sampling the multiple folding mechanisms of trp-cage in explicit solvent. *Proceedings of the National Academy of Sciences*, 103(43):15859–15864, 2006.

[203] Ryan Day, Dietmar Paschek, and Angel E Garcia. Microsecond simulations of the folding/unfolding thermodynamics of the trp-cage miniprotein. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1889–1899, 2010.

[204] Heleen Meuzelaar, Kristen A Marino, Adriana Huerta-Viga, Matthijs R Panman, Linde EJ Smeenk, Albert J Kettelarij, Jan H van Maarseveen, Peter Timmer-

man, Peter G Bolhuis, and Sander Woutersen. Folding dynamics of the trp-cage miniprotein: evidence for a native-like intermediate from combined time-resolved vibrational spectroscopy and molecular dynamics simulations. *The Journal of Physical Chemistry B*, 117(39):11490–11501, 2013.

[205] Ruhong Zhou. Trp-cage: folding free energy landscape in explicit water. *Proceedings of the National Academy of Sciences*, 100(23):13280–13285, 2003.

[206] Chen-Yang Zhou, Fan Jiang, and Yun-Dong Wu. Folding thermodynamics and mechanism of five trp-cage variants from replica-exchange MD simulations with rsff2 force field. *Journal of Chemical Theory and Computation*, 11(11):5473–5480, 2015.