Multiscale simulations of protein dynamics

THÈSE N^{O} 5841 (2013)

PRÉSENTÉE LE 15 NOVEMBRE 2013
À LA FACULTÉ DES SCIENCES DE LA VIE
UNITÉ DU PROF. DAL PERARO
PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Enrico SPIGA

acceptée sur proposition du jury:

Prof. P. Fraering, président du jury
Prof. M. Dal Peraro, directeur de thèse
Prof. P. De Los Rios, rapporteur
Dr B. Ensing, rapporteur
Dr R. Lavery, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2013

Abstract

> *Who can say in twenty words what can be said in ten, is also able of evils of all kinds*
>
> —Giosuè Carducci

Characteristic timescales associated with the function of biomolecules, like proteins, range from femtoseconds up to minutes, whereas their corresponding spatial extent ranges from few Å to μm when associating in large macromolecular complexes. Moreover, biomolecules are functional in a large variety of different physico-chemical conditions strongly dependent on pH, ionic strength, crowding agents, etc. This huge complexity is hard to be studied with an arbitrary level of resolution embracing all these spatial and temporal scales. Molecular simulation is a well established approach to gain mechanistic insights into the function of biomolecular systems, producing atomistically detailed models of *in vitro* and/or *in vivo* conditions. I present in this thesis two projects that aim at improving on the current limitations of multiscale molecular simulations, namely (i) the sampling of large systems, and (ii) the detailed representation and description of realistic physiological conditions.

Addressing the first issue, I propose a new coarse-grained model for proteins to be used in molecular dynamics simulations. This coarse-grained model is based on a more accurate description of protein electrostatics, which accounts for dipolar contributions. The parameterization of this force field is based on force-matching methods and on the use of a particle swarm optimization heuristic algorithm. The obtained results are encouraging being structural and electrostatic properties accurately reproduced with the coarse-grained model for a variety of protein folds. Moreover, the parameterization procedure can be straightforwardly applied to any protein, and can be extended to a larger dataset to generate a fully transferable coarse-grained force field, to be applied to any protein and any large macromolecular assemblies for which long all-atom simulations are still a challenge.

While the development and use of coarse-grained models are important to tackle the limited sampling of large systems, it is still important to use all-atom molecular dynamics simulation to investigate with high accuracy in physiological conditions protein dynamics. For this reason, I present here the results of state-of-art molecular dynamics simulations applied to study the influence of crowding agents on the internal dynamics of the protein ubiquitin. Their analysis allows to describe how ubiquitin dynamics is slaved by crowding agents.

This work demonstrates that the description of protein dynamics should take into account its intrinsic multiscale nature. The development and applications of coarse-grained models permit to simulate proteins at low computational cost, the use of atomistic simulations allows to accurately describe proteins in absence and presence of crowding agents, and both of them permit to highlight the essence of protein dynamics.

Keywords biomolecular modeling, proteins, multiscale simulation, molecular dynamics, coarse-grained model, electrostatics, crowding, ubiquitin.

Riassunto

> *Chi riesce a dire con venti parole ciò che può essere detto in dieci, è capace pure di tutte le altre cattiverie.*
>
> —Giosuè Carducci

Le tipiche scale temporali associate con il funzionamento di biomolecole, come le proteine, variano dal femtosecondo al minuto, mentre i corrispondenti moti spaziali variano da pochi Å ai micrometri, quando si associano in complessi supramolecolari. Inoltre le biomolecole funzionano in una grande varietà di differenti condizioni fisico-chimiche che dipendono dal pH, forza ionica, "agenti affollanti" ecc. È difficile studiare questa grande complessità con un unico metodo capace di descrivere tutte queste scale spaziali e temporali. Le simulazioni di dinamica molecolare sono un promettente metodo di studio che permette di fornire dettagli meccanicistici sulla funzionalità di macchine biomolecolari, esibendo modelli a risoluzione atomica di condizioni *in vitro* e/o *in vivo*. In questa tesi io presento due progetti il cui fine è di superare le seguenti limitazioni delle simulazioni di dinamica molecolare (i) "sampling" di grandi sistemi e (ii) la rappresentazione dettagliata, con annessa descrizione, di realistiche condizioni fisiologiche.

Per affrontare il primo problema, io propongo un nuovo modello semplificato di proteine da impiegare in simulazioni di dinamica molecolare. Questo modello semplificato è basato su una più accurata descrizione dell'elettrostatica delle proteine, che tiene conto di contributi dipolari. L'assegnazione dei parametri a questo "campo di forze" è basato sul metodo del *force-matching* e sull'uso di un algoritmo euristico basato sulla "ottimizzazione con sciami di particelle". I risultati sono incoraggianti perchè le simulazioni a livello atomistico sono consistenti con quelle fatte con il modello semplificato. Infatti proprietà strutturali ed elettrostatiche sono riprodotte in maniera accurata per una serie di diversi ripiegamenti proteici. Inoltre la procedura di assegnazione dei parametri può essere applicata facilmente a qualsiasi proteina e può essere estesa ad un più vasto set di proteine per generare un "campo di forze" semplificato trasferibile, che può essere usato per qualsiasi proteina e qualsiasi complesso supramolecolare, per i quali le simulazioni atomistiche risultano essere di difficile realizzazione.

Mentre lo sviluppo e uso di modelli semplificati sono importanti per affrontare il problema del "sampling" di grandi sistemi, è importante anche usare metodologie di dinamica molecolare a livello atomistico per studiare con grande accuratezza e in condizioni fisiologiche la dinamica delle proteine. Per questo motivo, io presento l'uso di tecniche di dinamica molecolare di ultima generazione al fine di studiare l'influenza di agenti affollanti sulla dinamica interna della proteina ubiquitina. Le analisi proposte per le traiettorie di dinamica molecolare permettono di fornire una descrizione dettagliata su come e perchè la dinamica dell'ubiquitina è influenzata dalla concentrazione di "agenti affollanti".

Questo lavoro di ricerca dimostra che la descrizione della dinamica delle proteine dovrebbe considerare la sua intrinseca natura a differenti scale. Lo sviluppo di modelli semplificati permettono la realizzazione di simulazioni di proteine a basso costo computazione, l'uso di simulazioni atomistiche permette di descrivere la dinamica dell'ubiquitina in assenza e presenza di "agenti affollanti", ed entrambi permettono di cogliere i fondamenti della dinamica delle proteine.

Parole chiave biomolecular modeling, proteins, multiscale simulation, molecular dynamics, coarse-grained model, electrostatics, crowding, ubiquitin.

Contents

1 Introduction **11**

2 Methods **17**

- 2.1 Molecular Dynamics Simulation: a Tool of Statistical Mechanics 17
  - 2.1.1 The Ergodic Hypothesis 18
  - 2.1.2 The Canonical Ensemble (NVT) 19
  - 2.1.3 The Isothermal-isobaric Ensemble (NPT) 20
  - 2.1.4 Integration of the Equations of Motion 21
  - 2.1.5 Some Other Tricks of Molecular Dynamics 22
- 2.2 Atomistic Representation of Biomolecules 24
  - 2.2.1 Functional Form of the Potential Energy 25
  - 2.2.2 Atomistic Models for Water 26
  - 2.2.3 Recent Improvements and Development 26
- 2.3 Coarse-Grained Representation of Biomolecules 27
  - 2.3.1 Atomistic to Coarse-Grained Mapping 28
  - 2.3.2 CG Energy Functional Form 29
  - 2.3.3 Solvent Description 30
  - 2.3.4 CG Parameterization Strategies 31
  - 2.3.5 Choice of a Realistic CG Time Step 35
- 2.4 Analysis of Protein Structure and Dynamics 36
  - 2.4.1 Structural Properties 36
  - 2.4.2 Structural fluctuation properties 37
  - 2.4.3 Principal Component Analysis 38
  - 2.4.4 Auto correlation functions 39
  - 2.4.5 Describing Solvation Dynamics 40

3 Development of a coarse-grained model for numerical simulations of proteins 43
3.1 Preface 43
3.2 Introduction 44
3.3 Methods 46
  3.3.1 Coarse-graining the atomistic structure and electrostatics of proteins 46
  3.3.2 Coarse-graining the potential function 48
  3.3.3 Integrating dipole dynamics 51
  3.3.4 Parameterizing the coarse-grained potentials 53
  3.3.5 Coarse-grained simulations and structural observables 57
3.4 Results and Discussion 58
  3.4.1 Structural and electrostatic coarse-grained properties for different protein families 58
  3.4.2 Towards a transferable coarse-grained force field for proteins . . 77
3.5 Conclusions 83
3.6 Future Directions and Development 85

4 All-atom simulations of crowding effects on protein dynamics 87
4.1 Preface 87
4.2 Introduction 88
4.3 Methods 90
  4.3.1 Molecular dynamics simulations 90
  4.3.2 Principal components analysis (PCA) and projection of the trajectories 90
  4.3.3 Calculation of free-energy landscapes 91
  4.3.4 Interactions between ubiquitin and other proteins in X-ray structures 91
  4.3.5 Calculation of NMR parameters from MD trajectories 91
  4.3.6 Root mean squared fluctuations of C_α atoms in protein ensembles 94
  4.3.7 Kinetic description of basins in the conformational landscapes . . 94
  4.3.8 Internal conformational diffusion 95
  4.3.9 Analysis of water and glucose residence times 95
4.4 Crowding effects on ubiquitin dynamics 95
  4.4.1 Simulation of ubiquitin dynamics in water 95
  4.4.2 Ubiquitin dynamics in 325 g/L glucose 98

4.4.3 Exploration of basins in both simulations 101
4.4.4 Extensive interactions between glucose molecules and ubiquitin . 102
4.5 Effect of crowding concentration and initial conditions on ubiquitin dynamics 105
4.5.1 Simulation in water and 108 and 325 g/L glucose, starting from the structure of free ubiquitin 106
4.5.2 Exploring alternative states of the ubiquitin conformational landscape 108
4.6 Discussion 118
4.7 Conclusions 119

5 Conclusion and perspectives 121

Bibliography 125

Chapter 1

Introduction

> *But, if you really want to understand the detailed molecular interactions that make it go in a particular direction, make certain contacts, break other contacts, hydrolyze GTP, you know, form bonds, etcetera, and do it all amazingly accurately, then you do need a high resolution picture of those states. But, that's not going to be enough. It's going to take a lot of work by biochemists, by computational people who do* ***molecular dynamics*** *and things like that to really, eventually, understand it in the sense that we would understand, say, a more typical reaction.*
>
> —Venkatraman Ramakrishnan

Why should one use or develop molecular simulation methods during a Ph.D. in bioengineering? Why should one use such theoretical methodology in order to tackle Life Sciences related problems? I have asked myself these questions several times during the past years and I have come to the following conclusion: Because most of the scientists working in Life Sciences, without telling explicitly, would like to be reassured that all the processes within the cell is the direct consequence of the dynamics of atoms, simply dictated by the laws of Physics. In fact the cell, as a single entity or in a culture, in stable thermodynamic conditions, is so complex that we would like to look for certainty among this overwhelming uncertainty. The mathematical tools that can be used to describe atomic dynamics come from analytic mechanics, statistical mechanics and numerical analysis. These tools have allowed us, with a lot of creativity and hard work, to write down the equations of motion, solve them numerically on high performance computers and analyze the results in a rigorous way. This is essentially the vision of Feynman, as stated in his famous *Lectures on Physics*. However, this does not clarify how hard it is to understand properly life sciences related problems using such a bottom-up theoretical approach. For example, in the late '50s Kendrew and Perutz solved for the first time by X-ray crystallography the structure of hemoglobin, but several theoretical and computational groups around the world are still working on hemoglobin and myoglobin to understand how they really work!

Not less importantly, the knowledge that is accumulated through theoretical and computational studies about the functional mechanism of biomolecules can be exploited to design and program novel functions for them. On a bigger scale, in principle, one could propose a complete atomistic picture of a cell that would help to identify weak points in our understanding and improve our capability in forcing cells to do whatever we need from them. Recently, in the Venter's group, the first cell with a synthetic genome has been created [1]. An atomistic model of such cell could allow to refine or improve its functionalities through the identification of stable and/or transient interactions among biomolecules composing it.

Therefore, in order to realize a faithful simulation of the cell, assuming to have access to sufficient computer resources, we would need to accumulate the following information:

1. Which are all the biomolecules composing the cell?
2. What are their concentrations?
3. Which are the starting conditions to solve the equations of motion?
4. What type of equations of motion should one use?
5. Which are the most appropriate energy potentials to describe their interactions?

The first two questions can be answered by biologists. Reply to the third one belongs to the expertise of biologically oriented experimental physicists and chemists, while the last two belong to the expertise of theoretical physicists and chemists, who can eventually and practically run the simulation of the whole cell. Toward the realization of this task there are many bottlenecks, such as the accumulations of this huge amount of experimental data, their interpretation and access to communities that "speak" different languages, and the development of new theoretical models and methods.

Without taking into account this whole complexity, theoreticians estimated the year when the simulation of a cell will be possible according with the Moore's law. van Gunsteren in 2006 made a prediction that one single copy of a cell composed by 10^{11} atoms will (could) be simulated by 2034 for 1 nanosecond [2]. This estimation is maybe too naive because it does not take into account that is highly improbable that by 2034 we will have all high-resolution structures for all the possible protein folds (whose rate of discovery is less than one per year), all the possible DNA (and RNA) structures, membrane compositions, etc. Therefore, it is maybe too optimistic to think that the community will be able to perform such a simulation in only twenty years. Yet, the simulation won't be able to cover much of the biologically relevant temporal scale experienced by the cell.

Nevertheless, there are many intermediate steps we can do in order to be prepared to approach a realistic simulation of a cell. For instance one possibility is to work toward the accurate modeling of all the biomolecular types and to understand the behaviour of small portions of the cell applying a *divide et impera* strategy. Proteins for instance are one of the most important actors; in E. Coli the entire collection of proteins represents the 55 % of total dry weight (with $\simeq 2.4{\cdot}10^6$ elements) of the cell [3]. Also, historically most of the attention has been focused on this class of biomolecules since the early days of biomolecular modeling. In my thesis, the main focus will also be on proteins studied using classical molecular dynamics simulations. However, the same attention should be paid to membranes, nucleic acids, carbohydrates and all the other small moieties composing the cellular environment.

Around 1965 the biomodeling era started in the group of Shneior Lifson, at the Weizmann Institute, where the first "force field" to study alkanes was developed [4]. Several of the most important scientists still active in this field met there, opening the era of biomolecular modeling. Arieh Warshel and Michael Levitt in Lifson's group started applying these ideas to protein [5]. At the same time Martin Karplus visited Lifson's group [4], ideas and codes were shared, allowing the first minimization of a protein structure in the 1969 [5; 6]. In Karplus' group later, thanks to the joint effort of Bruce Gelin and Andrew McCammon, and initial contribution of Warshel [4], it was possible to run the first simulation of a protein [7], paving the way to the further development of codes for molecular dynamics simulations and production of reliable force fields. This was a scientific milestone that allowed to replace the idea that proteins where just a collection of static atoms. During the same period when the first all-atom simulation of a protein came out, Levitt and Warshel proposed the first coarse-grained simulation of a protein [8].

Anyway, still in the late '70s molecular simulations suffered from the limited computational power that determined the use of very simple models, neglecting for example the explicit presence of water. In those years, thanks to the CECAM (CENTRE EUROPÉEN DE CALCUL ATOMIQUE ET MOLÉCULAIRE) people involved in biomodeling had the opportunity to share ideas, codes and perspectives. These meetings were the starting points for the creation of dedicated codes for biomolecular modeling like GROMOS, developed by Wilfred van Gunsteren and Herman Berendsen who benefited from the possibility to see and modify the CHARMM code [9]. Lately Wilfred van Gunsteren shared the early version of GROMOS to Paul Weiner and Peter Kollman that consequently wrote the first version of AMBER [9]. In the early ’80s the first reliable water models appeared, giving the opportunity to simulate proteins in a more realistic environment [10]. At the same time methodologies to perform free energy calculations were introduced [11]. Moreover, solid methods to perform simulations reproducing statistical ensembles were proposed providing the molecular simulation field with a solid theoretical foundation [12]. As soon as more computational power became available to push further the limit of biomolecular modeling a new problem came out, namely *in silico melting* of proteins due to the use of simple cut off for the treatment of the electrostatics of non bonded interactions. Therefore, the introduction of methods like the reaction field [2] and the particle mesh Ewald [13] permitted to perform, in the early ’90s, simulations in the multi nanosecond temporal scales. The simulations were also much improved from the physical point of view, because more reliable force fields were proposed, able to reproduce or rationalize experimental results [14].

Thus, in early ’90s all the recipes and ingredients to do realistic simulations of proteins were available to the community of biomolecular modelers. Nonetheless, new ”enemies” were waiting behind the corner. The enemies were, and still are, the continuos need for computational power and the intrinsic temporal scales of protein dynamics (*i.e.*, the sampling problem) [2; 15]. Moreover, in order to simulate proteins as close as possible to *in vivo* conditions all the other constituents of the cellular milieu needed to be proposed, developed and tested. Since the early years it was clear that it was necessary to work on three main fronts: (i) the creation of accurate models, (ii) the implementation of efficient new techniques, and (iii) the hardware integration. It was not by chance that the first force field was produced in a group that had access to one of the biggest super-computer available at that time [5].

The development of models of interactions for biomolecules has been always the crucial pillar to enrich the possibilities of modeling new systems. There are several types of all-atom force-fields available like Amber, CHARMM, OPLS, Gromos [14]. The majority of the cited force fields has been improved along the years in describing structural and dynamical properties of proteins and nucleic acids [16–18], allowing recently studies on protein folding [19–21]. Moreover, some of these force fields have been recently extended to carbohydrates, lipids, metabolites etc. [22–24] allowing now, under certain ranges of validity, to propose model systems that should better mimic *in vivo* and *in vitro* conditions.

Regarding novel techniques, a strategy used to enlarge the temporal scale of molecular simulations has been the development of enhanced sampling techniques, which maintaining an all-atom representation are able to accelerate conformational changes in proteins [25]. This type of approaches permits to maintain an accurate description while accelerating process thanks to the identification of reaction coordinates (*i.e.*, dimensionality reduction) that should be able in principle to describe complex conformational changes in proteins and in biomolecules in general. Another way which has been largely explored is to improve the efficiency of softwares to make use of highly parallel architectures on large clusters that recently disclose the possibility to perform multi-million atoms simulations [26; 27]. Similarly, codes have been written or tailored in order to be compatible with emerging commodity hardware, like GPUs (see for example codes like NAMD [28], Amber [29] and AceMD [30]). More *ad hoc* solutions have been available recently with the development of specialized computing units able only to run molecular dynamics, but producing a throughput able to overcome the millisecond barrier [31; 32].

Moreover, after 15 years from the pioneering studies of Levitt and Warshel [8], coarse-grained or simplified models have started to gain more and more importance as an alternative to all-atom representation in order to overcome sampling barriers intrinsic for large systems and longer temporal scales [33]. A plethora of simplified representations started to flourish first for surfactants, lipids and later for proteins and nucleic acids [34–37]. Moreover, several groups started developing systematic ways to produce force-fields for molecular dynamics packages [34].

Given these premises, we are today in the florid situation in which multiple models and techniques can be used to describe with high accuracy biological systems extending the analysis to very large systems for long timescales, which allows often a direct comparison with experimental measurements. Yet, I think it is of paramount importance to keep improving the present capabilities, especially if one wants to reach the ultimate goal of simulating the entire cellular environment. This thesis is meant to be a small contribution to the extension of biomolecular modeling boundaries. I decided to entitle this work *Multiscale Simulations of Protein Dynamics* because models and techniques at different levels of resolution have been adopted and developed to touch some of the limitations discussed above for the investigation of protein dynamics in realistic conditions. While on one side I think it is necessary to develop new simplified models to approach larger complexes, on the other hand atomistic simulations are still the most valuable approach to gain insights into the microscopic details of protein function in the cellular environment.

In particular, I first focused my attention on the development of a new coarse-grained model for proteins to be used in molecular dynamics simulations. The main ingredient of this new coarse-grained model for protein is the introduction of backbone and side chain electrostatic contributions, that allow for an accurate description of protein electrostatic properties. This improvement has important implications for the study of molecular recognition and protein-protein interactions at a quasi-atomistic level of resolution. The proposed coarse-grained model shows also a good compromise in the accuracy vs. speed ratio, being the results in agreement with the atomistic simulations while the computational cost is on average two orders of magnitude lower than finer-grained simulations.

While coarse-grained models are suitable to reach cellular dimensions, the finest details of protein dynamics still need an atomistic description to be captured. This is the case for instance of the effects exerted *in vivo* on protein dynamics and binding by the crowded environment of the cell. While coarse-grained models have been used to study the entropic effects of molecular crowding, recently it has become clear that the enthalpic contributions, due to repulsive and chemical interactions, play an important role that can be accurately modeled only if atomistic models are adopted. For this reason in the second part of my thesis I decided to investigate in detail the effect of small crowding molecules on the dynamic properties of ubiquitin using state-of-the-art large-sampling molecular dynamics simulations.

Therefore, the thesis is organized as follow:

- Chapter 2, where I present the theoretical foundation of the methods and analyses used and improved during my thesis.
- Chapter 3, where I present a newly developed coarse-grained model for molecular dynamics simulation of proteins.
- Chapter 4, where I present how small crowding agents influence protein dynamics.
- Chapter 5, where I will propose some ideas to further extend the boundaries of biomolecular modeling.

Chapter 2

Methods

> *[Anfisen] showed a film of the folding of a protein with "flickering helices forming and dissolving and coming together to form stable substructures". Of course, the film was purely imaginary, but it led to my asking him whether he had though of taking the ideas in the film and translating them into a quantitative model. He said that he did not really know how he would do this, but to me it seemed clear that such a model could be based on straightforward physical concepts [4].*
>
> Martin Karplus

In the present chapter I will describe the theoretical background and the techniques employed for the development of the research described in this thesis. I will report the foundations of biomolecular modeling as emerged from statistical mechanics, and I will give an overview of the state-of-the-art of molecular mechanics techniques for research in Life Sciences. In particular, I will start focusing on molecular dynamics simulation, then on empirical force fields currently used to simulate biomolecules and I will end presenting the analysis tools, used to benchmark or describe equilibrium properties of simulations presented in Chapters 3 and 4.

2.1 Molecular Dynamics Simulation: a Tool of Statistical Mechanics

Molecular dynamics is a technique for computing the equilibrium and transport properties of a many-body system for which a energy potential model for the interactions between system constituents (*i.e.*, particles) is assigned. In order to do that one has to solve the equations of motion able to reproduce a given statistical ensemble using an integrator, which propagates particle positions and velocities from time t to t +

δt. Eventually, relevant properties of the system, which can be directly or indirectly compared with experimental data, can be calculated from the trajectory [38; 39].

During molecular dynamics (MD) simulation different types of thermodynamic ensembles, characterized by the control of certain thermodynamic variables, can be realized. This possibility is of fundamental importance, especially for life-sciences related studies, because it allows to reproduce *in silico* the physical conditions that can be encountered in the *in vivo* or *in vitro* setting. In the following I will introduce the ergodic hypothesis at the basis of MD, and I will introduce the canonical (NVT) and isobaric-isothermal (NPT) ensembles (I will neglect the grand canonical ensemble despite its importance due to non-standard issues and implementation [40]). For both of them I will present the correspondent mechanisms of control used for the development of this research projects. Moreover, I will sketch the basic ideas behind the implementation of integrators and some other technical tricks routinely used in molecular dynamics simulations.

2.1.1 The Ergodic Hypothesis

A classical system is described by a classical Hamiltonian H, which is a function of both coordinates $\mathbf{r}$ and momenta $\mathbf{p}$. If the potential energy function is independent from time and velocity, the Hamiltonian is equal to the total energy:

$$H = H(\mathbf{r}, \mathbf{p}) \equiv K(\mathbf{p}) + U(\mathbf{r}) = \sum_i \frac{\mathbf{p}_i}{2m_i} + U(\mathbf{r}) \tag{2.1}$$

where $K(\mathbf{p})$ is the kinetic energy, $U(\mathbf{r})$ is the potential energy, $\mathbf{p}_i$ is the momentum of particle i, and m_i the mass of particle i. A microscopic state of the system is therefore characterized by the set of values $\{\mathbf{r}, \mathbf{p}\}$, which corresponds to a point in the space defined by both coordinates $\mathbf{r}$ and momenta $\mathbf{p}$ (commonly known as the phase space).

To obtain thermodynamic averages over a microcanonical ensemble, which is characterized by the macroscopic variables (N, V, E), is necessary to know the probability distribution $\rho(\mathbf{r}, \mathbf{p})$ of finding the system at every point (=state) in the phase space. Knowing this distribution function is in principle possible to calculate phase space averages of any dynamic variable $A(\mathbf{r}, \mathbf{p})$ of interest. Examples for dynamic variables are the position, the total energy, the kinetic energy, structural fluctuations, and any other function of $\mathbf{r}$ and/or $\mathbf{p}$. These averages:

$$\langle A(\mathbf{r}, \mathbf{p}) \rangle_Z = \int_V d\mathbf{r} \int_{-\infty}^{\infty} d\mathbf{p} \rho(\mathbf{r}, \mathbf{p}) A(\mathbf{r}, \mathbf{p}) \tag{2.2}$$

are called thermodynamic averages or ensemble averages because they take into account every possible microscopic state of the system (Z is the partition function of the system that counts the number of states the system can occupy). The drawback of this approach is that in order to calculate such thermodynamic averages, is necessary to simultaneously know the distribution probability for each and every state $\{\mathbf{r}, \mathbf{p}\}$.

An alternative strategy for calculating averages is to follow the motion of a single point (i.e., a single molecular state) through the phase space as a function of time by integrating the system's equations of motion, taking the averages only over those points that were visited during the trajectory. Averages calculated in this way are called dynamic averages. Dynamic averages of any dynamical variable $A(\mathbf{r}, \mathbf{p})$ are calculated along the trajectory as follows:

$$\langle A(\mathbf{r}, \mathbf{p}) \rangle_\tau = \frac{1}{\tau} \int_0^\tau A(\mathbf{r}(t), \mathbf{p}(t)) dt \tag{2.3}$$

where τ is the duration of the simulation.

The ergodic hypothesis claims that for a infinitely long trajectory the points generated by the equations of motion will cover the entire phase space, and in this limit ensemble averages are equivalent to dynamic averages, i.e.:

$$\lim_{\tau \to \infty} \langle A(\mathbf{r}, \mathbf{p}) \rangle_\tau = \langle A(\mathbf{r}, \mathbf{p}) \rangle_Z \tag{2.4}$$

While it is assumed that finite molecular dynamics trajectories are "long enough" in the ergodic sense, the longer the sampling the better.

2.1.2 The Canonical Ensemble (NVT)

Integral Control or Nosé-Hoover Thermostat

A way to sample the NVT ensemble within the framework of MD is based on extended system methods [41], where the Newton's equations of motion are modified by adding certain non physical variables. This methodology is used to control temperature using the following non-Hamiltonian equation called Nosé-Hoover equation [42; 43]:

$$H_{N-H} = \sum_i \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}) + \frac{p_\eta^2}{2Q} + Lk_B T \eta \tag{2.5}$$

from which one can derive the following equations of motion:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} \tag{2.6a}$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \frac{p_\eta}{Q}\mathbf{p}_i \tag{2.6b}$$

$$\dot{\eta} = \frac{p_\eta}{Q} \tag{2.6c}$$

$$\dot{p}_\eta = \sum_i \frac{\mathbf{p}_i^2}{m_i} - Lk_BT \tag{2.6d}$$

where $\{\mathbf{r}_i\}, \{\mathbf{p}_i\}$ are coordinates and momenta of the N particles with masses m_i, the forces $\mathbf{F}_i$ are derived from the N-particle potential and L is a parameter to be determined. The two non physical variables η and p_η in eq.2.6 regulate the fluctuations in the total kinetic energy of the physical variables, and can be thus regarded as an effective "thermostat" for the physical system. The parameter Q controls the strength of the coupling to the thermostat: high values result into a low coupling and viceversa. It has been shown that H_{N-H} allows to generate a canonical distribution of the physical degrees of freedom [43–45].

Stochastic Control

In a stochastic thermostat, all or a subset of the degrees of freedom of the system are subject to collisions with *virtual* particles. This method is based on a Langevin stochastic differential equation which describes the motion of a particle i subject to the thermal agitation of a heat bath:

$$\dot{\mathbf{p}}_i = -\nabla_i U - \gamma \mathbf{p}_i + \mathbf{F}^+ \tag{2.7}$$

where γ is a friction constant and $\mathbf{F}^+$ a Gaussian random force. The amplitude of $\mathbf{F}^+$ is determined by the second fluctuation dissipation theorem:

$$< \mathbf{F}_i^+(t_1)\mathbf{F}_j^+(t_2) >= 2\gamma k_BT\delta_{ij}\delta(t_1 - t_2) \tag{2.8}$$

A large value for γ will increase thermal fluctuations, while γ=0 corresponds to the microcanonical ensemble. It was proved that the stochastic thermostat generates a canonical distribution function [41].

2.1.3 The Isothermal-isobaric Ensemble (NPT)

In the framework of the extended Hamiltonian approach a molecular dynamics simulation at constant pressure and constant temperature is defined by the following extended

Hamiltonian:

$$H_{NPT} = \sum_i \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}) + \frac{p_\eta^2}{2Q} + Lk_BT\eta + \frac{p_\epsilon^2}{W} + P_{ext}V \tag{2.9}$$

The equations of motion that correspond to this Hamiltonian are the following:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} + \frac{p_\epsilon}{W}\mathbf{r}_i \tag{2.10a}$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \frac{p_\eta}{Q}\mathbf{p}_i - \frac{p_\epsilon}{W}\mathbf{p}_i \tag{2.10b}$$

$$\dot{V} = \frac{dVp_\epsilon}{W} \tag{2.10c}$$

$$\dot{p_\epsilon} = dV\big(P_{int} - P_{ext}\big) - \frac{p_\eta}{Q}p_\epsilon \tag{2.10d}$$

$$\dot{\eta} = \frac{p_\eta}{Q} \tag{2.10e}$$

$$\dot{p_\eta} = \sum_i \frac{\mathbf{p}_i^2}{m_i} + \frac{p_\epsilon^2}{W} - Lk_BT \tag{2.10f}$$

where the volume V of the system is incorporated in the equations of motion. The momentum p_ϵ is correlated to the variable $\frac{d\epsilon}{dt}$ that depends on the volume as $\epsilon = \frac{1}{d}ln\Big(\frac{V}{V_0}\Big)$ and d has a space dimensionality. The V_0 parameter is a reference arbitrary volume usually equal to the initial volume. The p_ϵ variable acts as a barostat driven by the fluctuations of the internal pressure P_{int} around to the external pressure applied in an isotropic manner on the walls of the simulation box. The internal pressure is given by the following:

$$P_{int} = \frac{1}{dV}\Big[\sum_i \frac{\mathbf{p}_i^2}{m_i} + \sum_i \mathbf{r}_i \cdot \mathbf{F}_i - (dV)\frac{\partial U}{\partial V}\Big] \tag{2.11}$$

where Q and W control the strength of the coupling to the thermostat and barostat, respectively.

It has been showed that the H_{NPT} allows to generate a isothermic-isobaric distribution of the degrees of freedom [44; 45].

2.1.4 Integration of the Equations of Motion

Assigned to the physical system a energy potential for intra- and intermolecular interactions, the integrator of the equations of motion is responsible for the accuracy of the results. Any finite difference integrator is an approximation for a system evolving continuously in time. The requirements for a good integrator are:

- accuracy, in the sense that it has to faithfully approximates the true trajectory
- stability, in the sense that it has to conserve energy avoiding instabilities
- robustness, in the sense that it allows for large time steps (but still physically realistic) in order to propagate the system efficiently through phase space

For simplicity I will describe here only the Verlet's algorithm for the canonical ensemble as an example of successful algorithm routinely used for the integration of the equations of motion in MD. Other algorithms, based on Liouville formulation, have been introduced to permit the use of multiple time steps [46]. I will just draw the main ideas behind an algorithm for the integration of the equations of motion without entering into details that go beyond the subject of the present research.

Verlet's Algorithm for Nosé-Hoover Equations

The simplest and most straightforward way to construct an integrator is by expanding the positions and velocities in a Taylor expansion. Starting from a backward and a forward Taylor expansion of $\mathbf{r}$ and $\mathbf{v}$ around t, for a small enough time step δt it is possible to write the following Verlet's equations [38]:

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\delta t + \left[\frac{\mathbf{F}_i}{m_i} - \eta(t)\mathbf{v}_i\right]\frac{\delta t^2}{2} + O(\delta t^4) \tag{2.12}$$

$$\mathbf{v}_i(t+\delta t) = \mathbf{v}_i(t) + \left[\frac{\mathbf{F}_i(t+\delta t) + \mathbf{F}(t)}{m_i} - \eta(t+\delta t)\mathbf{v}_i(t+\delta t) - \eta(t)\mathbf{v}_i(t)\right]\frac{\delta t}{2} + O(\delta t^4) \tag{2.13}$$

where η regulates the fluctuations in the total kinetic energy. Recently, these equations of motion have been improved in order to make them time reversible [46], allowing at the same time to increase accuracy and robustness of the integration.

2.1.5 Some Other Tricks of Molecular Dynamics

Periodic Boundary Conditions

To mimic the presence of an infinite bulk surrounding a N-particle model system, periodic boundary conditions are usually employed. The volume containing the N particles is treated as the primitive cell of an infinite periodic lattice of identical cells. With periodic boundary conditions a given particle i interacts with all other particles in this infinite periodic system. Assuming that all intermolecular interactions are pairwise additive, the total potential energy of the N particles in any periodic box is

$$U_{tot}(\mathbf{r}) = \frac{1}{2}\sum_{i,j,\mathbf{n}}{}' v(|\mathbf{r}_{ij} + \mathbf{n}L|) \tag{2.14}$$

where L is the diameter of the periodic box (here assumed cubic for simplicity) and **n** is an arbitrary vector of three integer numbers, while the prime on the sum indicates that the term with i=j is excluded when the **n**=**0**. This means that within periodic boundary conditions to simulate bulk behavior the potential energy is an infinite sum rather than a finite one. In practice, however, it is possible to divide the type of interactions in two categories: short and long range interactions.

Treatment of Short Range Interactions The force calculation is the most time-consuming part of molecular dynamics simulations. In fact if one considers a model system with pairwise additive interactions and does not truncate the interactions, for a system of N particles, N(N-1)/2 pair interactions need to be calculated. Efficient techniques exist for speeding up the evaluation of both short range and long range interactions in such a way that the computing time scales as $N^{3/2}$ and NlnN, rather than N^2.

In the case of short range interactions, like van der Waals interactions which decay at large r like $1/r^6$, the contribution to the potential energy beyond a certain cutoff r_c is truncated (applying at the same time the minimum image convention). The most used method is the Verlet's list, in which a second cutoff radius $r_v > r_c$ is introduced, and before to calculate the interactions, a list is made of all particles within a radius r_v of particle i. In the subsequent calculation of the interactions, only the particles in this list have to be considered. If the maximum displacement of the particles is less than r_v-r_c, only the particles in the list of particle i have to be considered (calculation of order N). As soon as one of the particles is displaced more than r_v-r_c, it is necessary to update the list (calculation of order N^2). The latter operation will dominate for a very large number of particles. Typical values of cutoff are between 10 and 12 Å.

Treatment of long range interactions When periodic boundary conditions are applied, the long range Coulombic interactions with particles in the central cell and with all periodic images must be taken into account. Formally the lattice sum to be evaluated is:

$$V_{Coul} = \frac{1}{2}\sum_{i,j=1}^{N}\sum_{\mathbf{n}}{}' \frac{q_i q_j}{|\mathbf{r}_{ij} - \mathbf{n}L|} \tag{2.15}$$

where $\mathbf{n}$ is a lattice vector and $\sum_{\mathbf{n}}{}'$ means that for $\mathbf{n}$=0 it is $i{\neq}j$. It is, however, a well known problem that this type of lattice sum is conditionally convergent and a method to overcome this limitation was invented [47]. The idea is to introduce a convergence factor into the sum 2.15, which depends on a parameter s, and then evaluate the sum for $s{\rightarrow}0$.

It is possible to demonstrate that 2.15 can be replaced by the following one:

$$V_{Coul} = \frac{1}{2}\left[\sum_{i,j=1}^{N}\sum_{\mathbf{n}}{}'\frac{q_iq_jerfc(\alpha|\mathbf{r}_{ij}-\mathbf{n}L|/L)}{|\mathbf{r}_{ij}-\mathbf{n}L|} + \frac{4\pi q_iq_j}{L^3}\sum_{k}\frac{1}{k^2}e^{i\mathbf{k}\mathbf{r}_{ij}}e^{-k^2/4\alpha^2} + \right.$$
$$\left.\frac{1}{L}\left[\sum_{\mathbf{n}\neq 0}\frac{erfc(\alpha\mathbf{n})}{|\mathbf{n}|} + \frac{e^{-\pi^2\mathbf{n}^2}/\alpha^2}{\pi\mathbf{n}^2} - \frac{2\alpha}{\sqrt{\pi}}\right]\sum_{i=1}^{N}q_i^2 + \frac{4\pi}{L^3}\Big|\sum_{i=1}^{N}q_i\Big|^2\right] \tag{2.16}$$

where the evaluation of the potential is split into four different terms, where the last two terms, called *self*- and *surface-terms*, are constant and may be calculated at the beginning of a simulation. The first two sums depend on the inter particle separation $\mathbf{r}_{ij}$ and need to be evaluated at each time step. Thus, the lattice sum is essentially split into a sum which is evaluated in real space and a sum over reciprocal space vectors, $\mathbf{k}$=$2\pi\mathbf{n}$/L. The first sum gives the potential of a set of point charges screened by an opposite charge of the same magnitude having a gaussian form factor with width α. The second sum subtracts this screening charge, but the sum is evaluated in reciprocal space. erfc(x)=1-erf(x) decays as e^{-x^2} for large x, so the first sum contains mainly short range contributions. On the other side, the second sum decays strongly for large k-vectors and thus contains mainly long range contributions. The typical implementation done and commonly used in molecular dynamics packages has been proposed by the Darden's group [13; 48]

2.2 Atomistic Representation of Biomolecules

The success of any computational approach for the study of biomolecular systems relies on the quality of the physical model used to calculate the energy of the system as a function of its structure. Empirical energy functions are usually employed in computational studies of biochemical and biophysical properties of systems [14; 15]. In conventional molecular dynamics simulations, the empirical energy functions are functions of nuclear coordinates only. The use of a single nuclear coordinate to represent atoms is justified in terms of the Born-Oppenheimer approximation [2; 15].

Here I will focus on the commonly used all-atom force fields neglecting the class of united-atom models, which have been shown to have however a similar accuracy [2].

2.2.1 Functional Form of the Potential Energy

The empirical energy functions usually consist of a large number of terms parameterized from experimental data and/or quantum mechanical studies of small molecules or fragments. It is assumed that such parameters may be transferred to simulate larger molecules of interest in different environment [14]. The set of functions along with the associated set of parameters is termed *force field.*

Here I will focus my attention on the AMBER force field [49; 50] because it has been demonstrated over the last years to be one of the best for the simulation of proteins and nucleic acids [16; 17], along with others like CHARMM [51] and OPLS [52]. For this reason this force field was also used for the development of the research presented in this thesis.

The functional form of the AMBER force field is based on the assumption of additivity and transferability of the contributions, and is given by the sum of bonded and non-bonded contributions:

$$V^{AA}_{total}(\mathbf{r}) = V^{AA}_{bonded}(\mathbf{r}) + V^{AA}_{nonbonded}(\mathbf{r}) \tag{2.17}$$

The bonded contributions comprise two-, three- and four-body terms

$$\begin{aligned} V^{AA}_{bonded}(\mathbf{r}) = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angle} K_\theta(\theta - \theta_0)^2 + \\ \sum_{dihedrals} K_\phi[1 + cos(n\phi - \phi_0)] + \sum_{improper} K_\nu(\nu - \nu_0)^2 \end{aligned} \tag{2.18}$$

where:

- the first sum models two-body interactions between consecutive atoms with force constant K_r and bond equilibrium length r_{eq}
- the second sum models three-body interactions to describe bond angles with force constant K_θ and equilibrium bond angle θ_0
- the third sum models four-body interactions due to proper torsions with force constant K_ϕ, multiplicity n and equilibrium torsion ϕ_0
- the fourth sum models out-of-plane vibration involving improper torsions with force constant K_ν and equilibrium value ν_0

The non-bonded interactions are the sum of two main contributions:

$$V^{AA}_{nonbonded}(\mathbf{r}) = \sum_{pairs} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{pairs}\frac{q_i q_j}{\epsilon_0 r_{ij}} \tag{2.19}$$

where:

- the first term, usually called Lennard-Jones potential, is used to model van der Waals interactions between atoms at the distance r_{ij}. ϵ_{ij} is the well depth, σ_{ij} is the distance at which the Lennard-Jones potential is zero. The Lorentz-Berthelodt combination rules [50] are used to obtain the parameters for each pair of different atoms.

- the second term is a Coulombic potential used to model electrostatic interactions between atoms with partial charges q_i and q_j at the distance r_{ij}, where ϵ_0 is dielectric permittivity of vacuo.

The AMBER force fields has been parameterized for proteins and nucleic acids using quantum mechanical calculations and the strategy of parameterization has been explicitly described in [22].

2.2.2 Atomistic Models for Water

Atomistic simulations of biomolecules are usually performed considering explicitly the presence of solvent molecules, such as water. Several models have been proposed for describing water at the atomistic level capable to reproduce the main thermodynamics features of water at room temperature [10; 53]. The majority of the water models are compatible with force field of proteins. Historically the family of SPC models has been developed to be compatible with GROMOS force field [54], whereas the TIPnP family is compatible with AMBER and CHARMM force fields [55–57]. Recent studies have shown that the mix of water models with different all-atom force fields does not raise major concerns for the accuracy of the simulations [58].

2.2.3 Recent Improvements and Development

The current generation of AMBER force field is able to reproduce experimental data reasonably well [59–61]. However, the force fields rely on a number of approximations and thus need continuous improvement as soon as disagreements with experimental data start to appear. For proteins, along this line of research, it has been recently proposed the AMBER99SB [16] refined set of parameters that, improving the parameters of the dihedral backbone angles, permits to better reproduce structural and relaxation data. More recently in the group of D. E. Shaw improvements of AMBER99SB have been proposed in order to better describe dihedral arrangements of several side-chain amino acids (*i.e.*, AMBER99SB-ildn [18]). For nucleic acids the Barcelona corrections have recently improved the reliability of DNA molecular dynamics simulations [17]. Still under development and refinement are more reliable parameters for RNA. Moreover, of paramount importance is the extension of force field to other categories of biomolecules. Recently, important extensions of the AMBER force field in this context are: Slipids [23; 62; 63] and GAFFLipid [64] for lipids; Glycam [24] for carbohydrates; Åqvist parameterization for ions [65].

2.3 Coarse-Grained Representation of Biomolecules

A coarse-grained (CG) representation of a biomolecule is a simplified representation with respect to its fine-grained description used as reference representation [8; 66]. Usually the reference representation is the all-atom models discussed above. In the previous sentences there is the essence of the coarse-grained procedure but no specific recipe to create robust models. In fact for a given simplified representation of a biomolecule a plethora of coarse-grained representations can be imagined and developed [34].

Some scientists think that also an all-atom representation can be considered a coarse-grained representation in which the electronic degrees of freedom have been integrated out [67]. I don't fully agree with this picture because the possibility to neglect the electronic degrees of freedom is guaranteed by the Born-Oppenheimer approximation under some range of validity. The electrons are 3 orders of magnitude faster than the fastest nucleus (the proton) thanks to fact that their mass is 3 orders of magnitude smaller. Such a clear size- and time-scale separation is not always valid in a coarse-grained representation. For these reasons I think that the theoretical justification of a coarse-grained model resides more on the theory of renormalization group, where the process of coarse-graining is made integrating out degrees of freedom in order to identify effective degrees of freedom [68; 69].

I decided to focus my attention on coarse-grained models constructed in order to match closely atomistic structures (or forces). I will discuss exclusively models developed to treat proteins in simulation, but similar models can be applied to nucleic acids [35] and membranes [70]. Thus, I will first discuss simplified representations based on primary structures of proteins, despite simplified representations based on secondary or tertiary structure exist [69]. I adopted this approach because I think that it is necessary to have an accurate quasi-atomic description of protein structures and dynamics to be put in correspondence with medium to low resolution structures produced nowadays by X-ray crystallography (between 2 to 5 Å) and cryo electron microscopy (up to 20 Å). This simplified representation permits to decrease the number of particles (interacting centers or beads) allowing in such a way a speed up of the force evaluation that, as described before, is the main bottleneck of a molecular dynamics simulation. Furthermore, this allows at the same time a sufficient level of accuracy in the description of physical-chemical properties of the biomolecule under investigation.

From my point of view, a coarse-grained representation of the primary structure is based on the definition of a mapping scheme that permits, for every possible amino acid composing a polypeptide chain, to pass from an all-atom representation to a simplified one having at least one bead per amino acid. Once the coarse-grained representation is defined, it remains to determine the functional form of the potential energy that should describe accurately, for that mapping, protein dynamics preserving secondary and tertiary structures.

I will present in the following a non exhaustive introduction to the main ideas behind the development of a coarse-grained model touching upon the type of mapping, functional forms for the potential energy, parameterization strategies, solvent description and choice of the time step. Relevant reviews and books have been written on this topic, where further information can be found [34; 68; 69; 71–77]. The specific CG force field developed in this thesis will be reported in detail in Chapter 3.

2.3.1 Atomistic to Coarse-Grained Mapping

The types of mapping for a coarse-grained model based on the primary structure have been classified by Tozzini [34] who identified 7 different classes. Here, I propose an extension of that classification in order to take into account coarse-grained models preserving an almost atomic description of the backbone (class 0).

Class 0 in which the backbone is described at all-atom level and 1 to 4 beads are used to describe side chains [78–81]

Class 1 in which an amino acid is described by one bead placed on the C_α [33; 82; 83]

Class 2 in which an amino acid is described by one bead placed on the C_β [84]

Class 3 in which an amino acid is described by two beads: one placed on the C_α an the other placed on the center of mass of the side chain [85]

Class 4 in which an amino acid is described by two beads: one placed on the center of mass of the backbone and the other in the center of mass of the side chain [86]

Class 5 in which an amino acid is described by 1 to 3 beads: one placed on the C_α and from 0 to 2 beads to represent the side chain [87; 88]

Class 6 in which an amino acid is described by 1 to 5 beads: one placed on the center of mass of the backbone and from 0 to 4 beads to represent the side chain [89]

Class 7 in which an amino acid is described by 3 beads: one bead placed on the C_α, one bead on the centroid of the rest of the backbone and on bead for the side chain [8; 66; 90]

All these mappings present advantages and disadvantages, which I won't touch in detail here. However, class 6 presents a good compromise in preserving the steric hindrance of side chains, although this mapping compared to others implies an higher number of beads. Models belonging to class 6 present moreover the best compromise between the two requirements I discussed before, namely speed up of the calculations and sufficient accuracy in describing the physicochemical properties of biomolecules. For this reason I adopted this mapping to develop the original coarse-grained model presented in Chapter 3.

2.3.2 CG Energy Functional Form

Apart for the UNRES force-field [90] the CG functional form that is usually adopted is reminiscent of that used in all-atom potential energy function and usually the general form is as follows:

$$V^{CG} = V^{CG}_{\substack{pseudo\\bonds}} + V^{CG}_{\substack{pseudo\\bendings}} + V^{CG}_{\substack{pseudo\\dihedrals}} + V^{CG}_{vdw} + V^{CG}_{ele} \tag{2.20}$$

where:

$$V^{CG}_{\substack{pseudo\\bonds}} = \sum_{\substack{pseudo\\bonds}} k_{p-b}(b-b_0)^2 \tag{2.21}$$

is used to describe harmonic pseudo bonds among two consecutive beads having force constant k_{p-b} and b_0 equilibrium values b_0;

$$V^{CG}_{\substack{pseudo\\bendings}} = \sum_{\substack{pseudo\\bendings}} \sum_{n=2}^{4} k_{p-a,n}(\omega-\omega_0)^n \tag{2.22}$$

is used to describe pseudo bending angles among three beads with ω_0 equilibrium values and $k_{p-a,n}$ constant forces;

$$V^{CG}_{\substack{pseudo\\dihedrals}} = \sum_{\substack{pseudo\\dihedrals}} \sum_{n=1}^{3} k_{p-d,n}[1+cos(n\chi-\chi_0)] \tag{2.23}$$

is used to describe pseudo-dihedrals having χ_0 equilibrium value, n as multiplicity and $k_{p-d,n}$ force constants;

$$V^{CG}_{vdw} = \sum_{pairs} 4\delta_{ij} \left[\left(\frac{\lambda_{ij}}{r_{ij}} \right)^{m} - \left(\frac{\lambda_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.24}$$

which is a generalized Lennard-Jones with repulsive part having an exponent m between 12 and 8 [89; 91; 92], which is used to mimic the apparent softness of coarse-grained beads[1];

$$V^{CG}_{ele} = \sum_{pairs} \frac{Q_i Q_j}{\epsilon_0 \epsilon_r r_{ij}} \tag{2.25}$$

is a simple Coulombic potential used to describe electrostatic interactions between beads of charge Q_i and Q_j (this is commonly the only simple electrostatic contribution modeled in this class of CG force field). Some time a relative dielectric constant ϵ_r is used in order to effectively screen charge-charge interaction due to a coarse representation of the solvent, or a more explicit solvent representation is used (see below).

2.3.3 Solvent Description

As in atomistic simulations also for CG models is important to consider an accurate treatment of the solvent effects. Here I will give a brief introduction of the approaches that are usually used for the description of the solvent in coarse-grained molecular dynamics simulations.

Implicit solvent representation To describe and reproduce solvent effects and evaluate electrostatic interactions in biopolymers, the concept of effective dielectric constant has been largely explored [93–98]. This is done by using linear or nonlinear distance-dependent dielectric functions, although the fact that they tend to ignore the dielectric boundary between protein and solvent. The linear distance-dependent dielectric function is usually given by the following formula:

$$\epsilon_{\mathbf{eff},L}(r) = 1 + kr \tag{2.26}$$

where k =4. This type of distance-dependent dielectric function has been used for several all-atom force fields, like the first generation of AMBER force field [99]. This approximation is thought to be too crude despite being extremely cheap from a computational point of view [97].

---

[1] Some groups use a Buckingham potential in order to mimic the apparent softness [87]

Nonlinear dielectric treatments are obtained using sigmoidal functions as the following:

$$\epsilon_{\mathbf{eff},S}(r) = \frac{\epsilon_3}{1 + (\frac{\epsilon_3}{\epsilon_2} - 1)e^{-\frac{r}{L_3}}} \tag{2.27}$$

where $\epsilon_2 \sim 5$, $\epsilon_3 \sim 80$, $L_3 \sim 20$ Å. This function approximates the low dielectric constant characteristic for biopolymers at short distances, whereas at large distances approaches the bulk dielectric constant of water. Other sigmoidal functions, for instance using cutoff values of ∼20 at short distances and ∼80 at large distances, have been used [96]. In order to take into account the effect of the dielectric boundary the Tanford-Kirkwood theory has been also used [95]. More accurate distance dependent dielectric functions have been proposed by the group of Mehler, that are connected to the classical dielectric theory of polar solvation [93]. Moreover numerical methods have been developed to describe the real geometry of proteins and solve the linearized Poisson-Boltzmann equation [100].

Explicit Solvent Representation An explicit coarse-grained representation of water molecules consists in grouping atoms belonging to one or more waters [101]. Usually the potential energy function used to describe the water solvent is given by the sum of a Lennard-Jones potential and a electrostatic contribution:

$$V_W^{CG} = V_{LJ} + V_{ele} \tag{2.28}$$

although some models do not make use of a Lennard-Jones (or use a modified Lennard-Jones potential) and the electrostatic contribution is neglected [92; 102]. The electrostatic contribution is mainly used to mimic the dielectric properties of the solvent.The most fruitful approaches used to parameterize this term consist in the creation of polarizable water models. In recent years, this has been done by using a variety of strategies, like the polarization density functional approach [103; 104], the inertial approach [105]; or by giving an internal structure (to the bead of water) with interaction sites free to vibrate and/or rotate around equilibrium values [106–108].

2.3.4 CG Parameterization Strategies

The calibration of coarse-grained parameters can be done in general based on matching strategies of:

1. quantities obtained from simulations performed with a finer-grained model;

2. quantities obtained from experiments;
3. a mix of experimental and computational quantities.

Here I will mainly focus on parameterization based on computational data since this is the approach that I mostly used during my thesis. I will sketch the main ideas that have been followed by other groups in the past, such as force-matching, relative entropy, reverse Monte Carlo and iterative Boltzmann inversion [1]. The main drawback of all these methods is that they assume the reliability of the underlying atomistic models, assumption that is not always true. On the other side, the main benefit is that these models are based on well established theoretical concepts.

Force-matching In the domain of protein simulations, the force-matching strategy has been broadly used mainly by the group of Voth, exploiting the idea proposed earlier to calibrate interatomic potentials from first principle calculations [109]. This method permits to determine an optimized force field describing the interactions between coarse-grained sites from atomistic force data [110; 111]. The assignment of the force-field parameters is done minimizing the residual function of N vector functions:

$$\chi[\mathbf{F}] = \frac{1}{3N} \langle \sum_{l=1}^{N} |\mathbf{F}_l(\mathbf{r}) - \mathbf{f}_l(\mathbf{r})|^2 \rangle \tag{2.29}$$

where: $\mathbf{F}$=$\{\mathbf{F}_1, ..., \mathbf{F}_N\}$ is an arbitrary set of N vector valued functions of the CG configuration on each CG site I; $\mathbf{f}_l = \sum_{i \in I_l} \mathbf{f}_i(\mathbf{r})$ is the net force on the atoms involved in the bead I; and the angular bracket indicates an equilibrium canonical ensemble average evaluated with the atomistic model. The force-matching approach has been further extended generalizing the Yvon-Born-Green theory [112–117]. It can be demonstrated, in the framework of this theory, that for two particles distant $|\mathbf{r}|$ belonging to a system composed by N particles, the total force acting on the first particle is given by two contributions: the direct force from the second particle; the correlated net force from the environment, which is decomposed into contributions from shells of particles at a distance $|\mathbf{r}'|$ away from the first particle [118]. While well founded theoretically, this method has not been able to produce a set of transferable parameters, despite some trials have been attempted [119; 120], determining the need to run always an all-atom simulation in order to perform the coarse-grained simulation.

---

[1] For simplicity I am not going to present cumulant-based expressions of multi-body terms used in the UNRES force-field [90]

Relative Entropy The concept of relative entropy has been exploited for the identification of one model system that best reproduces the features of an existing system called target system [121]. This requires to minimize the relative entropy given by:

$$S_{rel} = \sum_{i=1}^{N} p_{T,i} \ln \frac{p_{T,i}}{p_{M,i}} \tag{2.30}$$

where the summation is over all configurations, p_i is the probability of configuration i in an ensemble, and T and M indicate target and model, respectively. The minimization of the relative entropy is used to assign the values to a collection of adjustable parameters $\{\lambda_1, \lambda_2, ...\}$ from which a model potential energy function U_M depends, using an atomistic target function U_T. This method suffers, at the moment, of the same limitations encountered with the force-matching method.

Reverse Monte Carlo The reverse Monte Carlo has been introduced to renormalize the Hamiltonian of a molecular system given by:

$$H(\mathbf{r}) = \sum_{\alpha=1}^{N} k_\alpha S_\alpha(\mathbf{r}) \tag{2.31}$$

where $S_\alpha(\mathbf{r})$ are functions of particle coordinates $\mathbf{r}$ and k_α are constants defining the interaction potential [122]. By using an iterative procedure it is possible to assign numerical values to k_α using the following formula:

$$k_\alpha^{(n+1)} = k_\alpha^{(n)} + \Delta k_\alpha^{(n)} \tag{2.32}$$

where $\Delta k_\alpha^{(n)}$ is calculated starting from the knowledge of the differences of the trial values and the reference values $\Delta\langle S_\alpha^{(n)}\rangle = \langle S_\alpha^{(n)}\rangle - S_\alpha^*$ of the function one would like to reproduce, and inverting the equation

$$\Delta\langle S_\alpha^{(n)}\rangle = \sum_{\gamma=1}^{N} \frac{\partial\langle S_\alpha^{(n)}\rangle}{\partial k_\gamma} \Delta k_\gamma^{(n)} \tag{2.33}$$

where S_α^* is usually calculated at atomistic level.

Iterative Boltzmann inversion The technique of the iterative Boltzmann inversion is based on the assumption that the total energy V^{CG} can be separated into bonded V_B^{CG} and non-bonded V_{NB}^{CG} contributions as described before. The coarse-grained bonded interactions of the model are determined by sampling the coarse-grained degrees of freedom with atomistic simulations (or structural databases) and calculating the corresponding probability distributions. These distributions are characterized by pseudo-bonds $\{b\}$, pseudo-angles $\{\omega\}$ and pseudo-torsions $\{\chi\}$ giving a total distribution $P^{CG}(\{b\},\{\omega\},\{\chi\},T)$ which is temperature dependent. Assuming uncorrelation of the coarse-grained degrees of freedom the total distribution can be factorized:

$$\begin{aligned} P^{CG}(\{b\},\{\omega\},\{\chi\},T) = {} & P^{CG}(b_1,T)\cdot \ldots P^{CG}(b_{N_1},T)\cdot \\ & P^{CG}(\omega_1,T)\cdot \ldots P^{CG}(\omega_{N_2},T)\cdot \\ & P^{CG}(\chi_1,T)\cdot \ldots P^{CG}(\chi_{N_3},T) \end{aligned} \tag{2.34}$$

for N_1 pseudo-bonds, N_2 pseudo-angles and N_3 pseudo-torsions. The individual probability distributions are Boltzmann-inverted to obtain the corresponding potentials:

$$V^{CG}(b_i,T) = -k_B T \ln\left[\frac{P^{CG}(b_i,T)}{b_i^2}\right] + C_{b_i} \tag{2.35}$$

for the generic bond b_i,

$$V^{CG}(\omega_j,T) = -k_B T \ln\left[\frac{P^{CG}(\omega_j,T)}{\sin\omega_j}\right] + C_{\omega_j} \tag{2.36}$$

for the generic angle ω_j,

$$V^{CG}(\chi_k,T) = -k_B T \ln\left[P^{CG}(\chi_k,T)\right] + C_{\chi_k} \tag{2.37}$$

for the generic torsion χ_k.

The coarse-grained non-bonded potential instead is calculated starting from the knowledge of the atomistic radial distribution functions $g(\mathbf{r})$ associated with the system of interest using the following formula iteratively:

$$V^{CG}_{NB,i+1} = V^{CG}_{NB,i} - k_B T \ln\left[\frac{g_i(\mathbf{r})}{g(\mathbf{r})}\right] \tag{2.38}$$

Strictly speaking, the iterative Boltzmann inversion is mainly used to determine non-bonded potentials [123–126], while simple Boltzmann inversion approaches are used to have potentials describing bonded interactions in biomolecules.

To conclude, CG models have been also parameterized against experimental data. This consists in tuning the values of the model parameters using a test-and-trial scheme in order to reproduce some properties for which experimental data are available [92; 106; 107]. Typical properties than one would like to reproduce (mainly for surfactants) are density, surface tension, solvent accessible surface area, relative static dielectric permittivity and partition properties. One possible future approach to enhance the reach of these approaches could be to exploit efficient optimization strategies in order to automatize the parameterization step to reproduce system properties.

2.3.5 Choice of a Realistic CG Time Step

The choice of the time step is in general of paramount importance for any reliable molecular dynamics simulation. In the case of coarse-grained simulation the choice can be complicated by the different levels of coarse-graining one decide to adopt. This topic is still matter of debate and not clear and unique recipe has been adopted in literature despite some suggestions have been proposed by van Gunsteren and coworkers [107; 127–129]. Following these ideas I will discuss about the choice of the proper time step to use for a CG simulation, based on a very well known topic, the harmonic oscillator. In fact, every coarse-grained potential is in part a collection of harmonic oscillators having their own frequency of vibration given by:

$$\omega_{\mathbf{CG}} = \frac{2\pi}{T_{\mathbf{CG}}} = \sqrt{\frac{k_{\mathbf{CG}}}{\mu_{\mathbf{CG}}}} \tag{2.39}$$

where $T_{\mathbf{CG}}$ is the time of oscillation, $k_{\mathbf{CG}}$ the force constant and $\mu_{\mathbf{CG}}$ the reduced mass. Thus, the minimum oscillation time $T_{\mathbf{CG}}^{min}$ of a coarse-grained model should be compared with the minimum oscillation time $T_{\mathbf{AA}}^{min}$ of an all-atom potential. In the case of the AMBER force field this is given by the bond between carbon atoms and nitrogen atoms of purines (k=529 kcal$\cdot$mol$^{-1}\cdot$Å^{-2}, μ=6.47 amu), $T_{\mathbf{AA}}^{min}$=$0.34 \cdot 10^{-13}$s, thus that a δt_{AA}=$2 \cdot 10^{-15}$s is commonly used [1].

One can in principle calculate given a CG mapping, the $T_{\mathbf{CG}}^{min}$ of the fastest CG bond oscillation. The $T_{\mathbf{CG}}^{min}$ of the coarse-grained model presented in present thesis is given by the bond between the backbone bead and a tyrosine bead (k=82 kcal$\cdot$mol$^{-1}\cdot$Å^{-2}, μ=17.78amu), $T_{\mathbf{CG}}^{min}$=$1.43 \cdot 10^{-13}$s.

The scaling factor f_{CG} given by the fraction:

$$f_{\mathbf{CG}} = \frac{T_{\mathbf{CG}}^{min}}{T_{\mathbf{AA}}^{min}} \tag{2.40}$$

should provide a realistic multiplication factor to the all-atom time step to have a realistic time step to be employed in a coarse-grained simulation:

$$\delta t_{\mathbf{CG}} = f_{\mathbf{CG}} \cdot \delta t_{\mathbf{AA}} \tag{2.41}$$

---

[1] Whitout considering bonds involving hydrogen atoms

In the specific case of the coarse-grained presented here we thus obtain $\delta t_{\mathbf{CG}}$=$8.34 \cdot 10^{-15}$s, that should be considered as an upper limit [1]. If one uses a constraint algorithm like SHAKE [130], it could be acceptable to have $\delta t_{\mathbf{CG}}$=$10.0 \cdot 10^{-15}$s. The time steps used by other groups is between the two extreme cases of 1 fs [112; 121] and 40 fs [89]. The use of a time step of 1 fs seems to be too conservative considering that 2 fs is the all-atom time step. Others groups [79; 87; 90] use a time step of 5 fs that seems to be reasonable considering the level of adopted granularity of their model. Recently the group of van Gunsteren numerically demonstrated that the time step of 40 fs should be avoided in order to preserve the thermodynamic properties of the model system, proposing an upper limit of 10 fs for Lennard-Jones interacting particles having a mass of four time that of one single water molecule [127].

2.4 Analysis of Protein Structure and Dynamics

In this section, I will report the general aspects of some of the analyses performed on the MD trajectories at the atomistic and CG level. I will talk about analyses of structural properties, structural fluctuation properties, principal component analysis, autocorrelation functions and solvation dynamics. Altogether these analyses permit to give a description of protein dynamics giving an exhaustive representation of it.

2.4.1 Structural Properties

Among the structural properties I considered the root mean square deviation (RMSD) and gyration radius (R_g). I used these analyses to check the validity of my coarse-grained simulations (Chapter 3) and monitor the structural behaviour of atomistic MD simulations of ubiquitin (Chapter 4).

Root mean square deviation

The RMSD is generally used to quantify the structural deviation of the protein structure during molecular dynamics with respect to a reference structure that (usually) is experimentally resolved with X-ray crystallography or NMR techniques. The RMSD is given by the following formula:

$$RMSD(t_j, t_0) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i(\mathbf{r}_i(t_j) - \mathbf{r}_i(t_0))^2} \tag{2.42}$$

---

[1] In fact in the present thesis $\delta t_{\mathbf{CG}}$=$5.0 \cdot 10^{-15}$s

where

$$M = \sum_{i=1}^{N} m_i \tag{2.43}$$

is the total sum of the masses of the atoms/beads of the protein, and where $\{\mathbf{r}(t_j)\}$ represents the collection of atom/bead positions for the protein at time t_j, $\{\mathbf{r}(t_0)\}$ represents the collection of atom/bead positions of the reference structure.

Gyration radius

The gyration radius is calculated to quantify how compact is a globular protein. Its formula is given by:

$$R_g(t_j) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i(t_j) - \mathbf{r}_{com}(t_j))^2} \tag{2.44}$$

where $\mathbf{r}_{com}$ is the center of mass of the protein.

2.4.2 Structural fluctuation properties

Among the structural fluctuation properties I considered the root mean square fluctuation (RMSF) and S^2 order parameter that permit to quantify the level of dynamic flexibility of the protein. I used these analyses to check if my coarse-grained model was able to reproduce all-atom MD results (see Chapter 3).

Root mean square fluctuation

The RMSF is calculated usually for the C_α carbons using the following formula:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t_j=1}^{T}(\mathbf{r}_i(t_j) - \langle\mathbf{r}_i\rangle)^2} \tag{2.45}$$

where T is the total amount of time simulated, $\langle\mathbf{r}_i\rangle$ is the average position of C_α carbon i.

S^2 order parameter

The S^2 order parameter provides information about angular amplitude of the internal motion of proteins [131]. The S^2 order parameter for the backbone dipole moment given by the N and H atoms of the amide bond is considered. The S^2 order parameter is then calculated using the formula [132; 133]:

$$S_i^2 = \frac{1}{2}\Big[3\sum_{\alpha=1}^{3}\sum_{\beta=1}^{3}\Big\langle\mu_{i,\alpha}\mu_{i,\beta}\Big\rangle^2 - 1\Big] \tag{2.46}$$

where $\mu_{i,\alpha}$ (α=1, 2, 3) are the x, y and z components of the normalized dipole moment.

2.4.3 Principal Component Analysis

The principal component analysis (PCA), also called covariance analysis, is a mathematical technique for analyzing high-dimensional data sets, that allows for a reduction of the dimensionality of the space concentrating on the coordinates with larger spread or fluctuations. This procedure for N-dimensional data $\mathbf{r}$(t) is based on the construction of the covariance matrix C_{ij}, for atoms i and j, defined as

$$C_{ij} = \langle(\mathbf{r}_i - \langle\mathbf{r}_i\rangle)(\mathbf{r}_j - \langle\mathbf{r}_j\rangle)\rangle \tag{2.47}$$

where $\langle\rangle$ is the average over all data points. The symmetric NxN matrix C can be diagonalized with an orthonormal transformation matrix R:

$$R^T C R = diag(\lambda_1, \lambda_2, \ldots, \lambda_N) \tag{2.48}$$

where the column of R are the eigenvectors or principal modes. The eigenvalues λ_i are equal to the variance in the direction of the corresponding eigenvector. The original data can be projected on the reference system generated by the eigenvectors to give the so-called principal components p_i, $i = 1, \ldots, N$:

$$\mathbf{p} = R^T(\mathbf{r} - \langle\mathbf{r}\rangle) \tag{2.49}$$

In molecular dynamics of proteins, the data are the result of a dynamic process, so the principal components are a function of time and $p_1(t)$ is the principal component with the largest mean square fluctuation [134]. This type of analysis can be also used to study large structural ensembles [135].

Hess analysis of conformational friction

The analysis of the principal components allows a rough estimation of the effective conformational friction coefficient experienced by a protein during a molecular dynamics simulation [136; 137]. In practice, windows of the protein dynamics, extracted from MD trajectories, are identified on the basis of a structural property, or PCA for which the protein is exploring one single minimum. For the identified window one carries out PCA, calculating the principal components p_i, $i = 1, \ldots, N$, storing the correspondent eigenvalues and assuring that they are gaussian distributed. For each of these principal components one can estimate the harmonic force constant in the direction of the principal component i using the following formula:

$$k_i = \frac{k_B T}{\lambda_i} \tag{2.50}$$

where k_B is the Boltzmann's constant, T the chosen temperature of the simulation, λ_i the eigenvalue of the principal component i ($[\lambda]$=nm^2). Typical values of k are between 100 and 800 kJ mol^{-1} nm^{-2} [136]. For all the principal components one has to calculate the auto-correlation function using the usual definition:

$$C_{p_i}(t') = \frac{\frac{1}{t^{max}} \sum_{t_0=1}^{t^{max}} \frac{1}{N_{occ}} \sum_{j=1}^{N_{occ}} p_i(t_0 + t') \cdot p_i(t_0)}{\frac{1}{t^{max}} \sum_{t_0=1}^{t^{max}} p_i(t_0) \cdot p_i(t_0)} \tag{2.51}$$

that is calculated for all the possible starting time t_0 and where t' is temporal scale evaluated. The fit of 2.51 with a linear or exponential function permits then to give an estimation of the decay time τ of that principal component. Typical values of τ are between 15 to 10000 ps. Finally the estimation of the friction coefficient η for the particular minimum under study is given by the following formula:

$$\eta = k\tau \tag{2.52}$$

with η having values ranging from 10^3 to 10^7 $\frac{amu}{ps}$ [136; 137].

2.4.4 Auto correlation functions

Several auto correlation functions can be calculated to give a quantitative picture of protein dynamics. Here I will focus my attention on the conformational auto correlation function.

The internal conformational diffusion can be estimated calculating the following quantity:

$$< \Delta\mathbf{r}(t')^2 >= \frac{1}{t^{max}} \sum_{t_0=1}^{t^{max}} \frac{1}{N_{atoms}} \sum_{i=1}^{N_{atoms}} (\mathbf{r}(t_0 + t') - \mathbf{r}(t_0))^2 \tag{2.53}$$

that is calculated for all the possible starting time t_0. The fitting of this function permits the estimation of the effective diffusion coefficient in the internal conformational space of the protein and should deviate from the Einstein's relation in the sense that:

$$< \Delta\mathbf{r}(t')^2 > \propto t'^{\beta} \tag{2.54}$$

where β <1 which is called sub-diffusive regime [138]. I will show in Chapter 4 that this is the case for the proteins under study in the present thesis.

2.4.5 Describing Solvation Dynamics

The survival probabilities of water-protein contacts provide a quantitative estimation of the time-scales of their interactions. This estimation relies on the function $P_j(t_n, t)$, which takes the values of 1 if the jth water molecule is within a certain cut-off of the protein between time t_n and $t_n + t$, and zero otherwise [139]. Averaging out this function over the simulation time and all the water molecules, one can write the survival probability as follow:

$$N_w(t) = \frac{1}{N_t} \sum_{n=1}^{N_t} \sum_{j} P_j(t_n, t) \tag{2.55}$$

where N_t is the number of the simulation time-frames.

The protein hydration shell is approximated directly from the van der Waals parameters of the atomistic force field in use. Only water molecules with atoms within a distance less than R_{cut} of any protein's atom contribute to the first hydration shell. A different cut-off is chosen for any couple of water molecule and protein atoms, w and p, respectively, such that:

$$R_{cut} = f(r_w + r_p) \tag{2.56}$$

where r_w and r_p are the van der Waals radii of the atoms obtained from the force field. f is instead a coefficient, set to f=1.1, used to approximate R_{cut} to at least the first minimum of the protein-water pair correlation function. $N_w(0)$ gives the average number of hydration water molecules, or hydration number, $N_w(t_{sim})$ is instead the number of water molecules bound to the system for the whole simulation [140].

The fit of the calculated water survival probability is usually done by one stretched exponential combined with two or three simple exponential using the following formula [141]:

$$N_w(t) \simeq n_s e^{-\left(\frac{t}{\tau_s}\right)^{\gamma}} + \sum_{i=2}^{4} n_i e^{\left(-\frac{t}{\tau_i}\right)} \tag{2.57}$$

where the first exponential is called stretched exponential because usually $\gamma \leqslant 1.0$ and the residence time for it can be computed as average relaxation time:

$$\langle \tau_s \rangle = \frac{\tau_s}{\gamma} \Gamma\left(\frac{1}{\gamma}\right) \tag{2.58}$$

The temporal scales of the water survival on the surface of the protein are given by $\langle \tau_s \rangle$ and τ_i with i=2 to 4, and the corresponding number of waters that belong to that particular scale are n_s and n_i with i= 2 to 4. Typical temporal scales of water survival on the protein surface is somehow universal because it has been numerically demonstrated that: $< \tau_s > \leqslant$ 30ps, 35 $\leqslant \tau_2 \leqslant$ 200ps, 300 $\leqslant \tau_3 \leqslant$ 1000ps and $\tau_4 \geqslant$ 10 ns [140; 141]. This analysis was applied to the study of the effects of solvent and crowding agents for ubiquitin dynamics (see Chapter 4).

Chapter 3

Development of a coarse-grained model for numerical simulations of proteins

> *The idea of using a simplified model in computational studies of proteins dates back to Levitt & Warshel's (LW) simplified model for protein folding [77].*
>
> Arieh Warshel

3.1 Preface

In the present chapter I will present the development of a coarse-grained (CG) model that accounts for a simplified electrostatic description of soluble proteins. The motivation behind this choice is based on the fact that electrostatics is of fundamental importance for an accurate description of secondary, tertiary and quaternary structures of proteins. This CG model has been developed with the aim of satisfying two main requisites of a CG representation, namely to speed-up the simulations with respect to the all-atom representation, and to reproduce structural and dynamic properties in sufficient agreement with all-atom results.

The present chapter is adapted from an article accepted for publication in the *Journal of Chemical Theory and Computation* with the title ELECTROSTATIC-CONSISTENT COARSE-GRAINED POTENTIALS FOR MOLECULAR SIMULATIONS OF PROTEINS. Enrico Spiga[†], Davide Alemani[†], Matteo Thomas Degiacomi[†], Michele Cascella[‡], Matteo Dal Peraro[†] († Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne-EPFL, Lausanne, CH-1015, Switzerland; ‡ Departement für Chemie und Biochemie, Universität Bern, Freiestrasse 3, Bern, CH-3012, Switzerland)

3.2 Introduction

Fundamental biological processes at the molecular level involve macromolecular assemblies of the most different sizes, and can occur in a broad spectrum of time and length scales [142]. The most straightforward and accurate way to study these processes by computational approaches is to develop and use models at an atomistic level of resolution. Atomistic models for biomolecules have been successfully applied in the past decades [143], and are today at the most advanced frontier of biomolecular simulations [26; 28; 29; 31]. To date, the boundaries for atomistic simulations have reached the millisecond timescale and passed the million of atoms [19; 144–146]. Despite the current progress and success of all-atom simulations, the computational cost for this resolution remains challenging for the routine study of larger systems and for longer timescales. Following this objective, a variety of simplified models have been proposed in the last decades [8; 66]. In particular, coarse-grained (CG) Hamiltonians have been introduced to describe macromolecular systems. The first CG models focused on simple hydrophobic-polar interactions, which led to a series of models for surfactants-water or lipid-water mixtures [36; 37; 124; 147]. In more recent years, the same approach has been adopted for the development of multi-scale methods and CG models for proteins and nucleic acids, used for the investigation of a large variety of processes [34; 69; 71; 73; 75; 76; 79; 106; 148–152].

CG models for proteins are based on structural topologies that map the atomistic dimensionality to a given CG resolution, and on effective potentials able to reproduce the interactions of the original atomistic representation. The different strategies to derive CG potential parameters discussed in the literature are mostly based on mining degrees of freedom through Boltzmann inversion techniques, thermodynamic integration, force matching or cumulant-based descriptors [90; 111; 124; 153]. In fact, it is extremely difficult to derive a general, multi-resolution rigorous coarse-graining theory that would be able to generate a consistent and transferable CG force field at any given level of resolution. In the recent past, several steps towards this goal have been reported in the literature [111; 112; 121]. Force matching strategies have been particularly successful in determining effective coarse-grained potentials from atomistic simulations [110; 154]. This approach has been applied to the study of several problems ranging from the folding of small peptides [155] to the simulation of immature HIV-1 virion [156]. Moreover, tentatives to overcome the problem of transferability of the potential parameters have been done checking the correspondence of parameter values among systems [120] or adopting a "host and guest parameterization strategy" with consequent CG MD simulations to quantify the transferability of the parameters [119]. With the same objective, using the Yvon-Green-Born integral equation it has been possible to treat many-body structural correlations with the aim to determine more transferable potentials for folded proteins [116]. Using the concept of relative entropy to guide the parameterization procedure, promising results have been recently obtained for the study of large-scale fibrillar assembly [157]. Still, many issues afflict current CG schemes, which limit their general applicability to a large class of relevant biological problems. The functional form at CG level is not univocally defined, and in principle should explicitly treat many body effects [158] or polarization terms. Also, optimal mapping schemes able to ensure the accuracy of the potentials in reproducing particular properties of interest [159] should be applied. In practice, effective schemes able to overcome some of these problems producing CG force fields adapted to tackle specific biological problems are present in the literature. For example, the MARTINI force field for proteins and lipids, which is characterized by good transferability, has been developed using a combination of free energy based calculations and Boltzmann inversion [89; 153]. This model was successfully applied to membrane simulations and showed great potential for membrane proteins investigations [160–163]. One drawback of the MARTINI force field lies in the requirement of external biases to preserve secondary structure elements, limiting the possibility to explore phenomena associated to secondary structure transitions. Another transferable coarse-grained model with dipolar backbone contributions has been applied to small folded proteins, showing promising results for the description of structural fluctuation properties at this level of resolution [87]. The investigation of large conformational changes has been successfully addressed by cumulant-based approaches for the definition of effective multi-body potentials. Using such an approach it has been possible to quantify correlations between local and non-local interactions, creating an united-residue force-field [90; 164]. This force-field has been applied to the folding of α- and α/β-proteins [165] and the opening and closing of Hsp70 chaperones [166]. Preserving an atomistic description of the backbone, while only the side chains are coarse-grained, is an effective solution to explore and stabilize the secondary structure elements. Not surprisingly, CG models obtained using this strategy are widely used to study the aggregation of amyloidoigenic peptides, the folding of small peptides and refinement of protein structures [78–80; 88; 91; 167; 168]. On the other side, while improving on secondary structure stability, this approach departs from a uniform CG mapping and introduces additional degrees of freedom at the backbone level to create an almost atomistic representation.

My host group has recently proposed a strategy to describe the backbone contribution while preserving a single bead representation. The introduction of the explicit backbone dipole defined by three consecutive C_α beads along with the treatment of non-radial dipole-dipole interactions in dynamics allowed to obtain stable secondary structure elements of unspecific poly-peptides and to sample conformational transitions [83]. In this work, I introduce a description of side chains electrostatics and extend this model to real proteins. The electrostatic contribution is explicitly considered to the second order of the multipole expansion, so that the remaining part of the non-bonded interactions are responsible only for short-ranged contributions. As previously demonstrated [169], this approach has the benefit to better describe the actual electrostatic field at a CG level with implications to the treatment of molecular recognition in protein-protein interactions (PPIs). Moreover, as shown for short peptides [83], within this approach the secondary structure of a variety of folding motifs is naturally maintained.

Following a parameterization protocol that combines Boltzmann inversion schemes and force-matching methods, I tuned the bonded and non-bonded terms of the backbone and side-chain beads for a broad range of different structural protein folds, using a novel algorithm based on particle swarm optimization [170–174]. The resulting set of electrostatic-consistent potentials allowed simulating soluble proteins, and protein-protein complexes at the sub-microsecond scale in few days, preserving a large structural and dynamic agreement with respective all-atom simulations and experimental reference structures. Moreover, the proposed scheme of parameterization provides at the same time an inexpensive way to quickly derive CG parameters for protein systems, and can eventually contribute to the generation of more transferable parameters to be use in a general multipurpose transferable CG force field.

3.3 Methods

3.3.1 Coarse-graining the atomistic structure and electrostatics of proteins

The present CG model is based on an approximately four-to-one atoms-to-bead mapping, consistent to that used by other CG force fields (*e.g.* MARTINI [89; 92; 153]). On average, four heavy atoms are represented by a single interaction center, with the exception of aromatic side chains, where a higher resolution to map their geometric specificity has been used (Fig. 3.1). All the amino acids are composed by a bead representing the backbone and placed on top of the C_α atom, whereas one or more beads are used for the side chain, which are placed at the center of mass of their constituent heavy atoms. Alanine and glycine amino acids constitute the only exception, each being composed by a single bead. The mass of each bead is constituted by the total mass of all the respective atoms (Fig. 3.1, Tab. 3.1). Amino- and carboxy- terminal backbone beads are described bringing their respective zwitterionic charge, and with the corresponding different total mass with respect to normal backbone beads. Apart from the massive beads, this CG structure presents multiple electrostatic centers bringing a multipolar expansion of the corresponding all-atom electrostatic potential arrested to the dipolar term. In particular, it has been introduced mono-polar charges and/or permanent electrostatic dipoles at all charged/polar side-chains as well as at each peptide-bond of the backbone (Fig. 3.1 and 3.2). The backbone dipoles are embedded in the structure of the polypeptide chain following a previous work [83].

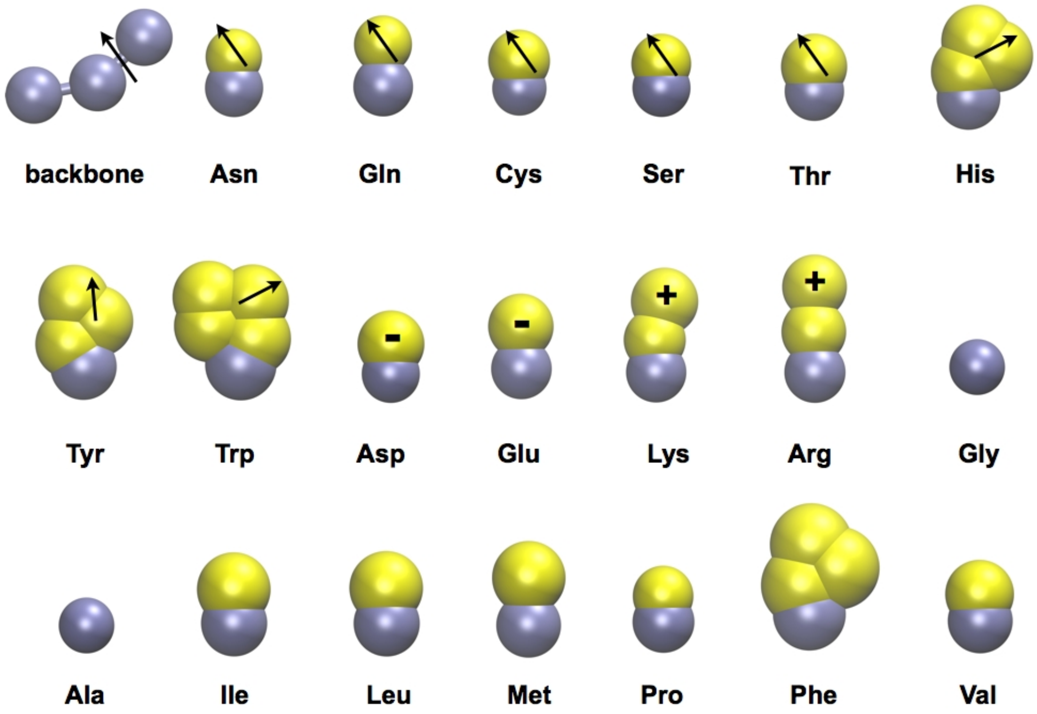

Figure 3.1: Coarse-grained representation of amino acids used in this work. Backbone beads are represented in ice-blue, side-chain beads are in yellow, arrows represent the electrostatic dipole moments associated with polar side-chains and the backbone. Acidic and basic amino acids carry net unitary charges.

Table 3.1: List of beads types

Bead type	description	mass [amu]
BA	alanine bead	71.142
BU	generic backbone bead	56.107
SR1	arginine side chain bead type 1	57.159
SR2	arginine side chain bead type 2	44.183
SN	asparagine side chain	58.123
SD	aspartate side chain	58.036
SC	cysteine side chain	47.095
SE	glutamate side chain	72.063
SQ	glutamine side chain	72.150
SH1	histidine side chain bead type 1	26.038
SH2	histidine side chain bead type 2	28.097
SI	isoleucine side chain	57.116
SL	leucine side chain	57.116
SK1	lysine side chain bead type 1	42.018
SK2	lysine side chain bead type 2	31.121
SM	methionine side chain	75.154
SF1	phenylalanine side chain bead type 1	26.038
SF2	phenylalanine side chain bead type 2	35.548
SP	proline side chain	42.081
SS	serine side chain	31.034
ST	threonine side chain	45.061
SW1	tryptophan side chain bead type 1	26.038
SW2	tryptophan side chain bead type 2	28.097
SW3	tryptophan side chain bead type 3	38.049
SY1	tyrosine side chain bead type 1	26.038
SY2	tyrosine side chain bead type 2	41.554
SV	valine side chain	43.089

3.3.2 Coarse-graining the potential function

For the CG protein representation it is adopted an additive potential function as typically used in all-atom Hamiltonians. This approach provides the best compromise be-

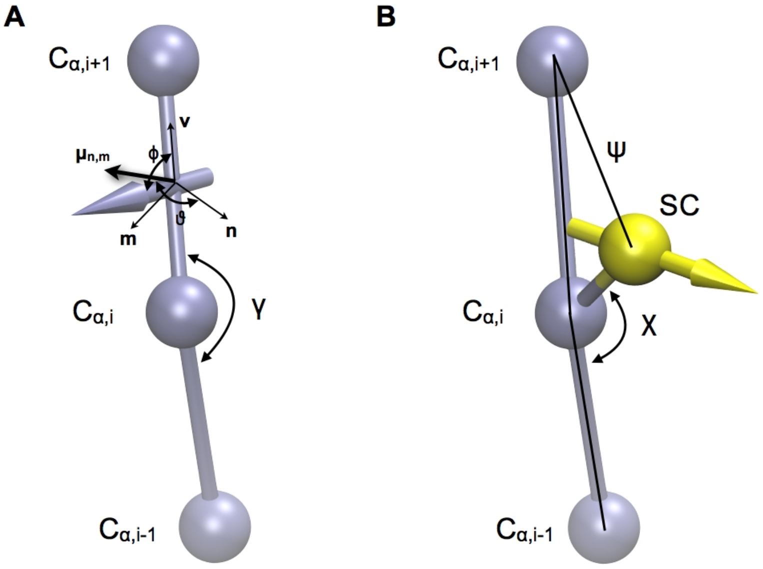

Figure 3.2: Representation of electrostatic dipole moments. (A) **v**, **n**, **m** are the vectors of the internal orthonormal basis used to reconstruct the backbone's dipole; $\mu_{m,n}$ is the projection of the dipole in the plane **m**, **n**; ϕ is the angle between the backbone dipole and the vector **v**; θ is the angle between the projection $\mu_{m,n}$ and the vector **n**; γ is the bending angle of the three consecutive C_α used to reconstruct the backbone dipole [83]. (B) SC is the side chain bead with the associated dipole drawn as an arrow; χ is the bending angle used to describe reorientation of the side chain; ψ is the improper torsion used to force the chirality of L-amino acids.

tween reasonable accuracy and computational efficiency for CG simulations [34]. The explicit introduction of electrostatic dipolar terms following [169] adds to the potential minimal many-body contributions, which enhances the stability of secondary structure elements without the use of *ad hoc* bias potentials [83]. The total potential function is given by:

$$
\begin{aligned}
V_{total} \quad &= \sum_{bonds} k_{\mathrm{b}}(|\vec{r}_{ij}| - r_0)^2 + \sum_{bendings} \sum_{n=2}^{4} k_{a,n}(\theta_{ijk} - \theta_0)^n + \\
&\sum_{dihe} \sum_{n=1}^{3} k_{d,n}[1 + \cos(n\phi_{ijkl} - \phi_0)] + \sum_{improper} \sum_{n=2}^{4} k_{i,n}(\psi_{ijkl} - \psi_0)^n + \\
&\sum_{pairs} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{|\vec{r}_{ij}|} \right)^{12} - \left(\frac{\sigma_{ij}}{|\vec{r}_{ij}|} \right)^{6} \right] + V_{\mathrm{el}}(\vec{r}_{ij}) \qquad (3.1)
\end{aligned}
$$

where the first four terms describe bonded interactions and the remaining are used to describe non-bonded ones. In particular, the first term describes pseudo-bonds between backbone beads and beads that belong to the same residue using a simple harmonic approximation, where r_0 is the equilibrium value and k_b is the constant force. The second term accounts for pseudo-bending for backbone and side chain beads [82], with θ_0 the equilibrium value and $k_{a,n}$ the constant forces.

The third for torsion potential of pseudo-dihedrals [82] for backbone and multi-bead side chains, with ϕ_0 as the equilibrium values and $k_{d,n}$ as the constant forces. The final term of the bonded potential describes improper torsion potentials used to force the L-chirality to the side chains or to force the planarity of aromatic side-chains where ψ_0 is the equilibrium value and $k_{i,n}$ the force constant.

The last two terms represent the non-bonded part of the total potential function: a common 6-12 Lennard-Jones potential is used to account for effective non-bonded interactions not explicitly included in the electrostatics potential term. The electrostatic potential, instead, reads:

$$V_{\mathrm{el}}(\vec{r}_{ij}) = C(|\vec{r}_{ij}|)\Big[V_{q_i q_j} + V_{q_i \mu_j} + V_{\mu_i q_j} + V_{\mu_i \mu_j}\Big] \tag{3.2}$$

where all charge-charge, charge-dipole, and dipole-dipole interactions are considered. An implicit solvent model with a distance dependent dielectric constant was used:

$$\epsilon(|\vec{r}_{ij}|) = 1 + k_{ij}(|\vec{r}_{ij}|) \tag{3.3}$$

where k_{ij}=4 [94]. This dielectric model has been chosen on the basis of its simplicity. However, more accurate methods have been proposed [93], and will be tested in the future (see Chapter 2).

Along with this simple distance-dependent screening, an explicit solvent model following the work of Warshel [103] and Borgis [104] was also implemented and tested. In this model four water molecules are mapped into one single water bead. The electrostatics problem of a dipolar or charged molecular solute immersed in a dielectric medium is described by a local non-equilibrium solvation free energy ΔF_{pol}, which is numerically integrated by discretizing the solvent region in "water" grains. Each water grain is associated with an induced dipole $\vec{p}_i$ and the electrostatics field $\vec{E}_{0i}$, generated by the solute. In this framework, the free-energy of solvation is given by:

$$\Delta F_{pol} = \sum_{i=1}^{N} \frac{\vec{p}_i^{\,2}}{2\alpha} - \sum_{i=1}^{N} \vec{p}_i \cdot \vec{E}_{0i} \tag{3.4}$$

thus, the local solvent model is obtained by minimizing the functional relative to all the $\vec{p}_i$, obtaining the equilibrium dipole moment, $\vec{p}_{i,eq}$

$$\vec{p}_i = \frac{p_{sat}}{|\vec{E}_{0i}|} L\big(3\alpha \frac{|\vec{E}_{0i}|}{p_{sat}}\big) \vec{E}_{0i} \tag{3.5}$$

where

$$L(x) = \coth(x) - \frac{1}{x} \tag{3.6}$$

is the Langevin function. At the minimum the electrostatic free-energy is given by the following:

$$\Delta F_{pol}^{min} = -\frac{p_{sat}^2}{3\alpha} \sum_{i=1}^{N} \ln \Big[\frac{\sinh(L^{-1}(3\alpha|\vec{E}_{0i}|/p_{sat}))}{L^{-1}(3\alpha|\vec{E}_{0i}|/p_{sat})} \Big] \tag{3.7}$$

Where $\alpha = 2.3$ is the polarizability and p_{sat}=1.5 D is the dipole saturation, that are model parameters. This model is computationally efficient: electrostatics calculations are limited to solvent-solute interactions and solvent-solvent interactions are short-ranged Lennard-Jones ones. It has been successfully applied by the Ha-Duong's group for the simulation of protein and protein-protein recognition [175–177].

3.3.3 Integrating dipole dynamics

Backbone dipoles are univocally defined by the C$_\alpha$ trace of the protein. The backbone dipole μ_i is associated to a triplet of consecutive C_α beads $(i-1, i, i+1)$-th, being located at the middle point between the second and the third bead, and its orientation determined by the angle of the bead triplet [83; 169]. At each time step during the MD integration loop, the forces on each backbone dipole are calculated and distributed on the beads of the corresponding triplet, affecting their position and the amplitude of the relative triplet bending angle [83] (Fig. 3.2 A).

The orientation of the dipoles associated to the polar side-chains (Fig. 3.2 B) are updated by solving the classical equation:

$$\frac{d\vec{\mu}}{dt} = \vec{\omega} \times \vec{\mu} \tag{3.8}$$

where $\vec{\mu}$ is the dipole moment and $\vec{\omega}$ is the angular velocity of the side-chain. The angular velocity of the side-chain is determined by its inertia tensor $\bar{I}$, derived from

Table 3.2: Bead types and their corresponding values of dipole moment $|\vec{\mu}|$, and ellipsoids of inertia I_x, I_y and I_z [178; 179]

Bead type	$\lvert\vec{\mu}\rvert$ [eÅ]	I_x [Å]	$\mathrm{I}_{y,z}$ [Å]
SN	0.845	2.68	2.03
SC	0.497	2.76	1.88
SQ	0.809	3.24	2.08
SH2	0.736	3.39	2.14
SS	0.381	2.28	1.77
ST	0.373	2.72	1.93
SW2	0.393	4.23	2.52
SY2	0.299	3.87	2.34

the respective all-atom representation, and the electrostatic torque experienced by the dipole (Tab. 3.2), following the equation:

$$\vec{\tau} = \bar{I}\frac{d\vec{\omega}}{dt} \tag{3.9}$$

For accuracy reasons, the equations 3.8, 3.9 are implemented in the MD loop using the quaternion formalism [39]. The side chain electrostatic dipole moments are treated as rigid bodies that can rotate around a fix point (the center of mass of the bead which they belong). The module of the electrostatic dipole moment represents the actual value as obtained from quantum chemical calculations [178] (Tab. 3.2). An analysis of a non redundant subset of NMR structures of the Protein Data Bank, once CG-mapped, permitted to find that there is no strong preferential orientation for the electrostatics dipole moment of the side chains like asparagine, glutamine, serine, threonine and cysteine. In the case of tyrosine instead two preferential orientations have been found with respect to the plane of the ring, while for histidine and tryptophan the dipole remains rigidly linked to the plane of the ring. Ellipsoids of rotations as calculated using all-atom molecular dynamics simulations [179] were used to describe the reorientation of the electrostatic dipole moment of the beads. In this model the electrostatic dipole moment is assumed to be aligned along the direction of the inertia axes with highest eigenvalue of the inertia tensor. The procedure for the backbone [83] and side chain dipolar dynamics has been implemented in the molecular dynamics code Lammps [180].

3.3.4 Parameterizing the coarse-grained potentials

A two-step procedure was adopted to derive reliable values of the parameters for the potential function discussed above. A first general set of bonded parameters is derived using a Boltzmann inversion approach [124] on structural ensembles extracted from the PDB and from MD simulations; then, bonded and non-bonded parameters are refined for a given protein using a force matching procedure [109; 111].

Boltzmann Inversion Based on the adopted CG mapping scheme (Fig. 3.1), from every atomistic degrees of freedom the relative CG conformational distributions of bonded terms are obtained. In particular, the individual CG distributions for bond lengths $\{r\}$, bending angles $\{\theta\}$ and torsions $\{\phi\}$, $P^{CG}(\chi_i, T)$ are Boltzmann inverted [124; 126] to obtain the corresponding potentials for the generic degree of freedom $\chi_i = r_i, \theta_i, \phi_i$, using the equation:

$$V^{CG}(\chi_i, T) = -k_B T \ln \left[\frac{P^{CG}(\chi_i, T)}{f(\chi_i)} \right] + C_{\chi_i} \tag{3.10}$$

where $f(\chi_i)$ is a function that takes into account the components of the Jacobian determinant.

This procedure was performed on a non-redundant subset of the PDB in order to identify all the possible coarse-grained degrees of freedom for the adopted mapping and possible preferential orientations of side-chain dipoles. The NMR part of the subset has been analyzed to identify possible preferential orientation of polar side chains. Since not all amino acids are equally represented in the PDB and some degrees of freedom could have poor statistics (*e.g.* degrees of freedom for tryptophan), additional probability distributions from all-atom MD simulations of all possible homo-nona-peptides in explicit solvent in both α- and β-conformation have been extracted. The conformational distributions obtained from the two sets, at least for the most represented degrees of freedom, are in qualitative agreement (*e.g.* position of the minima for the potentials).

In order to refine, and eventually converge to more transferable parameter values, this initial seeding set was optimized using a force matching procedure on a given protein to further tune the bonded and parameterize the non-bonded interatomic potentials based on the trajectories obtained from all-atom MD simulations. It is important to notice at this point that the Coulomb term for charge and dipole interactions is not affected by the parameterization procedure. Intramolecular electrostatic and Lennard-Jones interactions between charges and/or dipoles separated by one or two bonds (1-3 electrostatic interactions) have been excluded from our potential energy function.

Force matching The force matching procedure has been widely discussed in several publications [109; 110; 154; 155]. Here its main steps will be recalled. Let $\{\omega\}$ indicate the entire set of L parameters $\{\omega_1,\dots\ \omega_L\}$ used to define the potential function adopted for the coarse-grained representation. The optimal $\{\omega\}$ set defining the CG potential function is the one minimizing the fitness function $Z_F(\omega)$:

$$Z_F(\omega) = \sqrt{\left(3\sum_{k=1}^{M} N_k\right)^{-1} \sum_{k=1}^{M}\sum_{i=1}^{N_k} \left|F_{ki}(\omega) - F^0_{ki}\right|^2} \tag{3.11}$$

where M is the number of sets of atomic configurations available, N_k is the number of beads in configuration k, $F_{ki}(\omega)$ is the force on the i-th bead in set k obtained with parametrization ω, and F^0_{ki} is the reference force acting on the bead as given by the following formula:

$$F^0_{ki} = \sum_{j=1}^{L_i} F^0_{jki} \tag{3.12}$$

which is the sum of the forces acting on the atoms that belong to the i-th bead. All quantities are averaged for a large set of different configurations, sampled from a preceding all-atom MD run.

Particle Swarm Optimization The set of parameters $\{\omega\}$ that minimizes the fitness function $Z_F(\omega)$ are obtained using a Particle Swarm Optimization (PSO) heuristic method [181]. To do so, an ensemble of solutions (also called particles `p`) have their position ω(`p`) and velocity `v(p)` randomly initialized in the multidimensional search space identified by boundaries. Along the whole optimization process, every particle will keep track of the position ω(`p`) associated with the best objective (fitness) function value $Z_F(\omega(\texttt{p}))$. At the beginning of every discrete step, particles are updated about the swarm status, i.e. the current position of all particles, as well as their respective best found solution value and position. Subsequently, they will independently update their own velocity, which will be used to update their position. Velocity update is affected by three factors. The first, inertia, determines how a particle's trajectory is preserved along time. The second, personal best, attracts particles towards their own best solution. The third, global best, attracts particles towards the best solution found by neighbouring particles. Once velocity has been updated, a new position in which to evaluate the objective (fitness) function can be computed.

Table 3.3: List of backbone parameters

Parameter	Value
ΔV^{θ} (barrier of backbone bendings)	8 kcal·mol^{-1}
ΔV^{α} (barrier of backbone dihedrals)	7 kcal·mol^{-1}
σ	4.27 Å
ϵ	1.3 kcal·mol^{-1}

The boundaries associated with each parameter in the search space are guided by previous values obtained by Boltzmann inversion (as in the case of side chain bonded terms) or physically reasonable quantities preliminarily calculated, as for the case of the backbone bonded terms. In particular, the PSO approach was used to define effective bending potential parameters, able to correctly describe secondary structure conformations of an unspecific polypeptide (e.g. poly-alanine) arranged as α-helix and in β-turn conformations. For the purpose of tuning the non-bonded potential terms the adopted boundaries for the Lennard-Jones terms in the following range: 4.0 Å$< \sigma_{ij} <$5.0 Å and 0.4 kcal $\cdot$ mol^{-1}$< \epsilon_{ij} <$1.3 kcal $\cdot$ mol^{-1}.

Multiple runs of PSO showed how some parameters converged sooner that others, for instance bonded parameters for the side chains invariantly converged to the same values, permitting to fix them on following optimization cycles for tuning more fluctuating parameters like bonded terms for backbone and general non-bonded term. The backbone bonded and non-bonded parameters converged roughly to identical values (Tab. 3.3, Fig. 3.3). This already hinted to a partial set of parameters that can be transferable and used for a general CG force field.

Reference all-atom simulations The atomistic reference forces, used for the force-matching procedure on the set of proteins studied in this work, have been extracted from all-atom simulations carried out using NAMD simulation package [26] in explicit solvent and periodic boundary conditions, using a Langevin dynamics for the thermostat and a Nosé-Hoover-Langevin piston for the barostat. Simulations were carried out using smooth particle-mesh Ewald (SPME) [48] for the calculation of electrostatic interactions. All simulations were carried out using all-atom force field Amber99SB [16; 50] for the protein and TIP3P model [55; 56] for the water. The all-atom MD simulations were 100 ns long for five proteins belonging to different structural families and

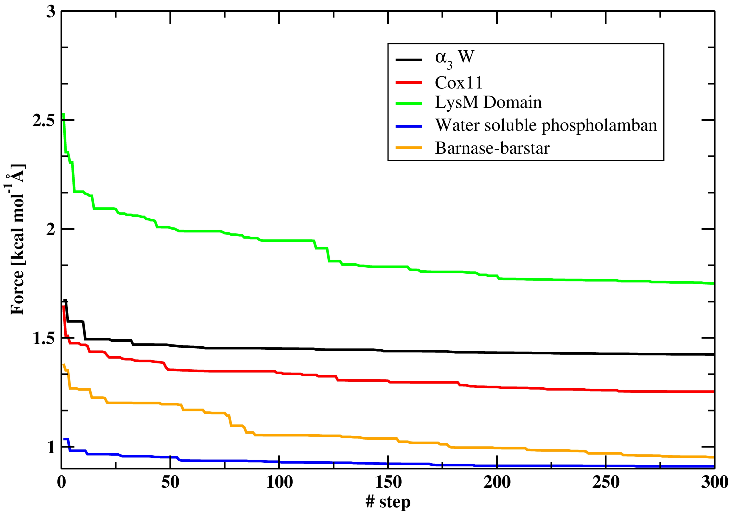

Figure 3.3: Convergence of the Particle Swarm Optimization during the force matching as a function of the number of steps

covering single molecule in solution and protein-protein complexes. Namely, α-, β-, α/β-proteins and small protein-protein complexes are selected to test the performance of our CG approach. The protein α_3W, with PDB entry 1lq7, is a *de novo* α-protein composed by 67 amino acids arranged as a clockwise bundle of three helices, whose structure has been obtained by NMR [182]. The Cox11 protein is the β-protein, PDB entry 1sp0, which structure has been obtained by NMR and is composed by 131 amino acids [183]. The LysM Domain, with PDB entry 1e0g, is the α/β-protein: it has been obtained by NMR, and is composed by 48 residues [184]. The coiled-coil protein is the engineered water soluble phospholamban, which is composed by four helical monomers of thirty amino acids in an anti-parallel arrangement, and which structure has been obtained by X-rays crystallography [185; 186]. Finally, the barnase-barstar complex solved by X-rays crystallography, is composed by a total of 189 residues [187].

For all the proteins the simulations time was set to 100 ns, from which 1000 structures were extracted to be used for the force-matching. For this purpose PSO was used with a setup of 20 particles with 3 consecutive repetitions of 300 optimization steps each. For all the particle swarm optimization runs, the difference between reference forces and the calculated one was in the order of 1 kcal$\cdot$mol$^{-1}\cdot$Å^{-1} for degree of freedom (Fig. 3.3). For a protein of $\sim$50 residues such a setup permits to have a refined set of parameters in less than 2 days on 4 CPUs.

The possibility to reduce the number of structures and the length of the required all-atom MD trajectory to be used for the particle swarm optimization runs has been explored. Using the Jarvis-Patrick clusterization method [188], as implemented in Gromacs [27], it is possible to reduce the number of structures of another order of magnitude. Preliminary force-matching calculations and consequent results, obtained from CG simulations, gave the same qualitative results showed in the present work (data not shown). For the α_3W protein a simulation in explicit solvent were carried out without tuning the water-protein Lennard-Jones parameters of interaction and setting them to the same value that are ϵ_{ij}= 0.8 $\frac{kcal}{mol}$ and σ=4.7Å whereas the water-water Lennard-Jones parameters of interaction are: ϵ_{ij}= 0.8 $\frac{kcal}{mol}$ and σ=4.6Å.

The possibility to define fully transferable non-bonded parameters as extracted from the specific parameterization of the 5 structurally different proteins studied in this work were explored. To do so, a simple averaging of the values of non-bonded ϵ and σ obtained for each pairs of beads were calculated and the resulting structural features from CG MD simulations compared with the previous specific CG parameterization. Such a parameterization has been used to simulate protein structures that do not belong to the training set, namely L25 and B1 immunoglobulin-binding domain protein [189; 190], with PDB entries 1b75 and 1pgb, respectively.

3.3.5 Coarse-grained simulations and structural observables

The coarse-grained molecular dynamics simulations for all the proteins were performed with the MD suite of programs Lammps, in the canonical NVT ensemble using the Langevin thermostat and an integration time step of 5 fs. The values of the harmonic spring constant of the CG models dictate the most convenient time steps [73; 107]. Calculating the ratio between the highest frequencies of harmonic springs between this CG and atomistic potentials, it has been estimated a convenient value for the CG time step were estimated to be up to 4 times bigger than for all-atom MD. Therefore, not using any constraints on the bonded degrees of freedom [130] it is possible to conservatively integrate the equations of motion with a time step of 5 fs. The use of an algorithm like SHAKE on all bonds potential terms will presumably allow to increase the time step to 10 fs, which appeared to give stable dynamics already within the current setup for most of the systems. All systems were first progressively heated from 100 K to 300 K for 0.5 ns, then equilibrated at this temperature for an additional 1 ns, and finally simulated for a production trajectory of 100 ns. For Lennard-Jones interactions the cut-off is 15 Å, whereas for electrostatics interactions it is 50 Å.

On average, for the proteins under study, the computational gain using this CG model is in the order of 200 times, without considering further optimization of our routines. The computational gain has been calculated dividing the total number of hours needed to run 100 ns with the all-atom force field by the total number of hours need to run 100 ns with our coarse-grained force field. For the explicit solvent the computational gain is in the order of 10 times.

For each simulated protein structural fluctuations and electrostatics properties were monitored to compare the results from CG and atomistic MD simulations. Among the structural properties it has been considered values of backbone bending and torsional angles, RMSD (root mean square deviation), gyration radius (R_g) were considered. Structural fluctuations were also considered as RMSF (root mean square fluctuations) and S^2 order parameters of backbone's dipoles. The S^2 order parameter quantifies the angular amplitude of N-H dipole internal motions, and quantification has been done for the backbone dipoles. The S^2 order parameter is calculated using the formula [133]:

$$S_i^2 = \frac{1}{2}\Big[3\sum_{\alpha=1}^{3}\sum_{\beta=1}^{3}\Big\langle \mu_{i,\alpha}\mu_{i,\beta}\Big\rangle^2 - 1\Big] \tag{3.13}$$

where $\mu_{i,\alpha}$ (α=1, 2, 3) are the x, y and z components of the normalized backbone's dipole moment at all-atom and coarse-grained level. The presence of coarse-grained monopoles and dipoles allows for the comparison the electrostatic features at the two levels of resolution. The electrostatics potentials of each protein have been calculated using APBS [100], and the results compared using the PIPSA 3.0 package [191; 192]. The all-atom results of RMSF, S^2, bending and dihedral curves were compared with coarse-grained results model using the cosine similarity to which it will be referred as a similarity index as it has been done in PIPSA [191]. The cosine similarity is a measure of the similarity of two vectors of a inner product space.

3.4 Results and Discussion

3.4.1 Structural and electrostatic coarse-grained properties for different protein families

The coarse-graining procedure has been tested using a set of proteins representative of distinct SCOP families. In particular, α_3W as an α-helical protein, Cox11 as a representative of full β proteins, and the LysM domain as a mixed α/β fold were simulated. The latter was also previously investigated by other CG approaches [193], allowing for a cross-comparison with my method.

All the proteins simulated at the CG level conserved their fold for 100 ns as during atomistic simulations, showing a very good agreement between all-atom and coarse-

Table 3.4: Summary of the structural and dynamic CG properties. RMSD: root mean square displacement in Å; R_g : gyration radius in Å; SI_{RMSF}, SI_{S^2}, SI_{bend} and SI_{dihed} are respectively the similarity indexes between all-atom and CG representation for RMSF, S^2, bending and dihedral angles quantities. In square brackets are reported the all-atom values for RMSD and R_g in Å.

Protein	RMSD	R_g	SI_{RMSF}	SI_{S^2}	SI_{bend}	SI_{dihe}
α_3W	2.6±0.3 [2.5±0.2]	11.3±0.2 [12.2±0.2]	0.95	0.95	0.99	0.83
Cox11	3.0±0.3 [3.0±0.3]	15.3±0.2 [16.8±0.2]	0.80	0.93	0.98	0.70
LysM Domain	2.6±0.3 [2.7±0.4]	8.8±0.2 [9.9±0.2]	0.93	0.92	0.99	0.85
Water soluble phospholamban	4.7±0.6 [2.7±0.5]	14.7±0.4 [16.4±0.2]	0.95	0.99	0.99	0.97
Barnase-Barstar	3.5±0.2 [1.1±0.2]	15.7±0.1 [17.2±0.1]	0.95	0.99	0.99	0.83

grained description (Tab. 3.4, Fig. 3.4). The secondary structures were conserved without the use of any additional *ad hoc* biases on the bending and torsional potential terms. Only minor discrepancies are observed on the loop regions connecting secondary structure elements like in α_3W and Cox11 (Fig. 3.4A and 3.4B). The agreement between the backbone bending and torsional angles calculated at the two levels of resolution is very good (Tab. 3.4; Fig. 3.5A and C; Fig. 3.6A and C; Fig. 3.7A and C). The cosine similarities between all-atom and coarse-grained values for backbone bending angles are between 0.98 and 0.99, while for backbone dihedrals are between 0.70 and 0.83, with the Cox11 protein being the least good.

The RMSD values reach convergence in around 5-10 ns (similarly to atomistic MD), fluctuating to values as low as 3 Å for 100 ns for all three proteins (Tab. 3.4; Fig. 3.8 A, B and C). The absolute values observed for RMSD are in line or lower than results reported using other CG models [193]. For instance, the LysM domain protein shows using this CG representation an RMSD as low as 2.7 Å, while simulations with the forcefield OPEP 4.0 [193] obtained a RMSD of 3.6 Å. The gyration radius is systematically slightly higher at all-atom level with respect to the coarse-grained representation, being

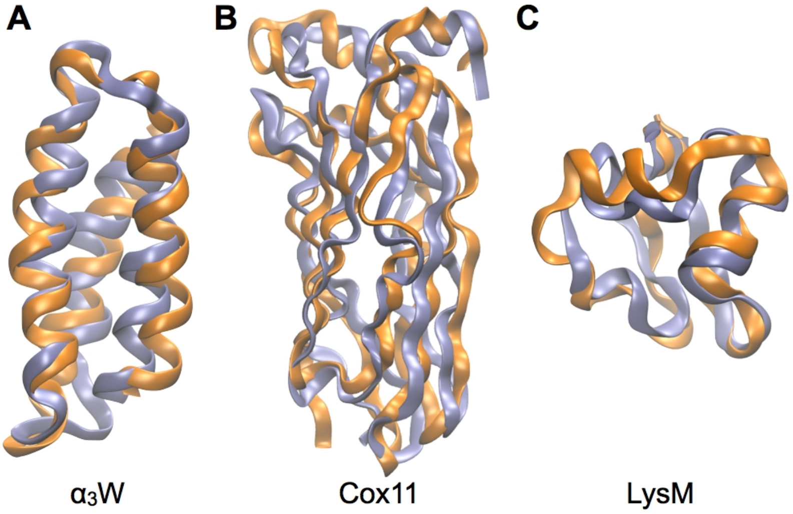

Figure 3.4: Structural comparison between all-atom and CG simulations of soluble proteins. Backbone superpositions of the last structure obtained from all-atom (in orange) and CG (ice-blue) MD simulations for (A) the α_3W protein, (B) Cox11 protein (C) LysM domain protein. Relative RMSD and gyration radius values are reported in Table 3.4.

the difference however in the order of 1 Å(Tab. 3.4). This slight collapse is likely to be intrinsically dependent on the coarse-grained representation, because the adopted mapping is not able to completely reproduce the steric effects of all the side chains, and buried cavities accommodating few water molecules cannot be filled by water beads having larger hindrance at CG granularity.

The general dynamic features are also in good agreement with the atomistic simulations. The RMSF calculated at CG level is systematically lower than for the all-atom one (Fig. 3.9), as already observed using other models [87]. The major differences are again on the loop regions. For example in the case of the α_3W protein the loops are composed by glycines that are very flexible, whereas this coarse-grained representation of the bending potential does not take into account in the current state a specific bending for glycines. RMSF peaks are not always well reproduced but the model correctly reproduces the trends of the fluctuations. The similarity cosines of the RMSF calculated at all-atom and coarse-grained level are 0.95 for α_3W, 0.79 for the Cox11 and 0.93 for the LysM domain. The decrease in flexibility observed for the RMSF is confirmed also when calculating the S^2 order parameter of the backbone. Anyway the

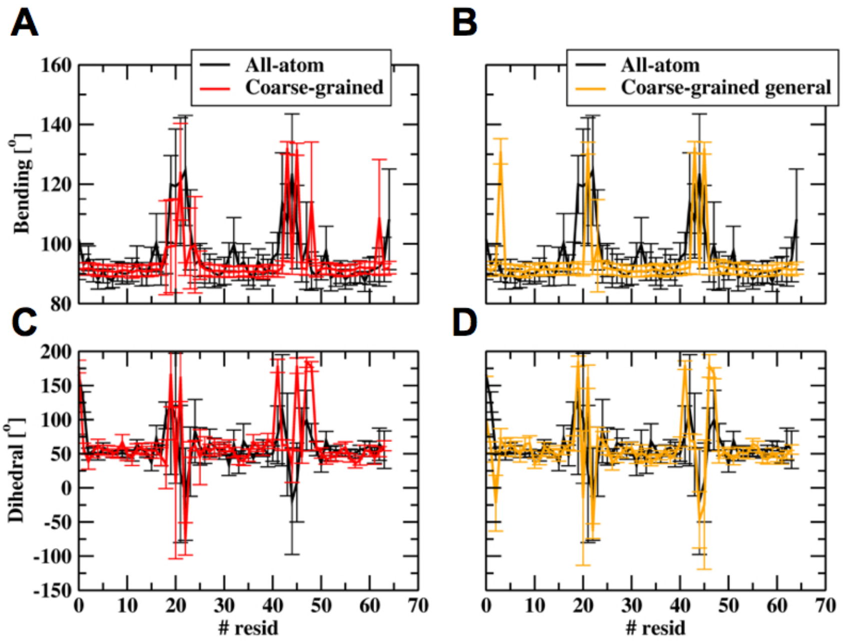

Figure 3.5: α_3W: Bending and dihedral results. (A) and (B) Bendings for specific and "general" parameterization respectively, (C) and (D) Dihedrals for specific and "general" parameterization respectively

agreement between the two levels of resolution is good: the similarity cosines of the S^2 calculated at all-atom and coarse-grained level are 0.95 for α_3W, 0.93 for the Cox11 and 0.92 for the LysM domain. The difference in flexibility observed for RMSF and S^2 has been attributed (i) to the simple fact that at the coarse-grained level the lower number of degrees of freedom does not intrinsically allow the complete description of the structural fluctuations, and (ii) to the potential form of the bending terms, which is not parametrized to be sequence-dependent, but has a general form which is meant to describe at the same time α, β and coil structures.

The monopole and dipolar terms for the backbone and side chains are able to well reproduce the electrostatic potential of the proteins. In fact, when comparing the CG and atomistic values of the electrostatic potential using PIPSA, quite high similarity indexes were obtained (Tab. 3.5; Fig. 3.10, 3.11, 3.12). The lowest value of the similarity index has been obtained for the α_3W protein (0.97): in fact at the surface of this protein there are fewer charged or polar residues than in the other cases.

Along with the results using the distance-dependent model of implicit solvent for the α_3W protein were obtained results with the explicit CG water model (Fig. 3.13, 3.14, 3.15). The RMSD is 2.2±0.2 Å, whereas the gyration radius is 11.7±0.2 Å,

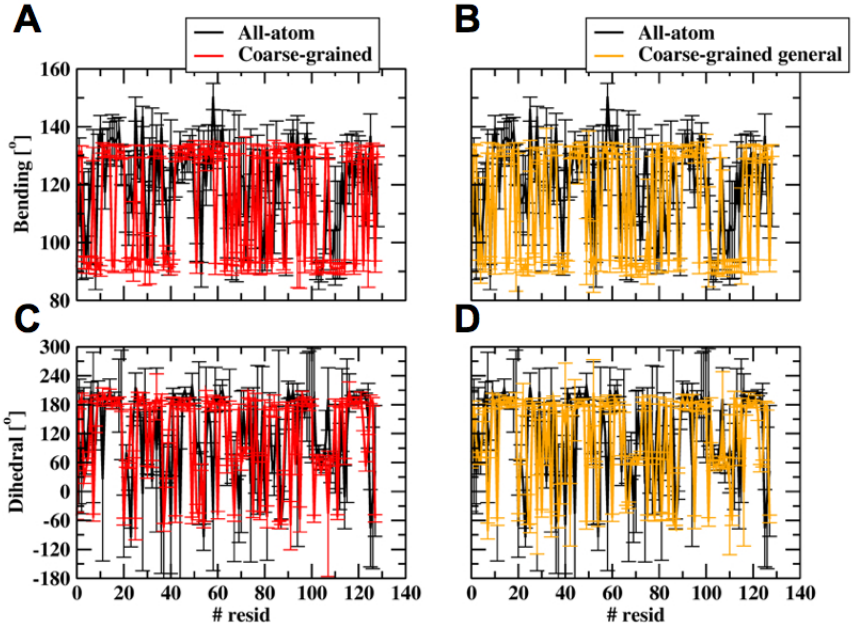

Figure 3.6: Cox11: Bending and dihedral results. (A) and (B) Bendings for specific and "general" parameterization respectively, (C) and (D) Dihedrals for specific and "general" parameterization

Table 3.5: Summary of the electrostatics properties. P (in eÅ) is the total electrostatic dipole moment for the entire protein, SI_{ele} is the similarity index between the all-atom and CG electrostatic potential as calculated with PIPSA. Compare with Fig. 3.10, 3.11, 3.12, 3.20, 3.21, 3.28, 3.29 for a 3D visualization of the electrostatic potential at the molecular surface calculated using APBS.

Protein	$\|P_{AA}\|$	$\|P_{CG}\|$	SI_{ele}
α_3W	88.6	119.3	0.97
Cox11	50.6	55.4	0.99
LysM	39.0	44.2	0.99
Water soluble phospholamban	630.7	630.2	0.95
Barnase-Barstar	200.2	218.8	0.93
L25	47.3	48.3	0.97
B1 Immunoglobulin-binding	63.8	94.6	0.93

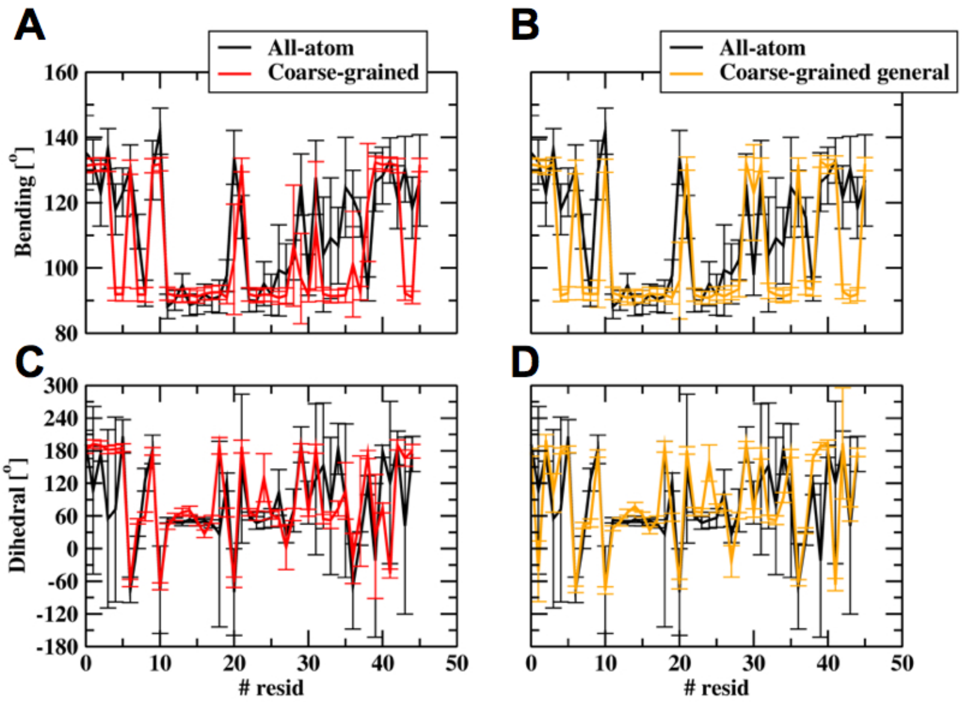

Figure 3.7: LysM domain: Bending and dihedral results. (A) and (B) Bendings for specific and "general" parameterization respectively, (C) and (D) Dihedrals for specific and "general" parameterization respectively

results that are in line with the all-atom ones. For the gyration radius the same type of collapse were not observed with the implicit solvent. A good agreement between all-atom and coarse-grained simulations with explicit solvent has been observed also for the others properties, namely RMSF, S^2, backbone's bending and dihedral (Fig. 3.14, 3.15). This indicates that some of the drawbacks observed using the implicit solvent could be partially solved by using an explicit CG model for water. The use of the explicit solvent will be further explored in next studies.

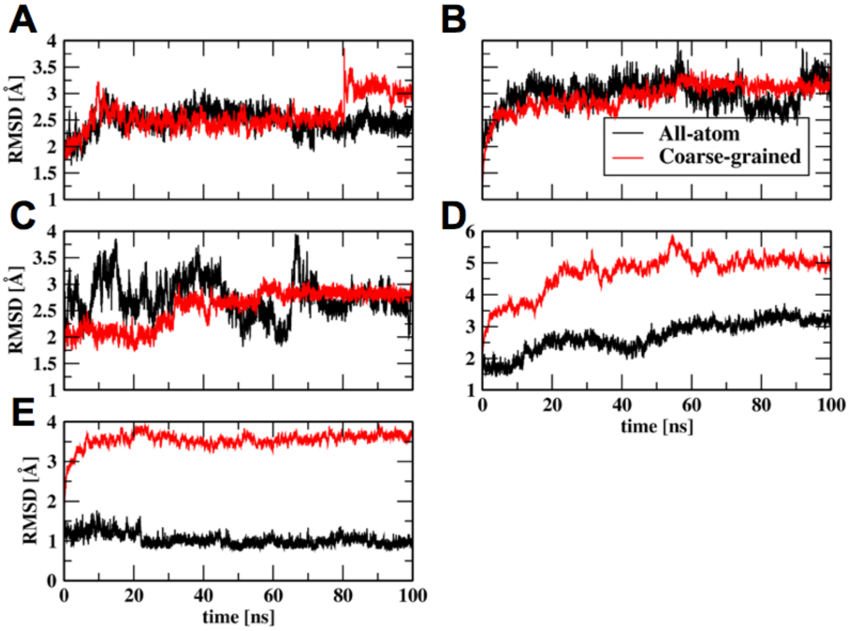

Figure 3.8: RMSD results of specific parameterization. (A) α_3W, (B) Cox11, (C) LysM domain, (D) Water soluble phospholamban, (E) Barnase-barstar complex

Summarizing, this section showed that it is possible to obtain, with a minimal amount of investment in terms of CPU time, a tailored parameterization for any soluble protein. The CG simulations produced results in very good agreement with the atomistic simulations and similar to results obtained using force fields adopting a comparable mapping topology [193].

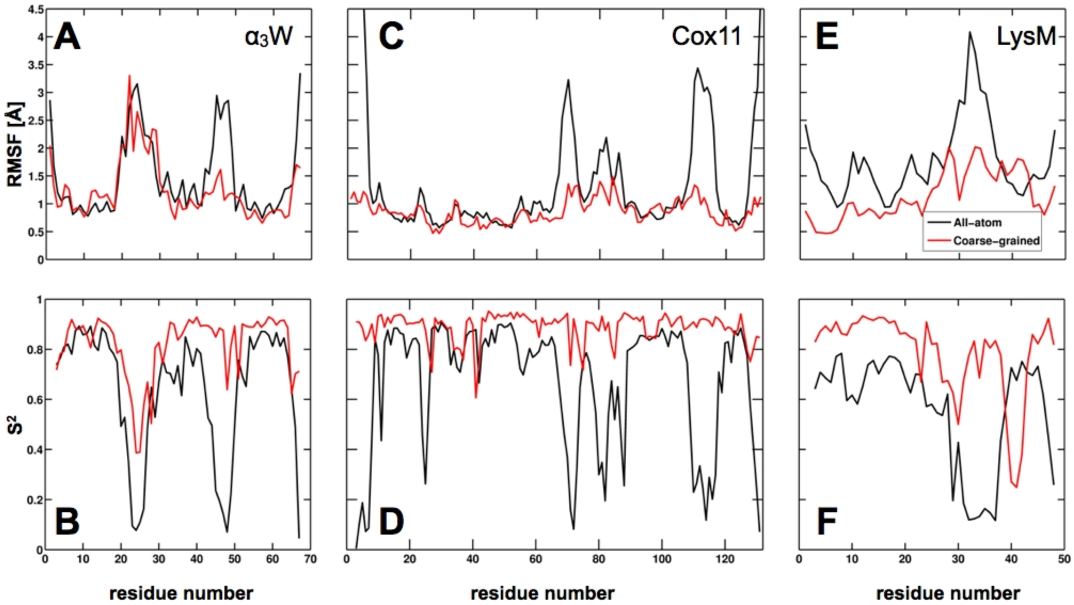

Figure 3.9: Comparison between dynamic properties of all-atom and CG MD simulations. RMSF and S^2 are reported for α_3W protein (A, B), Cox11 protein (C, D) LysM domain protein (E, F).

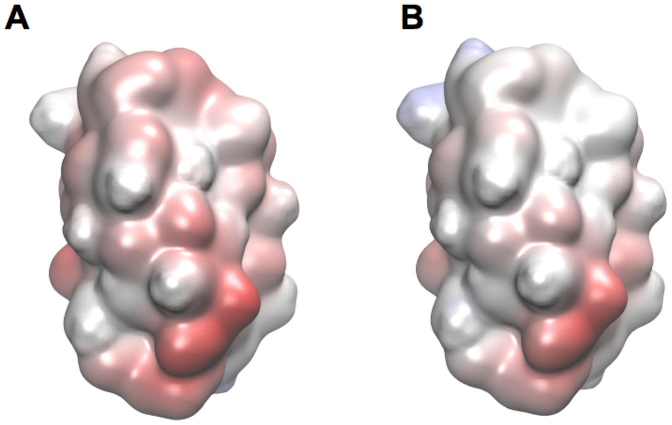

Figure 3.10: Electrostatics of α_3W protein. Comparison between the atomistic (in A) and CG (in B) electrostatic potential mapped on the protein molecular surface. Potential is reported in k_BT/e with red for negative values and blue for positive. [SI_{ele}=0.83]

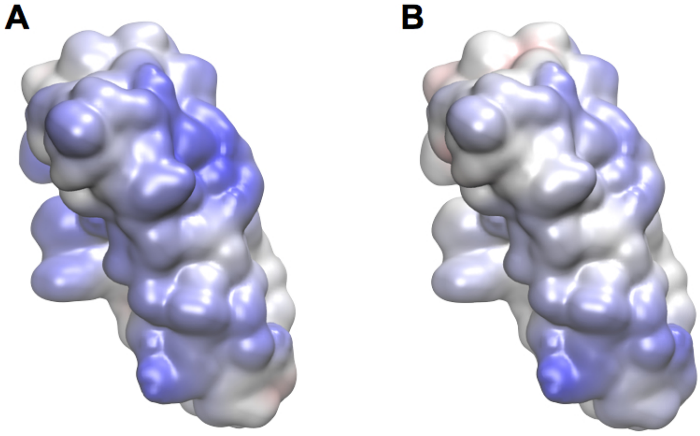

Figure 3.11: Electrostatics of Cox11 protein. Comparison between the atomistic (in A) and CG (in B) electrostatic potential mapped on the protein molecular surface. Potential is reported in k_BT/e with red for negative values and blue for positive. [SI_{ele}=0.95]

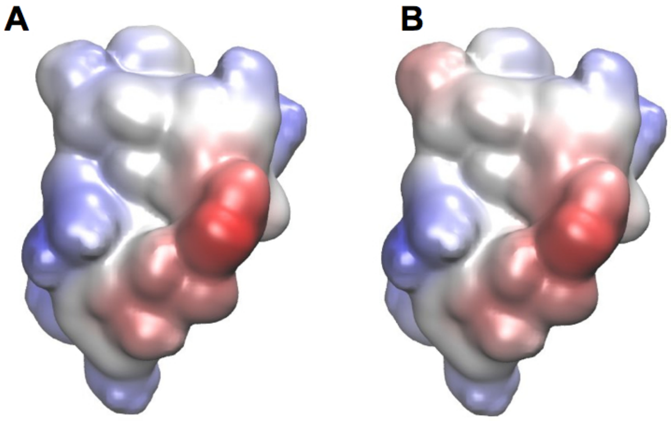

Figure 3.12: Electrostatics of LysM domain protein. Comparison between the atomistic (in A) and CG (in B) electrostatic potential mapped on the protein molecular surface. Potential is reported in $k_B T/e$ with red for negative values and blue for positive. [SI_{ele}=0.96]

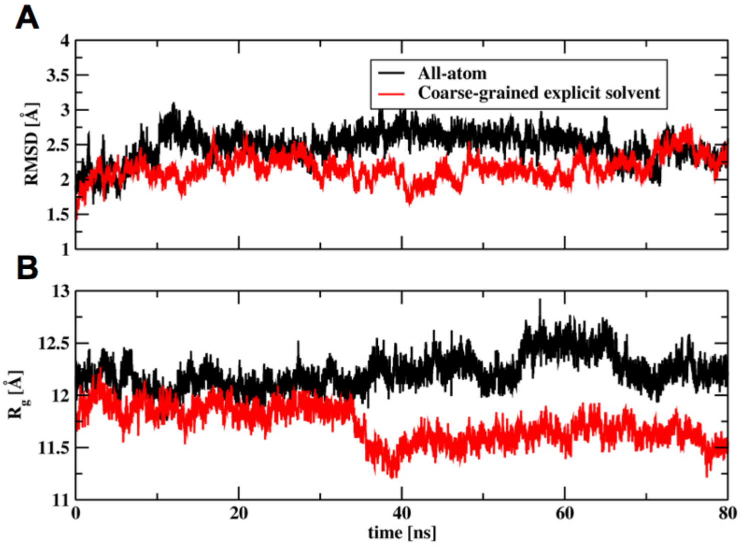

Figure 3.13: α_3W in explicit coarse-grained water: RMSD and gyration's radius results. (A) RMSD, (B) Gyration's radius

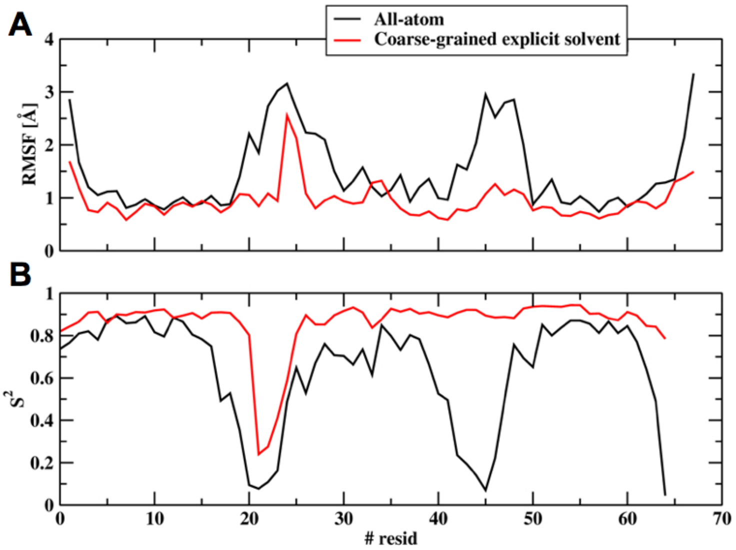

Figure 3.14: α_3W in explicit coarse-grained water: (A) RMSF and (B) S^2 results.

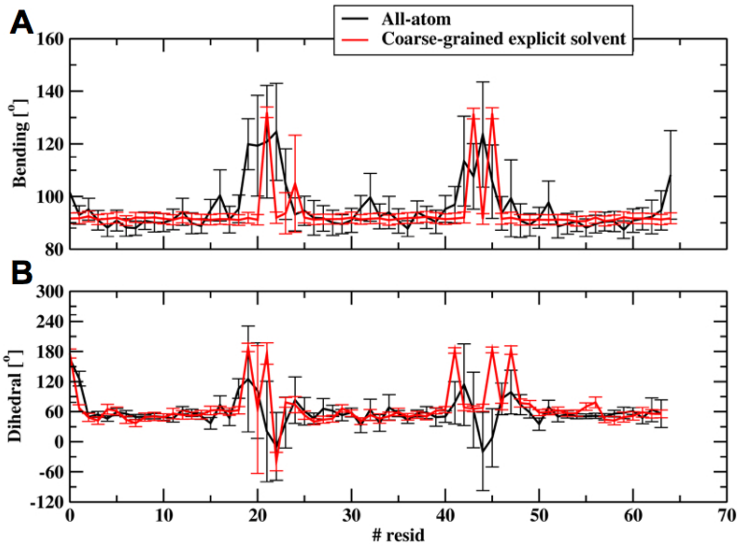

Figure 3.15: α_3W in explicit coarse-grained water: (A) bending and (B) dihedral results.

Structural and electrostatic coarse-grained properties for protein-protein complexes

Having shown that within the proposed approach it is possible to simulate soluble proteins reproducing structural, electrostatic and dynamic features of all-atom simulations, the investigation was extended to protein-protein complexes. First, the engineered soluble phospholamban protein-protein complex, chosen because it is among the simplest not covalently bonded helix bundles was studied [185]. The second complex investigated was the barnase-barstar complex [187], which is in principle more challenging because it is composed by different secondary structure elements with different reciprocal arrangements.

As already obtained for the single proteins, the secondary structures were strongly conserved during CG simulations with only some discrepancies on the loop regions. Comparing the last structures obtained at the two levels of resolution, the percentage of conserved secondary structures was on average around 70 %. The RMSD reached convergence after about 10 ns, in line with the atomistic simulations. The difference on the RMSD calculated at the two levels of resolution differ of about 2 Å (Tab. 3.5). In the case of the water soluble phospholamban the main structural differences are in correspondence of the protein termini, whereas in the case of the barnase-barstar complex similar differences are observed for both proteins composing the dimer. This is likely due to two main reasons: (i) the coarse-grained representation does not perfectly reproduce the steric effect of the side chains at the interface, and (ii) the implicit solvent model does not allow for optimal solvation of the regions at the interface between the two proteins. This can be particularly relevant in case that interstitial waters localize in the area (*e.g.*, CG water models would have the same problem), and it can lead to exploration of slightly different conformations at the interface. The gyration radius results confirm this hypothesis because the difference between the all-atom and the coarse-grained values is in the order 1.5 Å (Tab. 3.5).

Nonetheless, the protein-protein interfaces are well conserved for both complexes and key electrostatic interactions are mainly preserved. For example, in the water soluble phospholamban, electrostatics interactions that stabilize the complex such as Cys80-Cys110 and Cys80-Arg117 are maintained (Fig. 3.16B). In the case of the barnase-barstar complex interactions at the interface are also well preserved reproducing the majority of the contacts (e.g. Arg81 (barnase)-Asp147 (barstar), His100 (barnase)-Asp147 (barstar), Arg81 (barnase)-Gly151 (barstar), see Fig. 3.16D and Tab. 3.6).

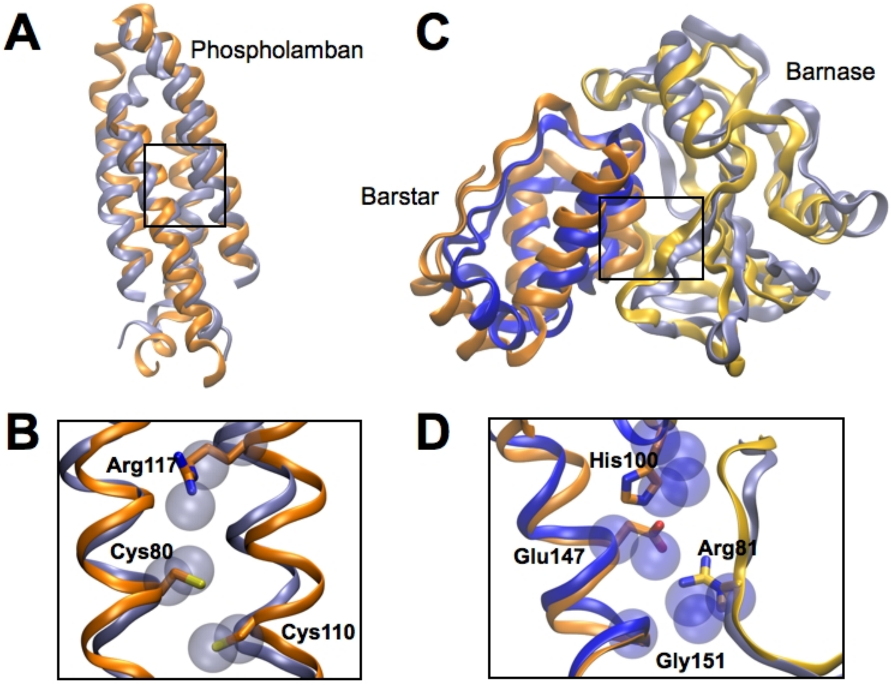

Figure 3.16: Structural comparison between all-atom and CG simulations of molecular complexes. Backbone superpositions of the last structure obtained from all-atom and CG MD simulations for (A) water soluble phospholamban, and (C) barnase-barstar complex (color code as in Figure 3 apart for all-atom barnase in light orange, CG barnase in orange, all-atom barstar in blue and CG barstar in ice-blue); (B) interface of water soluble phospholamban with all-atom residues in licorice representation and relative CG beads in transparent van der Waals representations; (D) interface of barnase-barstar with all-atom residues in licorice representation and CG ones in transparent van der Waals representations. Relative RMSD and gyration radius values are reported in Tab. 3.5

The CG RMSF and S^2 values are in good agreement with all-atom MD results (Fig. 3.17). The structural fluctuation properties are preserved and also the secondary structure elements, as seen from the superposition of the last structures obtained at the two levels of resolution (Fig. 3.16). This is strengthened by the agreement of the backbone bending and torsional angles for the two complexes (Tab. 3.5; Fig. 3.18 A and C; Fig. 3.19 A and C). Also in this case, a very good agreement were found the electrostatic potentials calculated at the atomistic and CG levels of resolution: the similarity indexes calculated with PIPSA are 0.95 and 0.93 for the water soluble phospholamban and the barnase-barstar complex, respectively (Tab. 3.5, Fig. 3.20 and 3.21).

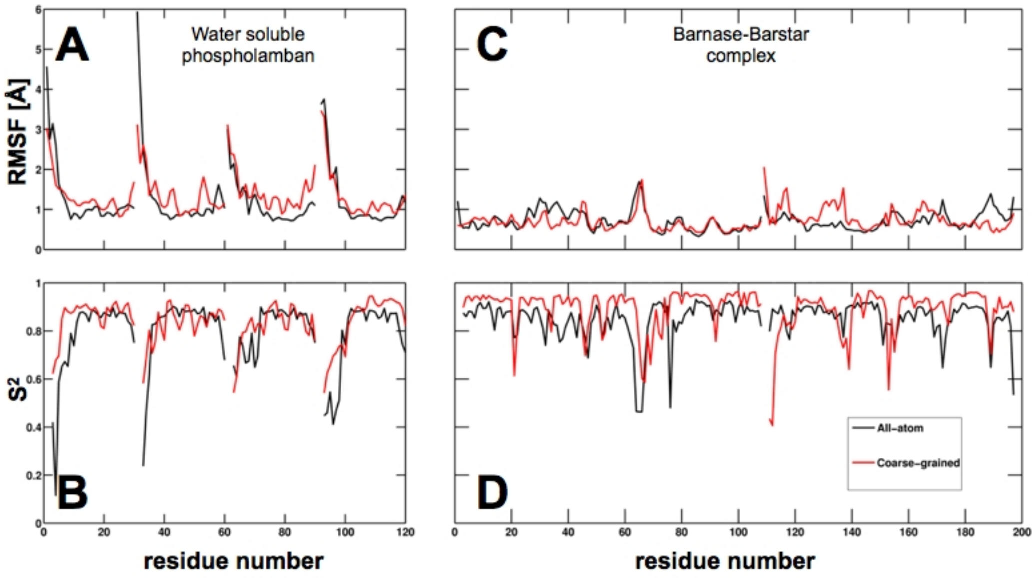

Figure 3.17: Comparison between dynamic properties of all-atom and CG MD simulations of molecular complexes. RMSF and S^2 are respectively reported for water soluble phospholamban (A, B) and barnase-barstar complex (C, D).

Table 3.6: Barnase-barstar interface: significant intermolecular distances are reported and compared in the two representations

Barnase residues/bead	Barstar residues/bead	All-atom [Å]	Coarse-grained [Å]
Arg81 BB	Tyr137 BS3	7.7±0.2	15.3±3.5
Asn82 BB	Tyr137 BS3	6.0±0.3	13.6±3.4
His100 BS2	Gly139 BB	4.4±0.2	8.3±0.4
His100 BB	Asn141 BS1	6.7±0.2	11.0±0.5
Glu58 BS1	Leu142 BB	5.1±0.6	5.0±0.7
Arg57 BB	Asp143 BS1	4.3±0.2	7.8±0.5
Arg81 BS2	Asp147 BS1	3.4±0.2	4.0±0.4
Arg85 BS2	Asp147 BS1	4.6±0.1	4.1±0.3
His100 BS3	Asp147 BS1	4.1±0.1	5.0±0.6
Lys25 BS2	Thr150 BS1	4.5±0.6	7.2±0.5
Arg81 BS2	Gly151 BB	4.9±0.3	4.9±0.3
Arg57 BS2	Glu184 BS1	3.7±0.3	5.0±0.7

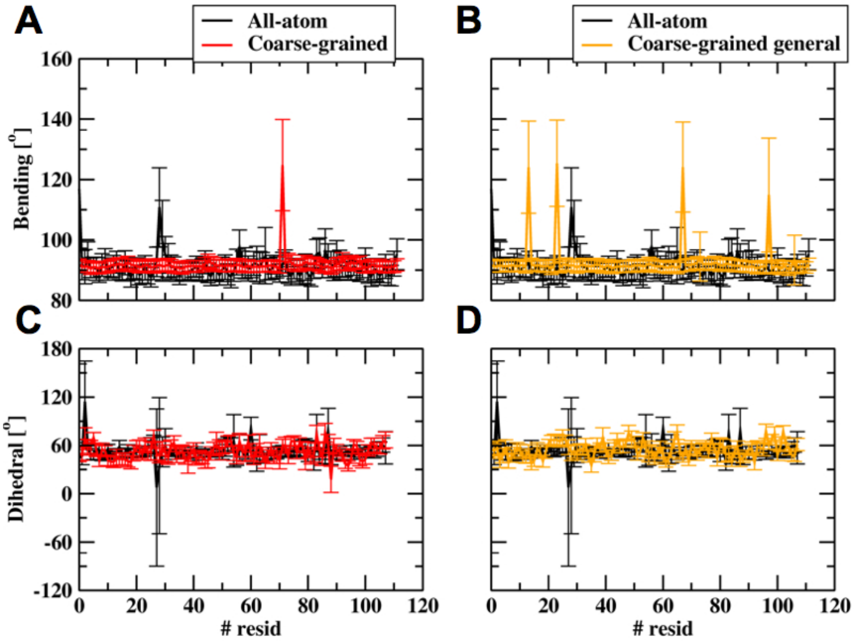

Figure 3.18: Water soluble phospholamban: Bending and dihedral results. (A) and (B) Bendings for specific and "general" parameterization respectively, (C) and (D) Dihedrals for specific and "general" parameterization respectively

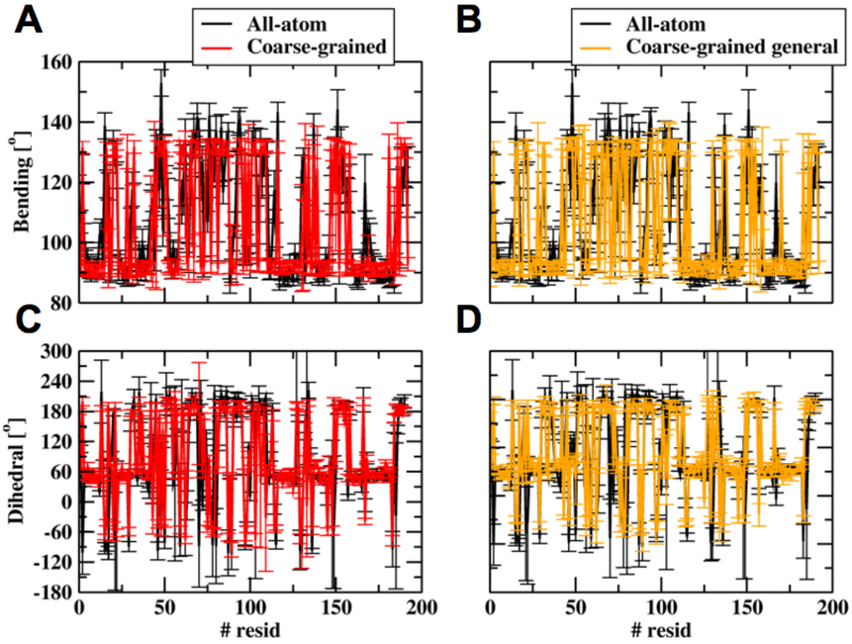

Figure 3.19: Barnase-barstar complex: Bending and dihedral results. (A) and (B) Bendings for specific and "general" parameterization respectively, (C) and (D) Dihedrals for specific and "general" parameterization respectively

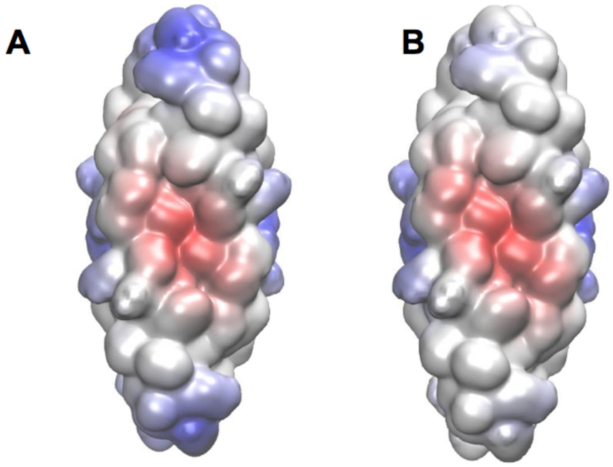

Figure 3.20: Electrostatics of water soluble phospholamban protein. Comparison between the atomistic (in A) and CG (in B) electrostatic potential mapped on the protein molecular surface. Potential is reported in k_BT/e with red for negative values and blue for positive. [SI_{ele}=0.95]

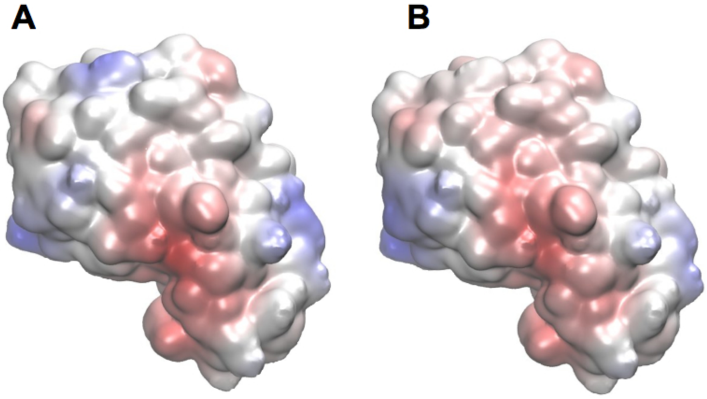

Figure 3.21: Electrostatics of barnase-barstar complex. Comparison between the atomistic (in A) and CG (in B) electrostatic potential mapped on the protein molecular surface. Potential is reported in k_BT/e with red for negative values and blue for positive. [SI_{ele}=0.93]

3.4.2 Towards a transferable coarse-grained force field for proteins

In the presented simulations, specific sets of parameters were optimized each time for each protein under study. Although the results show very good accuracy with respect to all-atom simulations and parameterization is very efficient, such approach lacks full transferability, as it requires re-optimization of some of the force field terms every time a new system is studied. In order to test the possibility of extending the approach towards the definition of a fully transferable potential, an averaged force field potentials were implemented. This new set of parameters was then used to run CG molecular dynamics of the 5 systems in the training set, as well as L25 and B1 immunoglobulin-binding proteins, two additional systems not included in the set employed for calibration.

Table 3.7: Summary of the structural and dynamic CG properties using a generalized CG parameterization. RMSD: root mean square displacement in Å; R_g : gyration radius in Å; SI_{RMSF}, SI_{S^2}, SI_{bend} and SI_{dihe} are respectively the similarity indexes between all-atom and CG representation for RMSF, S^2, bending and dihedral angles quantities. In square brackets are reported the all-atom values for RMSD and R_g in Å.

Protein	RMSD	R_g	SI_{RMSF}	SI_{S^2}	SI_{bend}	SI_{dihe}
α_3W	2.4±0.2 [2.5±0.2]	11.1±0.2 [12.2±0.2]	0.92	0.93	0.99	0.85
Cox11	3.0±0.4 [3.0±0.3]	15.1±0.2 [16.8±0.2]	0.81	0.94	0.98	0.73
LysM Domain	2.7±0.3 [2.7±0.4]	8.8±0.1 [9.9±0.2]	0.95	0.93	0.99	0.83
Water soluble phospholamban	4.5±0.4 [2.7±0.5]	16.3±0.2 [16.4±0.2]	0.93	0.98	0.99	0.97
Barnase-Barstar	3.4±0.2 [1.1±0.2]	15.9±0.1 [17.2±0.1]	0.93	0.98	0.99	0.82
L25	3.8±0.3 [3.5±0.8]	12.1±0.2 [13.1±0.3]	0.91	0.94	0.98	0.70
B1 immunoglobulin-binding	2.6±0.2 [1.1±0.3]	9.8±0.1 [10.4±0.1]	0.90	0.98	0.99	0.82

A good agreement for the CG simulations of the proteins from the training set has been obtained (Tab. 3.7). In particular, the percentage of preserved secondary structure elements for the proteins that belong to the training set are in line with what has been obtained with specific parameterizations (Tab. 3.7). The systems not included in the training set were instead described with slightly lower accuracy but in line with the results obtained for the others proteins (Fig. 3.23, 3.22, Tab. 3.7). Secondary structure elements are preserved, the main differences being on loop regions (Fig. 3.24).

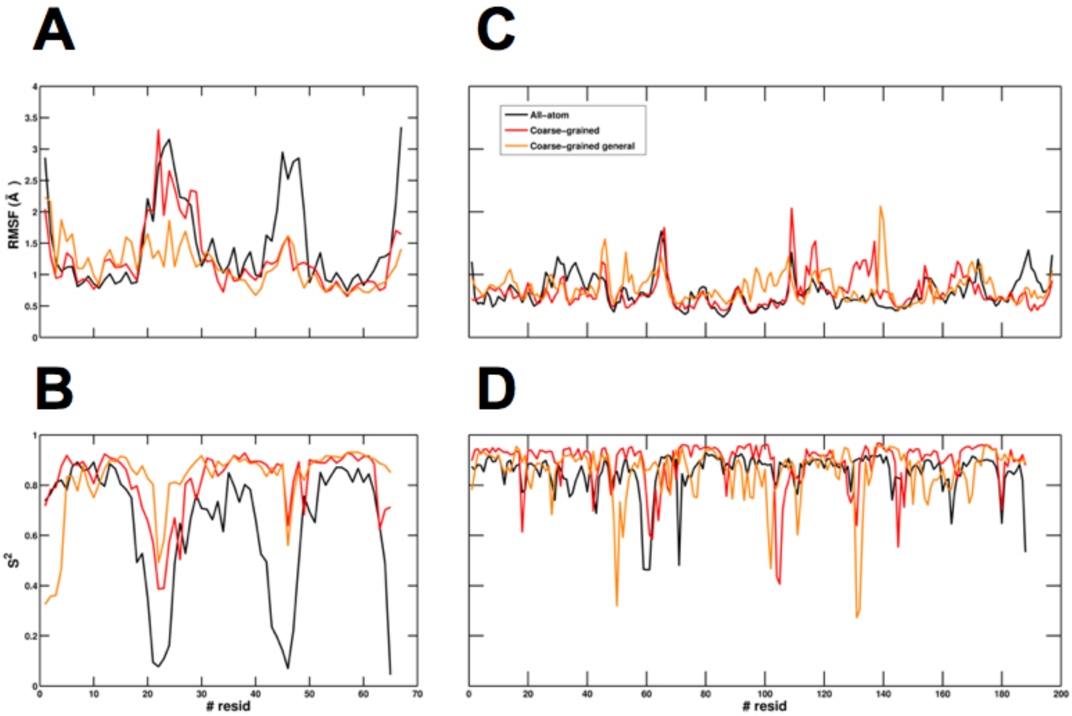

Figure 3.22: RMSF and S^2 calculations for "general" parameterization. (A) and (C) RMSF and S^2 for α_3W, (B) (D) for Barnase-barstar complex

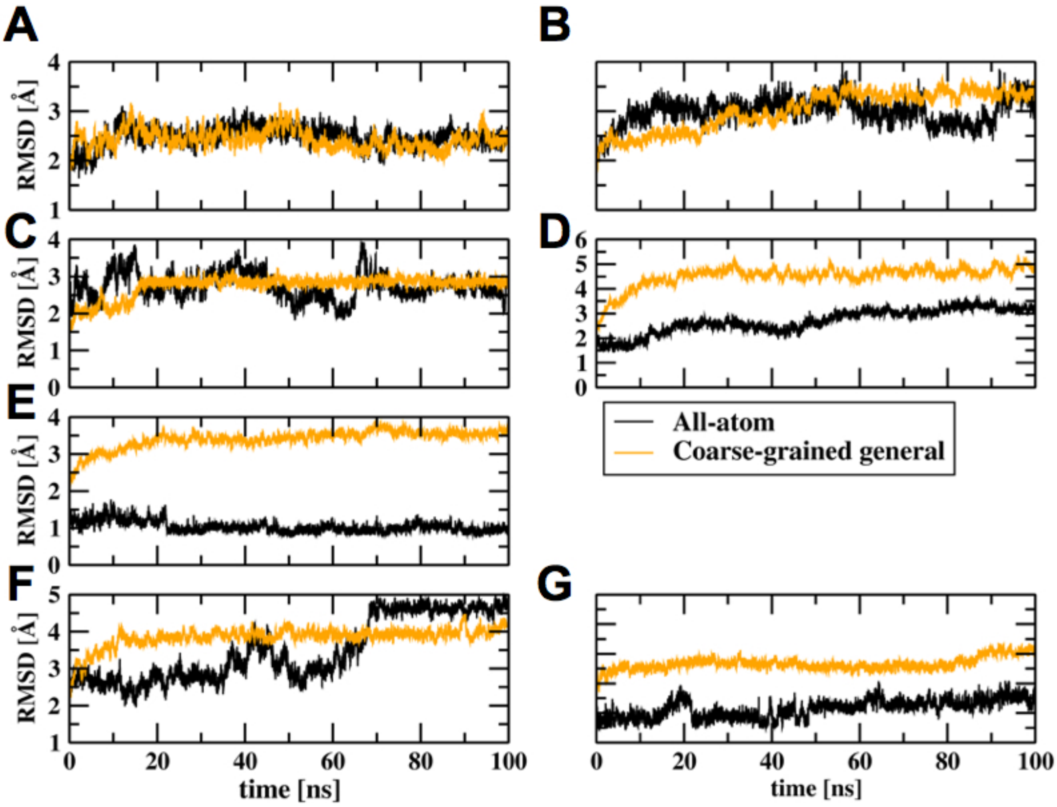

Figure 3.23: RMSD results for "general" parameterization. (A) α_3W, (B) Cox11, (C) LysM domain, (D) Water soluble phospholamban, (E) Barnase-barstar complex, (F) L25, (G) B1 immunoglobulin-binding

Superposition of the last structures obtained at the two levels of resolution gives RMSD values for L25 and B1 immunoglobulin-binding of 4.0 Å and 3.6 Å, respectively, whereas the percentage of preserved secondary structure elements is around 70 % for both for the last structures for the two level of resolution (Fig. 3.25, 3.26). RMSD and gyration radius (Tab. 3.7 and Fig. 3.23) obtained using the general CG force field are very similar to the relative atomistic values.

Both L25 and B1 immunoglobulin-binding proteins have been recently used to test two coarse-grained force-fields, *i.e.* the one developed by Ha-Duong and coworkers [87] and Opep4.0 [193]. A RMSD of 6 Å and 2.9 Å, respectively when using the former and the latter model was reported for L25. For the OPEP4.0 model it should be noted that only some parts of the protein have been selected for the calculation of the RMSD. B1 immunoglobulin-binding protein showed instead an RMSD of around 4 Å and 3.3 Å, respectively. Compared to the reported data, for both these proteins this coarse-grained

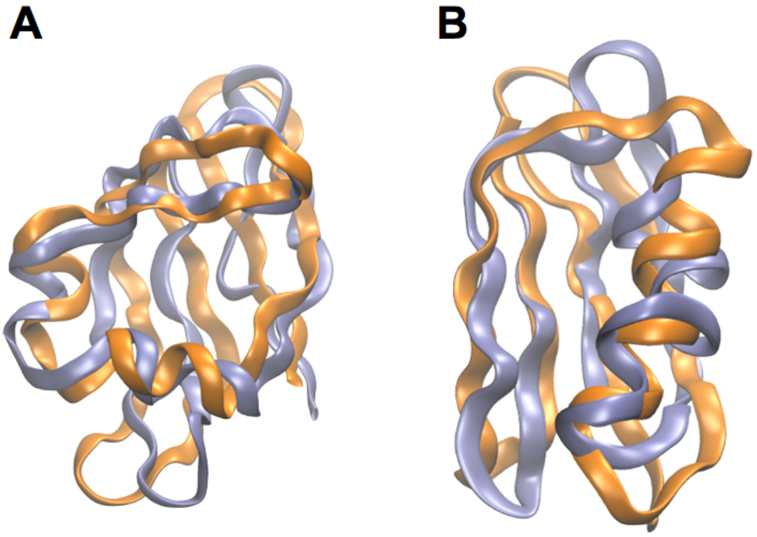

Figure 3.24: Comparison of atomistic and CG structures for the "general" parameterization. Backbone superimpositions of the last structure obtained from all-atom (orange) and CG (iceblue) simulation (A) L25 (B) B1 immunoglobulin-binding

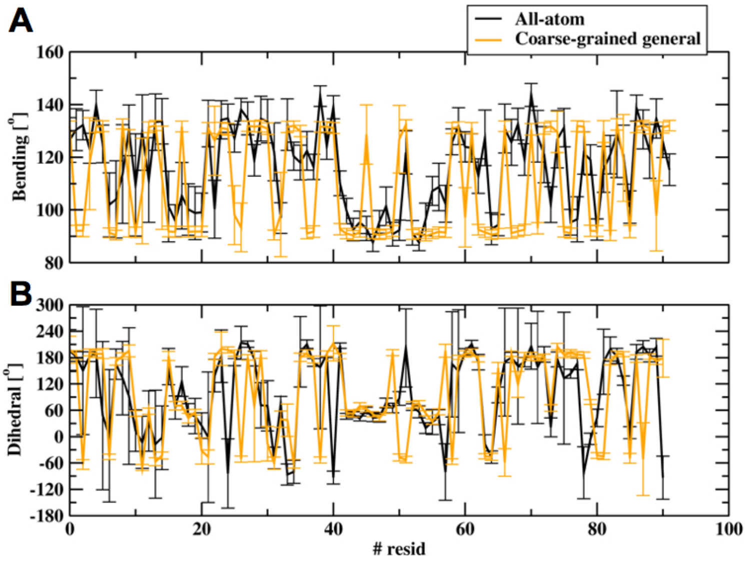

Figure 3.25: L25: bending (A) and dihedral (B) results.

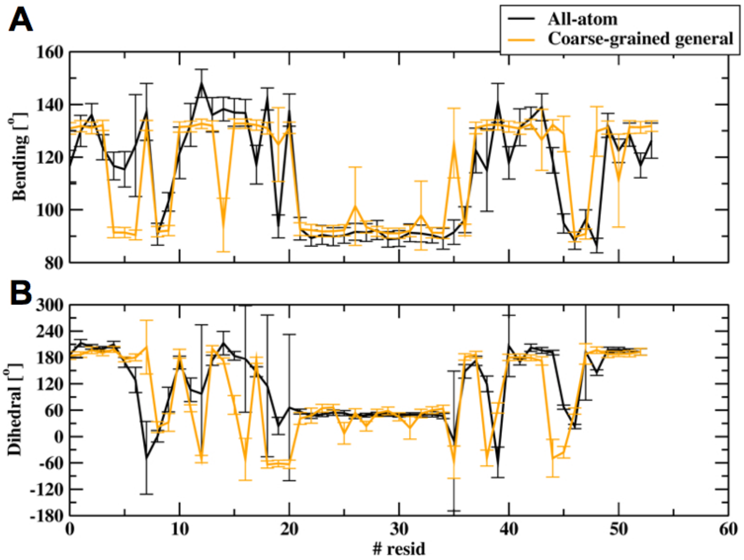

Figure 3.26: B1 immunoglobulin-binding: Bending (A) and dihedral (B) results.

approach performs reasonably well, despite the fact that there is always a discrepancy of around 1.5 Å with respect to the atomistic value for the RMSD and of around 1 Å for the gyration radius (Tab. 3.7).

The agreement of structural fluctuations with all-atom values is lower than for specific parameterizations (Tab. 3.7), however the similarity index for the properties calculated at the two levels of resolution for L25 and B1 immunoglobulin-binding proteins are quite high (Tab. 3.5 and 3.7, Fig. 3.25, 3.26, 3.27, 3.28, 3.29), showing that this preliminary version of a general set of electrostatic-consistent CG potentials goes in the right direction towards the development of a reliable transferable CG force field.

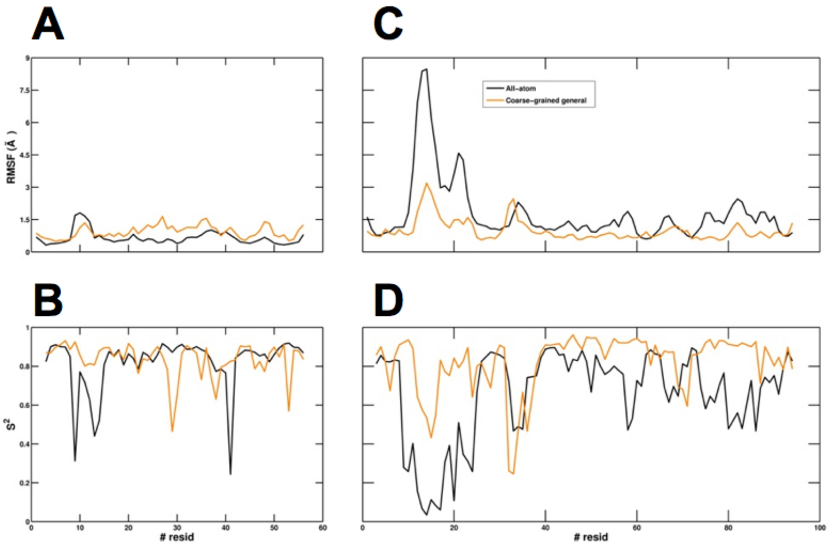

Figure 3.27: RMSF and S^2 calculations for "general parameterization". (A) and (C) RMSF and S^2 L25, (B) and (D) for B1 immunoglobulin-binding

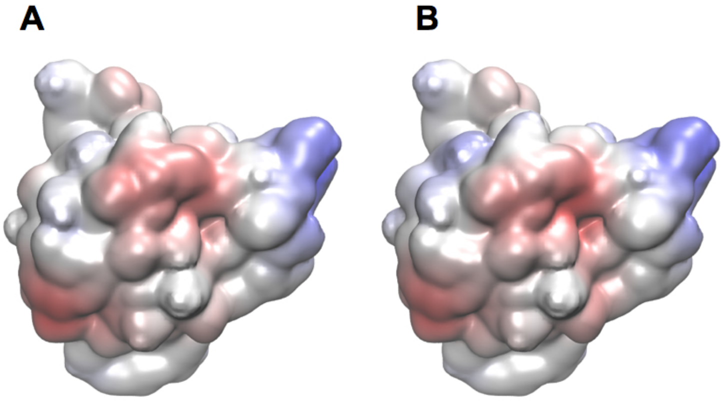

Figure 3.28: Electrostatics of L25 protein. Comparison between the atomistic (in A) and CG (in B) electrostatic potential mapped on the protein molecular surface. Potential is reported in k_BT/e with red for negative values and blue for positive. [SI_{ele}=0.97]

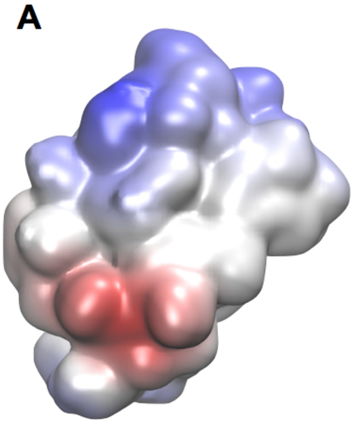

Figure 3.29: Electrostatics of B1 immunoglobulin-binding protein. Comparison between the atomistic (in A) and CG (in B) electrostatic potential mapped on the protein molecular surface. Potential is reported in $k_B T/e$ with red for negative values and blue for positive. [SI_{ele}=0.93]

3.5 Conclusions

In this work I presented a coarse-graining procedure to generate potentials for molecular simulation of soluble proteins incorporating explicit and detailed description of electrostatics. The adopted strategy for the parameterization of the coarse-grained potentials is based on Boltzmann inversion and a force-matching scheme relying on high-resolution protein structures and atomistic simulations. The derivation of the parameters is obtained using a new and robust global optimization algorithm based on particle swarm optimization, that handles the assignment of several hundreds parameters in a relatively short amount of time.

The combination of electrostatic terms at the backbone and polar side-chains to terms accounting for the steric hindrance of the CG beads produces stable protein tertiary structures, and maintains the global fold of a variety of soluble proteins. This approach produces also dynamically stable quaternary complexes, like in the case of the phospholamban and barnase-barstar systems. Despite the intrinsic limitations of any coarse-grained representation, these results demonstrate that CG potentials generated by this procedure produce a very good and consistent agreement with all-atom simulations, well reproducing the main structural and dynamic properties. Importantly, these

Table 3.8: Interface barnase-barstar for "general" parameterization

Barnase residues/bead	Barstar residues/bead	All-atom [Å]	Coarse-grained [Å]
Arg81 BB	Tyr137 BS3	7.7±0.2	9.8±0.7
Asn82 BB	Tyr137 BS3	6.0±0.3	8.6±0.7
His100 BS2	Gly139 BB	4.4±0.2	9.1±1.1
His100 BB	Asn141 BS1	6.7±0.2	10.1±0.6
Glu58 BS1	Leu142 BB	5.1±0.6	6.5±0.8
Arg57 BB	Asp143 BS1	4.3±0.2	5.6±0.6
Arg81 BS2	Asp147 BS1	3.4±0.2	4.3±0.5
Arg85 BS2	Asp147 BS1	4.6±0.1	4.1±0.2
His100 BS3	Asp147 BS1	4.1±0.1	6.6±0.8
Lys25 BS2	Thr150 BS1	4.5±0.6	8.1±0.8
Arg81 BS2	Gly151 BB	4.9±0.3	4.5±0.3
Arg57 BS2	Glu184 BS1	3.7±0.3	4.2±0.2

CG models are also able to describe the main interface interactions, producing stable protein complexes. Although not explicitly tested for all the existing folding families it is expected that the derivation of specific parameters obtained using this strategy would be as accurate to guarantee the description of the structure and dynamics of other proteins at this level of granularity.

These results are promising and suggest that electrostatic-consistent CG potentials can be efficiently used to explore protein-protein molecular recognition using molecular dynamics sampling. These results are in fact in good agreement with all-atom simulations and, when directly compared with previously reported CG force fields, showed similar or better performances in describing structural and dynamic determinants of soluble proteins. Moreover, this procedure can be straightforwardly extended for the parameterization of any protein. The extension of this optimization procedure to a larger dataset may prelude to the generation of a fully transferable CG force field that will be applied in principle to any protein or, more interestingly, any large macromolecular assemblies for which direct, long all-atom simulations may not be easily affordable.

3.6 Future Directions and Development

The first priority, which is currently ongoing, is to further advance on establishing a more robust transferability for the CG models developed in this thesis. I am currently following two strategies to overcome this limitation of the model by (i) extending the proteins set in order to have more statistics during the optimization of relevant parameters, and (ii) assigning and tuning residue-specific parameters for the backbone.

Moreover, the introduction of a correct treatment of electrostatics at this coarse-grained level of resolution should not be limited only to proteins but eventually extended to other biomolecules. In fact, being the ultimate goal to simulate the cellular environment, carbohydrates, phospholipids, nucleic acids and other small molecules should be described using a similar treatment of electrostatic contributions. Clearly, as done for proteins, all these models should be validated against experimental data and all-atom results in order to check their reliability. Another interesting extension of the present work would consist in the development of an hybrid all-atom/coarse-grained treatment following the same spirit of quantum mechanics/molecular mechanics hybrid methods. This approach would give the opportunity to still treat large systems in their cellular environment, considering regions where atomistic details are important, like ligand pockets or protein-protein interfaces, with higher resolution and accuracy.

In conclusion, the mapping proposed in the present work is not the only one possible and for the study of specific problems it could be the case to adopt mappings at lower resolution, based on secondary or even tertiary structure elements. Also for these types of mappings a simplified electrostatics can be proposed, preserving the main physical properties of biopolymers, such as macrodipoles or other simplified charge distributions. In this way one could study the dynamics mechanism of binding and aggregation between biomolecules with improved accuracy. In this context, the optimization strategy used in this work and based on particle swarm optimization could contribute to generate consistent models at lower levels of resolution based on properties extracted from the underlying high-resolution treatment of the molecular interactions.

Chapter 4

All-atom simulations of crowding effects on protein dynamics

> *È piccolo il mondo, eh?*
>
> *Si! Piccolo e affollato!*
> CONTINUAVANO A CHIAMARLO TRINITÀ
>
> *The world is small, right?*
>
> *Yes! Small and crowded!*
> TRINITY IS STILL MY NAME

4.1 Preface

In the present chapter I will present the work done to get insights, through the use of state-of-the-art molecular dynamics techniques, into the influence of crowding agents on the internal dynamics of proteins. In particular, ubiquitin has been chosen as model system to study the role of small crowding molecules. The initial part of the present chapter has been accepted for publication in *Physical Biology* with the title ALL-ATOM SIMULATIONS OF CROWDING EFFECTS ON UBIQUITIN DYNAMICS, Luciano Andrés Abriata[1], Enrico Spiga[1], Matteo Dal Peraro. The final part, reporting on the role of crowding concentration is instead currently under preparation for submission.

---

[1] The first two authors contributed equally to this work

4.2 Introduction

Cellular environments are crowded by solutes and macromolecules of all sizes, creating conditions far from those present in typical in vitro experiments [194]. The solute concentration in the cytoplasm of *E. coli* cells is estimated to be in the order of 300-400 g/L [194; 195]. This can be conceptualized visually and numerically in recent computational works [195; 196] to reach the conclusion that intracellular solutes must inevitably experience non-specific, unintended interactions with each other. Especially for macromolecules, such interactions are expected to introduce significant alterations in the thermodynamics and kinetics of the processes they are involved in. In the case of proteins, some of these alterations have been observed and quantified through experiments in which a property is quantified in the presence of increasing quantities of a crowding agent such as small sugar moieties, dextrane, PEG, Ficoll or other proteins. One of the most important and generalized findings is that the crowding agent usually improves the stability of globular proteins to the extent that it can even induce folded states on chemically denatured proteins and on natively unfolded proteins [196–203]. These stabilizing effects are typically in the order of a few kcal/mol, but they represent important contributions and are expected to be relevant in vivo because proteins are only marginally stable [197]. But on the other hand, structural stabilization and crowding itself are expected to alter the landscape of the conformational space accessible for motions, potentially compromising dynamic features important for protein function and regulation and for protein-protein interactions.

Although little is known about the underlying processes that drive crowding effects, most experiments suggest that they are rather non-specific, supporting the notion that repulsions and steric (*i.e.* entropic) effects are dominant. However, recent works reported approximately similar contributions from repulsions and chemical (*i.e.* enthalpic) interactions between proteins and crowders, although no specific interactions were observed with the crowders [198; 199]. Unfortunately, atomistic investigations of the impact of crowding on protein structure and dynamics are inherently difficult: in X-ray studies crystal packing does not leave space for crowding assays, whereas in NMR studies the increase in viscosity produced by high crowder concentrations broadens signals eventually rendering them undetectable. As an example, NMR studies could not go beyond 100 g/L of solutes, which would be desirable to achieve crowding levels as those inside cellular environments [198–200].

Like for many experimentally intractable phenomena, molecular dynamics simulations are a powerful alternative to study the problem of interest at atomistic level.

Indeed simulations have been carried out on a few proteins under cell-like crowded conditions, helping to understand some of the effects of crowding [196; 201–206]. Among these works, most treated the crowder molecules as coarse, big spheres that cannot account for the effects of chemical interactions. Only the most recent simulations have employed fully atomistic descriptions of the systems [196; 202; 203], starting to reveal the importance of enthalpic effects together with recent experiments [198; 199]. However, none of these works studied the effect of crowding on the internal protein dynamics. In order to start filling this gap, the effect of glucose crowding at 325 g/L on the dynamic features of ubiquitin through all-atom MD simulations has been studied here.

Ubiquitin stands as an interesting workhorse protein for studying the effect of crowding on protein structure and dynamics because both aspects have been vastly studied through experiments and simulation [135; 207–218]. Such studies have disclosed how structural fluctuations are intimately linked to the protein's capacity to interact with other proteins to achieve its primary function of targeting their fates in the cell, in-between what conformational selection and induced fit models predict [135; 213; 219–221]. Also, ubiquitin has been the subject of a few experimental studies under crowded conditions [198; 222; 223]; more specifically, it has been reported that wild type ubiquitin and three destabilized mutants of this protein gain stability in solutions crowded artificially with glucose or dextrose and that the amount of gained stability is similar for the wild type and destabilized mutants, and for glucose and dextrane, suggesting that the operating mechanism is unspecific [222]. Also, a very recent work revealed that crowding effects on ubiquitin are mediated by unspecific chemical interactions and repulsion effects in roughly similar amounts [198].

Our extensive knowledge on ubiquitin dynamics and its role in target recognition is a picture derived from experiments in dilute conditions and simulations in explicit water. However, no studies have assessed the effect of crowding on the protein's dynamic features at atomistic level. With this motivation in mind, simulations of ubiquitin in pure water and in 325 g/L of glucose as a starting point to study this matter were carried out and compared. Glucose at 325 g/L was chosen to mimic a crowded medium because (i) despite being still far from a true biological medium it is a popular solute crowder used in experimental studies, and (ii) because such studies have shown that it exerts several effects on proteins including stabilization of the protein fold. Moreover, (iii) accurate atomistic force fields for molecular dynamics simulations of glucose are available [24] allowing to obtain a realistic modeling of a simple crowded environment for ubiquitin.

4.3 Methods

4.3.1 Molecular dynamics simulations

Simulations were run with the NAMD [26] code using the amber99SB [16] force field for the protein, TIP3P [55; 56] for water and Glycam06 [24] parameters for α-D-glucose. The structure of human ubiquitin in PDB ID 1D3Z [224] (model 1) was used as a starting point for minimization, equilibration and production in both simulations. For the simulation of ubiquitin in water the molecule was solvated in a TIP3P box (315396 $Å^3$), whereas for the simulation under crowding conditions the molecule was first put together with 317 glucose molecules using the Packmol [225] program, and then solvated with TIP3P water molecules. After minimization and equilibration the volume of the box was 291135 $Å^3$ resulting in an average concentration of 325 g/L.

4.3.2 Principal components analysis (PCA) and projection of the trajectories

To build the reference frame for trajectory projection, PCA was carried out on the covariance matrix of C_α positions for residues 2-70 of 72 high-resolution X-ray structures of human ubiquitin aligned to model 1 of PDB ID 1D3Z [224]. This is the same strategy used by others [135; 219], but is based on a larger set of structures. This set includes free monomers, covalent oligomers and complexes with binding partners, all solved at a resolution equal or better than 2.5 Å (list of PDB codes and PDB titles in Tab. 4.1). Projection of the trajectories on the reference frame was also preceded by alignment to model 1 of PDB ID 1D3Z. The sum of squared PCA loadings (i.e. eigenvalues) 1 and 2 shows that the structural variability that spreads the X-ray structures on the reference frame arises from sequence segments 6-11, 33-36 and 46-49 (Fig. 4.2B) which show indeed different conformations in the set of structures (RMSF from structures in Fig. 4.1).

Table 4.1: PDB files used to build by PCA the reference frame used for projection of the simulations. All ubiquitin chains from these PDB were included

1P3Q	3BY4	3TMP	3NS8	1UBI	3IFW	2AYO
1UZX	1WRD	1S1Q	3ONS	3EHV	2XEW	2C7M
1XD3	2WWZ	2PHW	3N32	2W9N	3ALB	1CMX
2D3G	3B0A	3RUL	2IBI	3HM3	1UBQ	1NBF
2G45	3B08	3UGB	2O6V	3K90	2C7N	
2HD5	3MHS	3AXC	1TBE	3H7S	2FCQ	
2OOB	3NHE	3U30	2JF5	3H7P	1AAR	
2QHO	3PRP	3AUL	3EFU	3A33	1YD8	

4.3.3 Calculation of free-energy landscapes

Free-energy landscapes were built from the trajectories projected on the reference frame by binning the projection into a grid and counting the number of frames inside each cell of the grid (N_i) to compute its free energy according to:

$$\Delta G_i = -k_B T \ln\left(\frac{N_i}{N_0}\right) \tag{4.1}$$

where N_0 corresponds to the most populated bin, which thus sets the free energy offset. Strictly speaking this is not exactly a free-energy evaluation.

4.3.4 Interactions between ubiquitin and other proteins in X-ray structures

The average number of ubiquitin-partner interactions per residue was approximated as the average (through all X-ray structures that contained a protein interacting with ubiquitin) number of C_α atoms of ubiquitin-bound proteins that are within 5 Å of each C_α atom of ubiquitin.

4.3.5 Calculation of NMR parameters from MD trajectories

N-H order parameters (S^2) and residual dipolar couplings (RDC) were computed from the trajectories using equations reported in the literature. The trajectory is aligned to model 1 of PDB ID 1D3Z and the normalized N-H vectors μ are retrieved for each peptide bond in each frame of the aligned trajectory. The calculation of S^2 is done

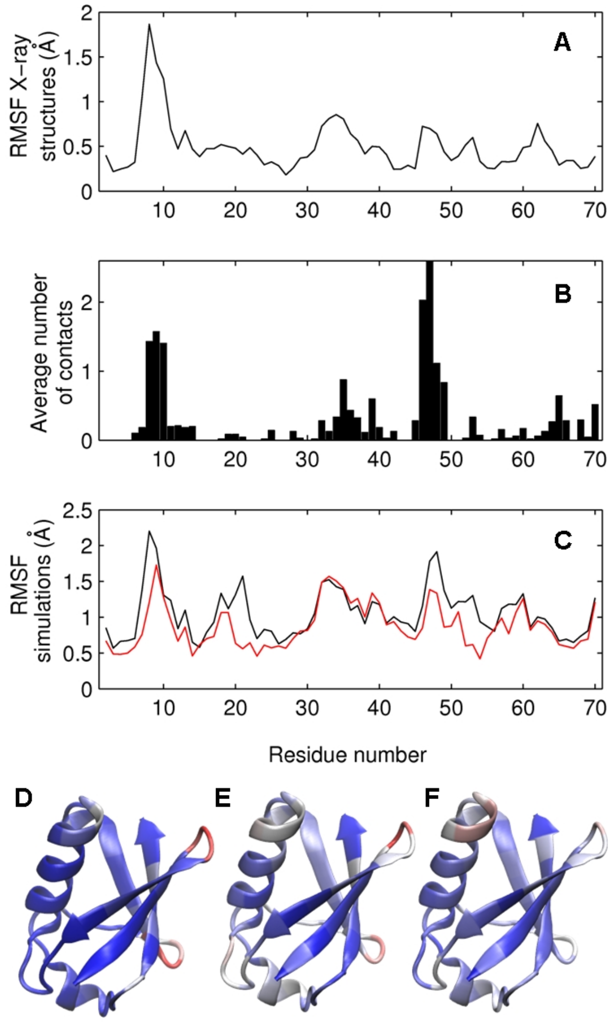

Figure 4.1: Functional flexibility of ubiquitin as derived from X-ray structures and MD simulations. (A) RMSF observed in the 72 X-ray structures on which PCA was carried out. (B) Average number of interactions observed for each ubiquitin residue with all other proteins in the 72 X-ray structures used for PCA. (C) RMSF derived from simulations of ubiquitin in water (black) and 325 g/L glucose (red). Panels D, E and F color-code the same information on the structure of ubiquitin (increasing from blue to red): average number of contacts (D), and flexible regions as determined from the RMSF in water (E) and in 325 g/L of glucose (F). The C-terminus (residues 71-76) is not reported because it is missing in some structures due to its high flexibility (trimmed in panels D-F).

as described elsewhere [226] assuming full decay of the autocorrelation function for μ

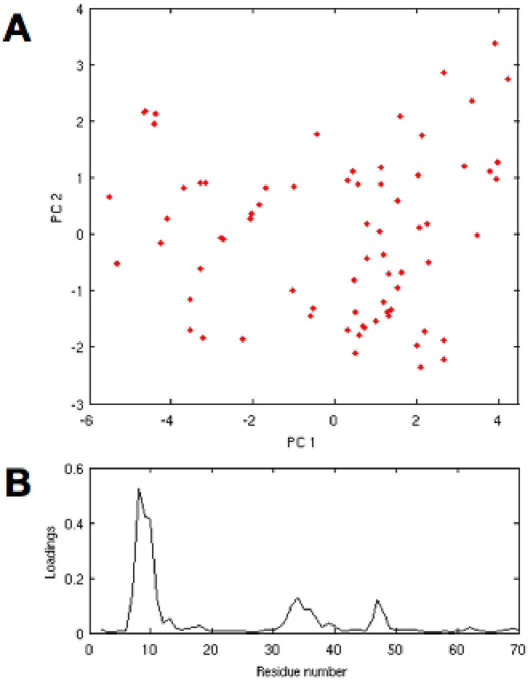

Figure 4.2: Construction of the reference frame for projecting simulations. (A) Principal components 1 against 2 (in Å), for the X-ray structures used to build these axes (each red point is a structure). (B) Squared sum of loadings 1 and 2.

during the time of the trajectory:

$$S_i^2 = \frac{1}{2}\left[3\sum_{\alpha=1}^{3}\sum_{\beta=1}^{3}\left\langle \mu_{i,\alpha}\mu_{i,\beta}\right\rangle^2 - 1\right] \tag{4.2}$$

where the average $\langle \mu_{i,\alpha}\mu_{i,\beta}\rangle$ is computed over the whole trajectory (see Chapter 2, section 2.4). The N-H residual dipolar couplings were computed for each alignment medium for which data are available, as described by Showalter et al [227]. Briefly, for each medium an alignment vector $\mathbf{A}$=(A_{xx}, A_{yy}, A_{xy}, A_{xz}, A_{yz}) is fit that minimizes the difference between the experimental RDC data (vector $\mathbf{D}_{exp}$ with as many elements as N-H peptide bonds for which data is available in that alignment medium) and the back-calculated RDC values ($\mathbf{D}_{back}$ of same size) in a least-squares sense. The back-calculated data is $\mathbf{D}_{exp} = \mathbf{M}\times\mathbf{A}$, where $\mathbf{M}$ is a matrix derived from the trajectory formed by one row of $\langle\mu_x^2\rangle - \langle\mu_y^2\rangle$, $\langle\mu_y^2\rangle - \langle\mu_z^2\rangle$, $2\langle\mu_x\mu_y\rangle$, $2\langle\mu_y\mu_z\rangle$ and $2\langle\mu_y\mu_z\rangle$ values for every N-H pair for which data are available. The least-squares problem ($\mathbf{M}\times$ $\mathbf{A}$ - $\mathbf{D}_{exp}$ $\rightarrow$ min) is solved by singular value decomposition as described in detail in that same work [227].

4.3.6 Root mean squared fluctuations of C_α atoms in protein ensembles

Root mean squared fluctuations (RMSF) were computed for each C_α atom as:

$$RMSF_i = \sqrt{\frac{\sum_j^N (\mathbf{r}_{i,j} - <\mathbf{r}_i>)^2}{N}} \tag{4.3}$$

where r_i denotes the position of C_α atom i, and j runs through the N frames of simulated time or X-ray structures. C_α atoms 2 to 70 were considered for the ensemble of X-ray structures and C_α atoms 1 to 76 for the simulations. Calculations were done after alignment to a reference structure (i.e., model 1 of 1D3Z).

4.3.7 Kinetic description of basins in the conformational landscapes

The formalism proposed by Hess [136; 137] was carried out on the basins observed in the conformational landscape of each simulation to describe them in terms of their deepness and roughness. First, ∼25 ns-long sections of the trajectories were identified in which the protein remained inside each basin according to the projection on the two-dimensional reference frame built from X-ray structures. For each sub-trajectory, PCA was carried out to obtain the most important fluctuations inside each basin, and the first and second eigenvalues (λ) were used to compute the force constant k that defines the harmonic well in the direction of the two most important eigenvectors, using the formula:

$$k = \frac{k_B T}{\lambda} \tag{4.4}$$

where k_B is Boltzmann's constant and T=300K. The autocorrelation functions were computed for the first principal components and their first decaying parts were fitted to obtain the reported τ parameters. The internal friction-like coefficient η is then:

$$\eta = k\tau \tag{4.5}$$

which measures the roughness of the basin completing its description together with the harmonic well constant k.

4.3.8 Internal conformational diffusion

The internal conformational diffusion were estimated calculating for the C_α atoms the following quantity:

$$< \Delta\mathbf{r}(t')^2 >= \frac{1}{t^{max}} \sum_{t_0=1}^{t^{max}} \frac{1}{N_{atoms}} \sum_{i=1}^{N_{atoms}} (\mathbf{r}(t_0 + t') - \mathbf{r}(t_0))^2 \tag{4.6}$$

The calculated functions were then fitted with a power-law function:

$$< \Delta\mathbf{r}(t')^2 > \simeq D_{eff} t'^{\alpha} \tag{4.7}$$

where D_{eff} is an effective internal diffusion coefficient and α measures the deviation from Brownian diffusion [138].

4.3.9 Analysis of water and glucose residence times

Interactions between water or glucose molecules and the protein were investigated in both simulations by computing survival probabilities for water-protein and sugar-protein contacts. Two atoms were considered to be in contact when the distance between them was lower than 1.1 times the sum of their van der Waals radii [139–141]. The survival probabilities were computed calculating the function:

$$N_w(t) = \frac{1}{N_t} \sum_{n=1}^{N_t} \sum_{j} P_j(t_n, t) \tag{4.8}$$

where $P_j(t_n, t)$ takes the values of 1 if the j^{th} water sugar molecule is in contact with the protein between time t_n and $t_n + t$, and zero otherwise, and N_t is the number of frames. The calculates survival probabilities were then fitted to a stretched exponential combined with two or more standard exponentials:

$$N_w(t) \simeq n_s e^{-\left(\frac{t}{\tau_s}\right)^{\gamma}} + \sum_{i=2}^{4} n_i e^{\left(-\frac{t}{\tau_i}\right)} \tag{4.9}$$

where n_s and n_i are the number of water/sugar molecules with residence times τ_s and τ_i on the surface of the protein.

4.4 Crowding effects on ubiquitin dynamics

4.4.1 Simulation of ubiquitin dynamics in water

The dynamic features of ubiquitin in dilute solution have been vastly studied by NMR, revealing nano- to micro-second timescale motions that are important for target recognition [135; 213; 219–221]. These experimental observations were complemented with molecular dynamic simulations in several works to further dissect the contributions of conformational selection and induced fit mechanisms on target binding [219; 220; 228]. Of particular importance, 0.5 to 1 μs-long simulations performed with amber99SB, a highly accurate force field that fairly reproduces NMR dynamic data, showed that free ubiquitin is constantly exploring conformational states that resemble those observed in X-ray structures of the protein complexed with different partners [219; 220]. Based on these works, it has been performed a reference 500 ns long simulation of free ubiquitin in water using the amber99SB force field. As reported for previous simulations, this simulation can also reproduce fairly well the experimental order parameters available from relaxation and RDC data [207] (Fig. 4.3A) and the residual dipolar couplings measured in [227] alignment media [210; 211; 217; 229–231] (Fig. 4.3B). The RMSF profile derived from the simulation (black curve in Fig. 4.1C and Fig. 4.1E) follows that observed in a set of X-ray structures of ubiquitin bound to different partners (Fig. 4.1A), both pointing at increased dynamics of the loops involved in the recognition of ubiquitin's binding partners (Fig. 4.1B and 4.1D).

All in all, this analysis shows that the amount of flexibility observed in the trajectory is consistent with experimental data in dilute solution, and that flexible regions of the protein are important for its ability to bind different partners. However, the analysis does not reveal any information about the conformational space accessible to the protein through collective motions as a result of such flexibility. In order to explore this, the simulation were projected on a two-dimensional reference frame built by mapping through principal components analysis the structural variability observed in 72 experimental X-ray structures of human ubiquitin under different scenarios (Fig. 4.2A, where each red point is an X-ray structure). This method has been recently introduced to analyze simulation trajectories and ensembles for which no specific reaction coordinate can be defined a priori [135; 219; 220]. Briefly, this method combines the collective fluctuations of the most variable regions of the protein into two new variables, i.e. the principal components 1 and 2, which account for as much fluctuation as possible. Projection of the C_α coordinates of our trajectory on this two-dimensional frame (Fig. 4.4A in blue, where only 1 frame was plot every 1 ns for simplicity) shows that the protein explores the whole range of conformations sampled by the X-ray structures, covering smoothly the conformational space. Fig. 4.4B is a simpler and more informative representation of the projection, in which the density of frames projected on a grid was converted into a free energy landscape relative to the deepest well. The largest barriers separating these states are around 1 kcal/mol, indicating a nearly flat

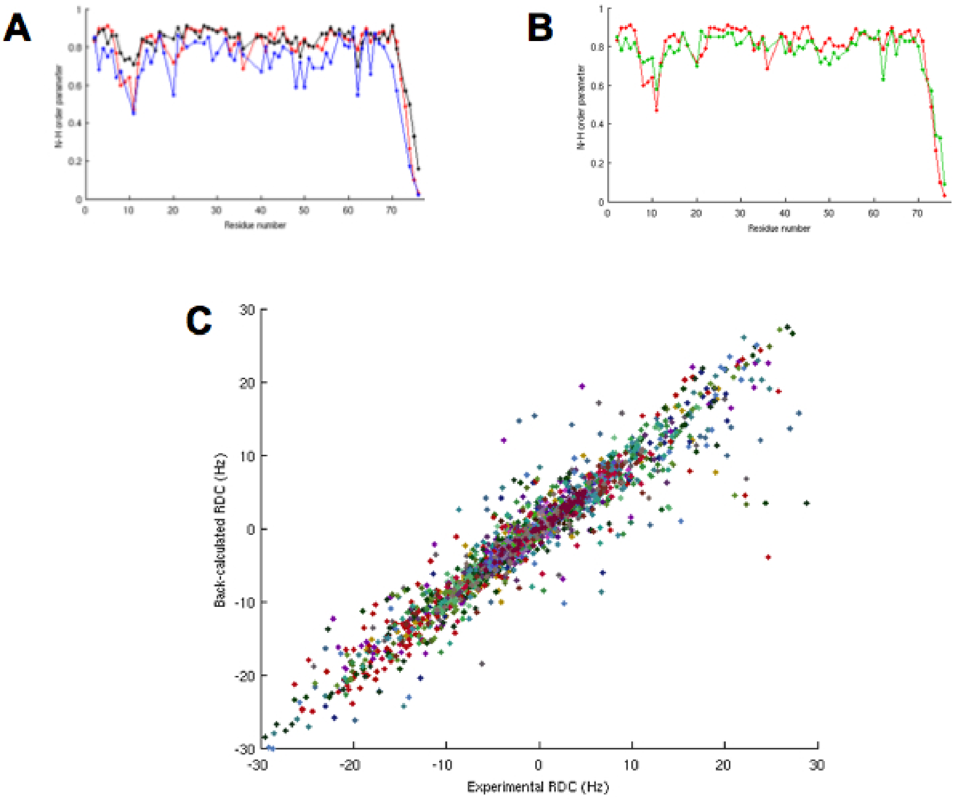

Figure 4.3: Assessment of the simulation in water against NMR data. (A) N-H order parameter computed from the 500 ns-long simulation of ubiquitin in water (red) compared to the values derived from NMR relaxation data (black) and from RDCs (blue). (B) N-H order parameter computed from the 500 ns-long simulation of ubiquitin in water (red) compared to the averaged order parameters from NMR relaxation and RDC data (green). (C) Correlation between experimental and back-predicted N-H residual dipolar couplings. Each color corresponds to a different alignment medium.

energy surface consistent with the very fast exchange observed between conformational states and with the fast dynamics determined by NMR.

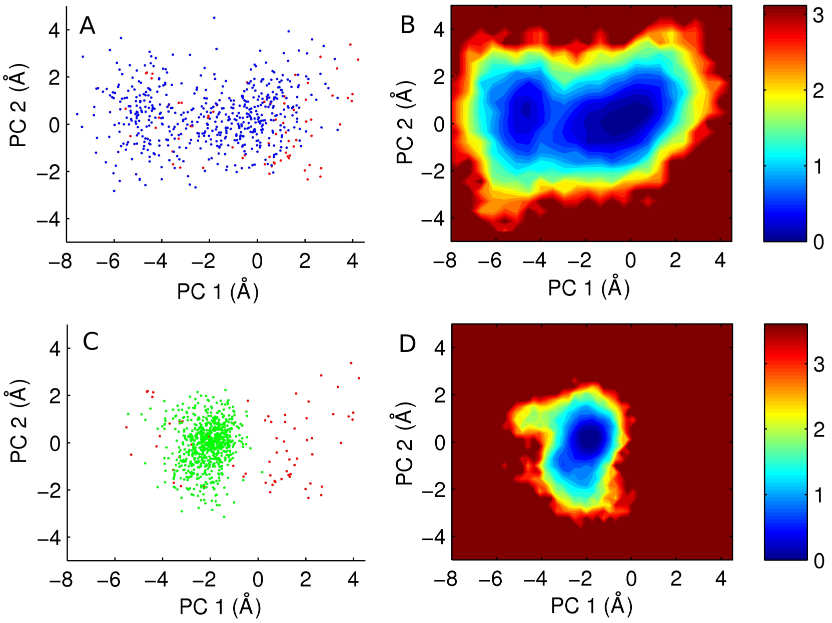

Figure 4.4: The conformational landscape sampled by ubiquitin in water and 325 g/L glucose. (A and C) 500 frames of the 500 ns trajectories in water (blue in A) and glucose (green in C) projected on the reference frame built through PCA of X-ray structures (red points). (B and D) Free energy landscapes computed from both MD trajectories relative to the lowest energy point in each of them, color-coded as shown on the bars in kcal/mol.

4.4.2 Ubiquitin dynamics in 325 g/L glucose

In the next step a 500 ns-long MD trajectory of ubiquitin in a solution crowded at 325 g/L of glucose were investigated. For this simulation the amber99SB force field [16] to describe the protein, and the Glycam06 [24] force field to describe glucose molecules, were combined. This setup leaves around the protein molecule a random distribution of water and glucose molecules in a ratio of about 20:1; details on system setup and simulation are given under Methods. Comparison of the RMSF in 325 g/L glucose against the values computed in water (red vs. black in Fig. 4.1C, respectively) shows that the amplitudes of loop motions are slightly reduced under crowded conditions during the timescale of the simulations. Notably, the effect on RMSF is not as drastic as that observed in a previous simulation of myoglobin under much more crowded, near-dry conditions at almost 90 % w/w sucrose [232; 233]. As a result of the mild reduction in loop flexibility, the RMSF and S^2 values of the loops become all similar in the crowded condition, indicating similar mobility at 325 g/L crowder concentration. Thus it seems that crowding restricts the mobility of loops in an amount proportional to the amplitude of their potential fluctuations. Importantly, and as shown in Fig. 4.5, this equalizing effect is also seen on the N-H order parameters, which can be obtained experimentally and compared to test our findings. In order to explore the impact of crowding on collective motions and the conformational landscape explored by the protein, the simulation was projected on the two-dimensional reference frame built above. Projection of 1 frame every 1 ns (Fig. 4.4C) shows that the protein has not explored the same conformational space as in water during the same time length. Instead, during the simulation the protein is restricted to small excursions slightly away from its initial conformation, always staying within the same basin. This is also evident in the time evolution of the conformational space sampled in Fig. 4.6. The finding of a strong restriction in the size of the explored conformational space with only a minor reduction in RMSF indicates that local fluctuations take place to similar extent under both conditions but collective motions (i.e. transitions between basins) have been compromised or more likely slowed down in the crowded condition defined by small sugar molecules.

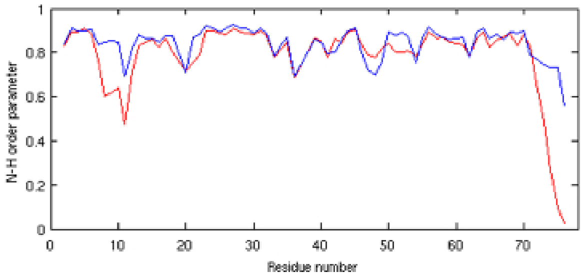

Figure 4.5: Effect of glucose crowding on the N-H order parameter. The order parameter for each non-proline residue computed from the MD simulations in water (red) and in 325 g/L glucose (blue).

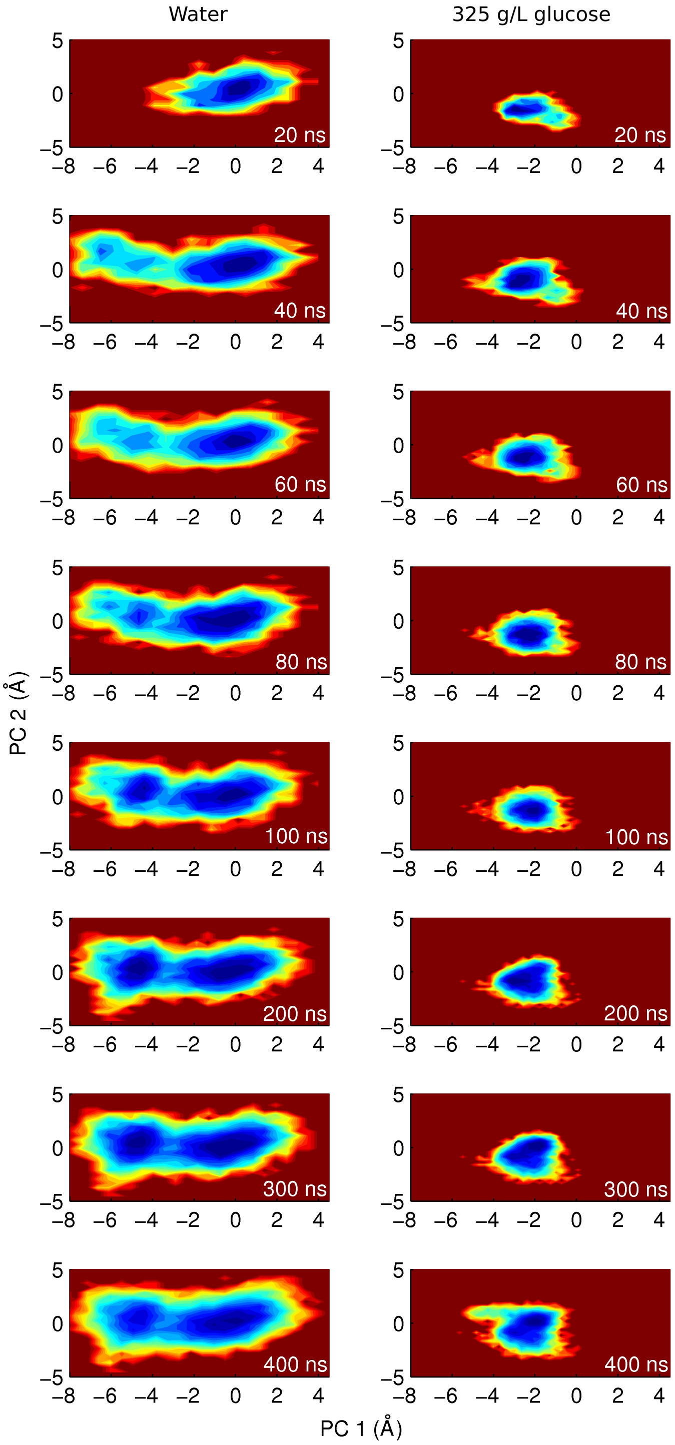

Figure 4.6: Crowding slows down the kinetics of exploration of the conformational space. Columns show the conformational space sampled at increasing time points of the simulations in water (left) and in 325 g/L glucose (right). The time corresponding to each free energy plot is indicated on the bottom left of each panel.

4.4.3 Exploration of basins in both simulations

In only 40 ns of simulation time the protein simulated in water has escaped from the initial energy well and has already explored most of the conformational space that it will sample during the rest of the simulation (Fig. 4.6, left). In contrast, the simulation in 325 g/L glucose is still trapped around its starting point at 300 ns (Fig 4.6, right) and has made only one unsuccessfull attempt to visit an alternative conformation at around 400 ns (Fig. 4.4 and 4.7). Moreover the speed at which the protein explores the conformational space is slower in the crowded condition than in water. In order to quantify this, the rate at which the conformational basins are explored has been explored by using the formalism proposed by Hess [136; 137] to model the diffusion of a protein inside a conformational basin in terms of internal friction coefficients (η) and harmonic force constants (k) characteristic of the basin. To do this ∼25 ns-long sections of the trajectories in which the protein remained inside each of the two basins of the simulation in water and in the single basin observed in glucose solution were analyzed. The parameters obtained for the two first principal components of motion inside each basin are given in Tab. 4.2. The two basins observed in water are characterized by similar harmonic force constants, whereas the basin observed in the crowded condition features a ∼5 times stronger k meaning a deeper and sharper energy well. The later is also rougher with η being an order of magnitude larger and meaning an increased "friction" and a much slower exploration of the basin in the presence of crowder. In

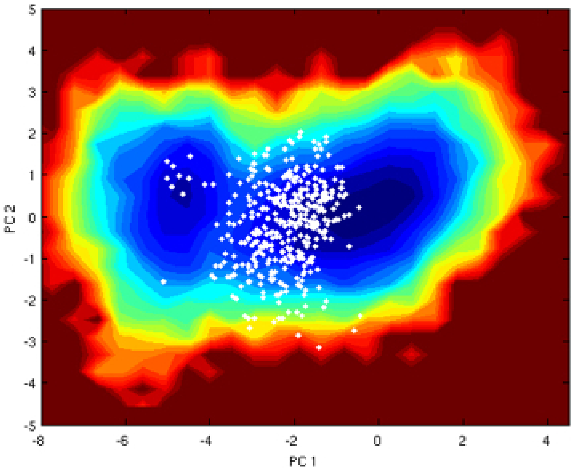

Figure 4.7: Conformational space explored by ubiquitin in water and in glucose. The free energy landscape of ubiquitin in water from Fig. 4.4B, on which 400 frames of the trajectory in 325 g/L glucose have been projected.

Table 4.2: Analysis of internal dynamics within each conformational basin

	λ [Å^2]	k [kJ$\cdot$ mol^{-1} nm^2]	τ [ps]	η [amu$\cdot$ps^{-1}]
Water (left basin)				
PC 1	8.84	28.2	9420	$2.66\cdot10^5$
PC 2	3.78	65.9	5860	$3.86\cdot10^5$
Water (right basin)				
PC 1	6.49	38.5	1350	$5.19\cdot10^4$
PC 2	3.46	72.1	681	$4.91\cdot10^4$
325 g/L glucose				
PC 1	1.46	170	7950	$1.36\cdot10^6$
PC 2	0.69	360	3700	$1.33\cdot10^6$

brief, these results predict that crowding slows down internal dynamics by exerting an increased "internal friction" and possibly by imposing higher energy barriers around the conformational basins reducing the rate of exchange between them. Importantly, the effect is more important for larger scale motions. Similar ideas implying that crowding slows down internal motions have been suggested by experimental observations in the last few years [234; 235]. Also, recent computational and experimental studies showed that the confinement of a protein inside a micelle reduces its internal dynamics due to interactions with the surfactant [236; 237]. Thus, little by little the notion that dynamics are affected by crowding is gaining further support and starting to be connected with the interactions that take place between crowders and the protein at its surface.

4.4.4 Extensive interactions between glucose molecules and ubiquitin

Considering the high concentration of glucose molecules in the simulation box, encounters between these molecules and the protein are very likely to happen. Indeed, visual inspection reveals extensive associations of several glucose molecules with the protein, resembling the multimolecular complexes observed in a recent all-atoms simulation of the cytoplasm of *E. coli* [196]. In the absence of crowders the protein interacts with an average of 236.6 water molecules at any given moment, whereas in 325 g/L glucose this decreases to an average of 80.3 water molecules due to an average of 27.2 glucose molecules that become in contact. This means that every ~156 water molecules that are excluded from the surface, ~27.2 glucose molecules are attached, resulting in a ratio of ~5.7 water/glucose present at the protein surface. This ratio is consistent with the larger area of glucose and shows how interactions between the crowder and the protein result in desolvation of its surface, as observed through simulations and experiments for near-dry conditions [232; 233; 238] but to a lesser extent. Notably, five glucose molecules never leave the protein surface along the whole trajectory. This does not imply constant interaction of the glucose molecules with the same protein residues; rather, the glucose molecules shuffle interactions to different atoms of the protein wandering around its surface. For example, there is a glucose molecule that interacts with the first loop of the protein switching hydrogen bonds between its hanging alcohol groups and different atoms of the protein during the whole simulation (Fig. 4.8D-E).

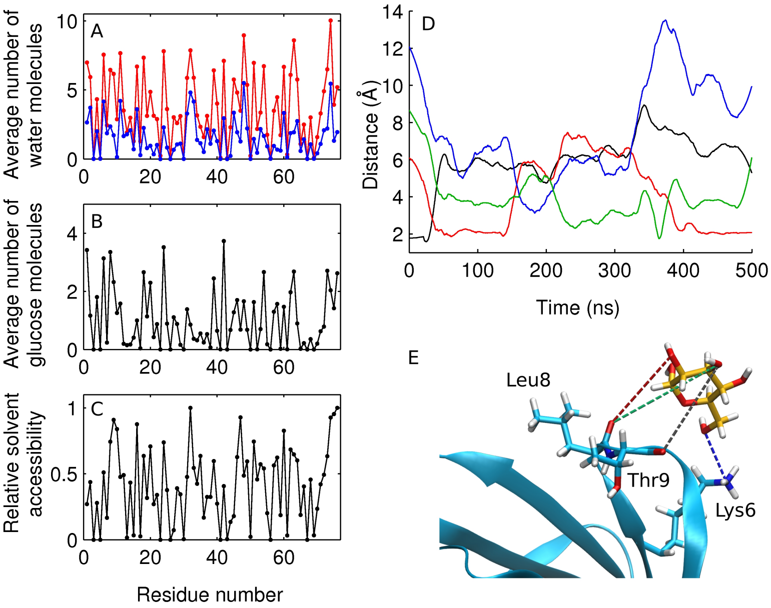

Figure 4.8: Interactions between ubiquitin and water/glucose molecules. (A) Average number of water molecules in contact with each residue at any given moment along the simulations with (blue) and without (red) crowders. (B) Average number of glucose molecules in contact with each residue in the simulation in 325 g/L glucose. (C) Relative solvent accessibility for each residue, as calculated with ASAView [239] (D and E) Typical sugar-ubiquitin interactions during MD. The time evolution of relevant distances between three different oxygen atoms of one glucose molecule and different ubiquitin atoms of the first loop is reported in panel D, while in E their spatial distribution is reported using the same color code.

Analyses per residue (Fig. 4.8A-C) reveal homogeneous desolvation and glucose binding along the exposed parts of the protein, suggesting extensive enthalpic perturbations that average smoothly on the protein surface. This picture is in line with the recent finding that unspecific interactions are an important component driving crowding effects on top of the already known steric effects. Regarding desolvation, perturbations of the hydration structure and dynamics have been predicted in a recent simulation meant to assess the effect of protein rather than small-molecule crowders [203]. In the case of glucose molecules, however, the perturbation of hydration structure and dynamics on the first water shell seems stronger, which can be attributed to the much smaller size of glucose and its higher density of polar groups. The results were interpreted thinking that glucose molecules bound to the protein and water molecules trapped in-between drag the exposed residues. Since the interactions are not specific, the effect is generalized and decorrelates collective motions making them less likely to happen; in other words, slowing them down. Further work is underway to better understand these effects of crowding on the internal protein dynamics, and to evaluate the role of protein-glucose, protein-water and glucose-water interactions (see next section).

4.5 Effect of crowding concentration and initial conditions on ubiquitin dynamics

In the previous sections it was shown how ubiquitin's internal dynamics is strongly slowed down under glucose crowding at 325 g/L compared to a reference simulation in water. A simulation in water could vastly explore in 40 ns all the states observed for ubiquitin in crystallographic (apo and bound) conformations and previously reported MD simulations [135; 213; 220], whereas the simulation in 325 g/L glucose could only fluctuate around the starting structure during a 500 ns simulation. However, a number of issues still remain open for the specific case of ubiquitin and likely in general for protein dynamics. First, does crowding affect diffusion through a conserved landscape, or does it alter the shape of this landscape? Second, what happens at an intermediate crowder concentration, or if the simulations are launched from different initial states of the conformational landscape observed in water? Third, is there an effect of crowding on the rate of exploration of the conformational space inside a basin, and how does this relate to the first two issues? In an attempt to answer these questions and better describe the effects of crowding on ubiquitin's internal dynamics, the effect of glucose concentrations and initial conditions were studied. Finally, the effect of crowding on the mobility of glucose and water molecules around the protein surface was addressed, trying to describe how interaction events at the protein surface connect to the protein's internal mechanics.

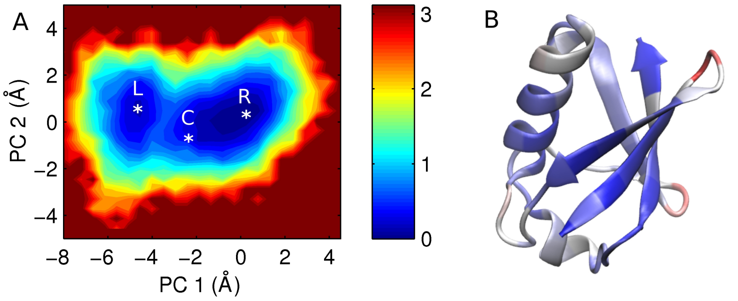

Figure 4.9: (A) Free energy landscape explored by ubiquitin in a 500 ns long simulation in water, as derived from its projection on the 2 first principal components that describe the variability observed in X-ray structures. Asterisks denote starting points for simulations under crowding conditions. (B) The most flexible segments of ubiquitin as revealed from the analysis of variation in X-ray structures, which is well reproduced by MD simulations as shown in several works.

In particular, looking at the conformational space of ubiquitin (Figure 4.9) two main basins can be identified: a smaller one centered at around [-4.5, 0.5] Å in PC1-PC2 space, and a bigger one centered at around [0, 0] Å. Based on the shape of this landscape and on the previous results, three relevant states were defined: point L (after left) that corresponds to the deepest point of the smallest basin, point R (after right) that corresponds to the deepest point of the biggest basin, and point C (after center) that corresponds to the starting structure of the simulation in water reported above. The additional two states (i.e., R and L) were used to start additional MD simulations and served to another purpose. Namely, some simulation studies suggest that this big basin might actually be two basins separated by a small energy barrier (with minima approximately on points C and R). In the previous section it was reported how ubiquitin in 325 g/L glucose was restricted to small fluctuations around the starting conformation (around point C) instead of exploring the whole conformational landscape defined by the X-rays structures or free MD in water. Therefore, simulations run at an intermediate glucose concentration of 108 g/L and at 325 g/L, each starting from these three distinct points were performed and analyzed.

4.5.1 Simulation in water and 108 and 325 g/L glucose, starting from the structure of free ubiquitin

Projection of the simulation in water at increasing time steps (Fig. 4.10, left) shows that in only 40 ns the protein has explored most of the conformational space it will explore in the rest of the simulation. Instead, the simulation in 108 g/L started from the same structure (i.e. that of free ubiquitin, point C of the conformational landscape in water) shows that after 500 ns the protein is stuck in what corresponds roughly to the rightmost basin of the simulation in water (Fig. 4.10, center). Comparison of the ending points of the simulations in water and in 108 g/L glucose suggests that the broad basin explored in the crowded condition matches fairly well with the rightmost basin of the landscape obtained in water. The simulation started from the same point in 325 g/L glucose (Fig. 4.10, right) reveals a more extreme situation, with only one tight basin that corresponds to the left area of the rightmost basin observed in water (or of the single basin observed in 108 g/L glucose).

Ideally, the two simulations under crowded conditions could be extended as long as needed expecting transitions to occur from the current basins to alternative basins. This could in principle allow to determine whether similar basins are available in the three conditions but separated by higher energy barriers as the crowding concentration increases, or if a radical perturbation of the landscape takes place in the presence

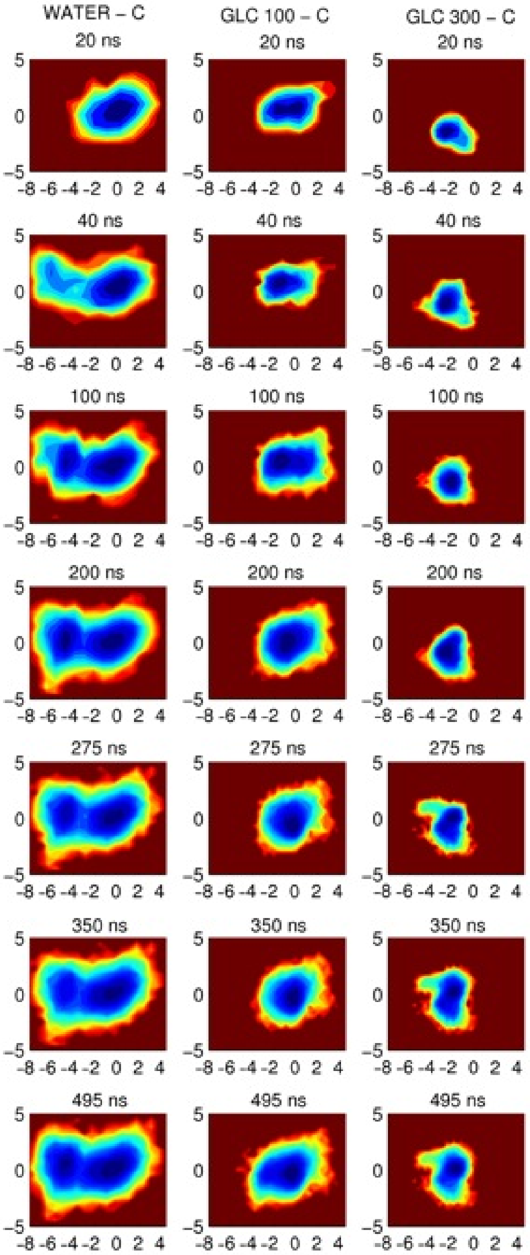

Figure 4.10: Projection of the C_α coordinates of ubiquitin at increasing times of simulation in water (left), 108 g/L glucose (center) and 325 g/L glucose (right), all started from the structure of free ubiquitin (point C of the conformational landscape observed in water, Figure 4.9)

of crowders removing and/or shifting the basins. However, a large number of such transitions would be required to build an actual free energy profile, becoming more computational demanding. As an alternative approach, MD simulations starting from points L and R of the conformational landscape observed in water, each at 108 and 325 g/L concentration of glucose as a crowder were performed.

4.5.2 Exploring alternative states of the ubiquitin conformational landscape

In MD simulations starting from points L and R at 325 g/L of glucose the protein remains stuck around each starting conformation during the whole simulation time (Figure 4.11). This shows that each conformation is very stable against interconversion into other conformations in the submicrosecond timescale, implying very high energetic barriers between them relative to those in water. The simulations in 108 g/L (Figure 4.11) show an intermediate effect, with only a few interconversion events taking place between the different basins. This precludes proper computation of the conformational landscapes for these simulations, but for sure indicates that at 108 g/L glucose the basins are separated by more modest barriers, slightly higher than those present in water but much lower than those in 325 g/L. This is even more evident calculating, using a Boltzmann inversion, from the density probabilities of PC1 the pseudo free energy profiles along this principal component (Fig. 4.12). From the figure 4.11 it is clear that the glucose concentration is determining an increase of the energetic barriers to be overcame to pass from a minimum to another (this ΔG estimation has to be considered qualitative, since more rigorous estimation can be obtained only by larger sampling or by using enhanced sampling techniques).

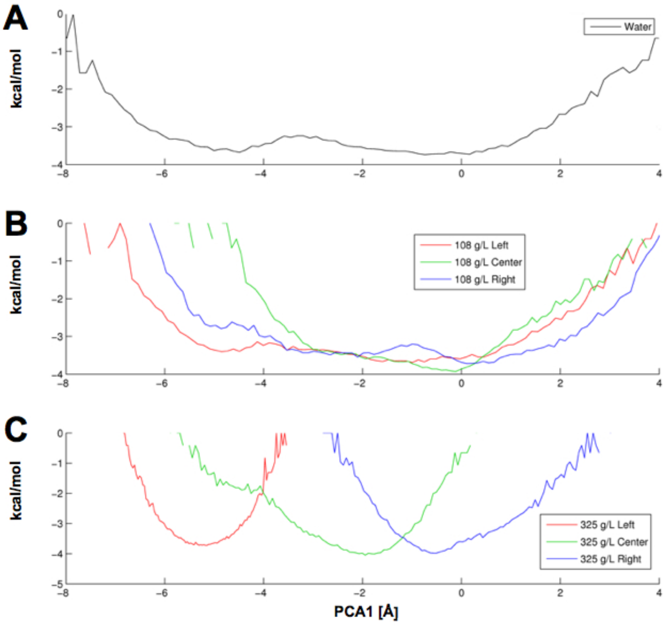

Figure 4.11: Pseudo free energy profiles of PCA1 calculated using Boltzmann's inversion. (A) Simulation in water, (B) Simulations in 108 g/L glucose concentration, (C) Simulations in 325 g/L glucose concentration

Structural fluctuation properties

From the calculation of structural fluctuation properties it is clear that the increasing glucose concentration generally decreases, in correspondence of the loops, the intensity of the fluctuations (Fig. 4.13). In fact the RMSF calculated from the simulations at 100 g/L glucose concentration are quite in line with the simulation carried out in water (Fig. 4.13A), whereas in the simulations performed at 325 g/L glucose concentration some peaks are almost abolished (Fig. 4.13C). For the S^2 order parameter the results are somehow different because passing from 108 g/L to 325 g/L the peaks are sufficiently reproduced (Fig. 4.13B and D). The different concentrations of sugar seem to modulate in different ways these two structural dynamics properties. In fact high concentration of sugar seems to decrease fluctuations around average position of C_α allowing at the same time the same angular amplitude movements of backbone dipoles.

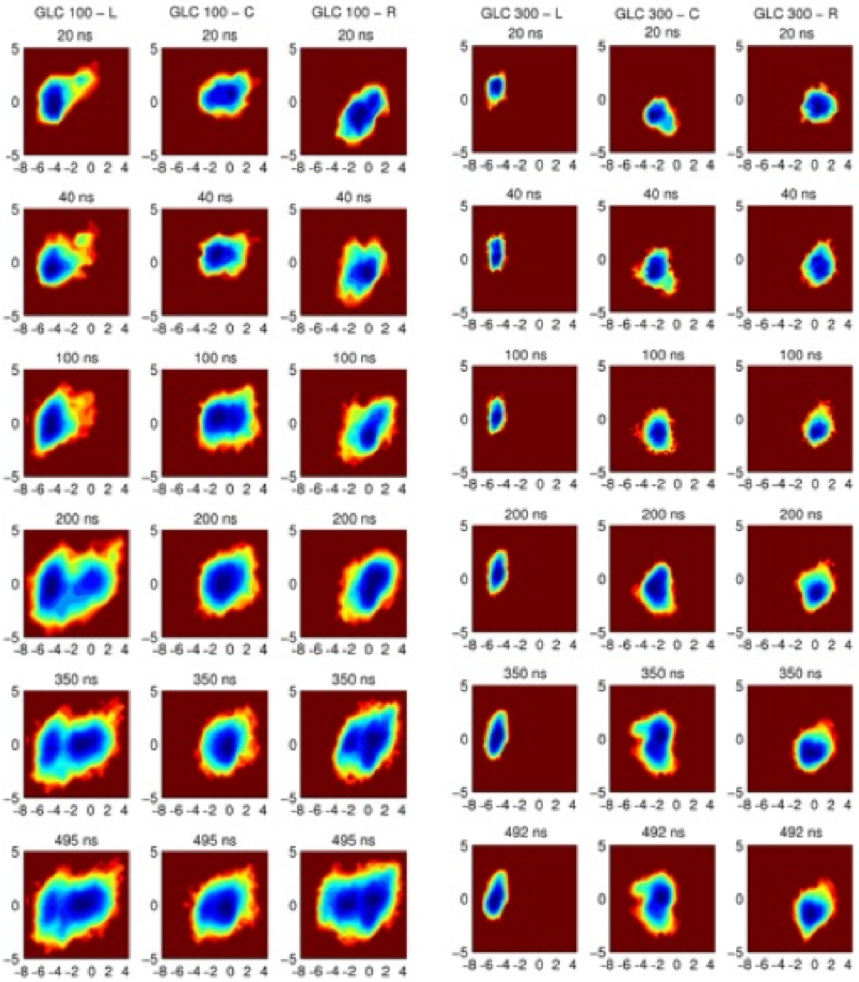

Figure 4.12: (Left) Projection of the C_α coordinates of ubiquitin at increasing times of simulations in 108 (left) and 325 (right) g/L glucose starting from points L, C and R of the conformational landscape

Internal conformational diffusion

In order to quantify the diffusion velocity in the conformational space in absence and presence of glucose the internal conformational diffusion was evaluated. This has been done calculating the effective diffusion coefficient in the conformational space of ubiquitin using the formula 4.6. From Fig. 4.14 it is clear that the diffusion on the conformational space depends on the concentration of glucose as already Fig. 4.9, 4.10 and 4.12 showed. Anyway this analysis permits to show that, despite different starting conditions, the same concentration of sugar is modulating of the same amount the

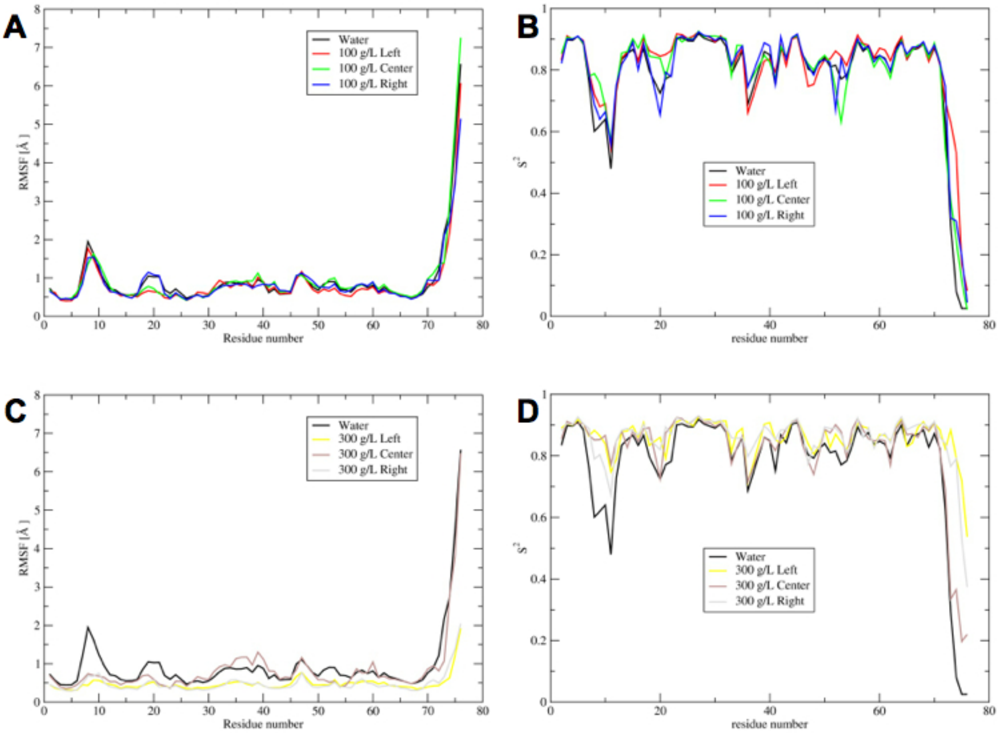

Figure 4.13: Comparison of structural fluctuations properties. (A) Comparison of RMSF between simulation in water and simulations at 100 g/L glucose; (B) Comparison of S^2 order parameter between simulation in water and simulations at 100 g/L glucose; (C) Comparison of RMSF between simulation in water and simulations at 300 g/L glucose; (D) Comparison of S^2 order parameter between simulation in water and simulations at 300 g/L glucose

diffusion in the conformational space (Tab. 4.3). The sugar strongly influences the effective diffusion in the conformational space of ubiquitin without altering significantly the sub diffusive regime, despite the fact that small differences exist in the parameter α (Tab. 4.3). In fact the increasing of glucose concentration corresponds to a decreasing of the effective diffusion coefficient passing from 0.272 in water to 0.049 in 325 g/L glucose concentration. The effective diffusion coefficients at 108 g/L glucose concentration are instead in between the value in water and the values in 325 g/L glucose concentration (Tab. 4.3, Fig. 4.14).

Table 4.3: Results of the fit of the internal conformation diffusion curves (Fig. 4.14) with equation 4.6

Type simulation	D_{eff}	α
Water	0.272	0.123
108 g/L (L)	0.123	0.138
108 g/L (C)	0.121	0.140
108 g/L (R)	0.093	0.172
325 g/L (L)	0.082	0.121
325 g/L (C)	0.049	0.190
325 g/L (R)	0.073	0.125

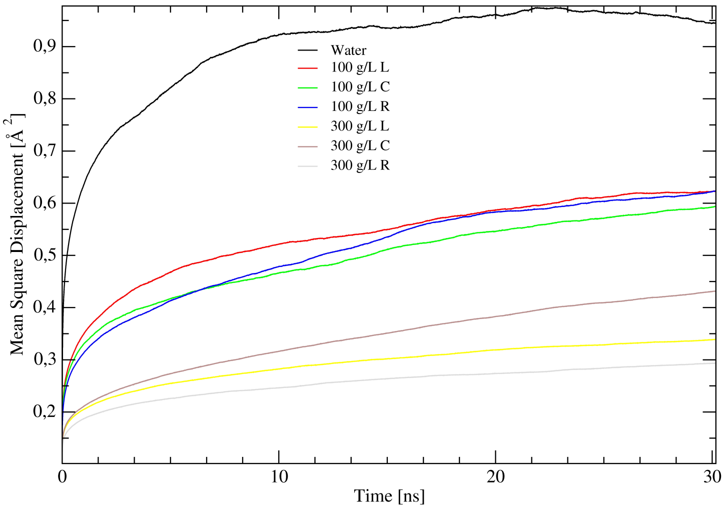

Figure 4.14: Comparison of internal conformational diffusion calculated from the different simulations

Kinetic description of the basins

In order to better understand the causes leading to a slow-down of atomic motions, a mechanical description of the basins observed in all the simulations presented above has been performed based on a formalism proposed by Hess [136; 137], which models the diffusion of the protein through the conformational basin in terms of internal friction coefficients (η) and harmonic force constants (k) for the principal components of motions inside the basin (see Chapter 2). For each basin in each simulation, the first principal components during ∼25 ns-long windows of the trajectories in which the protein remained inside the basin were analyzed.

Internal friction coefficients and harmonic force constants turned out to be very sensitive to glucose concentration, for the first principal component of motion. The case of component 1 is illustrated in Fig. 4.15, where both parameters are plotted as a function of glucose concentration. The force constant on component 1 grows exponentially with glucose concentration leading to sharper and more conformationally restricted basins. Assuming nearly fixed positions for the basins, this would lead to their crossing at higher energies lowering the probability of exchange between them, in agreement with the previous results. In parallel, internal friction grows roughly linearly with glucose concentration hampering diffusion inside the basins. In summary, this analysis reveals a double effect of the crowder on the topology of the basins.

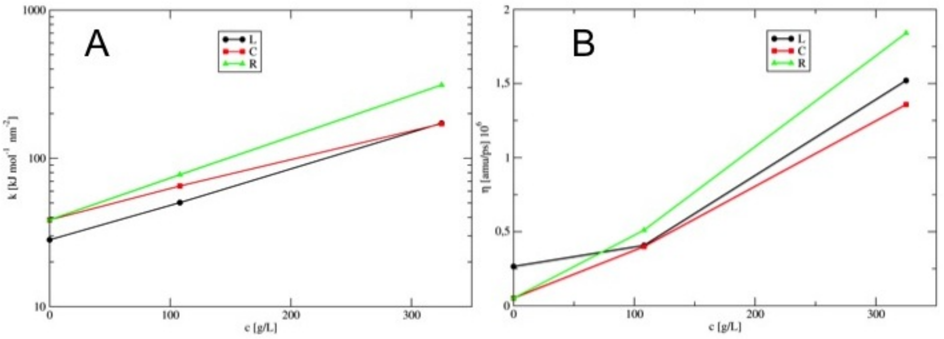

Figure 4.15: Plots of harmonic force constants k (A) and internal friction coefficients η (B) for component 1 of the main basins observed in the simulations as a function of glucose concentration

Interaction of water and glucose molecules with the protein surface

In order to quantify and better understand these phenomena, the residence times and diffusion properties of water and glucose molecules on the surface of the protein were analyzed. For this purpose it has been used Marchi's formalism for the calculation of survival probabilities for protein-solvent/solute interactions [140; 141]. Regarding water-protein interactions, the work by Marchi et al. identified three characteristic residence times for water molecules on the surface of proteins based on ∼10 ns-long simulations in pure explicit water. Following this approach, the MD trajectory in water has been analyzed, where additional (slower) time scales can be expected due to the much longer simulation time of the present MD simulations. The trajectories for the glucose, the survival probabilities of glucoses presence on the surface of the protein were estimated.

In the simulation of ubiquitin in pure water an average of 290 molecules interacting along the trajectory. Most of them (80 %) experience the fastest possible exchange regime with a residence time of 30 ps diffusing slightly off the Brownian limit (i.e., Kohlrausch's stretching parameter $\gamma = 0.899$ against a value of 1 for Brownian motion). Around 17 % of the molecules belong to the second time regime with an average residence time of 140 ps, and 2 % have an average residence time of 920 ps corresponding to the third regime. Very few molecules (0.35 %) have residence times longer than 1 ns, reaching an average of 13.4 ns. No water molecule remains bound to the protein for the entire simulation suggesting no exchange processes slower than 10 nanoseconds.

The presence of glucose in the solution induces strong non-Brownian diffusion on the fastest timescale of water motions, as revealed by the drop in Kohlrausch parameters. As shown in Fig. 4.16 glucose molecules interact with the protein removing water molecules from its surface. The effect is stronger at higher glucose concentrations and independent of the starting conformation; in particular, the observations at 325 g/L indicate that the number of attached glucose molecules and detached water molecules is similar in the three main conformations. More interestingly, Fig. 4.17 reveals that despite the drop in the total number of water molecules wetting the protein surface, the numbers of water molecules with the two longest residence times (n_3, n_4) increase with concentration. Fig. 4.18 further shows that the residence times of water molecules also increase steeply with glucose concentration. Altogheter, this evidence suggests that the binding of glucose molecules to the surface of the protein traps water molecules. Since the residence times of glucose molecules are at least 1-2 orders of magnitude larger than those of water molecules, the overall effect is dramatic even at only 100 g/L glucose.

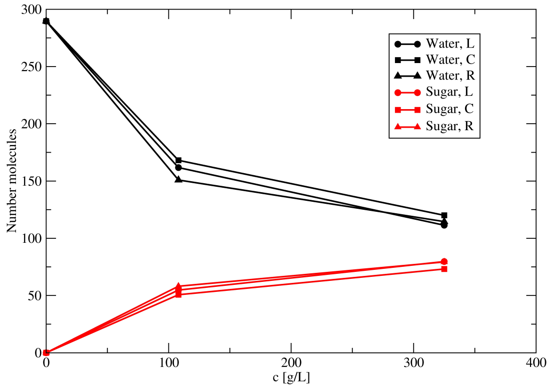

Figure 4.16: Number of water and sugar molecules on the protein surface as function of sugar concentration

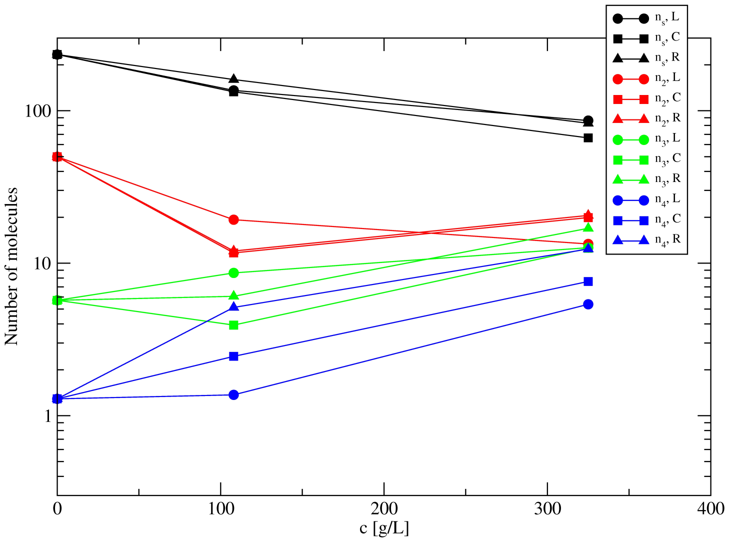

Figure 4.17: Number of water molecules exchanging interactions with the protein surface on different time scales: sub-nanosecond (n$_s$, in black), 0.1-100 ns (n$_2$ and n$_3$, in red and green), and the slowest process (n$_4$, blue) reaching up to 380 ns

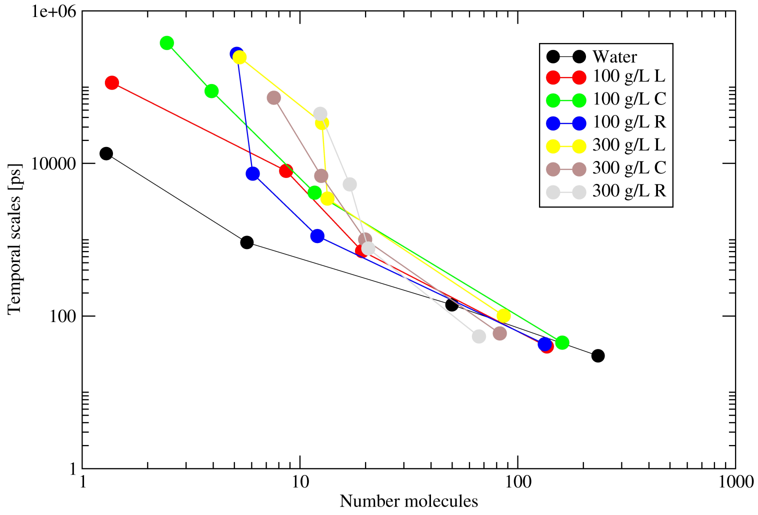

Figure 4.18: Correlation plots for the number of water molecules under each residence timescale in the seven simulations.

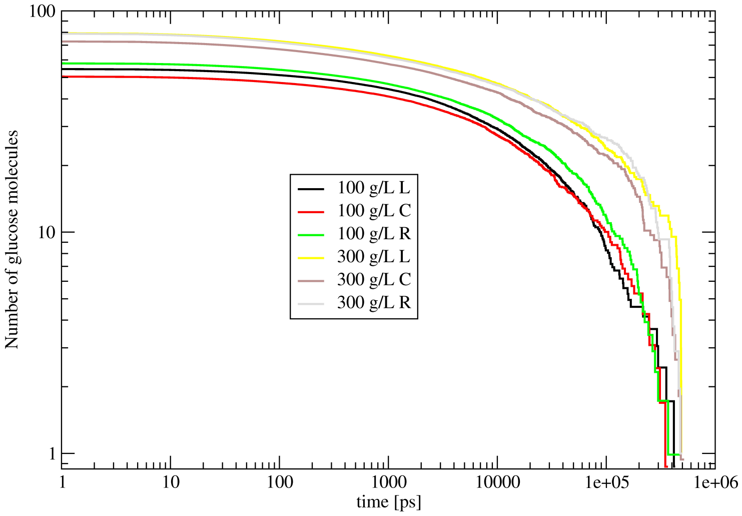

Figure 4.19: $N_S(t)$ for sugar molecules on the protein surface in dependence of sugar concentration

Table 4.4: Number of water and glucose molecules attached to the protein surface as calculated from simulation

Type simulation	$N_W(0)$	$N_S(0)$
Water	289.6	0.0
108 g/L (structure L)	161.7	54.7
108 g/L (structure C)	168.1	51.7
108 g/L (structure R)	150.9	59.1
325 g/L (structure L)	111.4	81.7
325 g/L (structure C)	120.0	78.1
325 g/L (structure R)	114.5	81.3

Fig. 4.19 shows the functions $N_S(t)$ as calculated from simulations at different glucose concentrations and starting conditions, whereas Tab 4.4 makes a summary about the number of water or glucose molecules always attached to the protein surface. It is clear that sugar molecules remain attached to the protein surface for all the simulation time forcing the protein to be dynamically slaved by glucoses at sub microsecond temporal scales. Moreover, the number of glucose molecules attached to the protein surface increases as the glucose concentration increases amplifying the effect of the dynamic slavery. The water and glucose molecules attached to the protein surface put weight on it slowing down the conformational exploration. In fact if only water is present the weight of waters on the protein surface is ~5200 Da (being 18.015 Da the water weight). At 108 g/L glucose concentration the weight of the attached molecules (waters and glucoses) is ~12.000 Da (being 180.156 Da the α-D-glucose weight), whereas at 325 g/L glucose concentration is ~23.000 Da. The weight is double passing from 0 g/L to 108 g/L and passing from 108 g/L to 325 g/L. These mass values together with the temporal scales of water and glucose permanence on the protein surface make clear the origin of the dynamic slavery (i. e. water and glucose fluctuations dominate protein dynamics [240]) due to the presence of the glucoses.

4.6 Discussion

The set of extended MD simulations carried out in this work allowed to describe and explain the influence of different concentrations of glucose on ubiquitin internal dynamics. The main finding, highlighted by the PCA analysis, is that the increment of the concentration of glucose, independently from the starting conditions, determines a slow down of the exploration of the conformational space. The evaluation of the effective diffusion coefficients permits to quantify the diffusion velocity in the ubiquitin conformational space. The results permitted to connect the slowing down of the conformational exploration to the glucose concentration.

The Hess analysis complements the previous findings showing that the internal dynamics of ubiquitin, when trapped in a minimum of free-energy, is largely affected by different concentration of glucose. The increment of glucose concentration determines that minima become at the same time steeper and rougher respect to the corresponding minima in water, being respectively k and η higher. The steeper and rougher nature of the minima has an effect on the structural dynamics properties as RMSF and S^2 order parameter. In fact higher concentrations of glucose correspond to a smoothing of the RMS fluctuations whereas S^2 order parameters seem affected but in a different way, showing that angular amplitudes of backbone dipoles are almost the same. This result is in line with a recent paper where ubiquitin where simulated as confined in a inverse micelle [236].

The evident slow down of the conformational space exploration is clearly due to the presence of glucose molecules. The presence of glucose molecules determines a more viscous medium, keeping trapped ubiquitin for longer time in stable and metastable conformations. The analysis of the contact survival probabilities protein-water and protein-glucose permitted to quantify the temporal scales of existence of such contacts. The glucose molecules, belonging to the first shell of contacts, interact with the surface of the protein on a temporal scale of several tens of nanoseconds. This temporal scale together with the observation that a glucose molecule is heavier than a water molecule, having at the same time more hydrogen bond acceptors and donors, permits to understand that the internal motions of ubiquitin are slowed down because the "effective mass" of the protein is increased. The presence of glucose molecules reduces the number of waters on the protein surface diminishing the accessible surface and at same time increases the water permanence time trapping them. The drastic alteration of the solvation mechanism does not permit a fast exchange of waters with the bulk. The long time permanence of glucose on the protein surface in combination with the slow exchange of waters turns out in the slow exploration of the ubiquitin conformational space.

Although the constituents of the cellular medium are chemically varied and not simply glucose rings, they contain exposed polar and charged groups capable of establishing electrostatic interactions, salt bridges and hydrogen bonds with protein surfaces, which are largely polar and charged. In our view, based on the present and others' results, all these transient, unspecific interactions average out to a constant enthalpic and entropic perturbation of the protein surface. This perturbation would imply both thermodynamic stabilization of the protein fold and a slow down in all large-scale kinetic processes thus slowing down unfolding rates too. In this regard, all the observations are in agreement with the experimental reports that crowding enhances the native structure of proteins [241; 242] and that part of this stabilizing mechanism acts on the folding pathways themselves [205]. On top of this, these finding that collective motions are slowed down allow to propose that crowding agents can stabilize proteins by simply trapping the native conformations, *i.e.* through a kinetic contribution. Similar ideas were suggested by different experimental and computational evidence [243; 244].

Notably, these effects would also stabilize protein complexes, as well as complexes composed of different kinds of macromolecules as reported in a recent all-atoms simulation of *E. coli*'s cytosol. This might be of little importance for strongly interacting proteins but would be important for proteins involved in interactions defined as "transient" by experiments. In particular, this can explain why some interactions predicted by genetic evidence are not confirmed by *in vitro* experiments. Stabilization of protein-protein interactions by crowding would have profound implications for many proteins, including ubiquitin itself. Moreover, crowding could have an effect on the relative weight of induced-fit and conformational selection mechanisms that guide protein-protein recognition.

Finally, one can speculate that the marginal stability observed for proteins in solution arises from the fact that they have evolved to work in crowded, stabilizing environments. Under this scenario, potential proteins that are too stable in dilute solution would be unnecessarily too stable in the cellular medium, and possibly too rigid to function properly.

4.7 Conclusions

Crowding effects are very important for life as we know it, where they influence the thermodynamics and kinetics of cellular chemistry, and in biotechnology where they can be exploited to stabilize protein samples and in protein refolding protocols. I have herein investigated the effect of glucose crowding on the internal dynamics of a protein, providing a starting point to better describe the interplay between protein-sugar, protein-water and sugar-water interactions at the protein surface at atomic level. These results back up the notion that crowding might be an element with which cells manipulate protein dynamics affecting their folding, function and regulation. As increasingly recognized by these findings and by several recent works, more realistic insights into the biological physics and chemistry of the cell might be obtained by including crowders in experimental and computational research.

These results also point out an important aspect in this field of research. While the entropic contribution of crowding agents can be modeled through coarse-grained models as already done in other works [245–248], the use of all-atom MD simulations is still necessary for taking into account the important enthalpic contributions of intermolecular interactions.

Despite the interesting outcomes of this work, there are several possible directions that I am exploring to extend the reach of this computational investigation. First, sampling is surely a limitation in this study and extending the duration of the simulations in order to reach at least the μs timescale should provide further insights. This could permit to observe transitions among the minima obtained at high concentration of glucose, allowing a more precise estimation of the free-energy profiles. In parallel, the use of enhanced sampling techniques, like replica exchange molecular dynamics [249], would also allow to have a more quantitative insight into the pseudo free energy landscape under different glucose concentrations. Furthermore, α-D-glucose is not the only type of crowder used in *in vitro* experiments and certainly is not the optimal crowder to mimic in vivo cellular conditions. In order to have a exhaustive picture of the influence of crowding agents, one should consider the effects of agents of different sizes and physicochemical properties (*e.g.* proteins, nucleic acids, metabolites etc.). In this study, glucose was chosen because its size allowed for an atomistic detailed analysis. Given the ever-growing sampling capability of atomistic MD simulations, the extension of this approach to more realistic physiological situations could be soon realized.

Chapter 5

Conclusion and perspectives

> *We choose to go to the moon. We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills, because that challenge is one that we are willing to accept, one we are unwilling to postpone, and one which we intend to win, and the others, too.*
>
> John F. Kennedy

In this thesis I presented two different contributions for the extension of the present boundaries of biomolecular modeling, namely the development of simplified coarse grained models for soluble proteins, and the application of state-of-the-art atomistic molecular dynamics to investigate the effect of molecular crowding on protein dynamics.

The development of a novel coarse-grained model for proteins aims at extending the possible sampling of molecular simulations retaining at the same time a sufficient accuracy in describing intermolecular interactions. The key point of this model is the introduction of electrostatics, which is of upmost importance in biology for driving molecular assembly and function. The model has been tested against finer-grained simulations giving structural and dynamic results that are encouraging for the simulation of large biomolecular complexes. While this contribution represents an effort in developing models with accuracy competitive with all-atom results, I am obviously aware that such model, as others of similar nature, won't be able to capture properties accessible only to atomistic simulations (one example is in fact reported in Chapter 4). However for the treatment of very large systems, their assembly and dynamics the present model offers some improvements over current methods, being able to preserve secondary structural elements and to reproduce more faithfully electrostatics. Nonetheless, this model is still open to several improvements making appealing the possibility to spend time on it in order to extend its capabilities. Among possible applications, I can envision: the study of protein-protein interactions in molecular dynamics, the use for macromolecular assembly scoring and the elucidation of fibril formation. At the moment the field of coarse-grained modeling is not as well established as atomistic simulations, but it is gaining more and more respect permitting the simulation of phenomena that are far from being accessible with all-atom molecular dynamics techniques. Thus, improving CG model accuracy and allowing hybrid treatments of the degrees of freedom (as for example in atomistic/CG schemes) will contribute to the routinely use of CG simulations as complement to finer-grained simulation to catch and attack biological problems in a multiscale fashion.

On the other side, especially with the ever-increasing power of accessible hardware, atomistic MD simulations is a fundamental technique for the study of the physicochemical properties of living systems. The chapter about the influence of crowding agents on the ubiquitin dynamics demonstrates that state-of-the-art molecular dynamics in combination with present atomistic force fields permits to quantify and estimate effects that are not easily accessible to experiments. Some estimated quantities, like effective internal diffusion coefficients and friction coefficients, can be considered theoretical predictions that deserve an experimental confirmation. This could permit to understand if the used model systems are realistic and if current force fields are accurate enough to describe this type of phenomena. This study represents an effort in describing and treating model systems, mimicking as close as possible at least *in vitro* conditions, with the aim in the future to reproduce *in vivo* conditions. In fact, often molecular dynamics simulations aimed at giving a microscopic picture of biological problems are not always performed having in mind the typical *in vitro* or *in vivo* conditions, apart for the ionic strenght. I think that for the future developments and directions of biomolecular modeling it is important to overcome also this barrier, given that computational approaches are sometime the only way to describe a system without heavily perturbing its native conditions.

Reconnecting with the introduction of this thesis, the previous chapters could be seen as small steps in the direction to mimic through molecular dynamics investigations a cell or at least parts of it. Anyway, I think that the road that brings to the simulation of a small cell or an organelle is still really long and extremely impervious. I think that not only a theoretical and methodological progress is needed, but also the responsible acting of single entities inside an organized and fully integrated community, where theoretical physicists and chemists can speak the same language of biochemists and biologists.

To reach these goals, I think that in the future the biomolecular dynamics community would need to focus on the following points:

1. Development and improvement of accurate atomistic force fields for all the possible categories of biomolecules
2. Development and improvement of coarse-grained models and all-atom/coarse-grained hybrid models
3. Extension of the structure-function paradigm accounting for dynamical properties of biomolecules

The accomplishment of the first point would permit to perform biomolecular simulations of model systems able to reproduce, under some ranges of validity, *in vitro* and/or *in vivo* conditions. The need to work on the second point is due to the fact that computational resources will be always somehow limited. Thus the development of coarse-grained models enabling the simulations or representation of large molecular complexes at nearly atomistic level is advisable in order to save computational power. In parallel, the development of hybrids models could permit to preserve an atomistic representation where and when is needed. Having now available molecular dynamics packages and all-atom force fields able to capture the main features of biomolecules it is time to be able to reinforce the role of dynamics within the structure-function paradigm.

All the proposed points can be achieved both by single groups or by clusters of groups, but I think that is becoming clearer and clearer that what the community of biomolecular modeling would like to do is too complex to be handled by a single group.

Thus, I think that what is missing in our community is the organization of joint efforts to accelerate the achievement of these goals. I clearly understand that the problem to organize joint efforts is mainly politic rather than scientific, but I think that with a reasonable amount of diplomacy this task can be accomplished. I have in mind that could be important for the european biomolecular modeling groups to create a community like SIMBIOS or IMP and Rosetta in USA. Inside this `European Biomolecular Modeling Community` it might be possible to create Divisions that are focused on specific fields (e.g., development of codes, models and force fields, etc.). Clearly the work of each division would not be self-consistent but would benefit from the contribution of the other divisions. I think, that the scientific achievements by an organized community should definitely boost the rationalization of experimental data and the predictive power of biomolecular modeling.

Bibliography

[1] D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. Hutchison, H. O. Smith, and J. C. Venter. **Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome**. *Science*, **329**(5987):52–56, 2010. 12

[2] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, Hünenberger P. H., M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt, and H. B. Yu. **Biomolecular modeling: goals, problems, perspectives**. *Angew. Chem. Int. Ed.*, **45**:4064–4092, 2006. 12, 14, 24

[3] B. Alberts. *Molecular biology of the cell.* Garland Science, 2002. 13

[4] M. Karplus. **Molecular dynamics of biological macromolecules: a brief history and perspective**. *Biopolymers*, **68**:350–358, 2003. 13, 17

[5] M. Levitt. **The birth of computational structural biology**. *Nat. Struct. Biol.*, **8**(5):392–393, 2001. 13, 14

[6] M. Levitt and S. Lifson. **Refinement of Protein Conformations Using a Macromolecular Energy Minimization Procedure**. *J. Mol. Biol.*, **46**:269–279, 1969. 13

[7] J. A. McCammon, B. R. Gelin, and M. Karplus. **Dynamics of folded proteins**. *Nature*, **267**:585–590, 1977. 13

[8] M. Levitt and A. Warshel. **Computer simulation of protein folding**. *Nature*, **253**(5494):94–98, 1975. 13, 15, 27, 29, 44

[9] P. H. Hünenberger, A. E. Mark, and H. J. C. Berendsen. **Wilfred van Gunsteren: 35 years of bimolecular simulation**. *J. Chem. Theory Comput.*, **8**(10):3425–3429, 2012. 14

[10] W. L. Jorgensen and J. Tirado-Rives. **Potential energy functions for atomic-level simulations of water and organic and biomolecular systems**. *PNAS*, **102**(19):6665–6670, 2005. 14, 26

[11] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. **Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models**. *Accounts of Chemical Research*, **33**(12):889–897, 2000. 14

[12] P. H. Hünenberger. **Thermostat Algorithms for Molecular Dynamics Simulations**. *Adv. Polym. Sci.*, **173**:105–149, 2005. 14

[13] T. Darden, D. York, and L. Pedersen. **Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems**. *J. Chem. Phys.*, **98**(12):10089–10092, 1993. 14, 24

[14] W. J. Ponder and D. A. Case. **Force fields for protein simulations**. *Adv. Protein Chem.*, **66**:27–85, 2003. 14, 24, 25

[15] S. A. Adcock and J. A. McCammon. **Molecular dynamics: survey of methods for simulating the activity of proteins**. *Chem. Rev.*, **106**(5):1589–1615, 2006. 14, 24

[16] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. **Comparison of multiple amber force fields and development of improved protein backbone parameters**. *Proteins Struct. Funct. Bioinf.*, **65**:712–725, 2006. 14, 25, 26, 55, 90, 98

[17] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatman III, C. A. Laughton, and M. Orozco. **Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers**. *Biophys. J.*, **92**(11):3817–3829, 2007. 25, 27

[18] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. **Improved side-chain torsion potentilas**

for the Amber ff99SB protein force field. *Proteins*, **78**:1950–1958, 2010. 14, 26

[19] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. **How fast-folding proteins fold**. *Science*, **334**:517–520, 2011. 14, 44

[20] P. L. Freddolino, B. C. Harrison, Y. Liu, and K. Schulten. **Challenges in protein-folding simulations**. *Nat. Phys.*, **6**:751–758, 2010.

[21] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele. **Absolute comparison of simulated and experimental protein-folding dynamics**. *Nature*, **420**(6911):102–106, 2002. 14

[22] T. Fox and P. A. Kollman. **Application of the RESP methodology in the parametrization of organic solvents**. *J. Phys. Chem. B*, **102**(41):8070–8079, 1998. 14, 26

[23] J. P. M. Jämbeck and A. P. Lyubartsev. **Derivation and systematic validation of a refined all-atom force field for phosphatidylcholine lipids**. *J. Phys. Chem. B*, **116**(10):3164–3179, 2012. 27

[24] K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeriño, C. R. Daniels, B. L. Foley, and R. J. Woods. **GLYCAM06: a generalizable biomolecular force field. Carbohydrates**. *J. Comput. Chem.*, **29**(4):622–655, 2008. 14, 27, 89, 90, 98

[25] T. Schlick. **Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules**. *F1000 Biology Reports*, **1**:1–9, 2009. 15

[26] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and Schulten K. **Scalable Molecular Dynamics with NAMD**. *J. Comput. Chem.*, **26**(16):1781–1802, 2005. 15, 44, 55, 90

[27] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindhal. **GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation**. *J. Chem. Theory Comput.*, **4**(3):435–447, 2008. 15, 57

[28] J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, and K. Schulten. **Accelerating molecular modeling applications with graphics processors**. *J. Comput. Chem.*, **28**(16):2618–2640, 2007. 15, 44

[29] L. C. T. Pierce, R. Salomon-Ferrer, C. A. F. de Oliveira, J. A. McCammon, and C. W. Walker. **Routine access to millisecond time scale events with accelerated molecular dynamics**. *J. Chem. Theory Comput.*, **8**(16):2997–3002, 2012. 15, 44

[30] M. Harvey, G. Giupponi, and G. De Fabritiis. **ACEMD: Accelerated molecular dynamics simulations in the microseconds timescale**. *J. Chem. Theory Comput.*, **5**(6):16321639, 2009. 15

[31] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles. **Milliscond-scale molecular dynamics simulations on Anton**. *Proceedings of the ACM/IEEE Conference on Supercomputing (SC09)*, 2009. 15, 44

[32] N. F. Endres, R. Das, A. W. Smith, A. Arkhipov, E. Kovacs, Y. Huang, J. G. Pelton, Y. Shan, D. E. Shaw, D. E. Wemmer, J. T. Groves, and J. Kuriyan. **Conformational Coupling across the Plasma Membrane in Activation of the EGF Receptor**. *Cell*, **152**(3):543 – 556, 2013. 15

[33] J. D. Honeycutt and D. Thirumalai. **Metastability of the folded states of globular proteins**. *PNAS*, **87**:3526–3529, 1990. 15, 28

[34] V. Tozzini. **Minimalist models for proteins: a comparative analysis**. *Q. Rev. Biophys.*, **43**(3):333–371, 2010. 15, 27, 28, 44, 49

[35] D. A. Potoyan, A. Savelyev, and G. A. Papoian. **Recent successes in coarse-grained modeling of DNA**. *WIREs Comput. Mol. Sci.*, **3**:69–83, 2013. 27

[36] R. Goetz, G. Gompper, and R. Lipowsky. **Mobility and elasticity of self-assembled membranes**. *Phys. Rev. Lett.*, **82**(1):221–224, 1999. 44

[37] M. Venturoli and B. Smit. **Simulating the self-assembly of model membranes**. *PhysChemComm*, **10**, 1999. 15, 44

[38] D. Frenkel and B. Smit. *Understanding molecular simulation.* Academic Press, 2002. 18, 22

[39] G. Sutmann. **Classical Molecular Dynamics**. *NIC Series*, **10**:211–254, 2002. 18, 52

[40] D. S. Bond and B. J. Leimkuhler. **Molecular dynamics and the accuracy of numerically computed averages**. *Acta Numerica*, **16**:1–65, 2007. 18

[41] C. H. Andersen. **Molecular dynamics simulations at constant pressure and/or temperature**. *J. Chem. Phys.*, **72**(4):2384–2393, 1980. 19, 20

[42] S. Nosé. **A unified formulation of the constant temperature molecular dynamics methods**. *J. Chem. Phys.*, **81**(1):511–519, 1984. 19

[43] W. G. Hoover. **Canonical dynamics: Equilibrium phase-space distributions**. *Phys. Rev. A*, **31**(3):1695–1697, 1985. 19, 20

[44] M. E. Tuckerman and Martyna G. J. **Understanding modern molecular dynamics: techniques and applications**. *J. Phys. Chem. B*, **104**(2):159–178, 2000. 21

[45] M. E. Tuckerman, Liu Y., Ciccotti G., and Martyna G. J. **Non-Hamiltonian molecular dynamics: generalizing hamiltonian phase space principles to non-Hamiltonian systems**. *J. Chem. Phys.*, **115**(4):1678–1702, 2001. 20, 21

[46] Martyna G. J., M. E. Tuckerman, Tobias D. J., and Klein M. L. **Explicit reversible integrators for extended systems dynamics**. *Mol. Phys.*, **87**(5):1117–1157, 1996. 22

[47] Gibbon P. and Sutmann G. **Long-Range interactions in many-particle simulation**. *NIC Series*, **10**:467–506, 2002. 24

[48] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. **A smooth particle mesh Ewal method**. *J. Chem. Phys.*, **103**(19):8577–8593, 1995. 24, 55

[49] P. Cieplak, W. D. Cornell, C. Bayly, and Kollman P. A. **Application of the multimolecule and multiconformational RESP methodology to biopolymers: charge derivation for DNA, RNA, and proteins**. *J. Comput. Chem.*, **16**(11):1357–1377, 1995. 25

[50] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Jr. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A.

Kollman. **A second generation force field for the simulation of proteins, nucleic acids and organic molecules**. *J. Am. Chem. Soc.*, **117**(19):5179–5197, 1995. 25, 26, 55

[51] A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wirkiewicz-Kuczera, D. Yin, and M. Karplus. **All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins**. *J. Phys. Chem. B*, **102**(18):3586–3616, 1998. 25

[52] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. **Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids**. *J. Am. Chem. Soc.*, **118**(45):11225–11236, 1996. 25

[53] B. Guillot. **A reappraisal of what we have learnt during three decades of computer simulations on water**. *J. Mol. Liq.*, **101**(1-3):219–160, 2002. 26

[54] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. **The missing term in effective pair potentials**. *J. Phys. Chem.*, **91**(24):6269–6271, 1987. 26

[55] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. **Comparison of simple potential functions for simulating liquid water**. *J. Chem. Phys.*, **79**(2):926–935, 1983. 26, 55, 90

[56] D. J. Price and C. L. Brooks. **A modified TIP3P water potential for simulation with Ewald summation**. *J. Chem. Phys.*, **121**(20):10096–10103, 2004. 55, 90

[57] H. W. Horn, W. C. Swope, J. W. Piter, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon. **Development of an improved four-sit water model for biomolecular simulations: TIP4P-Ew**. *J. Chem. Phys.*, **120**(20):9665–9678, 2004. 26

[58] B. Hess and N. F. A. van der Vegt. **Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular**

force fields and water models. *J. Phys. Chem. B*, **110**(35):17616–17626, 2006. 26

[59] F. O. Lange, D. van der Spoel, and B. L. de Groot. **Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data**. *Biophys. J.*, **99**(2):647–655, 2010. 26

[60] K. A. Beauchamp, Y. Lin, R. Das, and V. S. Pande. **Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements**. *J. Chem. Theory Comput.*, **8**(4):1409–1414, 2012.

[61] E. A. Cino, W. Choy, and M. Karttunen. **Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations**. *J. Chem. Theory Comput.*, **8**(8):2725–2740, 2012. 26

[62] J. P. M. Jämbeck and A. P. Lyubartsev. **An Extension and Further Validation of an All-Atomistic Force Field for Biological Membranes**. *J. Chem. Theory Comput.*, **8**(8):2938–2948, 2012. 27

[63] J. P. M. Jämbeck and A. P. Lyubartsev. **Another Piece of the Membrane Puzzle: Extending Slipids Further**. *J. Chem. Theory Comput.*, **9**(1):774–784, 2013. 27

[64] C. J. Dickson, L. Rosso, R. M. Betz, R. C. Walker, and I. R. Gould. **GAFFlipid: a general amber force filed for the accurate molecular dynamics simulation of phospholipid**. *Soft Matter*, **8**:9617–9627, 2012. 27

[65] J. Åqvist. **Ion–water interaction potentials derived from free energy perturbation simulations**. *J. Phys. Chem.*, **94**(21):8021–8024, 1990. 27

[66] M. Levitt. **A simplified representation of protein conformations for rapid simulation of protein folding**. *J. Mol. Biol.*, **104**(5494):59–107, 1976. 27, 29, 44

[67] C. Knight and Voth G. A. **Coarse-graining away electronic structure: a rigorous route to accurate condensed phase interaction potentials**. *Mol. Phys.*, **110**(9-10):935–944, 2012. 27

[68] T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen. **Multiscale modeling of emergent materials: biological and soft matter**. *Phys. Chem. Chem. Phys*, **11**:1869–1892, 2009. 27, 28

[69] M. G. Saunders and G. A. Voth. **Coarse-graining of multiprotein assemblies**. *Curr. Opin. Struct. Biol.*, **22**:1–7, 2012. 27, 28, 44

[70] S. V. Bennun, M. I. Hoopes, C. Xing, and R. Faller. **Coarse-grained modeling of lipids**. *Chem. Phys. Lipids*, **159**:59–66, 2009. 27

[71] S. G. Ayton, W. G. Noid, and G. A. Voth. **Multiscale modeling of biomolecular systems: in serial and in parallel**. *Curr. Opin. Struct. Biol.*, **17**:192–198, 2007. 28, 44

[72] S. Takada. **Coarse-grained molecular simulations of large biomolecules**. *Curr. Opin. Struct. Biol.*, **22**:130–137, 2012.

[73] S. Riniker, J. R. Allison, and W. F. van Gunsteren. **On developing coarse-grained models for biomolecular simulation: a review**. *Phys. Chem. Chem. Phys.*, **14**(36):12423–12430, 2012. 44, 57

[74] Meier K., Choutko A., Dolenc J., A. P. Eichenberger, S. Riniker, and W. F. van Gunsteren. **Multi-resolution simulation of biomolecular systems: a review of methodogical issues**. *Angew. Chem. Int. Ed.*, **52**:2–17, 2013.

[75] M. Cascella and M. Dal Peraro. **Challenges and perspectives in biomolecular simulations: from the atomistic picture to multiscale modeling**. *CHIMIA*, **63**(1/2):14–18, 2009. 44

[76] S. C. L. Kamerlin and A. Warshel. **Multiscale modeling of biological functions**. *Phys. Chem. Chem. Phys.*, **13**:10401–10411, 2011. 44

[77] S. C. L. Kamerlin, S. Vicatos, A. Dryga, and A. Warshel. **Coarse-grained (multiscale) simulations in studies of biophysical and chemical systems**. *Annu. Rev. Phys. CHem.*, **62**:41–64, 2011. 28, 43

[78] P. Derreumaux. **From polypeptide sequences to structures using Monte Carlo simulations and an optimized potential**. *J. Chem. Phys*, **111**(5):2301–2310, 1999. 28, 45

[79] M. Pasi, R. Lavery, and N. Ceres. **PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties**. *J. Chem. Theory Comput.*, **9**(1):785–793, 2013. 36, 44

[80] S. M. Gopal, S. Mukherjee, Y. M. Cheng, and M. Feig. **PRIMO/PRIMONA: a coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy**. *Proteins*, **78**:1266–1281, 2009. 45

[81] T. Bereau and M. Deserno. **Generic coarse-grained model for protein folding and aggregation**. *J. Chem. Phys*, **130**(235106):1–15, 2009. 28

[82] V. Tozzini, W. Rocchia, and J. A. McCammon. **Mapping all-atom models onto one-bead coarse-grained models: general properties and applications to a minimal polypeptide model**. *J. Chem. Theory Comput.*, **2**(3):667–673, 2006. 28, 50

[83] D. Alemani, F. Collu, M. Cascella, and M. Dal Peraro. **A Nonradial Coarse-Grained Potential for Proteins Produces Naturally Stable Secondary Structure Elements**. *J. Chem. Theory Comput.*, **6**(1):315–324, 2010. 28, 46, 47, 49, 51, 52

[84] J. A. McCammon and S. H. Northrup. **Helix-coil transition in a simple polypeptide model**. *Biopolymers*, **19**:2033–2045, 1980. 28

[85] I. Bahar and R. L. Jernigan. **Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation**. *J. Mol. Biol.*, **266**:195–214, 1997. 28

[86] J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth. **Coarse-Grained Peptide Modeling Using a Systematic Multiscale Approach**. *Biophys. J.*, **92**(12):4289–4303, 2007. 28

[87] Tap Ha-Duong. **Protein backbone dynamics simulations using coarse-grained bonded potentials and simplified hydrogen bonds**. *J. Chem. Theory and Comput.*, **5**(12):3211–3223, 2009. 28, 30, 36, 45, 60, 79

[88] M. Zacharias. **Protein-protein docking with a reduced protein model accounting for side-chain flexibility**. *Protein Sci.*, **12**:1271–1282, 2003. 28, 45

[89] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink. **The Martini coarse-grained force field: extension to proteins**. *J. Chem. Theory and Comput.*, **4**(5):819–834, 2008. 29, 30, 36, 45, 46

[90] A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. **Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field**. *J. Chem. Phys.*, **115**(5):2323–2357, 2001. 29, 32, 36, 44, 45

[91] M. Zacharias. **Combining coarse-grained nonbonded and atomistic bonded interactions for protein modeling**. *Proteins*, **81**(1):81–92, 2013. 30, 45

[92] R. DeVane, W. Shinoda, P. B. Moore, and M. L. Klein. **Transferable coarse grain nonbonded interaction model for amino acids**. *J. Chem. Theory Comput.*, **5**(9):2115–2124, 2009. 30, 31, 34, 46

[93] S. A. Hassan, F. Guarnieri, and E. L. Mehler. **A general treatment of solvent effects based on screened coulomb potentials**. *J. Phys. Chem.*, **104**(27):6478–6489, 2000. 30, 31, 50

[94] A. Rubinstein and S. Sherman. **Influence of the Solvent Structure on the Electrostatic Interactions in Proteins**. *Biophys. J.*, **878**(3):15441557, 2004. 50

[95] A. Warshel, S. T. Russell, and A. K. Churg. **Macroscopic models for studies of electrostatic interactions in proteins: limitations and applicability**. *PNAS*, **81**:4785–4789, 1984. 31

[96] J. Ramstein and R. Lavery. **Energetic coupling between DNA bending and base pair opening**. *PNAS*, **85**:7231–7235, 1988. 31

[97] P. E. Smith and M. B. Pettitt. **Modeling solvent in biomolecular systems**. *J. Phys. Chem.*, **98**(39):9700–9711, 1994. 30

[98] N. K. Rogers. **The modelling of electrostatic interactions in the function of globular proteins**. *Prog. Biophys. Molec. Biol.*, **48**:37–66, 1986. 30

[99] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, Jr. S. Profeta, and P. Weiner. **A new force field for molecular mechanical simulation of nucleic acids and proteins**. *J. Am. Chem. Soc.*, **106**(3):765–784, 1984. 30

[100] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. **Electrostatics of nanosystems: application to microtubules and the ribosome**. *PNAS*, **98**(18):10037–10041, 2001. 31, 58

[101] K. R. Hadley and C. McCabe. **Coarse-grained molecular models of water: a review**. *Mol. Sim.*, **38**(8-9):671–681, 2012. 31

[102] K. R. Hadley and C. McCabe. **On the Investigation of Coarse-Grained Models for Water: Balancing Computational Efficiency and the Retention of Structural Properties**. *J. Phys. Chem. B*, **114**(13):4590–4599, 2010. 31

[103] J. Florian and A. Warshel. **Langevin Dipoles Model for ab Initio Calculations of Chemical Processes in Solution: Parametrization and Application to Hydration Free Energies of Neutral and Ionic Solutes and Conformational Analysis in Aqueous Solution**. *J. Phys. Chem. B*, **101**(28):5583–5595, 1997. 31, 50

[104] T. Ha-Duong, S. Phan, M. Marchi, and D. Borgis. **Electrostatics on particles: phenomenological and orientational density functional approach**. *J. Chem. Phys.*, **117**(2):541–556, 2002. 31, 50

[105] M. Orsi and J. W. Essex. **The ELBA force field for coarse-grain modeling of lipid membranes**. *Plos. One*, **6**(12):1–22, 2011. 31

[106] L. Darré, M. R. Machado, P. D. Dans, F. E. Herrera, and S. Pantano. **Another Coarse Grain Model for Aqueous Solvation: WAT FOUR?** *J. Chem. Theory Comput.*, **6**(12):3793–3807, 2010. 31, 34, 44

[107] S. Riniker and W. F. van Gunsteren. **A simple, efficient polarizable coarse-grained water model for molecular dynamics simulations**. *J. Chem. Phys.*, **134**(084110):1–12, 2011. 34, 35, 57

[108] S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta, and S. J. Marrink. **Polarizable water model for the coarse-grained MARTINI force field**. *Plos. Comput. Biol.*, **6**(6):1–17, 2010. 31

[109] F. Ercolessi and J. B. Adams. **Interatomic potentials from first-principles calculations: the force-matching method**. *Europhys. Lett.*, **26**(8):583–588, 1994. 32, 53, 54

[110] S. Izvekov and G. A. Voth. **A multiscale coarse-graining method for biomolecular systems**. *J. Phys. Chem. B*, **109**(7):2469–2473, 2005. 32, 44, 54

[111] W. G. Noid, J. W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen. **The multiscale coarse-graining method. 1. A rigorous bridge between atomistic and coarse-grained models**. *J. Chem. Phys.*, **128**(244114):1–11, 2008. 32, 44, 53

[112] J. W. Mullinax and W. G. Noid. **Generalized Yvon-Born-Green theory for molecular systems**. *Phys. Rev. Lett.*, **103**(198104):1–4, 2009. 32, 36, 44

[113] J. W. Mullinax and W. G. Noid. **Extended ensemble approach for deriving transferable coarse-grained potentials**. *Phys. Rev. Lett.*, **131**(104110):1–13, 2009.

[114] J. W. Mullinax and W. G. Noid. **A generalized-Yvon-Born-Green theory for determining coarse-grained interaction potentials**. *J. Phys. Chem. C*, **114**(12):5661–5674, 2010.

[115] J. F. Rudzinski and W. G. Noid. **Coarse-graining entropy, forces, and structures**. *J. Chem. Phys.*, **135**(214101):1–15, 2011.

[116] J. W. Mullinax and W. G. Noid. **Recovering physical potentials from a model protein databank**. *PNAS*, **107**(46):19867–19872, 2010. 45

[117] J. F. Rudzinski and W. G. Noid. **The role of many-body correlations in determining potentials for coarse-grained models of equilibrium structure**. *J. Phys. Chem. B*, **116**(29):8621–8635, 2012. 32

[118] C. R. Ellis, J. F. Rudzinski, and W. G. Noid. **Generalized-Yvon-Born-Green model of toluene**. *Macromol. Theory Simul.*, **20**:478–495, 2011. 32

[119] I. F. Thorpe, D. P. Goldenberg, and G. A. Voth. **Exploration of transferability in multi scale coarse-grained peptide models**. *J. Phys. Chem. B*, **115**(41):11911–11926, 2011. 32, 45

[120] D. R. Jr Hills, L. Lu, and G. A. Voth. **Multiscale coarse-graining of the protein energy landscape**. *PLoS Comput. Biol.*, **6**(6):1–12, 2010. 32, 44

[121] M. Scott Shell. **The relative entropy is fundamental to multi scale and inverse thermodynamic problems**. *J. Chem. Phys.*, **129**(144108):1–7, 2008. 33, 36, 44

[122] A. P. Lyubartsev and A. Laaksonen. **Calculation of effective interaction potentials from radial distribution functions: a reverse Monte Carlo approach**. *Phys. Rev. E*, **52**(4):3730–3737, 1995. 33

[123] D. Reitz, M. Pütz, and F. Müller-Plathe. **Deriving effective mesoscale potentials from atomistic simulations**. *J. Comput. Chem*, **24**(13):1624–1636, 2003. 34

[124] W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, and O. Hahn. **Simulation of polymer melts. 1. Coarse-graining procedure for polycarbonates**. *Acta Polym.*, **49**:61–74, 1998. 44, 53

[125] W. Tschöp, K. Kremer, O. Hahn, J. Batoulis, and T. Bürger. **Simulation of polymer melts. 2. Coarse-graining procedure for polycarbonates**. *Acta Polym.*, **49**:61–74, 1998.

[126] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko. **Versatile object-oriented toolkit for coarse-graining applications**. *J. Chem. Theory Comput.*, **5**(12):3211–3223, 2009. 34, 53

[127] M. Winger, D. Trzesniak, R. Baron, and W. F. van Gunsteren. **On using a too large integration time step in molecular dynamics simulations of coarse-grained models**. *Phys. Chem. Chem. Phys.*, **11**:1934–1941, 2009. 35, 36

[128] S. J. Marrink, X. Periole, P. D. Tieleman, and A. H. de Vries. **Comment on "On using a too large integration time step in molecular dynamics simulations of coarse-grained models"**. *Phys. Chem. Chem. Phys.*, **12**:2254–2256, 2010.

[129] W. F. van Gunsteren and M. Winger. **Reply to the 'Comment on "On using a too large integration time step in molecular dynamics simulations of coarse-grained models' "**. *Phys. Chem. Chem. Phys.*, **12**:2257–2258, 2010. 35

[130] J. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. **Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes**. *J. Comput. Phys.*, **23**:327–341, 1977. 36, 57

[131] R. Ishima and D. A. Torchia. **Protein dynamics from NMR**. *Nat. Struct. Biol.*, **7**(9):740–743, 2000. 37

[132] I. Chandrasekhar, G. M. Clore, A. Szabo, A. M. Gronenborn, and B. R. Brooks. **A 500 ps molecular dynamics simulation study of interleukin-1b in water: correlation with nuclear magnetic resonance spectroscopy and crystallography**. *J. Mol. Biol.*, **226**:239–365, 1992. 38

[133] P. E. Smith, R. C. van Schaik, T. Szyperski, K. Wüthrich, and W. F. van Gunsteren. **Internal mobility of the basic pancreatic trypsin inhibitor in solution: a comparison of NMR spin relaxation measurements and molecular dynamics simulations**. *J. Mol. Biol.*, **246**:356–365, 1995. 38, 58

[134] A. Amadei, A. B. M. Linssen, and J. C. H. Berendsen. **Essential dynamics of proteins**. *J. Mol. Biol.*, **17**(4):412–425, 1993. 38

[135] O. F. Lange, N. A. Lakomek, C. Fares, G. F. Schroder, K. F. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger, and B. L. de Groot. **Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution**. *Science*, **320**(5882):1471–1475, 2008. 38, 89, 90, 96, 105

[136] B. Hess. **Similarities between principal components of protein dynamics and random diffusion**. *Phys. Rev. E*, **62**(6):8438–8448, 2000. 38, 39, 94, 101, 112

[137] B. Hess. **Convergence of sampling in protein simulations**. *Phys. Rev. E*, **65**(031910):1–10, 2001. 38, 39, 94, 101, 112

[138] K. Kämpf, F. Klameth, and M. Vogel. **Power-law and logarithmic relaxations of hydrated proteins: a molecular dynamics simulations study**. *J. Chem. Phys.*, **137**(205105):1–9, 2012. 40, 95

[139] R. W. Impey, P. A. Madden, and I. R. McDonald. **Hydration and mobility of ions in solution**. *J. Phys. Chem.*, **87**(25):5071–5083, 1983. 40, 95

[140] F. Sterpone, M. Ceccarelli, and M. Marchi. **Dynamics of hydration in hen egg white lysozyme**. *J. Mol. Biol.*, **311**:409–419, 2001. 40, 41, 113

[141] M. Marchi, F. Sterpone, and Ceccarelli. **Water rotational relaxation and diffusion in hydrated lysozyme**. *J. Am. Chem. Soc.*, **124**(23):6787–6791, 2002. 40, 41, 95, 113

[142] D. Russel, K. Lasker, J. Phillips, D. Schneidman-Duhovny, J. A. Velazquez-Muriel, and A. Sali. **The structural dynamics of macromolecular processes**. *Curr. Opin. Cell Biol.*, **21**:97–108, 2009. 44

[143] T. Schlick, R. Collepardo-Guevara, L. A. Halvorsen, S. Jung, and X. Xiao. **Biomolecular modeling and simulation: a field coming of age**. *Q. Rev. Biophys.*, **44**(2):191–228, 2011. 44

[144] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, J. K. Salmon, Y. Shan, and W. Wriggers. **Atomic-level characterization of the structural dynamics of proteins**. *Science*, **330**:341–346, 2010. 44

[145] K. Y. Sanbonmatsu and C. S. Tung. **High performance computing in biology: multimillion atom simulations of nanoscale systems**. *J. Struct. Biol.*, **157**:470–480, 2007.

[146] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten. **Molecular dynamics simulations of the complete satellite tobacco mosaic virus**. *Structure*, **14**:437–449, 2006. 44

[147] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein. **A coarse grain model for phospholipid simulations**. *J. Phys. Chem. B*, **105**(19):4464–4470, 2001. 44

[148] L. M. Klein and W. Shinoda. **Large-scale molecular dynamics simulations of self-assembling systems**. *Science*, **321**:798–800, 2008. 44

[149] G. A. Voth. *Coarse-graining of condensed phase and biomolecular systems.* CRC Press, 2009.

[150] B Kolinski. *Multiscale approaches to protein modeling.* Springer, 2011.

[151] P. D. Dans, A. Zeida, and S. Machado, M. R. an Pantano. **A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics**. *J. Chem. Theory Comput.*, **6**(5):17111725, 2010.

[152] S. O. Nielsen, P. B. Moore, and B. Ensing. **Adaptive multiscale molecular dynamics of macromolecular fluids**. *Phys. Rev. Lett.*, **105**:237802, 2010. 44

[153] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. **The Martini force field: coarse grained model for biomolecular simulations**. *J. Phys. Chem. B*, **111**(27):7812–7824, 2007. 44, 45, 46

[154] W. G. Noid, P. Liu, Y. Wang, J. W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth. **The multiscale coarse-graining method. 2. Numerical implementation for coarse-grained molecular models**. *J. Chem. Phys.*, **128**(244115):1–20, 2008. 44, 54

[155] I. F. Thorpe, J. Zhou, and G. A. Voth. **Peptide folding using multiscale coarse-grained models**. *J. Phys. Chem. B*, **112**(41):13079–13090, 2008. 44, 54

[156] G. S. Ayton and G. A. Voth. **Multiscale computer simulation of the immature HIV-1 virion**. *Biophys. J.*, **99**(9):2757–2765, 2010. 44

[157] S. P. Carmichael and M. Scott Shell. **A new multi scale algorithm and its application to coarse-grained peptide models for self-assembly**. *J. Phys. Chem. B*, **116**(29):8383–8393, 2012. 45

[158] L. Larini, L. Lu, and G. A. Voth. **The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials**. *J. Chem. Phys.*, **132**(164107):1–10, 2010. 45

[159] A. V. Sinitskiy, M. G. Saunders, and G. A. Voth. **Optimal number of coarse-grained sites in different components of large biomolecular complexes**. *J. Phys. Chem. B*, **116**(29):8363–8374, 2012. 45

[160] X. Periole, T. Huber, S. J. Marrink, and T. P. Sakmar. **G protein-coupled receptors self-assemble in dynamics simulations of model bilayers**. *J. Am. Chem. Soc.*, **129**(33):10126–10132, 2007. 45

[161] M. Louhivuoiri, H. J. Risselada, E. van der Giessen, and S. J. Marrink. **Release of content through mechano-sensitive gates in pressurized liposomes**. *PNAS*, **107**(46):19856–19860, 2010.

[162] L. V. Schaefer, D. H. de Jong, A. Holt, A. J. Rzepiela, A. H. de Vries, B. Poolman, J. A. Killian, and S. J. Marrink. **Lipid packing drives the**

segregation of transmembrane helices into disordered lipid domains in model membranes. *PNAS*, **108**(4):1343–1348, 2011.

[163] G. van den Bogaart, K. Meyenberg, H. J. Risselada, H. Amin, K. I. Willing, B. E. Hubrich, M. Dier, S. W. Hell, H. Grubmüller, U. Diederichsen, and R. Jahn. **Membrane protein sequestering by ionic protein-lipid interactions**. *Nature*, **479**:552–555, 2011. 45

[164] H. A. Scheraga. **Respice, adspice, and prospice**. *Annu. Rev. Biophys.*, **40**:1–39, 2011. 45

[165] C. Czaplewski, S. Kalinowski, A. Liwo, and H. A. Scheraga. **Application of multiplexed replica exchange molecular dynamics to the UNRES force field: tests with α and α+β proteins**. *J. Chem. Theory Comput.*, **5**(3):627–640, 2009. 45

[166] E. Gołaś, G. G. Maisuradze, P. Senet, S. Ołdziej, C. Czaplewski, H. A. Scheraga, and A. Liwo. **Simulation of the opening and closing of Hsp70 chaperones by coarse-grained molecular dynamics**. *J. Chem. Theory Comput.*, **8**(5):1750–1764, 2012. 45

[167] J. Maupetit, P. Tuffery, and P. Derreumaux. **A coarse-grained protein force field for folding and structure prediction**. *Proteins*, **69**:394–408, 2007. 45

[168] A. Barducci, M. Bonomi, and P. Derreumaux. **Assessing the quality of the OPEP coarse-grained force field**. *J. Chem. Theory Comput.*, **7**(6):1928–1934, 2011. 45

[169] M. Cascella, M. A. Neri, P. Carloni, and M. Dal Peraro. **Topologically based multipolar reconstruction of electrostatic interactions in multiscale simulations of proteins**. *J. Chem. Theory Comput.*, **4**(8):1378–1385, 2008. 46, 49, 51

[170] D. Besozzi, P. Cazzaniga, G. Mauri, D. Pescini, and L. Vanneschi. **A comparison of genetic algorithms and particle swarm optimization for parameter estimation in stochastic biochemical systems**. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 116–127, 2009. 46

[171] E. Elbeltagi, T. Hegazy, and D. Grierson. **Comparison among five evolutionary-based optimization algorithms**. *Adv. Eng. Inf.*, **19**(1):43–53, 2005.

[172] V. Namasivayam and R. Günther. **PSO Autodock: A fast flexible molecular docking program based on swarm intelligence**. *Chem. Biol. Drug Des.*, **70**(6):475–484, 2007.

[173] P. Angeline. **Evolutionary optimization versus particle swarm optimization: Philosophy and performance differences**. In *Evolutionary Programming VII*, pages 601–610. Springer, 1998.

[174] A. Abraham and H. Liu. **Turbulent Particle Swarm Optimization Using Fuzzy Parameter Tuning**. *Foundations of Computational Intelligence Volume 3*, pages 291–312, 2009. 46

[175] T. Ha-Duong, N. Basdevant, and D. Borgis. **A polarizable coarse-grained water model for coarse-grained proteins simulations**. *Chem. Phys. Lett.*, **468**:79–82, 2009. 51

[176] N. Basdevant, D. Borgis, and T. Ha-Duong. **A semi-implicit solvent model for the simulation of peptides and proteins**. *J. Comput. Chem.*, **25**(8):1015–1029, 2004.

[177] N. Basdevant, D. Borgis, and T. Ha-Duong. **Modeling protein-protein recognition in solution using the coarse-grained force field Scorpion**. *J. Chem. Theory and Comput.*, **9**(1):803–813, 2013. 51

[178] C. Chipot, J. G. Ángyán, B. Maigret, and H. A. Scheraga. **Modeling amino acid side chains. 2. Determination of point charges from electrostatic properties: toward transferable point charge models**. *J. Chem. Phys.*, **97**(38):9788–9796, 1993. 52

[179] F. Fogolari, G. Esposito, P. Viglino, and S. Cattarinussi. **Modeling of polypeptide chains as C_α chains, C_α chains with C_β, and C_α chains with ellipsoidal lateral chains**. *Biophys. J.*, **70**:1183–1197, 1996. 52

[180] S. Plimpton. **Fast Parallel Algorithms for Short-Range Molecular Dynamics**. *J. Comp. Phys.*, **117**:1–19, 1995. 52

[181] J. Kennedy and R. Eberhart. **Particle swarm optimization**. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, **4**, pages 1942–1948. IEEE, 1995. 54

[182] Q. H. Dai, C. Tommos, J. E. Fuentes, M. R. A. Blomberg, P. L. Dutton, and A. J. Wand. **Structure of a de novo designed protein model of radical enzymes**. *J. Am. Chem. Soc.*, **124**(37):10952–10953, 2002. 56

[183] L. Banci, I. Bertini, F. Cantini, S. Ciofi-Baffoni, L. Gonnelli, and S. Mangani. **Solution structure of Cox11, a novel type of β-immunoglobulin-like fold involved in Cu$_B$ site formation of cytochrome c oxidase**. *J. Biol. Chem.*, **279**(33):34833–34839, 2004. 56

[184] A. Bateman and M. Bycroft. **The structure of a LysM domain from E. Coli membrane-bound lytic murein transglycosylase**. *J. Mol. Biol.*, **209**:1113–1119, 2000. 56

[185] A. M. Slovic, S. E. Stayrook, B. North, and W. F. DeGrado. **X-ray structure of a water-soluble analog of the membrane protein phospholamban: sequence determinants defining the topology of tetrameric and pentameric coiled coils**. *J. Mol. Biol.*, **348**:777–787, 2005. 56, 69

[186] A. M. Slovic, C. M. Summa, J. D. Lear, and W. F. DeGrado. **Computational design of a water-soluble analog of phospholamban**. *Protein Sci.*, **12**:337–348, 2003. 56

[187] A. M. Buckle, G. Schreiber, and A. R. Fersht. **Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-Å resolution**. *Biochemistry*, **33**(30):8878–8889, 1994. 56, 69

[188] R. A. Jarvis and E. A. Patrick. **Clustering using a similarity measure based on shared near neighbors**. *IEEE Transactions on Computers*, **C-22**(11):1025–1034, 1973. 57

[189] M. Stoldt, J. Wöinert, M. Görlach, and L. R. Brown. **The NMR structure of *Escherichia Coli* ribosomal protein L25 shows homology to general stress proteins and glutaminyl-tRNA synthetase**. *EMBO J.*, **17**(21):6377–6384, 1998. 57

[190] T. Gallagher, P. Alexander, P. Bryan, and G. L. Gilliland. **Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR**. *Biochemistry*, **33**(15):4721–4729, 1994. 57

[191] N. Blomberg, R. R. Gabdoulline, M. Nilges, and R. C. Wade. **Classification of protein sequences by homology modeling and quantitative analysis of electrostatics similarity**. *Proteins Struct. Funct. Bioinf.*, **37**:379–387, 1999. 58

[192] R. C. Wade, R. R. Gabdoulline, and F. De Rienzo. **Protein interaction property similarity analysis**. *Int. J. Quantum Chem.*, **83**(3/4):122–127, 2001. 58

[193] Y. Chebaro, S. Pasquali, and P. Derreumaux. **The coarse-grained OPEP force field for non-amyloid and amyloid proteins**. *J. Phys. Chem. B*, **116**(30):8741–8752, 2012. 58, 59, 64, 79

[194] S. B. Zimmerman and S. O. Trach. **Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli**. *J. Mol. Biol.*, **222**(3):599–620, 1991. 88

[195] S. R. McGuffee and A. H. Elcock. **Diffusion, crowding and protein stability in a dynamic molecular model of the bacterial cytoplasm**. *PLoS Comput. Biol.*, **6**(3):e1000694, 2010. 88

[196] B. P. Cossins, M. P. Jacobson, and V. Guallar. **A new view of the bacterial cytosol environment**. *PLoS Comput. Biol.*, **7**(6):e1002066, 2011. 88, 89, 102

[197] N. Tokuriki and D. S. Tawfik. **Stability effect of mutations and protein evolvability**. *Curr. Opin. Struct. Biol.*, **19**:1–9, 2009. 88

[198] Y. Wang, M. Sarkar, A. E. Smith, A. S. Krois, and G. J. Pielak. **Macromolecular crowding and protein stability**. *J. Am. Chem. Soc.*, **134**(40):16614–16618, 2012. 88, 89

[199] L. A. Benton, A. E. Smith, G. B. Young, and G. J. Pielak. **Unexpected Effects of Macromolecular Crowding on Protein Stability**. *Biochemistry*, **51**(49):9773–9775, 2009. 88, 89

[200] W. D. Van Horn, M. E. Ogilvie, and P. F. Flynn. **Reverse micelle encapsulation as a model for intracellular crowding**. *J. Am. Chem. Soc.*, **131**(23):8030–8039, 2009. 88

[201] Q. Wang, K. C. Liang, A. Czader, M. N. Waxham, and M. S. Cheung. **The effect of macromolecular crowding, ionic strength and calcium binding on calmodulin dynamics**. *PLoS Comput. Biol.*, **7**(7):e1002114, 2001. 89

[202] M. Feig and Y. Sugita. **Variable interactions between protein crowders and biomolecular solutes are important in understanding cellular crowding**. *J. Phys. Chem. B*, **116**(1):599–605, 2012. 89

[203] R. Harada, Y. Sugita, and M. Feig. **Protein crowding affects hydration structure and dynamics**. *J. Am. Chem. Soc.*, **134**(10):48424849, 2012. 88, 89, 104

[204] D. Homouz, M. Perham, A. Samiotakis, M. S. Cheung, and P. Wittung-Stafshede. **Crowded, cell-like environment induces shape changes in aspherical protein**. *PNAS*, **105**(33):11754–11759, 2008.

[205] D. Homouz, L. Stagg, P. Wittung-Stafshede, and M. S. Cheung. **Macromolecular Crowding Modulates Folding Mechanism of α/β Protein Apoflavodoxin**. *Biophys. J.*, **96**(2):671–680, 2009. 119

[206] A. Dhar, A. Samiotakis, S. Ebbinghaus, L. Nienhaus, D. Homouz, M. Gruebele, and M. S. Cheung. **Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding**. *PNAS*, **107**(41):17586–17591, 2010. 89

[207] N. Tjandra, S. E. Feller, R. W. Pastor, , and A. Bax. **Rotational diffusion anisotropy of human ubiquitin from 15N NMR relaxation**. *J. Am. Chem. Soc.*, **117**(50):12562–12565, 1995. 89, 96

[208] S. Mangia, N. J. Traaseth, G. Veglia, M. Garwood, and S. Michaeli. **Probing slow protein dynamics by adiabatic R(1rho) and R(2rho) NMR experiments**. *J. Am. Chem. Soc.*, **132**(29):9979–9981, 1995.

[209] R. Ishima and D. A. Torchia. **Extending the range of amide proton relaxation dispersion experiments in proteins using a constant-time**

relaxation-compensated CPMG approach. *J. Biomol. NMR*, **25**(3):243–248, 2003.

[210] N. A. Lakomek, T. Carlomagno, S. Becker, C. Griesinger, and J. Meiler. **A thorough dynamic interpretation of residual dipolar couplings in ubiquitin**. *J. Biomol. NMR*, **34**(2):101–115, 2006. 96

[211] N. A. Lakomek, K. F. Walter, C. Fares, O. F. Lange, B. L. de Groot, H. Grubmuller, R. Bruschweiler, A. Munk, S. Becker, J. Meiler, and C. Griesinger. **Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics**. *J. Biomol. NMR*, **41**(3):139–155, 2008. 96

[212] O. F. Lange, D. van der Spoel, , and B. L. de Groot. **Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data**. *Biophys. J.*, **99**(2):647–655, 2010.

[213] N. A. Lakomek, O. F. Lange, K. F. Walter, C. Fares, D. Egger, P. Lunkenheimer, J. Meiler, H. Grubmuller, S. Becker, B. L. de Groot, and C. Griesinger. **Residual dipolar couplings as a tool to study molecular recognition of ubiquitin**. *Biochem. Soc. Trans.*, **36**(6):1433–1437, 2008. 89, 96, 105

[214] S. Esteban-Martin, R. B. Fenwick, and X. Salvatella. **Refinement of ensembles describing unstructured proteins using NMR residual dipolar couplings**. *J. Am. Chem. Soc.*, **132**(13):4626–4632, 2010.

[215] G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen, and M. Blackledge. **Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings**. *J. Am. Chem. Soc.*, **131**(49):17908–17918, 2009.

[216] J. Meiler, J. J. Prompers, W. Peti, C. Griesinger, and R. Bruschweiler. **Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins**. *J. Am. Chem. Soc.*, **123**(25):6098–6107, 2001.

[217] K. Ruan and J. R. Tolman. **Composite alignment media for the measurement of independent sets of NMR residual dipolar couplings**. *J. Am. Chem. Soc.*, **127**(43):15032–15033, 2005. 96

[218] P. Maragakis, K. Lindorff-Larsen, M. P. Eastwood, R. O. Dror, J. L. Klepeis, I. T. Arkin, M. O. Jensen, H. Xu, N. Trbovic, R. A. Friesner, A. G. Palmer, and D. E. Shaw. **Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins**. *J. Phys. Chem. B*, **112**(19):6155–6158, 2008. 89

[219] D. Long and R. Bruschweiler. **In silico elucidation of the recognition dynamics of ubiquitin**. *PLoS Comput. Biol.*, **7**(4):e1002035, 2011. 89, 90, 96

[220] J. H. Peters and B. L. de Groot. **Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection**. *PLoS Comput. Biol.*, **8**(10):e1002704, 2012. 96, 105

[221] T. Wlodarski and B. Zagrovic. **Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin**. *PNAS*, **106**(46):19346–19351, 2009. 89, 96

[222] T. Wlodarski and B. Zagrovic. **Destabilised mutants of ubiquitin gain equal stability in crowded solutions**. *Biophys. Chem.*, **128**(2-3):140–149, 2007. 89

[223] J. M. Yuan, C. L. Chyan, H. X. Zhou, T. Y. Chung, H. Peng, G. Ping, and G. Yang. **The effects of macromolecular crowding on the mechanical stability of protein molecules**. *Protein Sci*, **17**(12):2156–2166, 2008. 89

[224] G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax. **Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase**. *J. Am. Chem. Soc.*, **120**(27):6836–6837, 1998. 90

[225] L. Martinez, R. Andrade, E. G. Birgin, and J. M. Martinez. **Packmol: A package for building initial configurations for molecular dynamics simulations**. *J. Comput. Chem.*, **30**(13):2157–2164, 2009. 90

[226] I. Chandrasekhar, G. M. Clore, A. Szabo, A. M. Gronenborn, and B. R. Brooks. **A 500 ps molecular dynamics simulation study of interleukin-1 beta in water. Correlation with nuclear magnetic resonance spectroscopy and crystallography**. *J. Mol. Biol.*, **226**(1):239–250, 1992. 92

[227] S. A. Showalter and R. BrUschweiler. **Quantitative Molecular Ensemble Interpretation of NMR Dipolar Couplings without Restraints**. *J. Am. Chem. Soc.*, **129**(14):4158–4159, 2007. 93, 94, 96

[228] K. Okazaki and S. Takada. **Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms**. *PNAS*, **105**(32):11182–11187, 2008. 96

[229] M. Ottiger and A. Bax. **Determination of Relative N-H^N, N-C, Cα-C, and Cα-Hα Effective Bond Lengths in a Protein by NMR in a Dilute Liquid Crystalline Phase**. *J. Am. Chem. Soc.*, **120**(47):12334–12341, 1998. 96

[230] K. B. Briggman, , and J. R. Tolman. **De novo determination of bond orientations and order parameters from residual dipolar couplings with high accuracy**. *J. Am. Chem. Soc.*, **125**(34):10164–10165, 2003.

[231] J. R. Tolman. **A novel approach to the retrieval of structural and dynamic information from residual dipolar couplings using several oriented media in biomolecular NMR spectroscopy**. *J. Am. Chem. Soc.*, **124**(40):12020–12030, 2002. 96

[232] G. Cottone. **A comparative study of carboxy myoglobin in saccharide-water systems by molecular dynamics simulation**. *J. Phys. Chem. B*, **111**(13):3563–3569, 2007. 98, 103

[233] G. Cottone, S. Giuffrida, G. Ciccotti, and L. Cordone. **Molecular dynamics simulation of sucrose- and trehalose-coated carboxy-myoglobin**. *Proteins*, **59**(2):3563–3569, 2005. 98, 103

[234] I. Pozdnyakova and P. Wittung-Stafshede. **Non-linear effects of macromolecular crowding on enzymatic activity of multi-copper oxidase**. *Biochem. Biophys. Acta*, **1804**(4):740–744, 2010. 102

[235] M. P. Latham and L. E. Kay. **Is Buffer a Good Proxy for a Crowded Cell-Like Environment? A Comparative NMR Study of Calmodulin Side-Chain Dynamics in Buffer and E. coli Lysate**. *PLoS One*, **7**(10):e48226, 2012. 102

[236] J. Tian and A. E. Garcia. **Simulations of the confinement of ubiquitin in self-assembled reverse micelles**. *J. Chem. Phys.*, **134**(22):225101–11, 2011. 102, 118

[237] A. K. Simorellis and P. F. Flynn. **Fast local backbone dynamics of encapsulated ubiquitin**. *J. Am. Chem. Soc.*, **128**(30):9580–958, 2006. 102

[238] S. Giuffrida, G. Cottone, and L. Cordone. **Role of solvent on protein-matrix coupling in MbCO embedded in water-saccharide systems: a Fourier transform infrared spectroscopy study**. *Biophys. J.*, **91**(3):968–980, 2006. 103

[239] S. Ahmad, M. Gromiha, H. Fawareh, and A. Sarai. **ASAView: database and tool for solvent accessibility representation in proteins**. *BMC Bioinformatics*, **5**(51):1–5, 2004. 103

[240] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and F. G. Parak. **Slaving: Solvent fluctuations dominate protein dynamics and functions**. *PNAS*, **99**(25):16047–16051, 2002. 117

[241] M. Perham, L. Stagg, and P. Wittung-Stafshede. **Macromolecular crowding increases structural content of folded proteins**. *FEBS Lett*, **581**(26):5065–5069, 2007. 119

[242] M. Perham, L. Stagg, and P. Wittung-Stafshede. **Molecular crowding enhances native structure and stability of α/β protein flavodoxin**. *PNAS*, **104**(48):18976–18981, 2007. 119

[243] A. Christiansen, Q. Wang, A. Samiotakis, M. S. Cheung, and P. Wittung-Stafshede. **Factors defining effects of macromolecular crowding on protein stability: an in vitro/in silico case study using cytochrome c**. *Biochemistry*, **49**(31):6519–6530, 2010. 119

[244] E. Chen, A. Christiansen, Q. Wang, M. S. Cheung, D. S. Kliger, and P. Wittung-Stafshede. **Effects of Macromolecular Crowding on Burst Phase Kinetics of Cytochrome c Folding**. *Biochemistry*, **51**(49):9836–9845, 2012. 119

[245] M. S. Cheung, D. Klimov, and D. Thirumalai. **Molecular crowding enhances native state stability and refolding rates of globular proteins**. *PNAS*, **102**(13):4753–4758, 2005. 120

[246] D. L. Pincus and D. Thirumalai. **Crowding effects on the mechanical stability and unfolding pathways of ubiquitin**. *J. Phys. Chem. B*, **113**(1):359–368, 2009.

[247] A. Kudlay, M. S. Cheung, and D. Thirumalai. **Crowding effects on the structural transitions in a flexible helical homopolymer**. *Phys. Rev. Lett.*, **102**(118101):1–4, 2009.

[248] A. Dhar, A. Samiotakis, S. Ebbinghaus, L. Nienhaus, D. Homouz, M. Gruebele, and M. S. Cheung. **Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding**. *PNAS*, **107**(41):17586–17591, 2010. 120

[249] Y. Sugita and Y. Okamoto. **Replica-exchange molecular dynamics method for protein folding**. *Chem. Phys. Lett.*, **314**(27):141–151, 1999. 120

Acknowledgments

Further, science is a collaborative effort

—John Bardeen

I would like to start thanking Matteo, who gave me the unique opportunity to come to Switzerland to do the PhD at EPFL. He allowed me to explore different research roads. We faced difficult periods and together we managed to ride out. This has been one of our main important achievements!

Davide Alemani worked with me to the development of the coarse-grained model. I would like to thank him because he offered me all his support both from a scientific point of view and from the moral point of view. Marco Stenta shared with me the same office for almost three years giving me a lot of suggestions: thank you!

I acknowledge Matteo Ceccarelli for supporting me during these years giving me very useful advices.

I would like to acknowledge Dr. Bernd Ensing, Dr. Richard Lavery, Prof. Paolo De Los Rios and Prof. Patrick Fraering to have accepted to be part of the commission of my PhD private defense and for very useful suggestions.

I would like to thank Michele Cascella for useful discussions during the development of the coarse-grained model.

Matteo Degiacomi kindly permitted me to use his Particle Swarm Optimizer POW: thank you!

Matteo Degiacomi, Thomas Lemmin and Christophe Bovigny are kindly ackno**W**ledge for t**H**eir really kind supp**O**rt through smiles, laughs and fundamental techni**CA**l suppo**R**t during this y**E**ar**S**.

Thanks to Hassan Pezeshgi Modarres for very important cultural discussions and moral support.

Many thanks also to my new officemates Luciano and Martina that make a lot of noise organizing jokes and sharing projects. Many thanks to Luciano that worked with me at the ubiquitin project and helped me in proofreading articles.

Many thanks to Francesco Piazza for sharing ideas and results during the development of the ubiquitin project.

I would like to thank Phil and Gaurav that have been ”my students” for small projects: I had the opportunity to learn a lot from them!

Thanks to Shantanu, Matthieu, Basile, Anna, Benoîte, Veronica, Giorgio, Matteo and Vito that gave me a lot of different perspectives on several aspects of my life during my PhD.

Ringrazio tutta la mia famiglia per essermi stata accanto in questi anni passati all’estero.

Finally I would like to thank my girlfriend Francesca that helped, supported and tolerated me during this adventure in Switzerland, sharing bad and good experiences.

Curriculum vitae
Enrico Spiga

Chiamavamo noi stessi S'ARD*, che nell'antica lingua significa danzatori delle stelle.*

We called ourselves S'ARD*, that in the ancient language means stars dancers*

SERGIO ATZENI

Personal

- Born on August 01, 1981, Cagliari, Sardinia, Italy.

Education

- **Bachelor: Physics** (Università di Cagliari, Italy) 10.2000 – 11.2004

  Specialization: Physics; final grade: 104/110

  Thesis title: Simulation of structural properties of solvated tripeptides

  Supervisor: Prof. Matteo Ceccarelli

- **Master: Physics** (Università di Cagliari, Italy) 11.2004 – 02.2008

  Specialization: Computational physics; final grade: 110/110 cum laude

  Thesis title: Properties of antibiotic transport through the outer membrane of Gram-negative bacteria

  Supervisors: Prof. Matteo Ceccarelli and Prof. Paolo Ruggerone

- **PhD: Bioengineering and Biotechnology** (École Polytechnique Fédérale de Lausanne, Switzerland) 11.2008 – present

Specialization: Computational biophysics

Thesis title: Multi-scale simulations of protein dynamics

Supervisor: Prof. Matteo Dal Peraro

Research Experience

- **PhD position** (École Polytechnique Fédérale de Lausanne, Switzerland) 11.2008 – present

  Supervisor: Prof. Matteo Dal Peraro

  Topics: computational biophysics, coarse-grained models of proteins, all-atom simulations of proteins in crowded environment

  Project: Multi-scale simulations of protein dynamics

- **Research contract** (Università di Cagliari, Italy) 06.2008 – 10.2008

  Supervisors: Prof. Matteo Ceccarelli and Prof. Mariano Casu

  Topic: computational biophysics of human myoglobin

  Project: All-atom molecular dynamics simulation studies of human myoglobins

- **Internship for Master Thesis** (Università di Cagliari, Italy) 09.2005 – 02.2008

  Supervisors: Prof. Matteo Ceccarelli and Prof. Paolo Ruggerone

  Topic: computational biophysics of antibiotic resistance

  Project: All-atom molecular dynamics simulation studies of antibiotics translocations through the outer membrane protein OMPF

- **Internship for bachelor Thesis** (Università di Cagliari, Italy) 05.2004 –11.2004

  Supervisor: Prof. Matteo Ceccarelli

  Topic: computational biophysics of peptides

  Project: All-atom molecular dynamics simulation studies of solvated tripeptides

Work experience

- Summers from 1996 to 2005: Peach farming (picking up and packaging) (Sardinia, Italy)

- Summer 2009: Instructor for IGEM Team EPFL 2009
- Spring Semester 2009-2010: Teaching assistant for the course Biomolecular Mechanics (EPFL, Prof. Matteo Dal Peraro)
- Spring Semester 2010-2011: Teaching assistant for the course Biomolecular Mechanics (EPFL, Prof. Matteo Dal Peraro)
- Spring Semester 2010-2011: Teaching assistant for the course Biothermodynamics (EPFL, Prof. Matteo Dal Peraro and Prof. Paolo De Los Rios)
- Summer 2011: Co-advisor of students for Summer Research Program EPFL
- Spring Semester 2011-2012: Teaching assistant for the course Biothermodynamics (EPFL, Prof. Matteo Dal Peraro and Prof. Paolo De Los Rios)
- Accademic year 2011-2012: Co-advisor for master thesis EPFL

Other activities

- Co-organizer of the Symposium: ”Applied Computational Chemistry 2012: Computational Chemistry and Industry Simposium”

Training and contribution in scientific meetings

(P= poster, O=oral contribution, I= invited talk, SP= simple participation)

- 2013 Paul Scherrer Institute, Villigen, CH, [P]: *Swiss Soft Days X*
- 2012 Nestlé Research Center, Lausanne, CH, [I]: *Swiss Soft Days IX*
- 2012 EPFL, Lausanne, CH, [P]: *CECAM Workshop: Exploring Protein Interactions through Theory and Experiments*
- 2012 San Diego, California, USA, [O]: *ACS Spring 2012 National Meeting*
- 2012 EPFL, Lausanne, CH, [O]: *Lausanne Biomolecular Modelling Seminars*
- 2010 Crans Montana, CH, [O+P]: *EPFL Winter School: Studying biomolecules by experiments and theory*
- 2009 EPFL, Lausanne, CH, [P]: *CECAM Workshop: Linking Systems Biology and Biomolecular Simulations*

- 2009 Donostia, ES, [P]: *Summer School: Simulation approaches to problems in molecular and cellular biology*
- 2009 Espoo, FI, [SP]: *Winter School: Coarse Graining workshop*
- 2007 Sheffield, UK, [P]: *Summer School: CCP5 Methods in Molecular Simulation*
- 2006 Bremen, DE, [P]: *Summer School: Biosensing with channels: faster, smaller, smarter*

Computer skills

- Programming languages: Fortran, Python, Bash, Latex
- Operating systems: Linux, UNIX, OS/X, Windows
- Electronic structure packages: Gaussian
- Classical MD simulation packages: Amber, Gromacs, NAMD, LAMMPS
- Homology modeling Softwares: MODELLER
- Visualization Softwares: VMD, PyMol

Language competencies

Italian (native fluency); English (intermediate); French (intermediate)

Publications

Journal Articles (in preparation)

- **Spiga E.**, Abriata L. A., Piazza F., Dal Peraro M.; Crowding agents affect the dynamics of ubiquitin; in preparation
- **Spiga E.**, Alemani D., Abriata A. L., Audagnotto M., Degiacomi T. M., Lemmin T., Bovigny C. and Dal Peraro M.; Sparkling: a coarse-grained force field with dipolar contributions; in preparation

Journal Articles (published or submitted)

- **Spiga E.**, Alemani D., Degiacomi T. M., Cascella M. and Dal Peraro M.; Electrostatic-consistent coarse-grained potentials for molecular simulations of proteins; accepted in JCTC, DOI:10.1021/ct400137q

- Abriata L. A.*, **Spiga E.*** and Dal Peraro M.; All-atom simulations of crowding effects on ubiquitin dynamics, *Physical Biology*, accepted for publication **equally contributed*

- Scorciapino M. A., **Spiga E.**, Vezzoli A., Mrakic-Sposta S., Russo R., Fink B., Casu M., Gussoni M., Ceccarelli M.; Structure-function paradigm in Human Myoglobin: how a single-residue substitution affects proteins behavior depending on pO2; *Journal of the American Chemical Society*, **2013**, 135 (20), 7534-7544

- Mahendran K. R., Hajjar E., Mach T., Lovelle M., Kumar A., Sousa I., **Spiga E.**, Weingart H., Gameiro P., Winterhalter M. and Ceccarelli M.; Molecular Basis of Enrofloxacin Translocation through OmpF, an Outer Membrane Channel of *Escherichia coli* – When Binding Does Not Imply Translocation; *Journal of Physical Chemistry B* **2010**, 114 (15), pp. 5170-5179

- Collu F., **Spiga E.**, Kumar A., Hajjar E., Vargiu A.V., Ceccarelli M., Ruggerone P.; Drug design: Insights from atomistic simulations, *Nuovo Cimento C*, **2009**, 32 (2) pp. 67-71

- Mach T., Neves P., **Spiga E.**, Weingart H., Winterhalter M., Ruggerone P., Ceccarelli M. and Gameiro P.; Facilitated Permeation of Antibiotics across Membrane Channels – Interaction of the Quinolone Moxifloxacin with the OmpF Channel, *Journal of the American Chemical Society*, **2008**, 130 (40) pp. 13301-13309