Multiple-resolution simulations of biomolecules

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Raffaele Fiorentini

Dissertation

zur Erlangung des Grades

"Doktor der Naturwissenschaften"

am Fachbereich Physik, Mathematik und Informatik

der Johannes Gutenberg-Universität in Mainz

Angefertigt am Max-Planck-Institut für Polymerforschung

Mainz, June 2020

D77 (Dissertation Johannes Gutenberg-Universität Mainz)

Chairperson	*Prof. Dr. Wittig*
Committee	*Prof. Dr. K. Kremer* (1st supervisor) *Prof. Dr. R. Potestio* (2nd supervisor) *Prof. Dr. F. Schmid* *Prof. Dr. S. Weber*
Submitted	*16th March 2020*
Accepted	*24th March 2020*
Oral examination	*3th June 2020*

To my family

for always loving and supporting me

One day the machines will be able to solve any problem,
but none of them will raise one.

A. EINSTEIN

Detailed Contents

Detailed Contents ix

Figures xiii

Tables xvii

Abstract 1

1 Introduction 5

1.1 Biophysical background 6

1.1.1 What are proteins? 6

1.1.2 Protein structures 11

1.2 History and overview 13

1.3 Molecular Dynamics 16

1.3.1 Force Field 19

1.3.2 Integrators 23

1.3.3 NPT and NVT ensembles 26

1.4 Coarse-Grained models 27

1.4.1 Elastic Network Models (ENMs) 30

1.5 Multiscale Simulations 35

1.6 Adaptive Resolution Simulation (AdResS) 38

1.7 Outline 43

2 Using force-based adaptive resolution simulations to calculate solvation free energies of amino acid sidechain analogues 45

2.1 Introduction 46

2.2 Methodology 49

2.2.1 Adaptive Resolution Scheme and Thermodynamic Integration . 50

2.2.2 Thermodynamic Force 56

2.2.3 Simulation details 58
2.3 Results 61
2.4 Conclusions 66
2.5 Acknowledgements 68

3 Ligand-protein interactions in lysozyme investigated through a dual-resolution model 69
3.1 Introduction 70
3.2 Methods 73
3.2.1 Binding Free Energy calculation 74
3.2.2 Dual-Resolution protein model 77
3.2.3 Simulation details 79
3.3 Results and discussion 81
3.4 Conclusions 88
3.5 Supporting Information 89
3.5.1 Thermodynamic Cycle for binding free energy 89
3.5.2 Annihilation and Binding Free Energy 90
A. Results of Binding free energy calculation 91
3.5.3 Parametrization of the dual-resolution model 95
A. Determination of elastic constants between beads 96
B. Determination of WCA parameters ε and σ 96
3.6 Acknowledgements 98

4 Free energy landscapes calculation in 1BBA investigated through an improvement version of the ENM. 101
4.1 Introduction 102
4.2 Methodology 104
4.2.1 Finding boundaries between atomistic and CG part 106
4.2.2 Dual-resolution model 109
4.2.3 Different elastic constants in ENM 110
4.2.4 Free energy landscapes 113

4.2.5 Simulation details 115
4.3 Results and discussion 117
4.4 Conclusions 122

5 Simulating Adenylate Kinase through a Variable-Resolution model 125
5.1 Introduction 126
5.2 Methodology 129
5.2.1 Voronoi Tessellation 131
5.2.2 The CANVAS model 133
5.2.3 Simulation details 139
5.3 Result and Discussion 140
5.3.1 All-atom simulation 140
5.3.2 CANVAS simulations 142
5.3.3 Comparison 145
5.4 Conclusions 148
Appendix A: RMSD and RMSF 149

6 Conclusions 151
6.1 Summary 151
6.2 Outlook 156

Acknowledgements 162

Bibliography 167

Figures

1.1 Amino acid chain 8

1.2 Structure of an amino acid in its un-ionized form. 8

1.3 Amino acid representation showing the position of C_α, and C_β. 8

1.4 The naturally occurring 20 amino acids with complete name, abbreviations and chemical structures. 9

1.5 The different levels of protein organization: primary, secondary, tertiary and quaternary structures. 11

1.7 The two most common types of protein secondary structure, α-helices and β-sheets. 12

1.6 Generic fragment of a protein showing the two secondary structures: α-helix and β-sheet. 12

1.8 Tertiary structure of a protein showing disulphide bonds. 12

1.9 Quaternary structures of Human Aquaporin-4. This structure is a tetramer. . 13
1.10 Multi-scale nature of the matter 15
1.11 Representation of terms: U_{bond}, U_{tors}, U_{bend} and U_{impr} 21
1.12 Molecular structure of water molecule. The corresponding coarse-grained bead is shown with a transparent green bead. 27
1.13 Polyalanine structure in three different representations: all-atom, coarse grained of functional groups and coarse-grained of amino acid residues. . . 28
1.14 Adenylate Kinase in terms of fully atomistic and C_α-only representation. . . 34
1.15 Schematic representation of a solvated protein that interacts at its active site. 36
1.16 AdResS system in case of spherical and cuboid atomistic region 39
1.17 Schematic representation of AdResS approach in case of atomistic spherical region 40

2.1 Illustration of the AdResS approach. 51
2.2 Coulomb and LJ contribution to the free energy solvation for methanol and 3-methylindole. 62
2.3 Radial distribution function between water oxygen atoms and selected solute heavy atoms, compared to the fully atomistic reference. 65
2.4 Molecular fluctuations in spherical concentric bins of equal surface-to-volume ratio, as a function of distance from the solute atom. 66

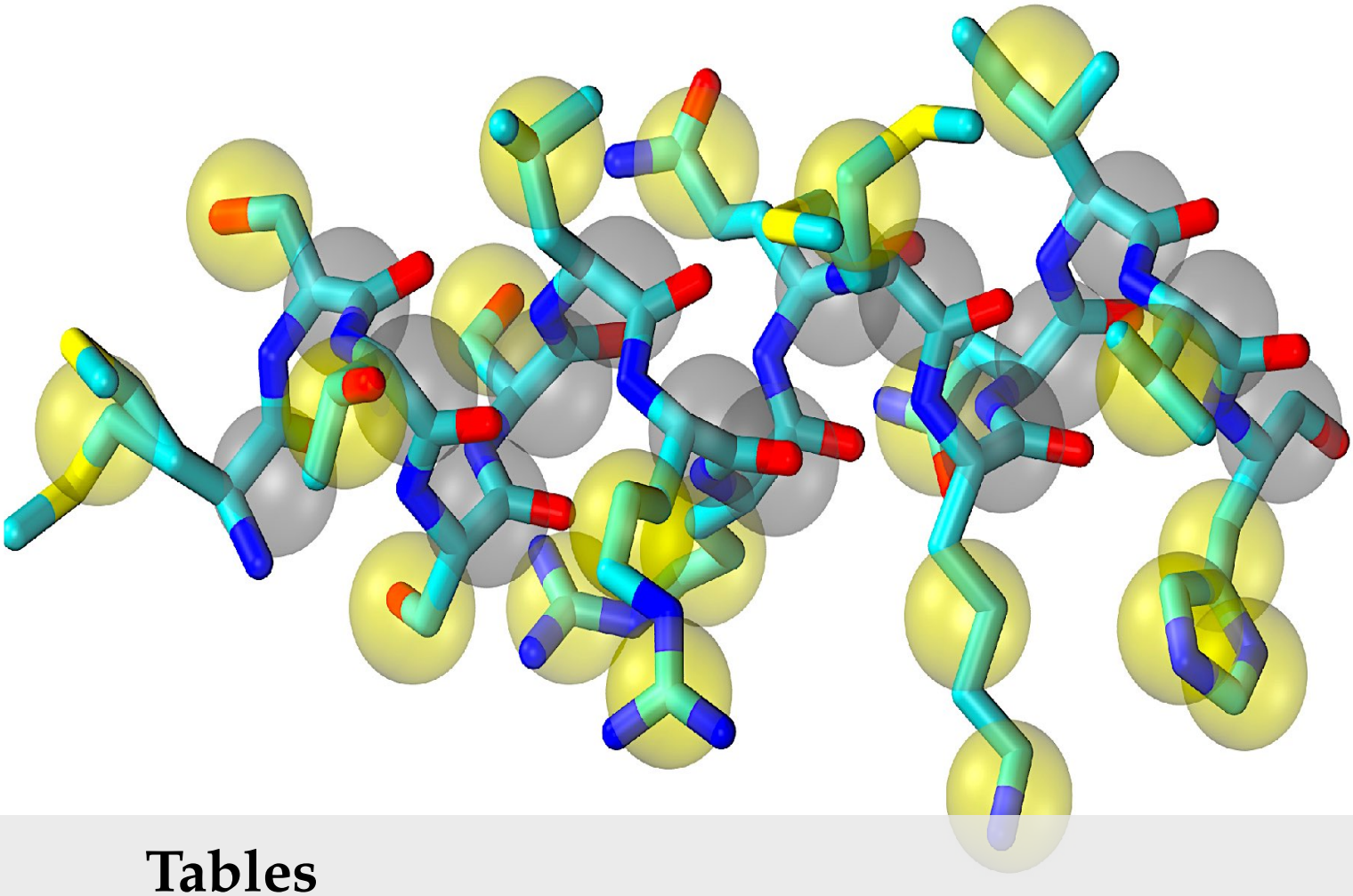
3.1 pictorial representation of thermodynamic cycle 76
3.2 Visualisation of the dual-resolution protein. 78
3.3 Coulomb, Lennard-Jones, restraint and total free energies in the protein-ligand complex, as a function of protein's residues number included in atomistic detail in the multi-resolution set-up. 83
3.4 Square root of quadratic deviation δ^2 vs the number of atomistic residues chosen. 85
3.5 Representation of lysozyme and ligand in different resolution 86
3.6 Binding free energies as a function of protein's residues included in atomistic detail in the multi-resolution set-up or fully atomistic set-up. 87

3.7 Comparison of the Thermodynamic Integration (TI) free energy derivative curves computed with ESPResSo++ and GROMACS for all atom protein. . . 94
3.8 Comparison of the Thermodynamic Integration (TI) free energy values computed with ESPResSo++ and GROMACS for all atom protein. 95
3.9 Bulk water density in the case of 8 atomistic residues for different value of *c*. 98
3.10 Representation of lysozyme and ligand in different resolution (from 3 to 10 atomistic residues) 99

4.1 Configuration of the atomistic gA peptide in CG DMPC lipid and water . . . 102
4.2 Visualization of 1BBA in dual-resolution protein. 109
4.3 Distance distribution between the C_α carbons 14 and 17 of 1BBA 111
4.4 Distance distribution between the C_α carbons 14 and 17 of 1BBA fitted with a Gaussian 113
4.5 Visualization of the two collective variables chosen to describe the system: *end-to-end* distance and degree of unfolding of the protein. 113
4.6 Comparison between the heat-map plots treated in terms of point probability and free energy 115
4.7 All-atom representation of Bovine Pancreatic Polypeptide (1BBA) after 1 ns equlibration in NPT ensamble, in terms of secondary structure. 115
4.8 Representation of Bovine Pancreatic Polypeptide in two, three, four, five, six, seven rigid blocks, by means of PiSQRD tool 118
4.9 Frequency histogram for each boundary in 1BBA 119
4.10 Schematic representation of 1BBA divided in 4 blocks 119
4.11 Free energy landscapes in six different cases, according to the presence or absence of salt in water, and the model used: all-atom or dual-resolution . . 120

5.1 Structure of Adenylate Kinase: LID, NMP, and CORE. 129
5.2 Voronoi Tessellation of a square region consisting of 20 points 133
5.3 Application of the Voronoi tessellation at the lowest-resolution part of a generic (schematic) protein 134

5.4 Representation of (a) a generic unstable CG network; (b) three tetrahedrons: each bead satisfies the tetrahedral condition. 135
5.5 Visualization of Adenylate Kinase (4AKE) in Variable Resolution 136
5.6 Visualization of a generic Voronoi cell in a protein 137
5.7 Fully atomistic representation of Adenylate Kinase (4AKE) after 125 ns equilibration in NPT ensemble in terms of secondary structure. 139
5.8 Fully atomistic representation of 4AKE in terms of primary structure. 140
5.9 RMSD of all 4AKE C_α in case the electrostatic used is reaction field or Particle Mesh Ewald (PME) 141
5.10 RMSF for each C_α of 4AKE in case of Reaction Field and Particle Mesh Ewald (PME) 142
5.11 RMSD of multi-scale resolution of 4AKE C_α in case the electrostatic used is reaction field or PME 143
5.12 RMSF of multi-scale resolution for each 4AKE C_α in case the electrostatic used is reaction field or PME 144
5.13 RMSF comparison between fully-atomistic and variable-resolution simulations 146
5.14 RMSF comparison between fully-atomistic (only open conformation) and CANVAS simulations 147
5.15 Schematic representation of 4AKE in variable resolution: arms and hinge are treated atomistically; the remainder in coarse-grained. 149



Tables

2.1 Comparison of the speedup in simulation time provided by AdResS simulations with respect to fully atomistic simulations. 53

2.2 Simulation box length, atomistic region radius, and number of atomistic or atomistic-like particles in the AdResS and fully atomistic systems used to perform free energy calculations. 59

2.3 Experimental solvation free energy values in kJ mol^{-1} compared to total solvation free energies ($\Delta G = \Delta G_{Coul} + \Delta G_{LJ}$). 64

3.1 Summary of the alchemical changes and the protein resolution dependence for each contribute of Binding free energy ΔG_{bind}. 77

3.2 Resulting values of Complex Free Energy and its components (Coulomb, Lennard Jones and Restraints) in fully-at system and varying the number of atomistic residues. 82

3.3 Representation of Free Energies values computed in ESPResSo++ and GROMACS in case of annihilation and decoupling. 87
3.4 Resulting values of Complex Free Energy and its components (Coulomb, Lennard Jones and Restraints) in fully-at system in case of annihilation. . . . 93
3.5 Resulting values of Ligand Free Energy and its components (Coulomb, Lennard Jones) in fully-at system in case of annihilation. 93
3.6 Representation of Free Energies values (ligand, complex and binding) computed in ESPResSo++ and GROMACS in case of annihilation. 94
3.7 Density found in the case of 8 atomistic residues for different value of c and comparison with the atomistic reference. 97
3.8 Bulk water's average density and percentage relative error in dual-resolution simulation with different atomistic residues from 3 to 10, keeping $c = 0.658$. 98

Abstract

The experiment plays a central role in science. It is the wealth of experimental results that provides a basis for the understanding of the chemical machinery of life. Experimental techniques, such as X-ray diffraction [1, 2], nuclear magnetic resonance (NMR) [3], or cryogenic electron microscopy (cryo-EM) [4], allow the determination of the structure and elucidation of the function of large molecules of biological interest. Yet, the experiment alone is often not sufficient to gain sufficient insight in the mechanisms and processes that take place at the molecular level. To this end, it is necessary to complement the direct experimental investigation with models and theories. Computer simulations have altered the interplay between experiment and theory. The essence of simulation is the use of the computers to model a physical system. Calculations implied by a mathematical model are carried out by the machine, and the results are interpreted in terms of physical properties. Since computer simulation deals with models, it may be classified as a theoretical method [5]. On the other hand, physical quantities can (in a sense) be measured on a computer, justifying the term *computer experiment* [6, 7]. The crucial advantage of simulations is the ability to expand the horizon of the complexity that separates 'solvable' from 'unsolvable'. Basic physical theories applicable to biologically relevant phenomena, such as quantum, classical and statistical mechanics, lead to equations that cannot be solved analytically (exactly), except for a few special cases. The quantum Schrödinger equa-

tion for any atom (or any molecule), except the hydrogen, or the classical Newton's equations of motion for a system of more than two-point masses can be solved only approximately. This is what physicists call the many-body problem. It is intuitively clear that less accurate approximations become inevitable with growing complexity. We can compute a more accurate wave function for the hydrogen molecule than for large molecules. It is also much harder to include explicitly the electrons in the model of a protein, rather than representing the atoms as classical point-like masses and the bonds as springs. The use of the computer makes less drastic approximations feasible. Thus, bridging experiment and theory by means of computer simulations makes it possible to test and improve our models using a more realistic representation of nature. It may also bring new insights into mechanisms and processes that are not directly accessible through the experiment.

However, the amount of available computational resources can be insufficient to simulate, for a physically meaningful time, even the simplest nontrivial macromolecule. Indeed, it is often the case that "interesting" phenomena occur on very long time-scales: a simple example of this is provided by the diffusion of a polymer in a melt [8, 9]; the same behavior can observed in conformational changes of proteins [10–15]. At the same time, in many cases the massive amount of data that are produced in a simulation is composed mostly of non-useful information. A relevant example is given by the solvent: the water molecules that solvate a protein or a membrane are usually discarded from the analysis that follows the simulation. In this case a large fraction of the computational effort is employed in

the integration of the equations of motion of degrees of freedom which are extremely relevant during the simulations, but are neglected afterwards. In order to overcome this limitation, coarse-grained models [16–19] have been developed, where the structure and interactions of the original system are replaced with simpler ones, which are easier to describe, model, simulate and understand. In recent years, systematic coarse graining approaches have gained more and more prominence. Currently, these models are often used in a multiscale simulation framework [20–27], where higher and lower levels of resolution are concurrently employed. In these approaches, the region of the system in which the chemical details play a crucial role is described by an accurate but computational expensive model, while the remainder has a lower resolution.

One of such multi-resolution techniques is the *force-based Adaptive Resolution Scheme* (AdResS): employed extensively in liquids and complex mixtures [28–34]. In this approach, the two resolutions (for instance, all-atom and coarse-grained) are simultaneously employed in different sub-regions: an essential feature of this method is that particles are allowed to diffuse from one region to the other freely.

Another class of multi-resolution models is known as *Dual Resolution* for proteins. The peculiarity of this model is that it is not adaptive; therefore, the resolution is fixed during the simulation. Several coarse-grained methodologies have been developed to describe the entire protein and afterwards employed in dual resolution models for treating the lower detailed part, e.g. the Gō Model [35], or the Elastic Network Model (ENM) [24, 36]. Since the

latter is one of the main objects of this thesis, it will be analyzed in detail in section 1.4.1: specifically, in this approach, only the C_α carbons of the protein chain are retained in the coarse-grained part, and connected one with other by harmonic springs.

In this work, we first make use of the Adaptive Resolution Scheme (AdResS) in combination with thermodynamic integration (TI) to calculate the solvation free energies of amino acid sidechain analogues in water. Then, we use the dual resolution method with a dual purpose:

- To compute the binding free energy of hen egg-white lysozyme (HEWL) with the inhibitor di-N-acetylchitotriose. Particular attention is posed to the impact of mapping, namely the selection of atomistic and coarse-grained residues on the binding free energy. The choice of the residues belonging to the protein active site modeled atomistically has a significant impact on such a value.
- To capture the dynamic properties of a small protein known as Bovine Pancreatic Polypeptide (PDB code 1BBA) in terms of free energy landscapes obtained after choosing two collective variables apt to describe the system. The original contribution in this thesis work is the refinement of the ENM because we use different elastic constants between CG beads, based on their distance distribution.

Lastly, we illustrate a novel multi-resolution approach dubbed *coarse-grained anisotropic network model for variable resultion simulations* (CANVAS), applied to a protein, Adenylate Kinase. It allows to smoothly couple virtually any desired degrees of coarse-graining within the same model.

1 Introduction

Molecular modeling and, in particular, Molecular dynamics (MD) has brought significant progress in a wide range of biological applications in the last decades due to the advancement of novel algorithms and high-performance computing. The gap between simulation and experimental timescales has been significantly reduced due to the concurrent advances in the corresponding techniques. Scientists can nowadays access microsecond-to-millisecond timescales with atomic detail, which is sufficient to characterize many critical biological processes, such as the folding dynamics proteins [37, 38].

However, classical molecular dynamics is often computationally expensive for many large-scale problems in molecular modeling. For example, the time scales on which most proteins fold cannot be reached with all-atom MD simulation: milliseconds are the norm and the fastest known protein folding reactions are complete within a few microseconds [38]. Similarly, biological membranes are often too large to allow for an atomistic description. Another challenge occurs in many complex polymer systems.

A solution is provided by coarse-grained (CG) models

[19, 39–42] and, in particular, multiscale simulation techniques developed in the last years [43, 44]. The latter are the main subject of this thesis. To set the work in the right context, we first provide an overview on proteins, since they are the main object of interest in this thesis work; subsequently we recapitulate in section 1.2 of this introductory chapter the birth and the development of computational molecular modeling in the last few decades. Next, in section 1.3, we discuss the molecular dynamics technique analyzing all required ingredients to perform a classical MD simulation. Section 1.4 illustrates coarse-grained models in general focusing, in particular, on the Elastic Network Model (ENM), widely used in this work. In section 1.5, we discuss the idea of multiscale modeling and review some of its applications. In this contest, the section 1.6 describes in full detail the Force-based Adaptive Resolution Scheme (AdResS), a computational method for the efficient multiscale simulation of molecular systems. Finally, in section 1.7, we provide an outline of the following chapters.

1.1 Biophysical background

1.1.1 What are proteins?

Proteins are large molecules consisting of amino acids. Our body structures, and functions, as well as the regulation of the body cells, tissues and organs are largely constituted by proteins. The human body's muscles, skin, bones and many other parts contain significant amounts of protein. In fact, protein accounts for 20% of total body weight [45]. Enzymes, hormones and antibodies are proteins. Proteins also work as neurotransmitters and carri-

ers of oxygen in the blood (hemoglobin). We can imagine proteins as machines making all living things (viruses, bacteria, butterflies, jellyfish, plants and human), functions. The human body is made up of approximately 100 trillion cells, each one having a specific function. Each cell has thousands of different proteins, which together make the cell do its job.

The enormous variety of protein functions is based on their high specificity for the molecules with which they interact. However, this specific relationship demands a fairly rigid spatial structure of the protein*. This is the reason why the biological functions of proteins are closely connected with the rigidity of their three-dimensional (3D) structures. A little damage to these structures is often the reason for the loss of or dramatic changes in protein activities. A knowledge of the 3D structure of a protein is thus necessary to understand how it functions.

Proteins are heteropolymers: they are built up by amino acids that are linked into a peptide chain, as discovered by E. Fischer at the beginning of the 20th century. In the early 1950s Sanger [49] showed that the sequence of amino acid residues (a *residue* is the portion of the amino acid that remains free after polymerization) is unique for each protein. The chain consists of a chemically regular backbone (*main chain*) from which various side chains (R_1, R_2, ... , R_N) project, as shown in Fig. 1.1.

The number M of residues in a protein chain ranges from a few dozen to many thousands. There are twenty main species of proteinogenic amino acid residues.

* The main exception to this rule are the intrinsically disordered proteins (IDPs) that lack a fixed or ordered three-dimensional structure [46–48].

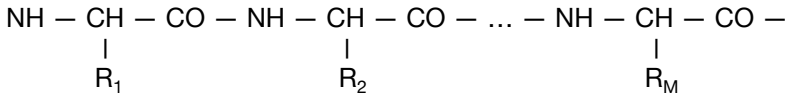


Figure 1.1: Amino acid chain. [Adapted from Ref. [50]].

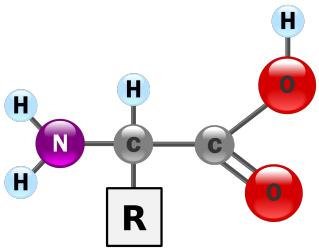


Figure 1.2: Structure of an amino acid in its un-ionized form.

In particular, the amino acids (Fig. 1.2) are biologically important organic compounds containing an amine (-NH_2) and a carboxylic acid (-COOH) functional groups, along with a side-chain that is specific to each amino acid. Fig. 1.4 displays the names and the structures of the naturally occurring amino acids.

The carbon that connects these two functional groups is called **alpha carbon** (or α-carbon or C_α), as shown in Fig. 1.3. It is the central point in the backbone of every amino acid, and it also serves as the point of attachment for the side chains of 19 out of 20 amino acids used in protein building. The sole exception is represented by the glycine, the only amino acid with no side chain (a hydrogen atom takes the spot where a side chain is attached to the C_α in the other amino acids).

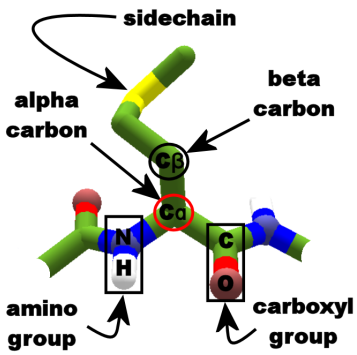


Figure 1.3: Amino acid representation showing the position of C_α, and C_β. In particular, the alpha carbon is the central point of all amino acids, while the beta carbon the first atom of the sidechain. Adapted from `Userblog:LociOiling/NewYear'sResolutions`

On the other hand, the **beta carbon** (β-carbon or C_β) is the first atom of the side chain in an amino acid. It is present in all twenty proteinogenic amino acids except for glycine.

To be able to perform their biological function, proteins fold into one or more specific spatial conformations driven by a number of non-covalent interactions such as hydrogen bonding, ionic interactions, Van der Waals forces, and hydrophobic packing [51]. To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure [51]. This is the topic of the scientific field of structural biology, which

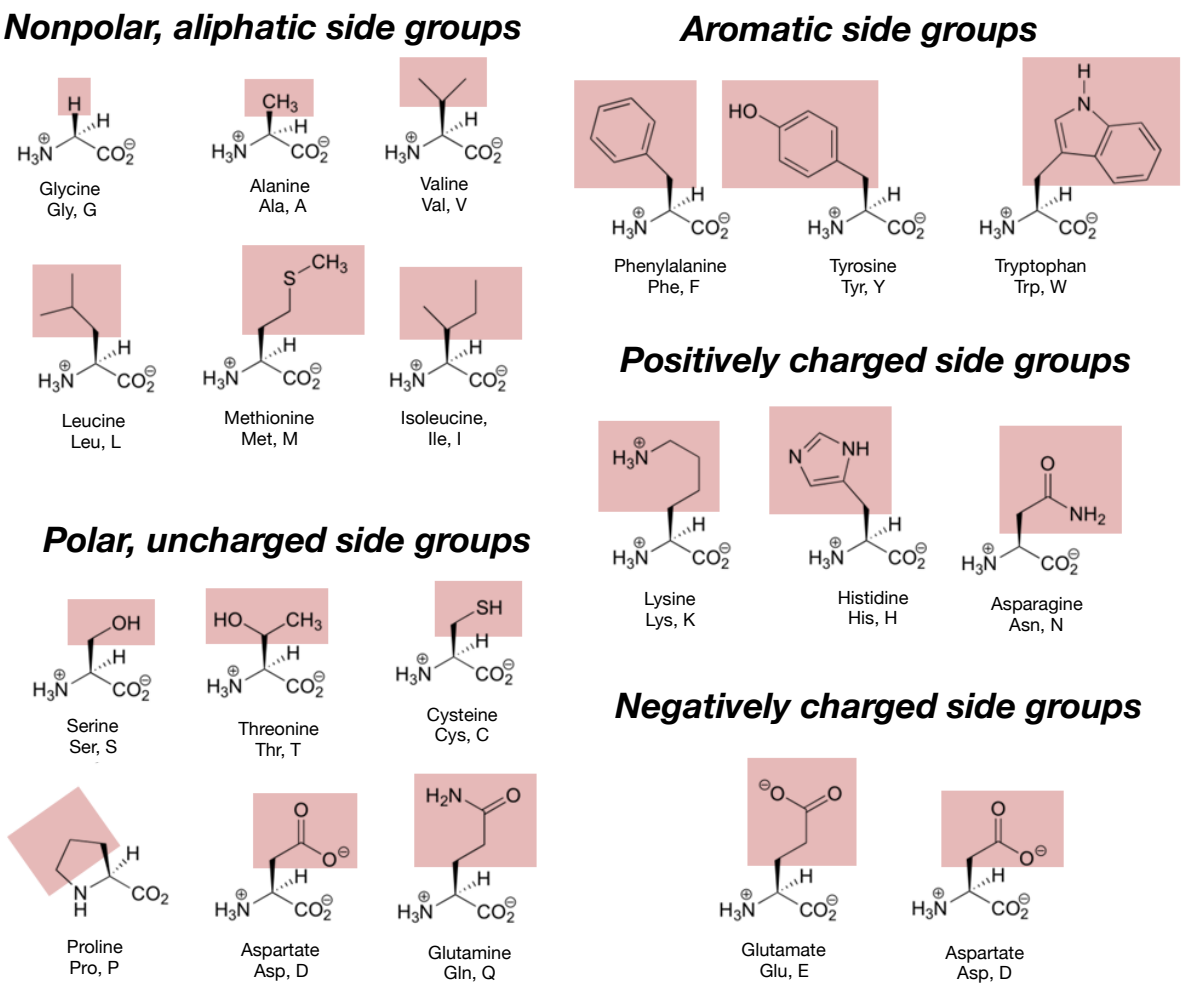


Figure 1.4: The naturally occurring 20 amino acids with complete name, abbreviations and chemical structures.

employs techniques such as X-ray crystallography [1, 2], or NMR spectroscopy [3]. In the late 1950s, Perutz and Kendrew [52] solved the first protein spatial structures and demonstrated their highly intricate and unique nature. Proteins perform their functions under various environmental conditions, which leave an obvious mark on their structures.

The less water there is around, the more valuable the hydrogen bonds are, and the more regular the stable protein structure ought to be. According to their environmental conditions and general structure, proteins can be roughly divided into three classes:

- **Fibrous proteins** form vast, usually water-deficient aggregates; their structure is usually highly hydrogen-bonded, very regular and maintained mainly by interactions between various chains.
- **Membrane proteins** are surrounded by lipids*: about a third of all human proteins falls into this class, and these are targets for more than half of all drugs. Nonetheless, determining membrane protein structures remains a challenge in large part due to the difficulty in establishing experimental conditions that can preserve the correct conformation of the protein in isolation from its native environment. Membrane proteins reside in a water-deficient membrane environment (although they partly project into water) and they have an amphiphilic nature: indeed, the lipids have one end that is soluble in water ('polar') and an ending that is soluble in fat ('nonpolar').
  Their intramembrane portions are extremely regular (like fibrous proteins) and vastly hydrogen-bonded, but restricted in size by the membrane thickness.
- Water-soluble (residing in water) **globular proteins** are less regular (especially small ones). Their structure is maintained by interactions of the chain with itself (where an important role is played by hydrophobic interactions between hydrocarbon groups that are far apart in the sequence but adjacent in space) and sometimes by chain interactions with co-factors.

The above classification is certainly extremely rough. Some proteins may comprise a *fibrous tail* and a *globular head* (like myosin, for example), and so on.

* membrane proteins need a membrane for structural stability and function

1.1.2 Protein structures

The structure of a protein is usually described in terms of **primary**, **secondary**, **tertiary** and **quaternary** structures (Fig. 1.5)

The *primary structure* of a protein refers to the linear sequence of amino acids in the polypeptide chain. It is held together by covalent (peptide) bonds, which are formed during the process of protein biosynthesis or translation. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity.

Secondary structures refers to highly regular local substructures in the polypeptide backbone chain. The presence of two main types of secondary structure, the α-helix and the β-sheet was proposed in 1951 by Linus Pauling and coworkers [54]. The former is often represented by helical ribbons (as shown in the grey part of Fig. 1.6), while the latter by arrows (red part of Fig. 1.6). Both structures are held in shape by hydrogen bonds, which form between the carbonyl O of one amino acid and the amino H of another, as represented in Fig. 1.7.

In an α-helix, the carbonyl O of one amino acid is hydrogen bonded to the amino H of an amino acid that is far down the chain. (E.g., the carbonyl of amino acid 1 would form a hydrogen bond to the N–H of amino acid 5.) This pattern of bonding pulls the polypeptide chain into a helical structure that resembles a curled ribbon, with each turn of the helix containing 3.6 amino acids. The R groups of the amino acids stick outward from the α-helix, where they are free to interact [55].

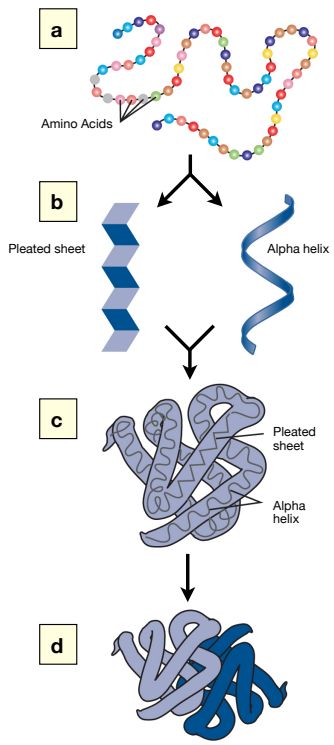


Figure 1.5: The different levels of protein organization: (a) primary, (b) secondary, (c) tertiary and (d) quaternary structures. The primary protein structure is the sequence of a chain of amino acids. The secondary protein structure occurs when the sequence of amino acids are linked by hidrogen bonds. The tertiary one occurs when certain attractions are present between α-helices and pleated sheets. Finally, the quaternary structure consists of more than one amino acid chain [Adapted from [53]].

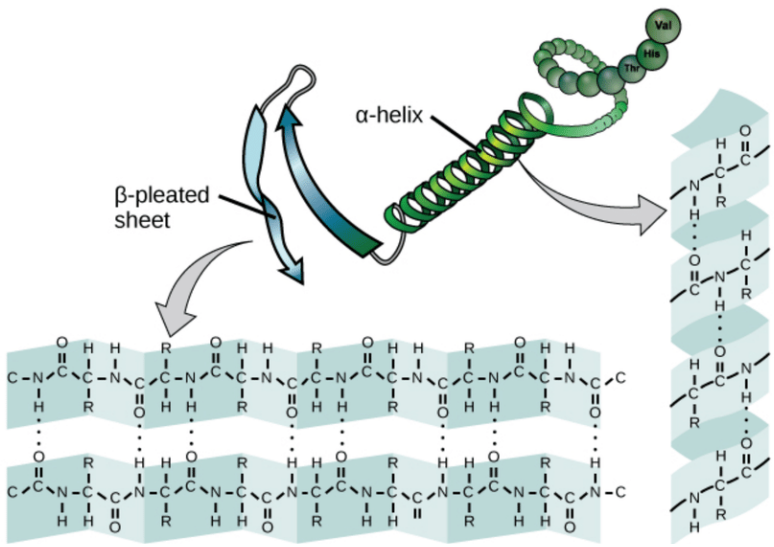


Figure 1.7: The two most common types of protein secondary structure, α-helices and β-sheets. that form because of hydrogen bonding (indicated by dots) between carbonyl and amino groups in the peptide backbone. Image from OpenStax Biology 2e / CC BY 4.0.

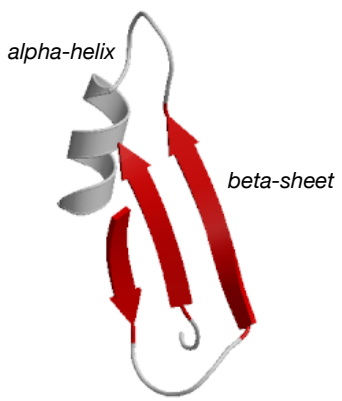


Figure 1.6: Generic fragment of a protein showing the two secondary structures: α-helix (grey helical ribbon) and β-sheet (red arrows).

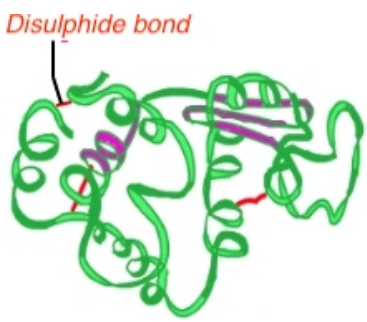


Figure 1.8: Tertiary structure of a protein. Disulphide bonds are highlighted by red lines.

In a β pleated sheet, two or more segments of a polypeptide chain line up next to each other, forming a sheet-like structure held together by hydrogen bonds. The hydrogen bonds form between carbonyl and amino groups of backbone, while the R groups extend above and below the plane of the sheet [55].

Secondary structures are characterized by a regular periodic shape (*conformation*) of the main chain with side chains assuming a variety of conformations.

The overall three-dimensional structure of a polypeptide is called its *tertiary structure*. In practice the α-helices and β pleated-sheets are folded into a compact structure. The folding is primarily due to interactions between the *R* groups of the amino acids that make up the protein. Also important are the non-specific hydrophobic interactions, but the structure is stable only when the parts of a protein domain are locked into place by specific tertiary interactions, such as salt bridges, hydrogen bonds, and disulphide bonds (Fig. 1.8).

Many proteins are made up of a single polypeptide chain

and have only three levels of structure. However, some of them contain multiple polypeptide chains, also known as subunits. When these subunits come together, they give the protein its *quaternary structure* (Fig. 1.9). Complexes of two or more polypeptides (i.e. multiple subunits) are called *multimers*. Specifically it would be called a dimer if it contains two subunits, a trimer if it contains three subunits, a tetramer if it contains four subunits, and a pentamer if it contains five subunits. Multimers made up of identical subunits are referred to with a prefix of "*homo-*" (e.g. a homotetramer) and those made up of different subunits are referred to with a prefix of "*hetero-*". The quaternary structure is stabilized by the same non-covalent interactions and disulphide bonds as the tertiary one.

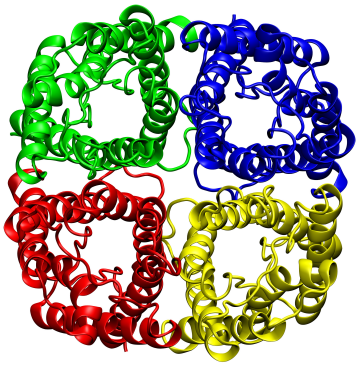


Figure 1.9: Quaternary structures of Human Aquaporin-4 (pdb code AQP4). This structure is a tetramer.

1.2 History and overview

Molecular simulation techniques have a relatively recent history. The first attempts to perform complex calculations using computers, as required to build the atomic bomb, date back to the early 1940s. However, the most significant progress in this field was achieved in the '50s, starting with the first simulations using Monte Carlo techniques for simple lattice systems. The method was published in 1953 [56] and became known as the Metropolis algorithm, the name of the first author. Simultaneously, the work of Fermi, Pasta, Ulam, and Tsingou [57] set the foundations of MD. The following years saw to the development of methods to calculate long-range interactions (*Ewald sum*), and various pioonering applications [51, 58]. This growing interest was favored by the technological developments

that brought computers to become more powerful, smaller, and cheaper.

In the '70s, the field of simulations was already so vast to encompass chemistry and molecular biology, thus building the basis of the research field of "*soft matter*"*. The first works that addressed the issue of a system out of equilibrium were also published in the '70s. Meanwhile, molecular simulations became increasingly related to the discipline of statistical mechanics, for two significant reasons. On the one hand, the need to calculate averages of different physical observables: these averages can be calculated along the trajectory of the system or as a result of proper sampling. In this sense, statistical mechanics is used as a tool to analyze the data produced by a computer. On the other hand, the increased performance of computers allowed simulating the behavior of systems composed of a large number of particles. Hence, the latter are ideal benchmarks for theories and approximations that properly pertain to statistical mechanics.

The development of molecular simulations at the quantum level of detail is more recent and can be traced back to 1985 with the first so-called *ab initio* calculations (that is attributable to first principles, such as the Schrödinger equation). Of particular importance in this context is the Car-Parrinello molecular dynamics method. Finally, in more recent years, the field of molecular simulations has seen the most significant progress in the development of free energy calculation and enhanced sampling techniques.

Nowadays, computer simulations play an essential role in molecular modeling. Molecular dynamics and related

* the field of "soft matter" research has been pioneered by Pierre-Gilles de Gennes [59].

techniques are applied in many fields of science: from chemical physics to material science, as well as in biophysics and biochemistry. They are used to discover new and efficient materials for organic electronics [60, 61], to investigate millions of chemical compounds in drug screening applications [62], to study the functionality of biological properties [21, 63–65].

However, despite ever-growing computer power, and modern parallelization techniques (e.g. by using GPUs), many applications are still a challenge: for instance, many proteins fold on timescales far beyond milliseconds, while fully atomistic protein folding simulations reach only a few milliseconds [66–68]. Moreover, other computational challenges occur on different time and length scales: they have driven the continuous development of advanced simulation methods for molecular modeling [69, 70]. In general, these techniques can be described in a hierarchy (see Fig. 1.10).

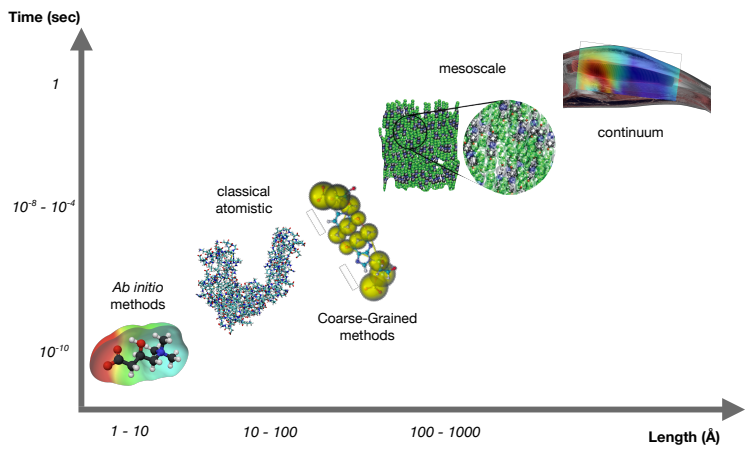


Figure 1.10: Multi-scale nature of matter. Depending on length and time scales of interest, a different approach should be used to describe a specific phenomenon.

In particular, depending on time and length scales of the phenomena, the approaches and methods are classified differently. Starting from the left-bottom corner, we find the *ab initio* methods, necessary at the electronic scale, in which the quantum mechanical effects are not negligible

and calculated by solving the *Kohn-Sham* equations [71, 72]. At the atomic scale, the molecular dynamics simulation method is used. In this approach, various numerical techniques are used to solve Newton's equation of motion. At the mesoscopic scales, coarse-grained molecular descriptions are employed. Eventually, the largest scales is the continuum level, in which the dynamics of the system is described by field equations, obtained by imposing local macro-scale conservation laws.

On the one hand, highly accurate techniques (e.g., *ab initio* simulations) are computationally expensive and thus applicable on small length and short time scales. On the other hand, numerically efficient but less accurate approaches, like effective, coarse-grained models, allow longer simulation of larger systems. In the following, we will discuss in further detail molecular dynamics and, in particular, the Coarse-Grained and Multi-Resolution methods: both are particularly crucial for the investigation of the system of interest in this thesis.

1.3 Molecular Dynamics

For nearly all systems of interest to us, the most transferable and fundamental description of matter is one that invokes quantum mechanics. At the highest level of accuracy, this requires to solve Schrödinger's equation for all of the subatomic particles in a system. Computationally, many approximations need to be made in order to use *ab initio* methods, and even these techniques are limited to small numbers of atoms. There are both practical and philosophical reasons for performing simulations on simpler

systems that do not entail a full solution of the quantum-mechanical equations. Practical reasons stem from the need to treat larger systems and run simulations for longer times than those that *ab initio* methods can achieve.

In this respect, **Molecular Dynamics** (MD) [51, 73] is a technique allowing the calculation of thermodynamic and dynamic properties of a large number of systems in different conditions. It is based on the applicability of the laws of classical mechanics to microscopic systems composed of molecules and atoms described as point particles. In general, many atoms are sufficiently massive to allow the description of their motion quite accurately by the laws of Newtonian mechanics. The setup of a classical MD simulation has several analogies with the setup of an experiment. A MD simulation essentially requires three basic ingredients:

- **A model for the interaction** between system constituents. Often, it is assumed that particles interact only pairwise, with the exception of bonded interactions. This assumption dramatically reduces the computational effort.
- **An integrator**, which propagates particles' positions and velocities from time t to $t + \Delta t$. It is a finite difference scheme that yields trajectories defined at discrete values of the time.
- **A statistical ensemble**, where thermodynamic quantities like temperature, pressure, and the number of particles are controlled. The natural choice for the ensemble is the microcanonical one "*NVE*" (i.e., number of particles, volume, and energy constants, respectively) since the Hamiltonian of the system is a con-

served quantity if external potentials are not present. Nevertheless, there are extensions to the equations of motion that also allow the simulation of different statistical ensembles.

In classical MD simulations, the time evolution of a set of interacting particles is usually calculated by numerically solving Newton's equations of motion of the particle belonging to the system. For a system of N point particles of masses m_i $(i = 1, \ldots, N)$ at position $\mathbf{r}_i(t) = (x_i(t), y_i(t), z_i(t))$ and velocity $\mathbf{v}_i(t)$ the force acting upon the i-th particle at time t can be computed as follows:

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

where $\mathbf{F}_i$ is obtained from the potential energy of the system $U(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ via the Eq. 1.2

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r}) = -\left(\frac{\partial U}{\partial x_i}, \frac{\partial U}{\partial y_i}, \frac{\partial U}{\partial z_i}\right) \tag{1.2}$$

Given the initial condition $\mathbf{r}_i(t_o)$ and $\mathbf{v}_i(t_o)$ at a certain time t_0, the solution of Eq. 1.1 provides the complete information on the motion of the system. Alternatively, Hamilton's equations of motion for the (generalized) momenta $\mathbf{p}_i$ and position $\mathbf{r}_i$ can be used to calculate the time evolution of the system:

$$\dot{\mathbf{r}}_i = \nabla_{\mathbf{p}_i}\mathcal{H} \qquad \dot{\mathbf{p}}_i = -\nabla_{\mathbf{r}_i}\mathcal{H} \tag{1.3}$$

where $\mathcal{H}$ is the Hamiltonian of the system:

$$\mathcal{H} = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}) \tag{1.4}$$

In Cartesian coordinates, Hamilton's equations become:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} \qquad \dot{\mathbf{p}}_i = -\nabla_{r_i} U(\mathbf{r}) = \mathbf{F}_i \tag{1.5}$$

Due to the complexity of the many-body problem, the only feasible way to solve Eq. 1.1 when N > 2 is by discretizing the time and solving it numerically through a computer. The positions and velocities are propagated with a finite time interval using numerical integrators, such as the Verlet algorithm (it will be described later). The position of each particle in space is defined by $\mathbf{r}_i(t)$, whereas the velocities $\mathbf{v}_i(t)$ or the momenta $\mathbf{p}_i(t)$ are used to get kinetic energy and temperature in the system. As the particles 'move', their trajectories may be displayed and analyzed, providing averaged properties. The dynamic events that may influence the functional properties of the system can be directly traced at the atomic level, making MD particularly valuable in molecular biology [51, 73, 74].

1.3.1 Force Field

In MD simulations [51, 73], the force is derived from the potential energies (U) that are defined by a **force field** (FF). A force field is a model that describes the interactions between atoms inside the molecular system of interest. The parameters used in the functional forms of the model are usually obtained by experimental data or quantum mechanics calculations such as the density functional theory (DFT) [71, 72]. There are many FF's, each one best suited to the description of a particular category of systems. For instance, the CHARMM [75], GROMOS [76], and AMBER [77] force fields describe proteins and biological systems,

while OPLS-AA [78] describes systems in the liquid state. The functional forms of the different FF are very similar. Here, we will describe the AMBER model used in this works. The potential is a function of the positions of all the atoms of the system, and it is given by the sum of two terms:

$$U = U_{bonded} + U_{non-bonded} \tag{1.6}$$

- U_{bonded} takes into account the interactions between the atoms chemically bonded by a covalent bond. Such interactions depend on the bond lengths, angles and rotations of the bonds in a molecule;
- $U_{non-bonded}$ takes into account the interactions between the atoms which are not chemically bonded or between atoms separated by three or more covalent bonds.

We can even separate the contribution to each of the two terms as follows:

$$\begin{aligned} U_{bonded} &= U_{bond} + U_{bend} + U_{tors} + U_{impr} \\ U_{non-bonded} &= U_{Van-der-Waals} + U_{Coulomb} \end{aligned} \tag{1.7}$$

The functional form for the AMBER force field is thus the following:

$$\begin{aligned} V = &\sum_{bonds} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \\ &+ \sum_{dihedrals} k_\theta [1 + cos(n\phi - \delta)] + \sum_{impr} k_\omega (\omega - \omega_0)^2 + \\ &+ \sum_{ij} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{ij} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \end{aligned} \tag{1.8}$$

The first term is the energy function accounting for the

bond stretches for a given bond where k_b is the bond force constant, b is the bond length, and b_0 is the equilibrium bond distance.

The second term in the equation accounting for the bond angles, where k_θ is the force constant, θ is the bond angle and θ_0 is the equilibrium angle formed by three bonded atoms. The third term describes dihedrals: k_ϕ is the dihedral force constant, n is the multiplicity of the function, ϕ is the dihedral angle, and δ is the phase shift. The fourth term accounts for the impropers, that is out-of-plane bending: k_ω is the force constant and $\omega - \omega_0$ is the out-of-plane angle (Fig. 1.11). It is also possible to notice that the functional form of the interaction is harmonic in all cases, except for dihedrals where it is expressed as a truncated fourier series.

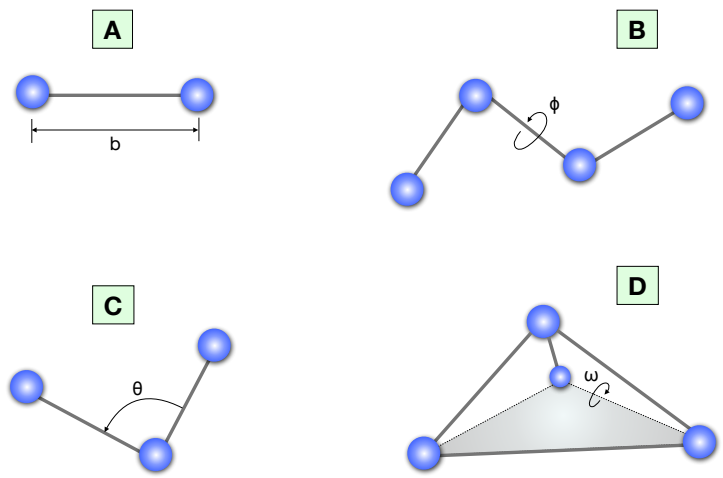


Fig. 1.11: Representation of terms: (A), U_{bond}. (B), U_{tors}. (C), U_{bend} (D), U_{impr}.

Non-bonded interactions are given by the sum of the last two energy terms in Eq. 1.8. In particular, the Lennard-Jones potential (fifth term) is used to model the *Van der Waals (VdW)* interactions:

- σ is a measure of how close two atoms can get. Thus

it corresponds to the Van der Waals radius of a given atom. σ_{ij} is the arithmetic mean between the radii of particle i and j respectively, namely $\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2}$.

- ϵ determines the magnitude of the attractive energy between atoms i and j. The parameters ϵ_{ij} are obtained with a geometric mean, namely $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$.

With the 6th and 12th powers, the Lennard-Jones potential decays very fast with the distance. Thus it is often treated by using the cutoff method, where the potential is truncated or smoothly switched to zero at a distance larger than a certain cutoff distance. The sixth term of Eq. 1.8 is the Coulomb potential, which is used to model the electrostatic interactions such as dipole-dipole, ion-dipole, and ion-ion interactions. In Eq. 1.8 $\mathbf{q}_i$ denotes the fixed partial charge associated to the i-th atom. The total charge of a molecule will be given by the sum of partial charges on its constituent atoms. In comparison with the Lennard-Jones potential, the Coulomb potential is long-ranged and decays very slowly. Therefore, the cutoff method cannot be applied to compute electrostatic interactions, thus increasing the computational cost of this contribution to the total potential energy. A popular method to compute the long-range electrostatic interactions is the Ewald summation [73]. In the Ewald summation, the slow-decaying Coulomb interaction is decomposed into a pair of fast converging terms: one can be directly computed in the Cartesian space, while the other one is calculated in the reciprocal space upon Fourier transformation. The reciprocal sum is made over an infinite number of periodic images: in fact, the Ewald summations were originally designed to compute the long-range interactions of crystals. Due to the periodic

boundary conditions, the Ewald summations can now be widely applied in MD simulations.

The Ewald summation is still computationally expensive because the cost of the reciprocal sum increases with $\mathcal{O}(N^2)$, where N is the number of particles in the system. Therefore, the application of the Ewald summation is limited to small systems. The performance of the reciprocal sum was improved by the development of the *Particle Mesh Ewald* (PME) method [79], in which the charges are assigned to grid-points using interpolating functions. The computational cost of the PME method scales as $N \cdot \log(N)$, and thus it has been widely used in simulating complex systems such as proteins.

1.3.2 Integrators

As previously discussed, solving Newton's equations of motion analytically is impossible due to the complex form of the potential energy function $U(\mathbf{r}_1, \ldots, \mathbf{r}_n)$ that depends on the positions of all the particles $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ in the system. Therefore, several numerical integration algorithms have been developed to solve Newton's equations of motion. All these algorithms are based on the Taylor expansions of positions*.

Verlet integrator

This integrator was developed by Verlet in 1967 [80]. In order to derive it, we first write down the Taylor expansion of $\mathbf{r}(t + \Delta t)$ for small Δt:

* Formally the Trotter decomposition of Liouville operator formulation is employed [73].

$$\mathbf{r}_i\left(t+\Delta t\right)=\mathbf{r}_i(t)+\mathbf{v}_i(t)\Delta t+\frac{1}{2}\frac{\mathbf{F}_i(t)}{m_i}\Delta t^2+\frac{1}{3!}\dddot{\mathbf{r}}_i(t)\Delta t^3+\mathcal{O}\left(\Delta t^4\right) \tag{1.9}$$

where $\mathbf{v}_i(t)$ is the velocity of the particle i and $\mathbf{F}_i$ is the force acting on the particle i at time i. The Taylor expansion of $\mathbf{r}\left(t-\Delta t\right)$ is:

$$\mathbf{r}_i\left(t-\Delta t\right)=\mathbf{r}_i(t)-\mathbf{v}_i(t)\Delta t+\frac{1}{2}\frac{\mathbf{F}_i(t)}{m_i}\Delta t^2-\frac{1}{3!}\dddot{\mathbf{r}}_i(t)\Delta t^3+\mathcal{O}\left(\Delta t^4\right) \tag{1.10}$$

Summing up both sides of Eqs. 1.9 and 1.10, we obtain:

$$\mathbf{r}_i\left(t+\Delta t\right)=2\mathbf{r}_i(t)-\mathbf{r}_i\left(t-\Delta t\right)+\frac{\mathbf{F}_i(t)}{m_i}\Delta t^2+\mathcal{O}\left(\Delta t^4\right) \tag{1.11}$$

The velocities do not appear in equation 1.11 but can be obtained by subtracting Eq. 1.10 from Eq. 1.9:

$$\mathbf{v}_i(t)=\frac{\mathbf{r}_i\left(t+\Delta t\right)-\mathbf{r}_i\left(t-\Delta t\right)}{2\Delta t} \tag{1.12}$$

Leap-frog algorithm

The Verlet algorithm requires knowledge of position at times t and $t-\Delta t$ to obtain the coordinates at time $t+\Delta t$. A fully equivalent alternative is provided by the *leap-frog algorithm*. This algorithm defines the speed at half time steps:

$$\mathbf{v}_i\left(t-\frac{\Delta t}{2}\right)=\frac{\mathbf{r}_i(t)-\mathbf{r}_i\left(t-\Delta t\right)}{\Delta t} \tag{1.13}$$

$$\mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)=\frac{\mathbf{r}_i\left(t+\Delta t\right)-\mathbf{r}_i(t)}{\Delta t} \tag{1.14}$$

From these definitions we find immediately, $\mathbf{r}_i\left(t+\Delta t\right)$ by

solving:

$$\mathbf{r}_i\left(t+\Delta t\right)=\mathbf{r}_i(t)+\Delta t\cdot\mathbf{v}_i\left(t+\frac{\Delta t}{2}\right) \tag{1.15}$$

Velocity Verlet algorithm

Although fully equivalent to the Verlet algorithm, the leap-frog algorithm yields coordinates and velocities at different instants of time. This implies that it is not possible to calculate the total energy of a system at any time since the kinetic and potential energies will be defined at semi-integer and integer multiples of Δt, respectively. An improvement of the leap-frog algorithm was later proposed by Swope et al. [81]. Considering the expansion of the coordinates up to the second order, we have:

$$\mathbf{r}_i\left(t+\Delta t\right)=\mathbf{r}_i(t)+\mathbf{v}_i(t)\cdot\Delta t+\frac{\Delta t^2}{2m_i}\mathbf{F}_i(t) \tag{1.16}$$

We can also start from $\mathbf{r}_i(t+\Delta t)$ and $\mathbf{v}_i(t+\Delta t)$ and then integrate back in time to $r_i(t)$ obtaining:

$$\mathbf{r}_i(t)=\mathbf{r}_i\left(t+\Delta t\right)-\mathbf{v}_i(t+\Delta t)\Delta t+\frac{\Delta t^2}{2m_i}\mathbf{F}_i\left(t+\Delta t\right) \tag{1.17}$$

Combining the Eqs. 1.16 and 1.17 and solving for $\mathbf{v}_i\left(t+\Delta t\right)$ we get:

$$\mathbf{v}_i\left(t+\Delta t\right)=\mathbf{v}_i(t)+\frac{\Delta t}{2m_i}\left[\mathbf{F}_i(t)+\mathbf{F}_i\left(t+\Delta t\right)\right] \tag{1.18}$$

The two Eqs. 1.16 and 1.18 allow calculating the time evolution of positions and velocities simultaneously. It can be easily shown that the *Velocity Verlet* algorithm is fully equivalent to the *Verlet* algorithm.

1.3.3 NPT and NVT ensembles

The direct application of the integrators introduced in section 1.3.2 will produce simulations in the microcanonical ensemble. However, in many cases, we wish to simulate the system at a constant temperature or in the canonical ensemble. Several methods have been proposed to achieve temperature control in MD simulations.

These methods mimic the effect of a large energy reservoir (thermostat) coupled to the system. The temperature in an MD simulation is obtained through the equipartition theorem using the instantaneous value of the total kinetic energy:

$$k_b T^* = \frac{1}{D} \sum_{i=1}^{N} m_i \mathbf{v}_i^2 \tag{1.19}$$

where D is the number of degrees of freedom of the system, and k_b is the Boltzmann constant. A first intuitive approach would be to rescale all velocities at each time step to keep T constant. This approach is wrong for two reasons: in fact, it introduces sudden jumps of the kinetic energy of each particle trajectories at certain points in time. This discontinuity fits badly with the approach of molecular dynamics. Moreover, this algorithm cancels in whole (or in part if the correction is applied only when the temperature exceeds a certain degree) the fluctuations of the kinetic energy, which are instead typical of a canonical ensemble. To overcome the first problem, the *Berendsen thermostat* (1984) [82] introduces an additional differential equation for the kinetic energy :

$$\dot{K} = \frac{2K - 3Nk_b T^*}{\tau} \tag{1.20}$$

where τ is a time constant. The Berendsen thermostat cannot produce a proper canonical ensemble since it suppresses the fluctuations of the kinetic energy. However, it has the advantage of easy tuning of the coupling strength: it is thus recommended in the equilibration phase of a simulation. A later study fixed some of the issues in the Berendsen thermostat by introducing an additional stochastic term that ensures the correct fluctuations of the kinetic energy. This improved Berendsen scheme is also called the *velocity-rescaling thermostat* [83].

Other methods widely used in the physics community achieve temperature control in MD simulation by modeling the interaction with a heat reservoir in a stochastic fashion through introducing random perturbations and friction acting on the particles [84, 85].

1.4 Coarse-Grained models

Atomistic molecular dynamics simulations (MD) with an all-atom force field provide deep and broad insights into molecular-scale phenomena. Nevertheless, all-atom simulations are limited to tiny systems and nanosecond time scales. Therefore the development of simplified or coarse-grained (CG) molecular models has become an active field of research in the past few decades.

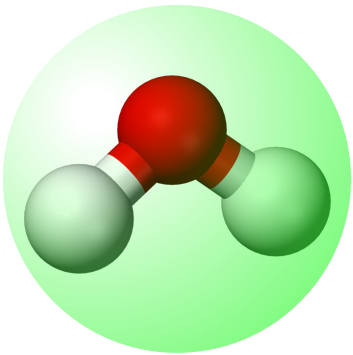


Figure 1.12: Molecular structure of water molecule. The corresponding coarse-grained bead is shown with a transparent green bead.

Coarse-grained models are a reduced representation of all-atom models: in this approach, several atoms are grouped together and described jointly as a single point-like particle (see Figs. 1.12 and 1.13 for a visualization). This significantly reduces the number of particles in the system and hence also the computational cost allowing the

simulation of large-scale biological systems.

Other three aspects are prominent: first, the interaction potentials in CG models are typically much softer than atomistic force fields; second, larger time steps can be chosen when integrating the equations of motion, further alleviating the numerical effort; third, the potential energy surface on which the molecules move is smoothed, leading to an acceleration of the molecular dynamics [86].

However, it is not only their computational efficiency that makes CG models attractive though. In many large-scale applications, we are not interested in the microscopic details of the system anyhow. In a sense, CG approaches automatically average out these details and focus only on the relevant length scales [19].

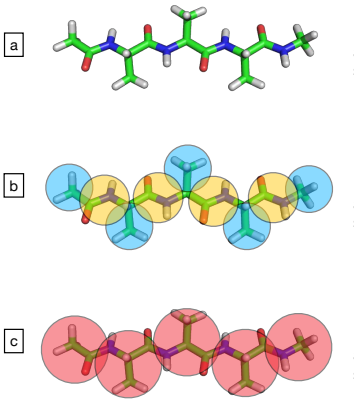


Figure 1.13: Polyalanine structure in three different representations: (a) all-atoms, (b) coarse grained of functional groups, (c) coarse-grained of amino acid residues.

As stated above, CG models typically entail **pseudoatom sites** that are designed to represent combined groups of atoms. Usually, pseudo atoms are defined as groups of atoms of common chemistry, like methyl or carbonyl groups. Alternatively, they can contain many functional units, as represented in the Fig. 1.13. The latter, in particular, shows the polyalanine in three different representations: (a) all-atom, (b) coarse-grained of functional groups, and (c) coarse-grained of amino acid residues.

Such a description of a system requires the definition of a mapping function. The latter takes as input a set of atomistic coordinates in the fully atomistic system and maps this to a unique bead in the CG system. Mathematically, this means the CG coordinates $\mathbf{R}$ are constructed from the atomistic coordinates $\mathbf{r}$ via

$$\mathbf{R} = \mathscr{M}\mathbf{r} \tag{1.21}$$

where $\mathcal{M}$ is an $n \times N$ matrix (n and N being the number of particles in the atomistic and CG system, respectively).

However, the main challenge of coarse-graining lies in the derivation of CG models that correctly catch only the relevant features of the system, neglecting the unnecessary details. A wide variety of approaches for the parametrisation of CG model exists that can be divided into two main groups: *bottom-up* and *top-down*.

Bottom-up coarse-graining approaches employ information from a more detailed model (usually all-atom reference) to systematically fit the potential for a CG model of the same system. The most common techniques are **Iterative Boltzmann Inversion (IBI)** [87, 88] and **Inverse Monte Carlo** [89, 90], which aim at preserving reference pair correlation functions, **Force Matching** [91, 92], which tries to reproduce the multi-body potential of mean force, and **Relative Entropy based methods** [93, 94] which minimize the information loss between the CG and the reference system. In top-down models, the interactions are parametrized without explicit consideration of a more detailed model. Usually, the interactions are determined either based on physico-chemical properties to reproduce some structural or thermodynamic feature that is observed on larger scales. MARTINI force field [95, 96] falls into this category*.

Another example of a CG model derived from the top-down approach is the classical Elastic Network Model (ENM) [36, 97–102] in which the interactions between the CG beads are parametrized based on a reference structure, but without any knowledge of the real forces acting be-

* More precisely, MARTINI is parametrised "top-down" regarding the non-bonded interaction; on the other hand, the bonded interactions are defined "bottom-up", parametrised starting from all-atom or ab-initio simulations.

tween the atoms. Hereafter we discuss in further detail such a method.

1.4.1 Elastic Network Models (ENMs)

Generalities

Elastic Network models for proteins were introduced, for the first time, by Monique Tirion [36] as a simplified approximation of the potential energy function of a system (e.g., macromolecules) near equilibrium. In particular, she proved that the all-atom force field of a protein could be replaced by local springs, reproducing with great fidelity the protein's low-energy vibrational spectrum. In other words, the very accurate potential of a realistic model, i.e., bonds, angles, torsion, bending potentials, Van der Waals forces, and electrostatic interaction, can be substituted by an effective potential, whose form is:

$$V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r}) = \frac{1}{2}K \sum_{i<j} C_{ij}(r_{ij} - r_{ij}^0)^2 \tag{1.22}$$

where:

- r_{ij} is the scalar distance between the particles i and j computed as the absolute value of the distance vector i.e. $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$;
- r_{ij}^0 corresponds to the same quantity but evaluated in the reference conformation: $r_{ij}^0 = |\mathbf{r}_i^0 - \mathbf{r}_j^0|$;
- K is the spring constant;
- C_{ij} is called *contact matrix* and it is defined as:

$$C_{ij} = \begin{cases} 1, & \text{if } r_{ij}^0 \leq R_c \\ 0, & \text{otherwise} \end{cases}$$

where R_c is the cutoff distance within which two atoms must be located in the reference structure to interact.

Thus, despite the ENMs are a coarse-grained model, the number of degrees of freedom is not reduced.

It is important to realize that the potential energy function in the Eq. 1.22 is not quadratic in the coordinates $\mathbf{r}_i$ because the distance r_{ij} involves the calculation of a square root, by definition. Nevertheless, it is possible to expand the Eq. 1.22 in terms of the displacements $\mathbf{r}_i - \mathbf{r}_i^0$ from the reference structure, according with the Taylor formula:

$$\begin{aligned} V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r}) = V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r}^0) + \sum_i \left.\frac{\partial V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r}^0)}{\partial \mathbf{r}_i}\right|_{\mathbf{r}^0} \left(\mathbf{r}_i - \mathbf{r}_i^0\right) + \\ + \frac{1}{2}\sum_{i,j} \left.\frac{\partial^2 V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r})}{\partial \mathbf{r}_i \partial \mathbf{r}_j}\right|_{\mathbf{r}^0} \left(\mathbf{r}_i - \mathbf{r}_i^0\right)\left(\mathbf{r}_j - \mathbf{r}_j^0\right) + \mathcal{O}\left(\mathbf{r} - \mathbf{r}^0\right)^3 \end{aligned} \tag{1.23}$$

The constant term $V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r}^0)$ can be neglected because it is just a shift of the potential, and therefore can be considered as an irrelevant constant. Moreover, the first derivative in the previous equation vanishes at $\mathbf{r}^0$ because of the extremality condition: indeed, in correspondence of the minimum, the first derivative of a function is zero. Thus, the first non-zero contribution to the potential is given by its second derivative. All these considerations lead to:

$$\begin{aligned} V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r}) &\approx \frac{1}{2}\sum_{i,j} \left.\frac{\partial^2 V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r})}{\partial \mathbf{r}_i \partial \mathbf{r}_j}\right|_{\mathbf{r}^0} \left(\mathbf{r}_i - \mathbf{r}_i^0\right)\left(\mathbf{r}_j - \mathbf{r}_j^0\right) \\ &= \frac{1}{2}\sum_{i,j} \Delta\mathbf{r}_i^{\dagger}\mathbf{H}_{ij}\Delta\mathbf{r}_j \end{aligned} \tag{1.24}$$

where we used the substitutions $\Delta\mathbf{r}_i = \mathbf{r}_i - \mathbf{r}_i^0$, while $\mathbf{H}_{ij}$ is

the Hessian matrix defined as:

$$\mathbf{H}_{ij} = \left.\frac{\partial^2 V_{\mathrm{ENM}}^{\mathrm{AT}}(\mathbf{r})}{\partial \mathbf{r}_i \partial \mathbf{r}_j}\right|_{\mathbf{r}^0} \tag{1.25}$$

The elastic network model described in the Eq. 1.22 is also called *Anisotropic Elastic Network Model* (ANM) since the energy cost associated with the displacement for an atom depends on its direction: the information about the orientation of each interaction with respect to the global coordinates system is considered within the Force Hessian matrix $\mathbf{H}_{ij}$.

On the contrary, in the so-called *Gaussian ENM* (or GNM) [98, 103, 104], the potential is given by:

$$V_{\mathrm{GNM}} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}(\Delta R_j - \Delta R_i)^2\right] = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\Delta R_i \Gamma_{ij} \Delta R_j\right] \tag{1.26}$$

where γ is a force constant uniform for all springs and Γ_{ij} is the ij^{th} element of the Kirchhoff (or connectivity) matrix of inter-residue contacts, defined by:

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \quad \text{and } R_{ij} \leq r_c \\ 0, & \text{if } i \neq j \quad \text{and } R_{ij} > r_c \\ -\sum_{j,j\neq i}^{N} \Gamma_{ij}, & \text{if } i = j \end{cases} \tag{1.27}$$

where r_c is a cutoff distance for spatial interactions and taken to be 7 Å for amino acid pairs, represented by their α-carbons.

Since in the Eq. 1.26 the pairwise interaction is proportional to the square vector displacement $(\Delta R_j - \Delta R_i)^2$, the energy cost associated at a displacement does not depend on the direction in which it is performed, unlike the ANM.

From now on, except if explicitly declared, we will focus on ENM, always intended as anisotropic models.

From atomistic to Coarse Grained

As explained in Sec. 1.4, the approaches to coarse-graining can be enclosed in two main categories: bottom-up and top-down. In particular, the ENMs fall in the latter group, as the interactions between the CG beads are parametrized based on a reference structure, but without any knowledge of the real forces acting between the atoms.

However, the construction of a low-resolution ENM (or CG-ENM) requires, as a first thing, the choice of a smaller set of new degrees of freedom, and second the definition of effective interactions among them. The first step can be expressed by using a *mapping* between the atoms described in the high-resolution level (all-atom representation) and the smaller number of CG sites in the lower resolution.

Usually, the mapping is such that the CG coordinates (R_I) can always be expressed as a linear combination at atomistic coordinates (r_i). In this case, the mapping function becomes a **mapping matrix** $\mathcal{M}$:

$$\mathbf{R}_I = \mathcal{M}_{Ii}\mathbf{r}_i \tag{1.28}$$

where the convention of summation over repeated indices is employed.

When constructing CG-ENMs, the most common choice for mapping is to retain only the C_α for each amino acid, whose position is precisely the same of the C_α in the fully-atomistic representation. It leads to a quasi-uniform mass distribution along the protein backbone. This strategy is

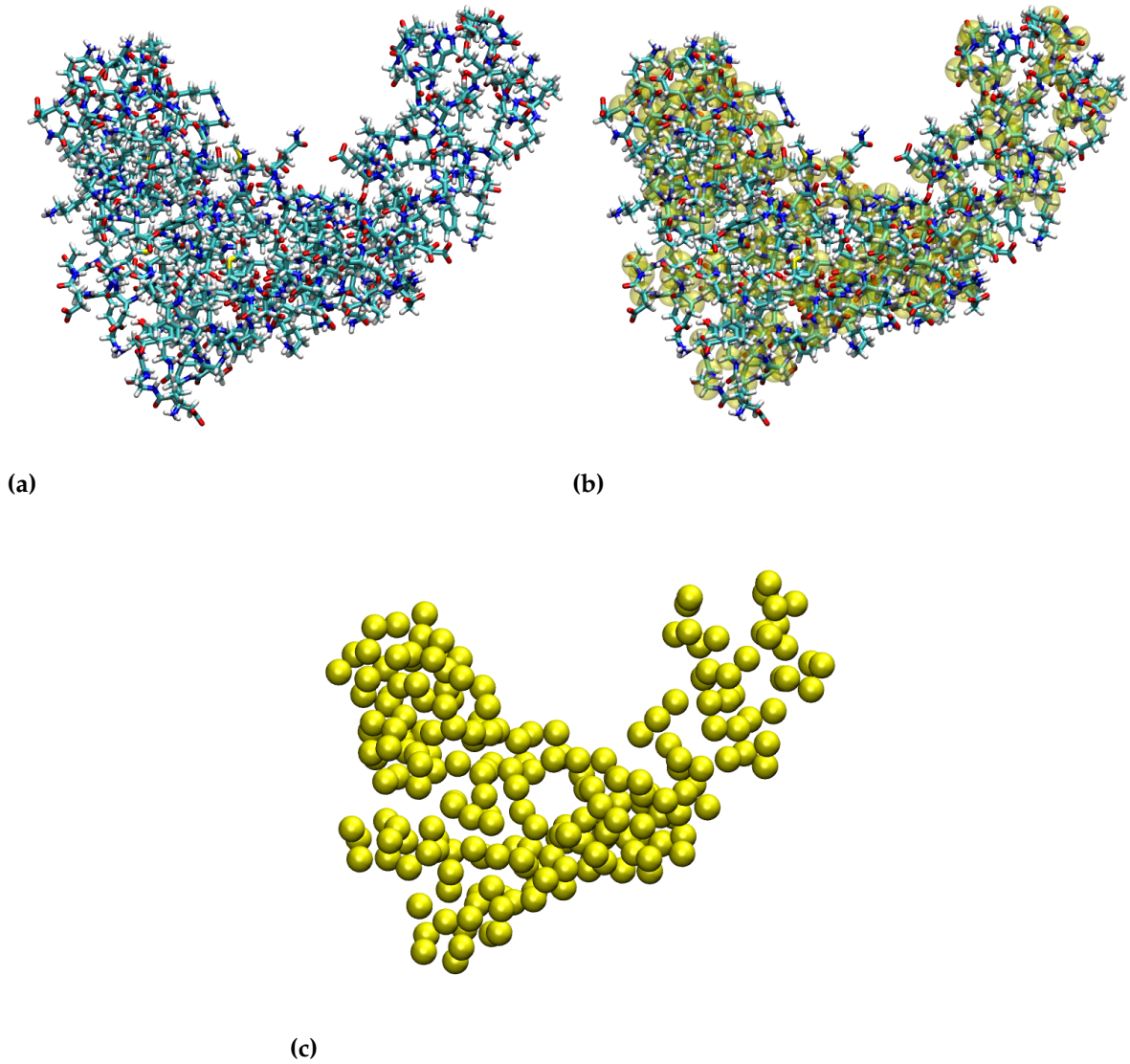


Figure 1.14: Adenylate kinase (4AKE) in case of: (a) all-atom representation, (b) mapping of each residue onto its C_α colored in light transparent yellow (c) C_α-only for each residue representation colored in yellow.

graphically shown in Fig. 1.14. In particular, in (a) and (c) is represented the all-atom, and the C_α-only representation of a protein Adenylate Kinase respectively; in (b) is shown the intermediate step between the two, mapping each residue onto its C_α. Other mappings may be possible, for instance, retaining both C_α and C_β for each residue, but it is less frequent.

Once the mapping has been established, interactions have to be defined. Usually, the interaction is harmonic,

as in the Eq. 1.22. In analogy, in our specific CG case, we could write that:

$$V_{\text{ENM}}^{\text{CG}}(\mathbf{R}) = \frac{1}{2}\sum_{I<J} K_{IJ}(R_{IJ} - R_{IJ}^{0})^2 \tag{1.29}$$

where K_{IJ} is the spring constant between the site I and J. Its value is 0 in case there is no spring between two sites. This model as well can be expanded in Taylor series. In analogy with the Eqs. 1.22, 1.23 and 1.24, it turns out that:

$$V_{\text{ENM}}^{\text{CG}}(\mathbf{r}) = \frac{1}{2}\sum_{I,J} \Delta\mathbf{R}_I^{\dagger}\mathbf{H}_{IJ}\Delta\mathbf{R}_J \tag{1.30}$$

where the Hessian Matrix is given by:

$$\mathbf{H}_{ij} = \left.\frac{\partial^2 V_{\text{ENM}}^{\text{CG}}(\mathbf{R})}{\partial\mathbf{R}_I\partial\mathbf{R}_J}\right|_{\mathbf{R}^0} \tag{1.31}$$

Two applications of the Elastic Network Models [36, 105, 106] in the context of protein modelling can be found in chapters 3 and 4 representing the coarse-grained part of the Multiscale Simulations. The latter are the subject of the next section.

1.5 Multiscale Simulations

Although using a coarse-graining technique (see Sec. 1.4) makes it possible to characterize the relevant properties of a system at a cheaper computational costs, it is not able to answer the questions related to the systems in which the chemical details of a small region have major effects on the system behavior.

Let us consider, for instance, the case of a solvated protein that interacts at its active site with a ligand (as proposed in Chapter 3 and Ref. [24]) schematically shown in the Fig. 1.15: in this case, on one hand, the computational cost increases by employing high resolution simulation (atomistic model) of all regions and, on the other hand, the system's properties at the region of interest are largely distorted or cancelled by simulating this part with low resolution (coarse-grained model). The solution to this dilemma is given by the multiscale models in which both atomistic and coarse grained resolutions are concurrently employed. Specifically, in the example reported in Fig. 1.15, the high-resolution treatment is limited only to the protein's active site and the ligand and the surrounding water (red square in the figure), while the rest is modeled at a lower resolution level, sufficient to capture the large-scale structure and thermodynamics.

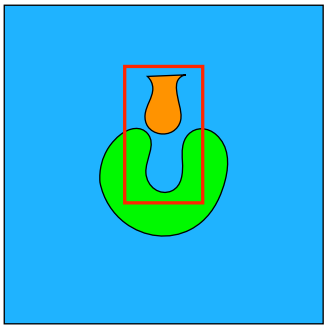


Figure 1.15: Schematic representation of a solvated protein (in green) that interacts at its active site (in green). The box of water is schematically shown with a light blue. The region of interest, namely the active site of the protein in which the ligand-binding and catalytic reactions occur, is shown with a red rectangle.

In general, these multi-resolution approaches are very useful when in a small region of the system the chemical details play a crucial role, such that no simplification of the description is feasible: thus, it requires a high-resolution modeling; the remainder on the other hand, allows a lower resolution treatment.

More generally, the term *multiscale modeling* is widely used to describe a hierarchy of simulation approaches to treat systems across different scales. For a given length and time scale of interest, one picks a method capable of simulating the systems. A common way of graphically representing this approach is a multiscale diagram, as shown in Fig. 1.10 (see Sec. 1.2). As we have already seen, when one moves to larger scales, a coarse-grained model is re-

quired to make simulations feasible.

During the last few years, many methodologies have been developed in order to couple multiple resolution methods. However, linking fully-atomistic simulations with coarse-grained modelling is a challenging process [20, 39, 107, 108], which allows in principle to describe the behavior of the system at multiple scales.

One of the most known concurrent multiple-resolution schemes is the quantum mechanic/molecular mechanic (QM/MM) method [109–112]. It allows a connection between *ab initio* resolution and classical all-atom models. In particular, in a small domain forces acting on atoms are obtained through quantum calculations, while in the rest classical atomistic force fields are employed. Such a scheme is widely used in studying enzymatic chemical reactions [113, 114].

Another class of multi-resolution schemes focuses on the connection between atomistic and CG models simultaneously [115–118]. In practice, this idea lies in a smooth spatial interpolation on the atomistic and CG force field. Several methods have been proposed in the past few years, which can be classified into two main classes: on the one hand, some methods interpolate on the forces acting on the particles; on the other hand, some methods interpolate the interaction potential. In the former category falls a very popular technique known as *Adaptive Resolution Simulation* (AdResS) [115], while in the second class we find the *Hamiltonian Adaptive Resolution Simulation* (H-AdResS) [116] based on a well defined Hamiltonian, as the name suggests.

Both schemes have advantages and disadvantages de-

pending on the application of interest, and none of the two is somewhat better in the absolute sense. In short, when an exact fulfillment of Newton's third law is essential, it could be better to use force-based AdResS, whereas, for all applications that require an Hamiltonian formulation, H-AdResS is preferable. Indeed, both methods have been successfully used in soft matter systems, such as solvated proteins, DNA, macromolecules, and so on [20, 28, 30, 116, 119–122] in comparison with all-atom simulations.

Another class of multiple resolution scheme is known as *Dual Resolution Model*. At difference with the AdResS method, such a model is not adaptive; thus, the resolution is fixed during the simulation. In particular, the region of the system that plays a pivotal role is treated at high-resolution level, while the remainder is coarse-grained, for example as an Elastic Network Model (ENM).

Since part of this research work (e.g. Chapter 2) focuses on the force-based Adaptive Resolution (AdResS) methodology, the latter will be discussed in further detail in the next section, while the Dual Resolution method is described in chapter 3 and 4.

1.6 Adaptive Resolution Simulation (AdResS)

The force-based AdResS methodology was proposed in 2005 by Praprotnik et al. [20, 115]. It allows to simulate a system where two different models, for instance, an all-atom one and a coarse-grained one, are simultaneously employed in different sub-regions of the simulation do-

main. An important feature of this scheme is that particles are allowed to diffuse from one region to the other freely. The atomistic region can have different, but regular geometries, such as a spherical one (Fig. 1.16(a)) or a cuboid one (Fig. 1.16(b)) or a cylindrical one.

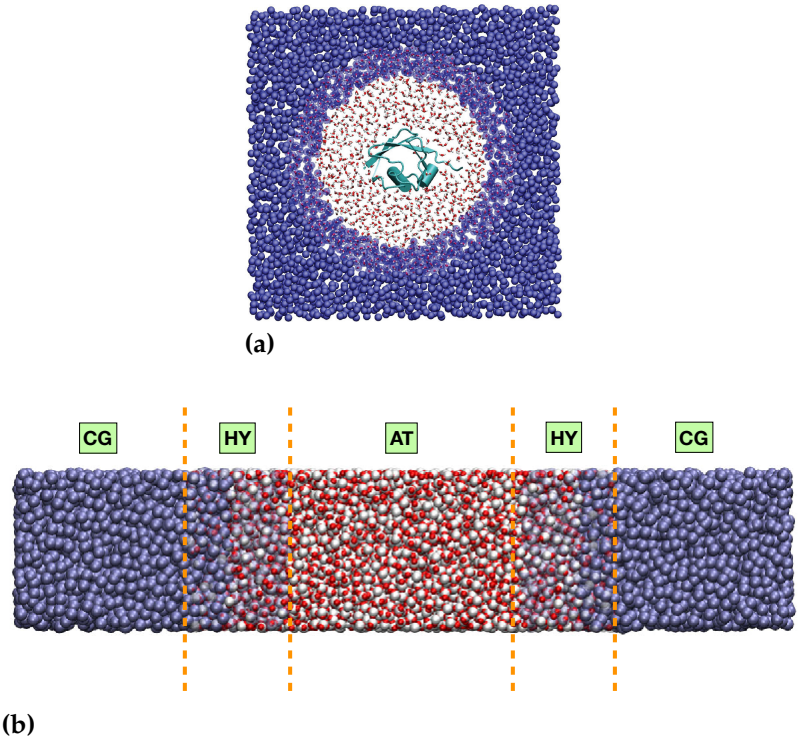


Figure 1.16: AdResS system in case of: (a) spherical atomistic region [119] and (b) cuboid atomistic region [123]. In particular, in both cases, the atomistic region (AT) is shown in red and white, the hybrid one (HY) in red, blue and white, and finally the coarse-grained domain (CG) is colored in blue.

Between the atomistic region (**AT**) and the coarse-grained one (**CG**), a hybrid (or transition) region (**HY**) is employed in which the coupling between different levels of resolution occurs. In particular, the non-bonded force $\mathbf{F}_{\alpha\beta}$ acting between two particles α and β is given by:

$$\mathbf{F}_{\alpha\beta} = \lambda(\mathbf{r}_\alpha)\lambda(\mathbf{r}_\beta)\mathbf{F}^{AT}_{\alpha\beta} + \left[1 - \lambda(\mathbf{r}_\alpha)\lambda(\mathbf{r}_\beta)\right]\mathbf{F}^{CG}_{\alpha\beta} \tag{1.32}$$

where:

$$\mathbf{F}^{AT}_{\alpha\beta} = \sum_{i\in\alpha}\sum_{i\in\beta}\mathbf{F}^{AT}_{ij} \tag{1.33}$$

Here, $\mathbf{F}^{AT}_{ij}$ is the interaction between atoms i and j using the atomistic force-field, while $\mathbf{F}^{CG}_{\alpha\beta}$ is the interaction

between molecules α and β using the coarse-grained force-field. Finally, λ is a transition function varying smoothly and monotonically between 1 and 0. In particular, it assumes the value 1 in the atomistic region and 0 in the coarse-grained domain. Actually, from the Eq. 1.32 it turns out that:

$$\mathbf{F}_{\alpha\beta} = \begin{cases} \mathbf{F}^{AT}_{\alpha\beta}, & \text{if } \lambda = 1 \\ \mathbf{F}^{CG}_{\alpha\beta}, & \text{if } \lambda = 0 \end{cases} \tag{1.34}$$

In the hybrid region λ assumes intermediate values between 0 and 1: the precise shape of λ can vary, but it is essential that it guarantees a *smooth* transition between the force fields. To this end, squared cosine functions are commonly used [115–117, 119, 122, 124]. For instance, in case of spheric atomistic regions [28, 119], λ has the following shape:

$$\lambda(r) = \begin{cases} 1, & r < r_{at} \\ \cos^2\left(\frac{\pi}{2d_{hy}}(r - r_{at})\right), & r_{at} < r < r_{at} + d_{hy} \\ 0, & r_{at} + d_{hy} < r \end{cases} \tag{1.35}$$

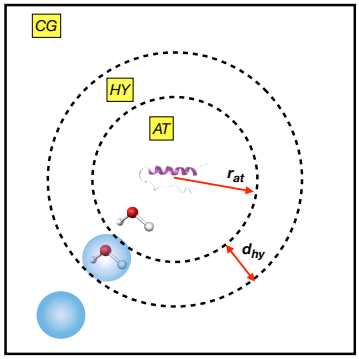


Figure 1.17: Schematic representation of AdResS approach in case of atomistic spherical region. The figure also shows the hybrid and coarse-region. The water molecules, the protein in the center, and the regions are not in scale.

where d_{hy} is the diameter of the hybrid region, whereas r_{at} is the radius of the atomistic part, as schematically shown in the Fig. 1.17

Another important aspect is that the force interpolation scheme described by Eq. 1.32 conserves Newton's third law, but it does not admit a Hamiltonian formulation. In fact, the requirement of having Newton's third law satisfied everywhere in the system is not compatible with an energy interpolation. Therefore, the consequence is that

the energy is not conserved, and excess heat is produced in the transition region. This surplus of energy can be removed by using a thermostat, such as Langevin thermostat, establishing thermal equilibrium [22, 122] everywhere.

Thermodynamic Force

In general, the coarse-grained potential does not reproduce all thermodynamical properties of the atomistic potential, which it is supposed to represent [125–127]. The pressure of the CG and AT potential differs significantly from each other, and it also leads to a non-uniform density profile. This undesirable thermodynamic imbalance can be corrected by using the so-called *Thermodynamic Force* (TF), a compensatory force that is applied within the hybrid region, ensuring a flat density profile along the direction of resolution change [118].

The thermodynamic force can be obtained with an iterative procedure via the following expression:

$$F_{TH}^{i+1} = F_{TH}^{i} - \frac{1}{\rho_0 \kappa_T} \nabla \rho^i(r) \tag{1.36}$$

where ρ_0 is the molecular density reference, κ_T is the system's isothermal compressibility, and $\rho^i(r)$ is the molecular density profile as a function of the position, whose direction is orthogonal to the CG-AT interface. In the beginning, we impose that $F_{TH}^0 = 0$, and the initial density profile is calculated. The protocol converges by construction once the density profile is flat, namely when the condition $\nabla \rho(r) = 0$ is satisfied.

Computational cost in AdResS

From the numerical/computational point of view, the possibility of treating a system with a reduced number of degrees of freedom except where it is strictly necessary making use of AdResS scheme, represents an advantage, since a much smaller number of force calculations are required in the coarse-grained region. This is particularly true for parallel MD codes such as GROMACS [128], where a dynamical decomposition of the simulation box allows one to subdivide the box with a finer grid in the AT and HY region, while a smaller number of processors is assigned to the CG region. For example, for a water system with an AT region covering 1/6 of the total simulation box, simulated with GROMACS on a 16-cores processor, the speed-up is about a factor three. This factor is nonetheless small compared to what can be achieved with other simulation packages, such as ESPResSo++ [129, 130]: in fact, water simulation in GROMACS is extremely optimized, and any modification of the standard code can introduce a bottleneck. Other results obtained with AdResS approach and their relative computational cost can be found in Chapter 2: in the simple test system studied (methanol and 3-methylindole in a cubic box of water), we observed that this method provides a substantial reduction in simulation time with respect to a fully atomistic simulation; the speed-up is about a factor three.

AdResS or not?

As aforementioned, in this scheme, a Hamiltonian formulation is not possible, but it is not a problem: equilibrium and canonical sampling can be enforced by using a

Langevin thermostat. Moreover, the thermodynamics of the system is under control introducing an external field – the thermodynamic force – in the hybrid region to compensate for the density imbalance. Nevertheless, the lack of Hamiltonian can have negative consequences: for instance, microcanonical simulation is not feasible, and no Monte Carlo scheme can be implemented. The solution is using another method called H-AdResS [116] where **H** stands for Hamiltonian. As always, everything has a cost: Newton's third law is satisfied only on average in the hybrid region. However, this is another story [131, 132], whose discussion goes beyond the scope of this introduction.

1.7 Outline

In the following, a brief overview of Chapters 2-5 is given. In particular:

- The second chapter focuses on the force-based Adaptive Resolution Scheme and its use, in combination with Thermodynamic Integration (TI) [133], for the calculation of free energy solvation of amino acid sidechain analogs. All simulations have been performed with the ESPResSo++ package [129, 130].
- The third chapter describes another multi-resolution scheme, in which the CG part is modeled in Elastic Network Model ENM [36, 97–102]. In particular, this scheme has been employed to calculate the binding free energy of egg white lysozyme (HEWL) with the inhibitor di-N-acetylchitotriose. Particular attention is paid in the selection of the atomistic and the coarse-grained part: indeed, the active site is modeled with

different numbers of residues treated all-atom.

- The fourth chapter focuses, once again, on the Dual Resolution scheme proposed in Chapter 3, applied on a small protein called Bovine Pancreatic Polypeptide (or 1BBA in short). The first part has the purpose of computing the free energy landscapes in terms of collective variables that describe the solvated system, comparing atomistic and Dual Resolution simulations. The second part, on the other hand, proposes a further refinement of the ENM part by using different elastic constants between CG beads.
- The fifth chapter illustrates a novel multi-scale resolution scheme dubbed CANVAS or coarse-grained anisotropic network model for variable resolution simulations. The model is implemented in a python script, that generates GROMACS input files. The mapping function of a group of atoms onto a CG site is determined by a Voronoi-like partitioning of the structure. The parametrization of the CG interactions is based on simple averaging rules of the properties of the group of atoms which map on a given CG site. Each survived atom has average properties of the entire block, at which it belongs. The chapter shows the first attempts to simulate the protein Adenylate Kinase by means of this new model in GROMACS [128] with the purpose of characterising the model's performance, advantages, and limits, and to identify possible modifications to improve its accuracy based on these preliminary results.

Using force-based adaptive resolution simulations to calculate solvation free energies of amino acid sidechain analogues

2

This chapter is a research article that has been published in *The Journal of Chemical Physics.*

Raffaele Fiorentini, Kurt Kremer, Raffaello Potestio, Aoife C. Fogarty

Using force-based adaptive resolution simulations to calculate solvation free energies of amino acid sidechain analogues.

The Journal of Chemical Physics **146**, 244113 (2017)

DOI: 10.1063/1.4989486

Published by the *American Institute of Physics.*

The calculation of free energy differences is a crucial step in the characterization and understanding of the physical properties of biological molecules. In the development of efficient methods to compute these quantities, a promising strategy is that of employing a dual-resolution representation of the solvent, specifically using an accurate model in proximity of a molecule of interest and a simplified description elsewhere. One such concurrent multi-resolution simulation method is the Adaptive Resolution Scheme (AdResS), in which particles smoothly change their resolution on-the-fly as they move between different sub-

regions. Before using this approach in the context of free energy calculations, however, it is necessary to make sure that the dual-resolution treatment of the solvent does not cause undesired effects on the computed quantities.

Here, we show how AdResS can be used to calculate solvation free energies of small polar solutes using Thermodynamic Integration (TI). We discuss how the potential-energy-based TI approach combines with the force-based AdResS methodology, in which no global Hamiltonian is defined. The AdResS free energy values agree with those calculated from fully atomistic simulations to within a fraction of k_BT. This is true even for small atomistic regions whose size is on the order of the correlation length, or when the properties of the coarse-grained region are extremely different from those of the atomistic region. These accurate free energy calculations are possible because AdResS allows the sampling of solvation shell configurations which are equivalent to those of fully atomistic simulations.

The results of the present work thus demonstrate the viability of the use of adaptive resolution simulation methods to perform free energy calculations, and pave the way for large-scale applications where a substantial computational gain can be attained.

2.1 Introduction

One of the most challenging applications of computational methods in biochemistry is the accurate calculation of solvation and binding free energies. A prototypical example is provided by *in silico* drug design, where one needs to obtain, by means of computational experi-

ments, quantitative information about the effectiveness of a new molecule or set of molecules in promoting or inhibiting a given enzyme. It is often the case that the number of viable candidates to become usable drugs is too large for experimental screenings, where the complexity of the processes under examination makes it difficult to dissect the observed system properties into its different components.

Computer simulations represent a valuable tool, as they enable the pre-screening of large numbers of different systems and the comprehension of their properties at the molecular and atomic level. This detailed information can prove crucial to identify the most promising molecules, thus allowing experimental research to focus on a reduced subset of case studies.

However, the detailed determination of ligand-enzyme binding free energies still remains a daunting task in most cases, due to the large size of the molecules under examination. In particular, a considerable bottleneck can be the simulation of the solvent, which might represent a substantial fraction of the computational cost.

A promising way of mitigating the computational overhead due to the explicit solvent molecules is to employ concurrent multi-resolution simulation methods. These use a combination of computationally expensive high-resolution potentials and cheaper low-resolution potentials simultaneously in order to facilitate the study of systems in which a large range of time and length-scales play a role. The accurate high-resolution model is used to describe those parts of the system where fine-grained or chemically detailed processes take place, while use of the less expensive

coarse-grained (CG) potential in the rest of the system allows bigger system sizes and longer simulations.

One such multi-resolution method is called adaptive resolution scheme (AdResS) [108], in which the simulation box is divided into atomistic (AT) and coarse-grained regions, with particles [134] smoothly changing their resolution on-the-fly as they move between regions. This resolution change is achieved by the interpolation of energies [116] or forces [115] across a transition region. The AdResS methodology allows a significant reduction in the number of degrees of freedom simulated atomistically, while still reproducing the properties of a sub-region of a fully atomistic simulation [108]. In the past decade, most works using the AdResS approach have concentrated on the study of structural and sometimes dynamical properties, as well as basic thermodynamic quantities such as density, pressure, chemical potential or compressibility [24, 115, 119, 121, 124, 135–137]. So far, less attention has been paid to how well free energies can be computed within an AdResS set-up. Recent explorations of the thermodynamics of AdResS include Refs. [31] and [138]. In particular, Agarwal *et al.* compared chemical potentials calculated as an intrinsic side-product of their Grand Canonical AdResS setup to free energies of solvation calculated in fully atomistic systems [139].

Here we introduce the combination of the force-based AdResS approach and Thermodynamic Integration [133] (TI) to calculate free energies. We obtain solvation free energies of amino acid sidechain analogues in water, a set of classic systems studied notably by Shirts *et al.* [140], and also recently employed in an exploration of a non-adaptive

multi-resolution technique, in which Kuhn *et al.* discussed the influence of density deviations and orientational edge effects on the solvation free energy in that approach [23].

In our AdResS setup, we describe the solute molecule and surrounding solvent molecules using an atomistic potential, while the rest of the system is modelled at a cheaper, coarse-grained level. We explore the influence of atomistic region size, coarse-grained potential and density control on the free energy. We also discuss how the potential-energy-based Thermodynamic Integration approach combines with the force-based AdResS methodology, in which no global Hamiltonian is defined. We show that, because AdResS allows the sampling of atomistic configurations which are equivalent to those of fully atomistic simulations in the equivalent ensemble, we can nevertheless accurately calculate free energy values with this approach.

These results demonstrate that the usage of the force-based AdResS method in tandem with Hamiltonian-based free energy calculations is viable and quantitatively sound. This validation paves the way to large-scale applications involving large macromolecules and, therefore, large amounts of explicit solvent to be treated at dual resolution.

2.2 Methodology

In this work, we calculate the solvation free energy of amino acid sidechain analogues methanol and 3-methylindole (analogues of serine and tryptophan, respectively). These two molecules were chosen because they have significantly different sizes: methanol has a fairly similar size to water and a molar mass of 32.04 g mol^{-1}, while

3-methylindole has a molar mass of 131.18 g mol^{-1}. The radii of gyration, 0.08 nm for methanol and 0.21 nm for 3-methylindole, give an indication of the size difference. Each simulation system used contains one solute molecule in aqueous solution. We perform fully atomistic reference simulations with a range of box sizes, and AdResS simulations with a range of different atomistic region sizes and two different coarse-grained potentials for water. The first coarse-grained potential used is derived via the systematic coarse-graining procedure Iterative Boltzmann Inversion (IBI) [87, 88, 141] to reproduce as closely as possible the atomistic water centre-of-mass structure. In the second case, the coarse-grained region contains a gas of non-interacting particles, i.e., an ideal gas, which can be seen as the most extreme possible coarse-grained "potential" [142].

2.2.1 Adaptive Resolution Scheme and Thermodynamic Integration

In the AdResS methodology (illustrated in Figure 2.1(a)), the simulation box is divided into different regions: the atomistic (AT) region, where non-bonded interactions are modeled using an atomistic force field, and the coarse-grained (CG) region, where a coarse-grained force field is used. Between them is a hybrid (HY) region where particles smoothly change their resolution between atomistic and coarse-grained. In this work, the AT region is a sphere of radius r_{at} centered on an atom with coordinates $\mathbf{r}_{centr}$ at the center of the solute molecule. With this definition, the AT region follows the solute and moves together with it: this strategy, which relies on the translational invariance of the

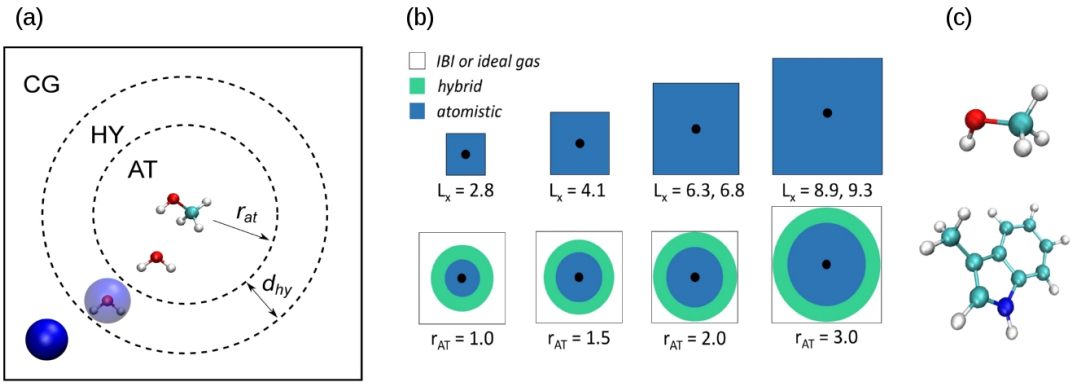


Figure 2.1: (a) Illustration of the AdResS approach, showing the atomistic, hybrid and coarse-grained regions (not to scale) (b) Schematic representation of fully atomistic and AdResS systems used. The methanol or 3-methylindole solute is represented by a black dot in the centre of the box, (c) Up: methanol chemical structure, Down: 3-methylindole chemical structure. Atomistic details are shown in red, blue, cyan and white (O, N, C and H atoms).

uniform solvent with periodic boundary conditions, makes it unnecessary to restrain the molecule in a particular point of the simulation box.

The HY region is then a spherical shell of width d_{hy}, and the remainder of the system is the CG region. Water molecules diffuse freely between regions, changing resolution as a function of their instantaneous position. Particle resolution is described using a function w that varies smoothly and monotonically across the HY region, from a value of 1 in the AT region to 0 in the CG region. For a molecule α whose center of mass $\mathbf{r}_\alpha$ is at a distance $r = |\mathbf{r}_{centr} - \mathbf{r}_\alpha|$ from the center of the AT region, it has the form:

$$w(r) = \begin{cases} 1, & r < r_{at} \\ \cos^2\left(\frac{\pi}{2d_{hy}}(r - r_{at})\right), & r_{at} < r < r_{at} + d_{hy} \\ 0, & r_{at} + d_{hy} < r \end{cases} \tag{2.1}$$

Note that in many works which use the AdResS methodology, the resolution function w is referred to using the

symbol λ. We write w here to avoid confusion with the order parameter λ used in Thermodynamic Integration.

Non-bonded interaction forces are then modeled using a force-interpolation scheme, in which the intermolecular force between the centers of mass of molecules α and β is given by

$$\mathbf{F}_{\alpha\beta} = w(\mathbf{r}_\alpha)w(\mathbf{r}_\beta)\mathbf{F}^{AT}_{\alpha\beta} + [1 - w(\mathbf{r}_\alpha)w(\mathbf{r}_\beta)]\mathbf{F}^{CG}_{\alpha\beta} \tag{2.2}$$

where

$$\mathbf{F}^{AT}_{\alpha\beta} = \sum_{i\in\alpha}\sum_{j\in\beta}\mathbf{F}^{AT}_{ij} \tag{2.3}$$

where $\mathbf{F}^{AT}_{ij}$ is the atomistic non-bonded interaction between atoms i and j, and $\mathbf{F}^{CG}_{\alpha\beta}$ is the coarse-grained non-bonded interaction between molecules α and β. In this scheme, the forces interacting between two atomistic water molecules simplify to $\mathbf{F}^{AT}_{\alpha\beta}$ and between two coarse-grained water molecules to $\mathbf{F}^{CG}_{\alpha\beta}$. Water-water interactions across the resolution boundaries are treated using the interpolation Eq. 2.2. This scheme allows simulations which are momentum-conserving but not energy-conserving. The global system Hamiltonian corresponding to Eq. 2.2 is not defined [143], and a local thermostat must be applied to deal with heat production in the HY region [32, 115].

The dual-resolution treatment of the solvent allows a reduction of the computational cost of the simulations. The systems under examination here have a relatively small size and, most importantly, a relatively large *atomistic-to-total* volume ratio, meaning that the fraction of volume where molecules are treated at the atomistic level is relatively large. Because of this, in the simple test systems studied here, the computational gain is not large, and indeed a

R	Ideal sim speedup	AdResS sim speedup
methanol		
1.0	5.6	3.1
1.5	3.0	2.4
2.0	1.8	1.7
3-methylindole		
1.0	7.1	3.6
1.5	3.7	2.6
2.0	2.3	2.3

Table 2.1: Comparison of the speedup in simulation time provided by AdResS simulations with respect to fully atomistic simulations of the same size (namely, 6.3 nm side for the methanol and 6.8 nm side for 3-methylindole) run on a single core. These data are obtained from 19 ps long runs, in order to minimize the idle time employed in non-run processes (system setup, memory allocation etc.). These speedups are also compared to the ideal ones, defined as the inverse of the fraction of AdResS system volume where the calculation of atomistic forces takes place.

major speedup is not the goal of the present investigation. Nonetheless, we could observe that the AdResS method provided a substantial reduction in simulation time with respect to a fully atomistic simulation. This gain, quantitatively reported in Table 2.1, is defined as the inverse of the ratio of atomistic-to-total volume. The latter is obtained as the volume of a sphere of radius $R = r_{at} + d_{hy}$ divided by the simulation box volume. The obtained speedup is somewhat lower than the corresponding ideal value; however the discrepancy diminishes as the volume where atomistic forces are computed increases. This behavior stems from the approximation on which the definition of ideal AdResS simulation time relies, namely, that the only computational cost is due to the calculation of forces, and that this takes place only in the atomistic and hybrid regions. This assumption willfully neglects surface and finite size effects, hence the deviations for systems with small atomistic regions. For the setups with the smallest atomistic regions, the speedup is between $\simeq 3$ (methanol) and $\simeq 3.6$ (methylindole).

We calculate solvation free energies using the Thermodynamic Integration (TI) method [133]. For any two states A and B in which the solute-solvent interaction differs, we write the solute-solvent interaction potential U_{sw} as a function of an order parameter λ which takes values between 0 and 1, defining a pathway from state A to B. The free energy difference between the states is then given by

$$\Delta G = \int_0^1 \left\langle \frac{dU_{sw}(\lambda, q)}{d\lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{2.4}$$

In practice this is done by discretising λ and sampling $dU_{sw}(\lambda, q)/d\lambda$ for a series of different λ values between 0 and 1.

We now address one perceived possible problem. TI involves derivatives of the potential energy with respect to the parameter λ, while in force-based AdResS no global Hamiltonian is defined [143]. For the calculation of solvation free energies, the energy derivative required is that of the potential energy of the interaction between solute and solvent, since all other energy terms in the system are independent of λ. This is defined in AdResS as long as all atoms in all pairs contributing to $dU_{sw}/d\lambda$ fall within the AT or HY regions. Moreover, the value of $\langle dU_{sw}/d\lambda \rangle_\lambda$ will be the same in the fully atomistic and AdResS systems as long as two conditions are fulfilled: (i) all interaction pairs contributing to U_{sw} fall within the AT region (i.e., the interaction cutoff plus the solute size is less than r_{at}), and (ii) both systems sample the same ensemble of configurations in the atomistic region. We will show below that this is indeed the case.

There also exists a formulation of AdResS (called H-AdResS) based on the interpolation of energies instead of

forces [116]. In this case, a global Hamiltonian is defined and simulations are energy-conserving. We anticipate that this formulation of AdResS can also be used without problems in TI calculations. However, using H-AdResS with moving atomistic regions is inadvisable because the forces in H-AdResS involve a term which is a derivative of the resolution function; in particular cases, that force term could create additional spurious forces on that atom. This would happen if the position of the atomistic region were made mathematically dependent on the instantaneous position of a given atom, which would be necessary to have an isolated Hamiltonian with no external forces. This problem could be circumvented; however it could overshadow the issues specific to the usage of Kirkwood TI in the context of adaptive resolution simulations. Because of these reasons, and since in the long term we are interested in complex applications such as protein-ligand binding which will require the use of moving AT regions, we decided to validate the TI/force-based AdResS combination.

We note also at this point that in H-AdResS and in the auxiliary Hamiltonian approach of Agarwal *et al.* [139], free energies (excess chemical potentials) can be obtained automatically as a by-product of the standard process of system set-up. However, this "by-product" approach applies only in the case of simple interactions between small molecules and can no longer be used for the calculation of free energies in more complex situations such as protein-ligand binding or interactions involving solids.

2.2.2 Thermodynamic Force

In general, coarse-grained potentials cannot necessarily reproduce all thermodynamic properties of the atomistic reference potential which they are intended to represent [125–127]. In this work, we use a CG potential derived via Iterative Boltzmann Inversion, and also simulate a system in which particles in the CG region are modelled using a gas of non-interacting particles. In both cases, the pressure of the CG potential differs significantly from that of the AT potential, and would lead to an undesirable density difference between AT and CG regions. In order to avoid this, we use a thermodynamic force [118] F_T, a compensatory force which is applied within the HY region, ensuring a flat density profile along the direction of resolution change. F_T is generally obtained via an iterative procedure based on the gradient of the density profile along the direction of resolution change [118]. How straightforward it is to obtain this tabulated force depends on factors such as the thermodynamic difference between AT and CG potentials, atomistic region size and geometry, and concentration of different particle types in multicomponent systems [144].

For the current purpose of calculating free energies, a very accurate density is required in the AT region. In order to reach this level of accuracy, even in the most difficult conditions (such as very small spherical atomistic region or large differences between AT and CG potentials in terms of thermodynamical properties such as pressure or compressibility), we have developed an upgraded algorithm to compute the thermodynamic force. Specifically, we include an additional term in the previously established procedure for obtaining F_T in tabulated form, and define the thermo-

dynamic force at iteration $i+1$ as:

$$F_T(r)_{i+1} = F_T(r)_i - \kappa_1 \nabla \rho_i(r) + \kappa_2 (X_{ref} - X_i) \nabla w(r) \qquad (2.5)$$

where $\rho_i(r)$ is the density profile along the direction of resolution change, calculated from a simulation using $F_T(r)_i$. In the newly added term $\kappa_2(X_{ref} - X_i)\nabla w(r)$, $w(r)$ is a function that goes smoothly from 1 to 0 across the region in which F_T is applied. We use the same functional form as given in Equation 2.1 for $w(r)$ which defines the resolution change in AdResS, but this is just for convenience and there is no fundamental theoretical connection between them. The term X_i is a measure of the density throughout the atomistic region. It must be a well-defined value which can be determined with very high accuracy. In the current work, since the atomistic region is centered on an atom, the density profile is equivalent to a radial distribution, and we define X_i as the height of the first-solvation-shell peak. X_{ref} is the corresponding value in the fully atomistic reference system. Other measures of the density are possible and equally valid, for example the average number of particles in the AT region. We found that the measure we used here [height of the first solvation-shell peak in the radial distribution function (RDF)] converged fastest as a function of simulation trajectory length and was therefore easiest to work with.

Finally, κ_1 and κ_2 are prefactors which can be varied to aid convergence. A useful procedure is to start with $\kappa_1 \neq 0$, $\kappa_2 = 0$ and perform many iterations using relatively short, inexpensive simulations in order to rapidly obtain a good approximation of F_T. One can then set $\kappa_1 = 0$, $\kappa_2 \neq 0$ and perform iterations with simulations long enough to de-

termine X with high accuracy, continuing these iterations until the density in the atomistic regions is as close as desired to the reference density. These two steps can then be repeated as necessary.

Since finite-length simulations inevitably yield a density profile containing statistical noise which is then transmitted to the tabulated thermodynamic force, it can be helpful to use some procedure to smoothen the density profile $\rho(r)$, such as replacing each atom (which is a delta function, i.e. is located at one defined point in space) by a triangle or Gaussian function, to smooth out its mass over several bins [123].

We note in passing that it is also possible to obtain via IBI a coarse-grained potential with the same pressure as the atomistic reference [88]; however this is at the cost of having the wrong compressibility in the coarse-grained region. Here, we chose to work with the non-pressure-corrected IBI potential, which has the same compressibility and structure as for the atomistic potential. This provides a strong contrast to the other coarse-grained "potential" we use, the fluid of non-interacting particles, in which the structure, compressibility and pressure all differ from the atomistic reference.

2.2.3 Simulation details

Fully atomistic and AdResS systems containing methanol or 3-methylindole were constructed using the simulation box sizes and atomistic region radii summarised in Table 2.2 and illustrated in Figure 2.1(b). The box sizes range from a little over twice the non-bonded interaction cutoff to almost eight times the cutoff. The amino acid forcefield

used was Amber94 [145] (we note that the non-bonded parameters are the same as in the more recent Amber force-fields, which were mostly focussed on improvements in backbone parameters, not relevant here).

AdResS				fully atomistic	
r_{at}	$\langle \sum w_i \rangle$	≈ box length		≈ box length	# molecules
methanol					
1.0	608	6.3		2.8	694
1.5	1330	6.3		4.1	2189
2.0	2486	6.3		6.3	8212
3.0	5915	8.9		8.9	23399
3-methylindole					
1.5	1330	6.8		6.8	10164
2.0	2486	6.8		-	-
3.0	5915	9.3		-	-

Table 2.2: Simulation box length, atomistic region radius (r_{at}) and number of atomistic or atomistic-like particles in the AdResS and fully atomistic systems used to perform free energy calculations. Distances are given in nm, while w_i is defined in Equation 2.1. As a function of its position in the HY region, each HY particle has a weight between 0 and 1 (the closer it is to the AT region, the bigger w_i is), whereas w_i is 0 in the CG region and 1 in the AT region. $\langle \sum w_i \rangle$ in the second column is the summation of all the weights averaged over the whole trajectory.

The water model used was TIP3P [146]. Side chain analogue force fields were constructed from Amber94 amino acid residue force fields using the procedure of Ref. [140]: the backbone atoms were replaced by a hydrogen of the same atom type and with the same charge as other hydrogen atoms connected to the β-carbon, and the β-carbon charge was adjusted so that the molecule was neutral overall. All other parameters were exactly as in the amino acid residue force field. The IBI coarse-grained potential was obtained using the VOTCA package [147].

In free energy calculations using TI, the alchemical change was performed in two steps: first switching off Coulombic solute-water interactions (ΔG_{Coul}), then Lennard-Jones (ΔG_{LJ}). The Coulomb step had a linear dependence of U_{sw} on λ, while for the Lennard-Jones step we used the soft-core potential of Ref. [148] with parameters $\alpha = 0.5$ and $p = 1.0$ to avoid possible singularities from overlapping atoms

during the alchemical change.

The temperature was kept constant at 298K by a Langevin thermostat with a friction constant γ of 15 ps^{-1}. The non-bonded cutoff was 1.2 nm. The integration time step was 2 fs. Electrostatics were treated using the reaction field method with a dielectric constant $\epsilon = 80$ and a cutoff of 1.2 nm; these parameters provide a good compromise between accuracy and speed, as it was verified in [140]. The SETTLE [149] and RATTLE [150] algorithms for rigid water and rigid bonds to hydrogen were used.

Each system was prepared using fully atomistic minimization with the steepest descent method, 500 ps NPT equilibration and 500 ps NVT equilibration. All free energy calculations used 21 λ values per ΔG_{Coul} value and 40 equidistant λ values (with a separation of 0.025) per ΔG_{LJ}, with 1 ns of simulation per λ value, of which the first 100 ps were discarded as equilibration. Free energy calculations were performed in the NVT ensemble throughout, i.e. we approximate the Gibbs by the Helmholtz free energy, after initially verifying that the difference is negligible for systems at these concentrations. This approximation was validated through the comparison of system density with and without the solute molecule. The change in concentration is in fact on the order of 0.01 – 0.1%, thus suggesting that the amount of solvent in the system is sufficiently large to absorb the effective volume change due to the decoupling from the solute. Finally, production runs for studying system properties with full solute-solvent interaction were 6 ns long each. All AdResS and most fully atomistic simulations used the ESPResSo++ simulation package [129], in which we have implemented TI. Some preliminary fully

atomistic equilibration simulations used the GROMACS simulation package [128].

All error bars shown were calculated using the Student t distribution [151] at the 95% confidence limit, via standard deviations obtained using block averaging in which all trajectories were divided into five blocks of equal length.

2.3 Results

Figure 2.2 shows the solvation free energy values for methanol and 3-methylindole, comparing fully atomistic systems with different simulation box sizes to AdResS systems with different atomistic region sizes and different CG potentials. The systems are those visualised in Figure 2.1(b). The Coulomb (Figure 2.2(a),(c)) and Lennard-Jones (Figure 2.2(b),(d)) contributions to the free energy are plotted as a function of the number of atomistic or atomistic-like molecules in the system. For fully atomistic systems this is simply the total number of molecules. For AdResS systems this is the sum of the w values as defined in Equation 2.1, i.e., each fully atomistic molecule contributes 1 to the sum, each fully coarse-grained molecule contributes 0, and water molecules in the hybrid region contribute in accordance with their degree of atomistic character.

In Figures 2.2(a) and (b), for methanol, the four fully atomistic values (black diamonds) correspond to the four different simulation box sizes. Previous studies have shown that in fully atomistic systems there are no detectable finite-size effects for solvation free energies of neutral solutes as a function of simulation box size [152], and we make the same observation here. The four values for the AdResS

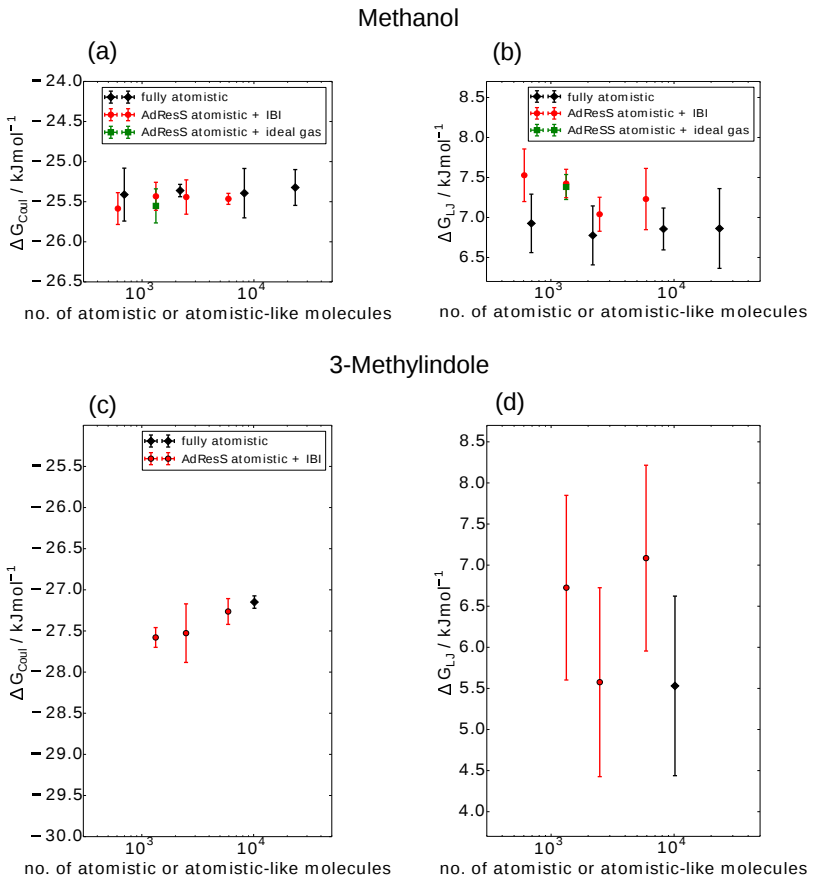


Figure 2.2: (a),(c) Coulomb and (b),(d) LJ contributions to the free energies of solvation for (a),(b) methanol and (c),(d) 3-methylindole, fully atomistic versus AdResS with an IBI CG potential and with an ideal gas CG region. (a) and (b) subplots are plotted such that the y-axis covers a range of 2.5 kJ mol^{-1}, or approximately k_BT, instead (c) and (d) are plotted so that y-axis covers a range of 5.0 kJ mol^{-1}. Note that the x-axes use a logarithmic scale. The quantity plotted on the x-axis is defined in the text. The color legend in (d) is the same as in (c).

systems using the IBI coarse-grained potential (red circles) correspond to the four different atomistic region sizes (r_{at} = 1.0 to 3.0 nm), while the value for the AdResS system using a coarse-grained reservoir of non-interacting particles (blue square) has an atomistic region with r_{at} = 1.5 nm. In all cases, the AdResS free energy values agree with the fully atomistic reference to within at least 0.6 kJ mol^{-1}, or $0.2k_BT$ at 298 K, k_B being the Boltzmann constant and T the temperature. This is the case even when the radius of the atomistic region r_{at} is 1.0 nm, somewhat less than the non-bonded interaction cutoff 1.2 nm, and some of the water molecules contributing to $dU_{sw}/d\lambda$ fall within the HY region. The use of such a small atomistic region is possible because the interpolation-based AdResS approach creates a smooth transition from AT to CG regions. The water molecules within the HY region close to the AT region

have w values close to 1.0 (Equation 2.1), and therefore considerable atomistic character and atomistic-like properties. Nevertheless, in practice and taking into account the non-bonded cutoff, a prudent choice for the minimum atomistic region would be closer to 1.5 nm, or the solute radius of gyration, R_g (0.08 nm for methanol), plus a 1.2 to 1.4 nm thick layer of atomistic water. This is in accordance with the rule of thumb we suggested previously based on a consideration of structural and dynamical properties, which was (R_g + 1.3) nm [119].

Similarly, for 3-methylindole (Figures 2.2(c) and (d)), fully atomistic and AdResS free energy values agree to within at least 1.5 kJ mol^{-1}, or 0.6 k_BT. The three values for the AdResS systems using the IBI coarse-grained potential (red circles) have atomistic region sizes r_{at} = 1.5, 2.0, 3.0 nm. Of course the minimum advisable AT region size is bigger for 3-methylindole (radius of gyration = 0.21 nm) than for the smaller molecule, methanol. Error bars are larger for 3-methylindole than for methanol because the solvation shell of the larger molecule has a more complex configurational space. Moreover, error bars are larger for Lennard-Jones than for Coulomb contributions to the free energy because the linear dependence of the Coulomb energy on λ produces a smoother, more easily integrated curve than the non-linear softcore potential used for the Lennard-Jones alchemical step.

Finally, Table 2.3 summarizes the comparison between experimental solvation free energy values and those calculated in this work. We note that simulated solvation free energy values for these amino acid sidechain analogue systems are known to differ by roughly 1 kcal mol^{-1} or 4

	methanol	3-methylindole
experimental [140, 153]	-21.1 to -21.5	-24.6
fully atomistic, this work	-18.5 to -18.6	-21.6
AdResS + IBI, this work	-18.0 to -18.4	-20.2 to -22.0
AdResS + ideal gas, this work	-18.1	-

Table 2.3: Experimental solvation free energy values in kJ mol^{-1} compared to total solvation free energies ($\Delta G = \Delta G_{Coul} + \Delta G_{LJ}$) calculated in this work. In the second and third row of the table we report the values for methanol and 3-methylindole obtained in fully atomistic reference simulations and in adaptive resolution simulations with IBI CG potential, respectively. The last row, shows the value of ΔG for methanol obtained with ideal gas CG potential. It is useful to point out that the latter model has in fact been employed only in the case of methanol for testing purposes and therefore the value of total free energy solvation for AdResS with ideal gas CG potential for 3-methylindole was not computed.

kJ mol^{-1} from experimental values [140], something we also see here. Simulated free energy values also depend sensitively on the force field chosen and the method used to treat non-bonded interactions [140]. We stress that our main goal here is the comparison of AdResS free energy values to the equivalent fully atomistic reference, for a given force field and set of simulation parameters, and that for this comparison the differences are within the statistical error bars of the simulations, and well below $k_B T$.

The AdResS approach can yield such accurate free energy values relative to the fully atomistic reference because in the atomistic region the AdResS simulations sample configurations from the same ensemble as the fully atomistic simulations. We now examine some of the structural and thermodynamic properties of the atomistic region in the methanol system.

Fig. 2.3 shows the radial distribution functions (RDF) of water oxygen atoms around selected solute heavy atoms, comparing the various AdResS systems to the fully atomistic reference. In every case, the structure of the solute's solvation shell is perfectly reproduced in the AdResS systems, as has been shown before for a variety of other

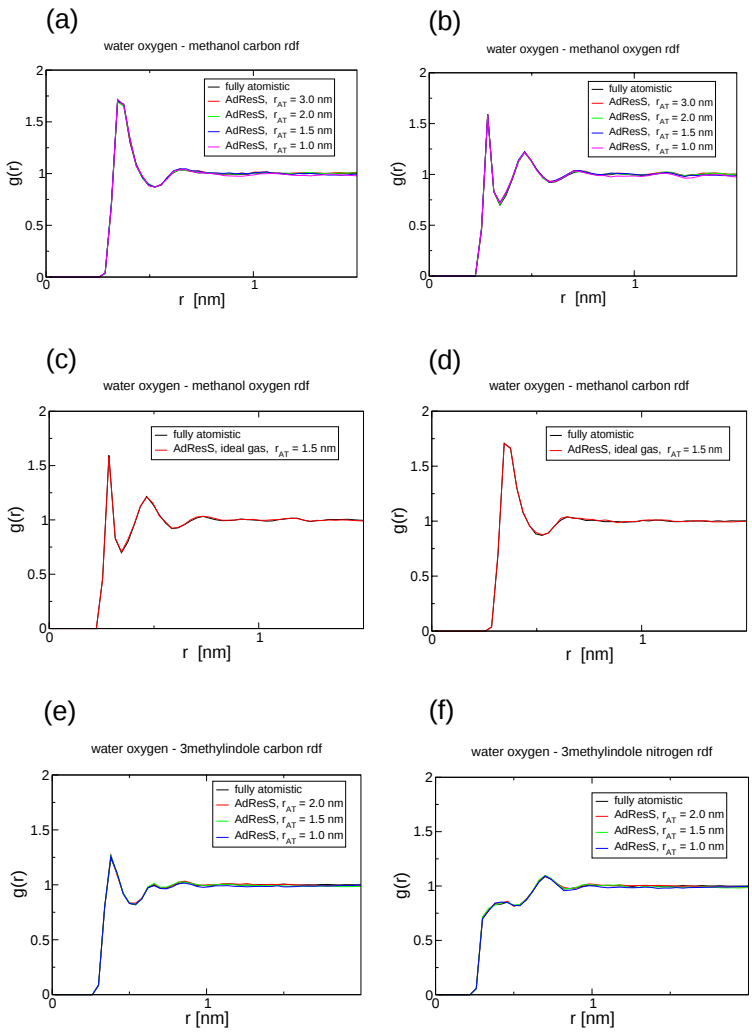


Figure 2.3: Radial distribution functions between water oxygen atoms and selected solute heavy atoms, compared to the fully atomistic reference: (a) and (b) methanol, fully atomistic versus AdResS with IBI coarse-grained potential, (c) and (d) methanol, fully atomistic versus AdResS with ideal gas coarse-grained region, (e) and (f) 3-methylindole, fully atomistic versus AdResS with IBI coarse-grained potential: in particular in figure (e) we used the carbon atom with sp3 hybridization as solute heavy atom.

solutes [30, 119, 121]. In Fig. 2.4 we plot the molecular fluctuations $\left(\langle N^2\rangle - \langle N\rangle^2\right)/\langle N\rangle$ along the direction of resolution change, where N is the instantaneous number of particles in a given bin, and all bins have the same surface-to-volume ratio. The molecular fluctuations are proportional to the compressibility. Figures 2.4(a) and (c) show the AdResS systems using the IBI CG potential for methanol and 3-methylindole respectively, which is parametrised to have the same structure, and therefore the same compressibility, i.e., the same molecular fluctuations, as the atomistic reference. In these AdResS systems, therefore, the molecular fluctuations across the entire system including atomistic and coarse-grained regions correspond to those

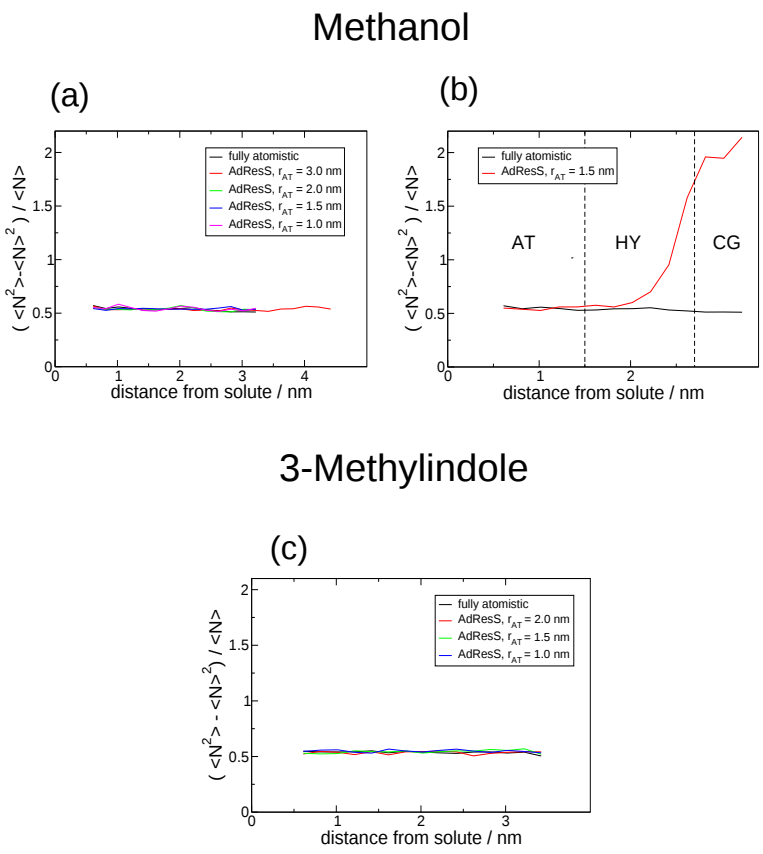


Figure 2.4: Molecular fluctuations in spherical concentric bins of equal surface-to-volume ratio, as a function of distance from the solute atom defining the center of the atomistic region, (a) and (c) AdResS with IBI coarse-grained potential for methanol and 3-methylindole respectively, (b) AdResS with ideal gas coarse-grained region for methanol.

measured in the fully atomistic system. More striking is the case shown in Figure 2.4(b) for the system where the coarse-grained region contains a fluid of non-interacting particles (ideal gas) only for methanol. The molecular fluctuations there are considerably larger than in the atomistic model and the coarse-grained fluid is completely structureless. Nevertheless, even in this extreme case the properties of the atomistic region remain unperturbed and hence the atomistic solvation free energy values are still reproduced in this system.

2.4 Conclusions

We have shown how the force-based adaptive resolution approach can be used to calculate solvation free energy values, even when using a coarse-grained region or reservoir

with extremely different thermodynamic properties. The free energy values obtained in the AdResS setup are accurate to within a fraction of $k_B T$ compared to fully atomistic reference values. These calculations highlight one of the strengths of the AdResS approach, in that it allows accurate control of the atomistic region density, and a smooth transition between atomistic and coarse-grained regions, with no perturbation of the structural and thermodynamic properties of the solute and its solvation shell even for atomistic regions whose size is on the order of the correlation length. We also discussed how the energy derivative is defined in the case of a system with no global Hamiltonian.

The speed-up obtained via the AdResS approach compared to fully atomistic simulations depends on the ratio of the coarse-grained and atomistic region volumes, and the relative computational cost of atomistic and coarse-grained potentials. In this work we studied relatively small systems where the atomistic region occupies a large proportion of the total simulation box, and where fully atomistic simulations are also feasible. Studying these small systems allowed us to validate the AdResS approach via comparison to fully atomistic reference values. Our long-term goal is the calculation of free energies in large, complex systems where fully atomistic simulations are unfeasible because of system size or indeed because not all system components have been characterised to within atomistic resolution [24]. This includes, for example, ligand binding processes in high-molecular-weight proteins, ligand intercalation in DNA, or small molecule-surface interactions. In such systems, the AdResS approach can be used to simulate at an atomistic level only those solvent molecules

in the vicinity of the process of interest, thus significantly reducing the number of atomistic degrees of freedom in the system. The current work forms the basis for such calculations.

2.5 Acknowledgements

K.K., R.F. and A.C.F. acknowledge research funding through the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 340906-MOLPROCOMP. We thank the John von Neumann Institute for Computing at the Jülich Supercomputing Center for allocating computer time on JURECA. We are grateful to Tristan Bereau and Maziar Heidari for a critical reading of the manuscript.

Ligand-protein interactions in lysozyme investigated through a dual-resolution model

3

This chapter is a research paper that has been submitted in *Proteins: Structure, Function and Bioinformatics* during the submission of the thesis.

Raffaele Fiorentini, Kurt Kremer, and Raffaello Potestio.

Ligand-protein interactions in lysozyme investigated through a dual-resolution model.

A fully atomistic modelling of biological macromolecules at relevant length- and time-scales is often cumbersome or not even desirable, both in terms of computational effort required and *a posteriori* analysis. This difficulty can be overcome with the use of multi-resolution models, in which different regions of the same system are concurrently described at different levels of detail. In enzymes, computationally expensive atomistic detail is crucial in the modelling of the active site in order to capture e.g. the chemically subtle process of ligand binding. In contrast, important yet more collective properties of the remainder of the protein can be reproduced with a coarser

description. In the present work, we demonstrate the effectiveness of this approach through the calculation of the binding free energy of hen egg white lysozyme (HEWL) with the inhibitor di-N-acetylchitotriose. Particular attention is posed to the impact of the mapping, i.e. the selection of atomistic and coarse-grained residues, on the binding free energy. It is shown that, in spite of small variations of the binding free energy with respect to the active site resolution, the separate contributions coming from different energetic terms (such as electrostatic and van der Waals interactions) manifest a stronger dependence on the mapping, thus pointing to the existence of an optimal level of intermediate resolution.

3.1 Introduction

One of the most relevant challenges of computational biochemistry and biophysics is the accurate calculation of binding free energies [154–156], which represents one of the key steps in the identification of pharmacological targets as well as in the development of new drugs [157–159]. However, the large sizes of the molecules under examination (often above the hundred of residues), as well as the necessity to screen through large datasets of potential candidate molecules, make this effort onerous in terms of time and computational resources.

A promising way to mitigate these limitations is the use of multiple-resolution models of the protein, that is, representations in which different parts of the molecule are concurrently described at different levels of resolution [24–27, 29, 108, 160–162]. The chemically relevant part of

the protein, e.g. the active site, is modelled at the highest level of detail, typically atomistic. For the remainder, on the contrary, a simplified representation is used, where several atoms are lumped together in effective interaction sites. The working hypothesis underlying these methods is that only a relatively small part of the molecule requires an explicitly atomistic treatment; the remainder, in fact, is mainly responsible for large-scale, collective fluctuations whose function-oriented role is well recognised and prominent [15, 24, 163–165], however also prone to be accurately reproduced by lower-resolution representations [36, 166–170]. Hence, the resulting model favourably joins the accuracy of an atomistic (AT) description where needed and the computational efficiency of a coarse-grained (CG) one where possible.

In order to take full advantage of the dual-resolution approach to protein modelling, though, one has to solve a few key open issues: first, the definition of the appropriate coarse-grained model to employ in the low-resolution part [106, 170–177]; second, the coupling between high- and low-resolution models, which has to be performed so as to guarantee that the appropriate observables are reproduced with respect to the reference provided for example by a fully atomistic simulation. This issue entails a further one, namely the identification of the correct observables apt to quantify the fidelity with which the behaviour of the system is reproduced by the dual-resolution model. Third, the selection of the subpart of the molecule that *requires* a high-resolution modelling. In the present work we will focus specifically on this third aspect.

Various methods and approaches have been developed

in the past few years to describe proteins in dual resolution [25–27, 160–162]. In general, the high-resolution part is modelled at the all-atom level, making use of one of the several atomistic force fields available. The coarse-grained representations employed, on the other hand, range from simple bead-spring elastic networks [24, 36, 168] to more sophisticated Gō-type models [160]. Recently, we have proposed a dual-resolution model [24] where, in the CG part, only the C_α carbons of the protein chain are retained and connected one with the other by harmonic bonds. This model has been employed in the present work with the aim of assessing the accuracy of a hybrid atomistic/coarse-grained description of a protein for binding free energy calculations. The system under examination is hen egg-white lysozyme in explicit water, bound to a sugar substrate, di-N-acetylchitotriose. We carried out calculations of the binding free energy of the ligand in the active site, with a twofold objective. In fact, not only we aimed at verifying that the computed quantity in the dual-resolution model matches a reference, all-atom calculation; but rather we also investigated the impact of different choices in the definition of the high-resolution subdomain. This aspect bears the highest prominence, as it is becoming increasingly more evident that a crucial component in the construction of accurate and effective low-resolution models for biological and soft matter systems is represented by the mapping [24, 106, 177], that is, the particular selection of collective variables employed to describe the system. Here, we provide novel evidence of this general property in the context of a dual-resolution model of a biomolecule, and describe a transferable strategy to tackle this issue.

3.2 Methods

The system under examination in the present work is hen egg-white lysozyme (HEWL) in aqueous solution. In this model, the binding site of the enzyme and the substrate molecule, the inhibitor di-N-acetylchitotriose, are represented with atomistic detail. The protein model employed is not adaptive, that is, the resolution of a given residue is fixed – either atomistic or coarse-grained – and does not change throughout a simulation. However, at difference with other works [29, 108, 160], several values of the number of protein residues treated at high resolution have been explored and employed in independent calculations. The impact of choosing different numbers of active site residues to model at the atomistic level is a central aspect of this study. The coarse-grained model employed to describe the low-resolution part of the protein is a simple bead-spring representation where the selected sites (namely the C_α atoms) are connected by elastic bonds penalising the deviations from the distances that interacting atoms have in the reference conformation. Two values of elastic constants employed, one for C_α's along the chain, and one for all other bonds. Water molecules are described in atomistic detail throughout the whole simulation box: the interaction with the high-resolution part of the protein takes place through the standard all-atom force field, while the interaction with the coarse-grained beads is mediated by a purely repulsive potential acting on the sole oxygen atom.

Hereafter we provide a detailed description of the model. We first discuss the calculation of the binding free energy ΔG_{bind}, then we outline the dual-resolution model and its

coupling to the atomistic part, and finally report information about the simulation setup. Further details are made available in the *Supporting Information*.

3.2.1 Binding Free Energy calculation

One of the key points of this work is the calculation of the protein-ligand binding free energy ΔG_{bind}, which quantifies the affinity of a molecule towards a protein [154–156]. As such, it plays a prominent role in the investigation of the biochemical function and activity of enzymes and similar biomolecules, and in the development of effective drugs.

ΔG_{bind} is defined as the difference between the free energy of the system in the configuration in which the ligand is bound to the active site (G_b) and the corresponding value when the ligand is absent (G_{ub}):

$$\Delta G_{bind} = G_b - G_{ub} \tag{3.1}$$

This value, in the specific case under examination, changes according to the number of active site residues modelled with atomistic resolution, as we will see in Sect. 3.3.

The free energy difference between two states is here computed by means of thermodynamic integration (TI) [133]. Specifically, a scalar $\lambda \in [0,1]$ is defined which parametrises the potential energy of the system as

$$U_\lambda(\mathbf{r}) = \lambda U_A(\mathbf{r}) + (1-\lambda) U_B(\mathbf{r})$$

connecting the states A and B. The sought quantity is given by:

$$\Delta G = \int_0^1 \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{3.2}$$

Since the free energy is a state function, the nature of the path is unimportant, and one can choose a thermodynamic cycle that connects the bound and unbound states through several intermediate ones, as illustrated in Fig. 3.1. In particular, we can identify two main terms: the insertion of the ligand from vacuum to water ΔG_{lig}, and the decoupling from the protein ΔG_{compl}. A further step is the removal of the restraints that keep the ligand in proximity of the protein during the damping of the ligand-protein interactions, ΔG_{r_off}; this latter calculation can be carried out analytically without the need to run simulations. Hence, ΔG_{bind} is the algebraic sum of the previous three terms:

$$\Delta G_{bind} = \Delta G_{compl} + \Delta G_{lig} + \Delta G_{r\_off} \tag{3.3}$$

According to the previous definitions of each term, neither ΔG_{lig} nor ΔG_{r_off} changes with the protein resolution: indeed, the former corresponds to the solvation free energy of the ligand, which is always treated at the atomistic level; likewise, the calculation of the restraint removal free energy is analytic [156]. The unique term that varies depending on the number of active site residues modelled in high resolution is the free energy change of the protein-ligand complex between the bound state and the state where the ligand is removed, that is, the variation of ΔG_{bind} is equal to the variation of ΔG_{compl}.

The alchemical change in the calculation of ΔG_{compl} is performed in three steps (in the following, the subscripts c and ℓ stand for complex and ligand, respectively). First, one adds a set of restraints between protein and ligand

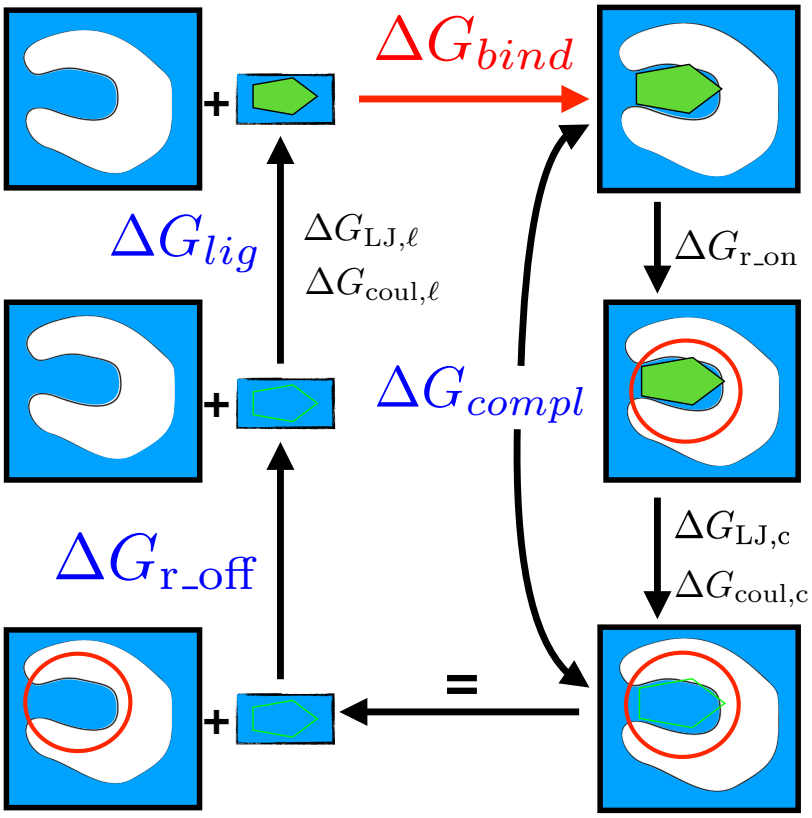


Figure 3.1: pictorial representation of thermodynamic cycle. Starting from the top-right corner of the figure, we decouple the ligand from the protein (ΔG_{compl}, which also includes a set of restraints between ligand and protein) and subsequently introduce it in water (ΔG_{lig}). A further step is the restraints removal (ΔG_{r_off}) whose calculation is analytical.

(ΔG_{r_on}) in order to avoid the problem of the ligand leaving the binding pocket when interactions are being removed. The presence of restraints is indicated in the cycle scheme of Fig. 3.1 with a red circle: it represents the fact that the ligand is confined in a certain volume. For this work we use the set of restraints described by Boresch [156]. Second, Coulomb interactions are switched off ($\Delta G_{coul,c}$); third, the Lennard-Jones potentials modelling van der Waals interactions are removed ($\Delta G_{LJ,c}$). Likewise, the alchemical change in the ligand free energy ΔG_{lig} is performed in two steps: first switching on Coulomb interaction ($\Delta G_{coul,\ell}$), and then Lennard-Jones ($\Delta G_{LJ,\ell}$). The last contribution to the binding free energy, ΔG_{r_off}, derives from restraint removal: its calculation is analytical and therefore it does not require alchemical changes. These transformations are summarised in Fig. 3.1 and Tab. 3.1. Further details can be

found in the *Supporting Information* in the section relative to the thermodynamic cycle.

	alchemical changes	**prot. res. dependence**
$\Delta\mathbf{G_{compl}}$	$\Delta G_{\mathrm{coul,c}} + \Delta G_{\mathrm{LJ,c}} + \Delta G_{\mathrm{r\_on}}$	YES
$\Delta\mathbf{G_{lig}}$	$\Delta G_{\mathrm{coul},\ell} + \Delta G_{\mathrm{LJ},\ell}$	NO
$\Delta\mathbf{G_{r\_off}}$	Analytical	NO

Table 3.1: Summary of the alchemical changes and the protein resolution dependence for each contribute of Binding free energy ΔG_{bind}.

The calculation of ΔG_{compl} can be carried out in two different ways, namely decoupling and annihilation. Decoupling refers to turning off the interaction between the molecule and its environment, while maintaining the potentials among atoms constituting the molecule; annihilation, on the other hand, implies turning off the interaction between the molecule and the environment *as well as* the intramolecular interaction. Here we consider the values of ΔG obtained through ligand decoupling, since this process is more intuitive with respect to annihilation; furthermore, the ligand is always treated at fully atomistic detail, therefore it is not involved in the change of free energy while varying the protein resolution. In Tab. 3.3 and Fig. 3.6 (and with greater detail in the *Supporting Information*, annihilation section) we provide data showing that the values of binding free energy obtained using decoupling and annihilation are consistent within the error bars.

3.2.2 Dual-Resolution protein model

In this work the solvent is treated with all-atom detail, while the protein has a fixed (i.e. position- and time-independent) dual-resolution. The binding site is modelled with atomistic resolution, whereas the rest of the protein is coarse-grained. To describe the lower-resolution part

we employ an elastic network model (ENM) [24, 36], in which each residue is mapped onto a bead whose position corresponds to the C_α atom in the atomistic description. These beads are connected by harmonic springs as shown in Fig. 3.2.

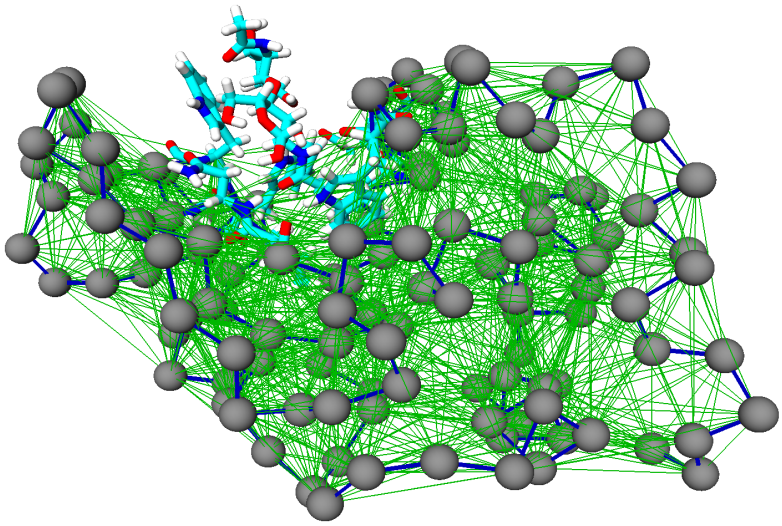


Figure 3.2: Visualisation of the dual-resolution protein. The residues included in atomistic detail are shown in red, blue, cyan and white (O, N, C and H atoms). The grey spheres are ENM nodes, the stiff backbone springs are shown as dark blue lines and all others (weaker) springs are shown in green. Adapted from [24].

The potential energy is given by:

$$E = \sum_i \sum_j k_{ij} \left(r_{ij} - r_{ij}^0\right)^2 \theta(r_c - r_{ij}^0) \tag{3.4}$$

with spring constants k_{ij}, equilibrium distance r_{ij}^0, a cut-off distance r_c, i and j are the node index, and $\theta(r)$ is a Heaviside theta function taking value 1 if $r > 0$ and 0 otherwise. In this model we made use of two different elastic constants: a very stiff spring (k_b) for consecutive beads, represented in blue in Fig. 3.2; and a weaker spring k_{nb} for not consecutive beads whose distance in the reference (native) conformation lies below a fixed cutoff (in green).

The ENM used here is parametrised to reproduce the conformational fluctuations of the reference all-atom model, these being quantified by the root mean square fluctuations (RMSF) of the all C_α atoms of the system [24]. The residues

in direct contact (H-bonding or hydrophobic contact) with the substrate are modelled with all-atom detail; in order to select the other binding site residues to be described at the atomistic level, we sorted them by increasing distance of their the center of mass from the closest ligand atom.

The water-CG protein interaction consists in a simple excluded volume, modelled *via* a Weeks-Chandler-Anderson (WCA) potential [178]. The details about the procedure followed to determine the ENM elastic constants and the excluded volume interaction are provided in the *Supporting Information*, while the numerical values of the resulting parameters are reported hereafter.

3.2.3 Simulation details

The reference model is given by the 2 ns equilibrated PDB structure 1HEW in the NPT ensamble (the Parrinello-Rahman barostat [179] with a time constant of 2.0 ps and 1 bar was used). Both fully atomistic and dual-resolution models of HEWL are solvated in water and placed in a cubic simulation box of 7.06 nm side. The force field employed is Amber99SB [77], whereas the water model is TIP3P [146]. The inhibitor, which was always atomistic, had GLYCAM forcefield parameters consistent with Amber99SB [180]. The TI binding free energy calculation consists of 3 different steps: ΔG_{compl}, ΔG_{r_off}, ΔG_{lig}:

- The protein-ligand complex free energy (ΔG_{compl}) calculation uses 11 λ values per $\Delta G_{restr\_on,c}$, 5 evenly spaced λ values per $\Delta G_{LJ,c}$ (with separation 0.20) and 15 λ values per $\Delta G_{coul,c}$, with 600 ps of simulation per λ in the fully atomistic case, and 4000 ps in the dual-resolution case to improve the statistics.

- The restraint removal free energy (ΔG_{r_off}) calculation is analytical (details on *Supporting Information*).
- The ligand solvation free energy (ΔG_{lig}) calculation uses 5 evenly spaced λ values per $\Delta G_{coul,\ell}$ (with separation 0.20) and 16 λ values per $\Delta G_{LJ,\ell}$, with 600 ps of simulation of each λ-value.

In the thermodynamic integration we employ the soft-core potential of Ref. [148] with parameters $\alpha = 0.5$ and $p = 1.0$ to avoid possible singularities in the Lennard-Jones terms from atoms overlapping during the alchemical change. The temperature is kept constant at 298 K by means of a Langevin thermostat with a friction constant $\gamma = 15\ ps^{-1}$. The integration step is 1 fs. The calculation of electrostatic interaction is performed using the reaction field method with a dielectric constant $\epsilon = 80$ and a cut-off of 1.2 nm. These parameters are a good compromise between speed and accuracy, as verified in Ref. [140]. The SETTLE [149] and RATTLE [150] algorithms for rigid water and rigid bonds to hydrogen have been used. Each system is prepared using fully atomistic minimisation with steepest descent and 6 ns of equilibration in NVT (for both ligand-free and ligand-bound systems). All simulations (both fully atomistic and dual-resolution) are carried out with the ESPResSo++ simulation package [129, 130], in which we have implemented TI (except in case of annihilation, for which all steps are performed in both ESPResSo++ and GROMACS [128]). Some preliminary fully atomistic equilibration simulations use GROMACS. The error bars shown are calculated using the Student t at 95% confidence limit [151], via standard deviations obtained using block averaging in which all trajectories are divided into four

blocks of equal length.

In the dual-resolution model the spring constant between consecutive C_α nodes along the backbone (k_b) has a stiff value of $5 \cdot 10^4$ kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}, whilst all the other ones (k_{nb}) have a value of 160 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}, until 1.2 nm as cutoff, parametrised by minimising the average root mean square error in the C_α RMSF. Moreover, a WCA interaction is applied between C_α nodes and all solvent molecules center of mass. In the WCA potential, ϵ has a value of 0.34 kJ $\cdot$ mol^{-1} arbitrary chosen as the value for carbon in the atomistic forcefield, and $\sigma_i = R_{g,i} \cdot c$ where $R_{g,i}$ is the radius of gyration of a given residue i where c is the same for all amino acids. The value of c is tuned to give the correct bulk water density of reference for a protein-water system. The c value found is 0.658. Further explanations about c can be found in the *Supporting Information*.

3.3 Results and discussion

We performed the calculation of ΔG_b of lysozyme modelled in dual-resolution, varying the number of atomistic residues constituting the binding site and comparing the results with a fully atomistic reference simulation. Recall that the binding free energy calculation consists of three steps: restraint removal, ligand ΔG, and ligand-complex ΔG; of these, only the latter depends on protein resolution, that is, only ΔG_{compl} assumes different values for different numbers of active site residues described at the all-atom level.

As explained in the previous section, the contribution coming from the restraints can be analytically computed

at res	$\Delta G_{Coul,c}$	$\Delta G_{LJ,c}$	$\Delta G_{Restr\_on,c}$	ΔG_{compl}
fully-at	145.2 ± 3.5	44.2 ± 5.2	3.6 ± 0.4	193.0 ± 9.1
aa-3	125.5 ± 7.0	50.4 ± 6.3	8.3 ± 1.1	184.2 ± 14.4
aa-4	141.4 ± 4.9	39.7 ± 9.4	7.2 ± 1.0	188.3 ± 15.3
aa-5	140.2 ± 2.8	48.7 ± 4.5	7.5 ± 1.2	196.4 ± 8.5
aa-6	147.0 ± 1.9	41.7 ± 5.4	5.1 ± 0.5	193.8 ± 7.8
aa-7	144.5 ± 0.8	38.4 ± 3.8	5.0 ± 0.2	187.9 ± 4.8
aa-8	148.0 ± 1.4	33.6 ± 1.9	6.4 ± 1.8	188.0 ± 5.1
aa-9	143.4 ± 4.7	38.1 ± 5.3	5.1 ± 0.3	186.6 ± 10.3
aa-10	145.9 ± 2.2	38.2 ± 1.0	4.4 ± 0.3	188.5 ± 3.5

Table 3.2: In this table are reported the resulting values of free energy of Complex Free Energy (4th column) and its components (Coulomb, Lennard Jones and Restraints respectively in the first three columns) in fully atomistic system and varying the number of atomistic residues. All the values are in kJ · mol^{-1} and performed with Thermodynamic Integration. Moreover, all simulations are carried out in ESPResSo++. In particular, for each value of λ, the dual-resolution simulations with different number of atomistic residues last 4 nsec; the atomistic simulation, instead, lasts 0.6 ns (600 ps).

and amounts to $\Delta G_{r\_off} = -31.3$ kJ · mol^{-1}. Likewise, the Coulomb and Lennard-Jones contributions to the ligand free energy ΔG_{lig} are the following:

$$\Delta G_{coul,\ell} = -142.8 \pm 1.7 \ \text{kJ} \cdot \text{mol}^{-1}$$
$$\Delta G_{LJ,\ell} = -9.1 \pm 6.3 \ \text{kJ} \cdot \text{mol}^{-1}$$

Hence:

$$\Delta G_{lig} = -151.9 \pm 8.0 \quad \text{kJ} \cdot \text{mol}^{-1}$$

The final step is the calculation of ΔG_{compl}, whose results, including the comparison between dual-resolution model and fully atomistic reference, are shown in Tab. 3.2 and illustrated in Fig. 3.3.

The first three columns of the table describe the Coulomb, Lennard-Jones, Restraints contributions to free energy, respectively, while the last one corresponds to the value of the total ligand-protein complex free energy. All the values are expressed in kJ · mol^{-1}. In Fig. 3.3, the atomistic refer-

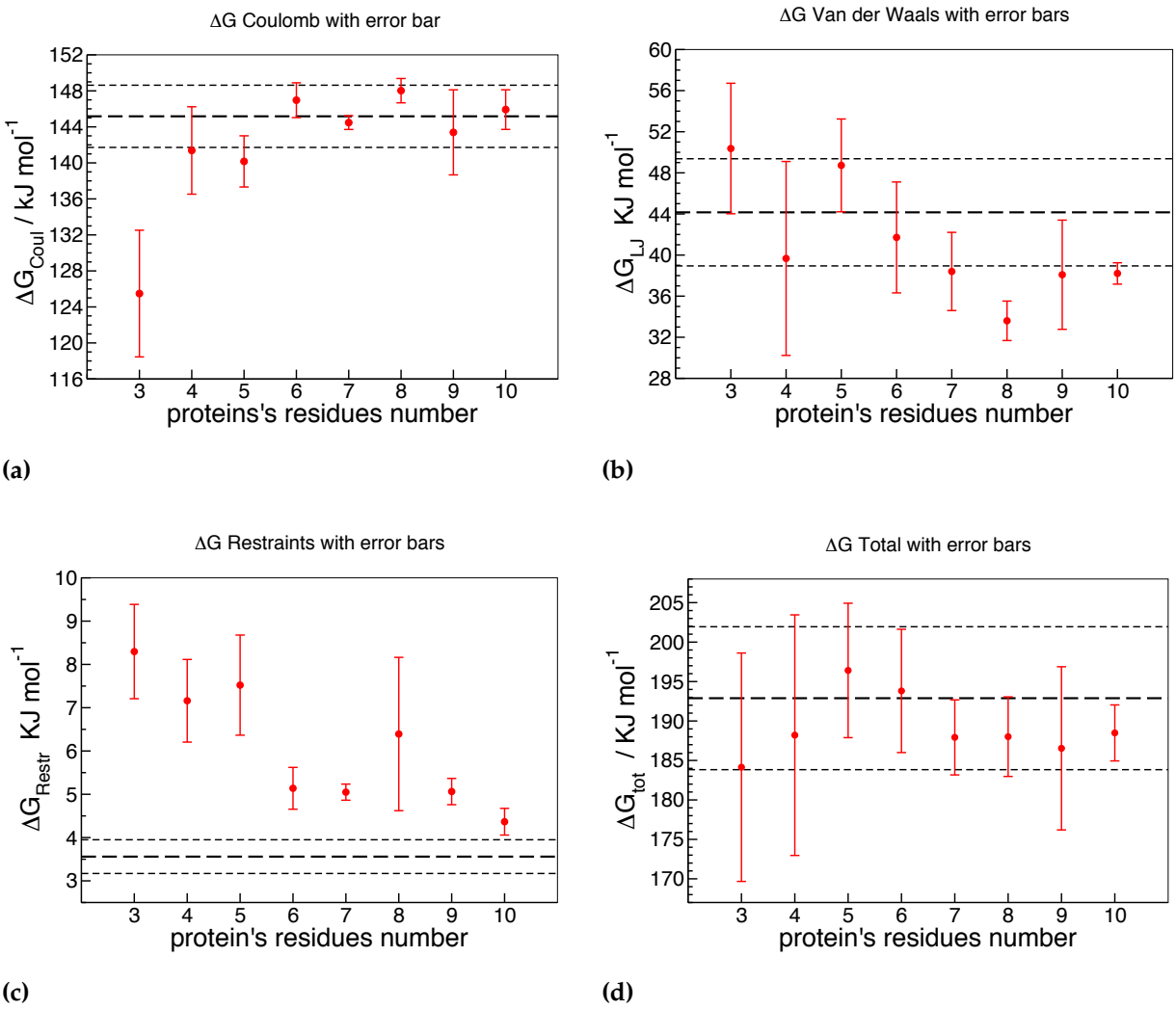


Figure 3.3: (a) Coulomb, (b) Lennard-Jones, (c) restraint and (d) total free energies in the protein-ligand complex, as a function of protein's residues number included in atomistic detail in the multi-resolution set-up. The heavy dashed black horizontal lines are the reference values from fully atomistic simulations, and the lighter dotted black horizontal lines are the error bars for those values. These simulations use decoupling, not annihilation. y-axes do not cover the same energy range.

ence is represented with a dash black line with its error bar. In particular, panels (a), (b) and (c) show the three components that contribute to the total complex free energy, reported in panel (d). Looking at these values as a function of the number of all-atom active site residues, we notice that there are important deviations of the free energy from the reference, especially in the case of 3 and 4 atomistic residues. On the contrary, the total value of the binding free energy agrees with the reference within the error bar

in all cases.

Furthermore, we observe that the trend of free energy values, in comparison to the reference, is essentially the same: starting from 3 amino acids it approaches the reference until reaching 6, both in its components and in total. In contrast, going from 6 to 8 atomistic residues the value deviates from the reference, even though the total remains close to it. Finally, from 8 to 10, ΔG converges again. Hence, increasing the number of atomistic residues does not introduce necessarily an improvement of the computed free energy, at least as long as the various free energy components are considered separately.

In order to gain further, quantitative insight into these results, we computed the quadratic deviation from the reference, δ^2, defined as:

$$\begin{aligned} \delta_i^2 &= \delta_{i-Coul}^2 + \delta_{i-LJ}^2 + \delta_{i-Restr}^2 = \\ &= (\Delta G_{Coul\_i} - \Delta G_{Coul-at})^2 \\ &+ (\Delta G_{LJ\_i} - \Delta G_{LJ-at})^2 \\ &+ (\Delta G_{Restr\_i} - \Delta G_{Restr-at})^2 \end{aligned} \tag{3.5}$$

where the index $i = 3...10$ runs over atomistic residues. Fig. 3.4 reports δ^2 as a function of the number of active site amino acids modelled with atomistic detail.

The plot shows that the binding free energy computed in the dual-res model approaches the reference as the number of atomistic active site residues increases, and most importantly this approach takes place for each component up 6 residues. Beyond this value, though, the trend stops and the deviation becomes larger, peaking at 8 residues and decreasing when further atomistic amino acids are added.

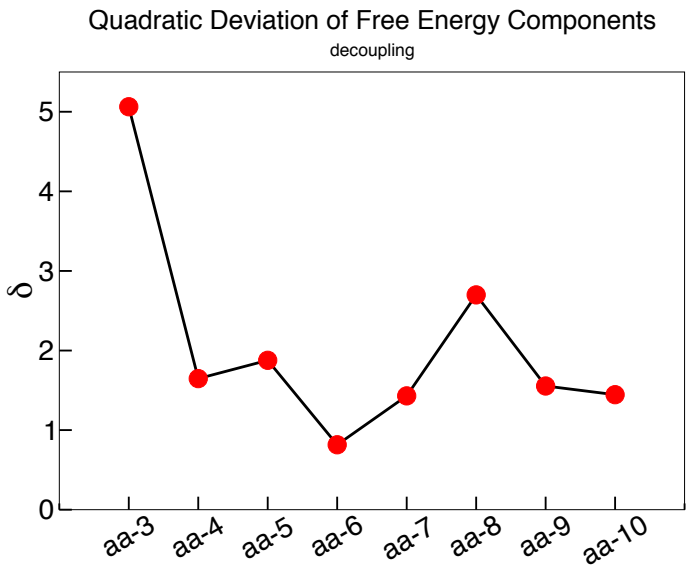


Figure 3.4: Square root of quadratic deviation δ^2 vs the number of atomistic residues chosen. The plot shows that in the case of 6 atomistic residues, the value of quadratic deviation is the lowest one and hence it means that such a number leads the best result of free energy. Moreover the black line shows the trend of FE values as discussed in section 3.3.

These results highlight a non-monotonic dependence of the free energy on the mapping, that is, the number of retained atomistic residues. If, on the one hand, the overall value of the binding free energy (Fig. 3.3 panel (d)) levels to the reference with as few all-atom residues as 4, the separate components oscillate and reach the plateau only for larger numbers. The existence of a minimum in the standard deviation of all three contributions pinpoints a particular number of atomistic active site residues for which the accuracy of the computed free energy is the highest and the economy of the high-resolution subpart the largest. Including more than 6 atomistic residues counterintuitively worsens the result –when the various contributions are looked at– and the previous accuracy is only recovered when more residues are included. This behaviour suggests that the total free energy undergoes an error cancellation which hides the deviations of the separate terms.

A possible explanation for this nontrivial behaviour is that when 6 active site residues are modelled with all-atom accuracy (Fig. 3.5(b)) the ligand is stable in the catalytic

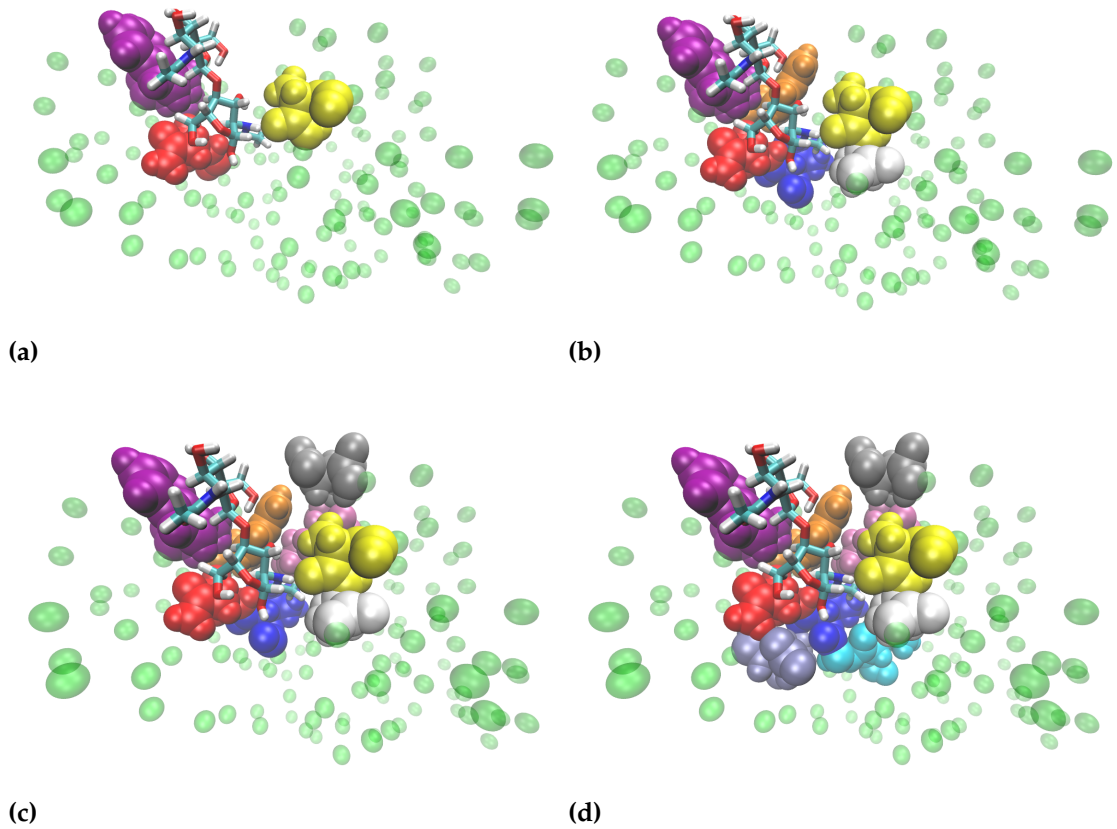


Figure 3.5: VMD representation of lysozyme and ligand in different resolution: (a) three, (b) six, (c) eight, (d) ten atomistic residues. The complete set can be found in *Supporting information* The ligand is always atomistic and it is represented in Licorice. In green are represented the ENM beads. With the other colors are represented, instead, the various atomistic residues which surround the ligand.

site, namely it is surrounded by a complete shell of atomistic residues. The addition or deletion of other residues (Figs. 3.5(c) and 3.5(a) respectively) leads to a worsening of ΔG: in the first case, the two added residues (in pink and grey) are located behind the first shell of amino acids (far away from the ligand) and start to form a second, incomplete shell; in the second case, only three atomistic amino acids take part in the direct interaction with the ligand: therefore, the first layer is still incomplete and important interactions are missing; in order to improve the free energy value one has to add further amino acids in order to complete the second shell. We emphasise that the impact on the deviation from the reference is inversely

proportional to the distance of the added/removed amino acid. Thus, the farther the atomistic amino acid is from the ligand, the more negligible its effect is. In the *Supporting Information* we provide detail about the other numbers of all-atom residues not reported here. Finally, the values of binding free energy (also for the case of annihilation whose calculations are reported in the *Supporting Information*) are summarised in Tab. 3.3 and illustrated in Fig. 3.6.

	Ligand	**Complex**	**Binding**
annihilation			
atom, espp	-1275.3 ± 11.2	1315.2 ± 16.3	8.6 ± 27.5
atom, grom	-1259.0 ± 5.9	1314.8 ± 13.2	24.5 ± 19.1
decoupling			
atom, espp	-151.9 ± 8.0	193.0 ± 9.1	9.8 ± 17.1
aa-3, espp	-151.9 ± 8.0	184.2 ± 14.4	1.0 ± 22.4
aa-4, espp	-151.9 ± 8.0	188.3 ± 15.3	5.1 ± 23.3
aa-5, espp	-151.9 ± 8.0	196.4 ± 8.5	13.2 ± 16.5
aa-6, espp	-151.9 ± 8.0	193.8 ± 7.8	10.6 ± 15.8
aa-7, espp	-151.9 ± 8.0	187.9 ± 4.8	4.7 ± 12.8
aa-8, espp	-151.9 ± 8.0	188.0 ± 5.1	4.8 ± 13.1
aa-9, espp	-151.9 ± 8.0	186.6 ± 10.3	3.4 ± 18.3
aa-10, espp	-151.9 ± 8.0	188.5 ± 3.5	5.3 ± 11.5

Table 3.3: Representation of Free Energies values computed in ESPResSo++ and GROMACS (respectively *espp* and *grom* using a short notation on the table) in case of annihilation and decoupling. The table is divided in three column: from left to right are represented the ligand, protein-ligand complex and binding FE. The latter is the algebraic sum of ΔG_{compl}, ΔG_{r_off} and ΔG_{lig}. The results are in $\mathrm{kJ} \cdot \mathrm{mol}^{-1}$.

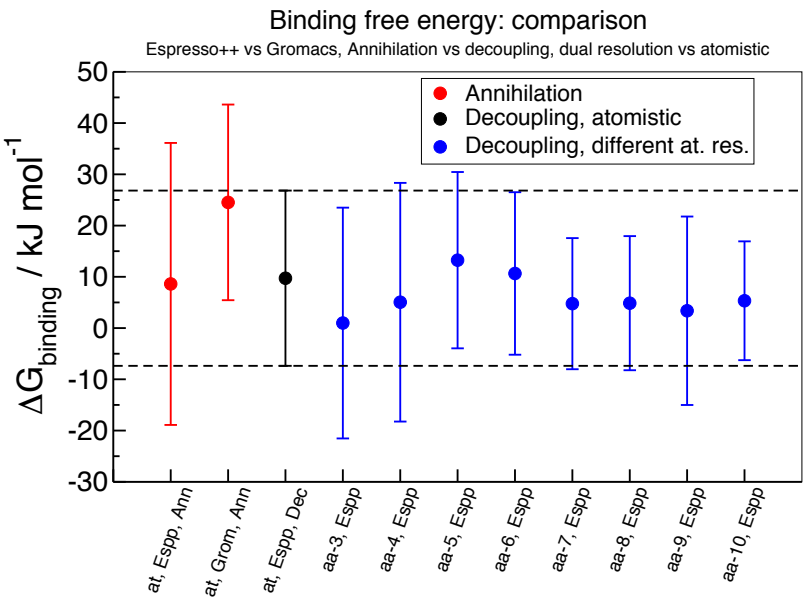


Figure 3.6: Binding free energies as a function of protein's residues included in atomistic detail in the multi-resolution set-up or fully atomistic set-up. The heavy dashed black horizontal lines and black point are the reference values from fully atomistic simulations obtained in ESPResSo++ with decoupling, and the lighter dotted black horizontal lines are the error bars for those values. In red are represented binding free energies values in ESPResSo++ and GROMACS in case of annihilation. In blue is represented the binding FE value in dual resolution simulation changing the number of atomistic residues.

3.4 Conclusions

In this work we have shown how the dual resolution model employed, constituted by an all-atom subregion coupled to an elastic network model remainder, can be used to calculate the binding free energy of an enzyme-substrate complex with atomistic accuracy. Furthermore, and most importantly, we have highlighted the impact that different choices of the model resolution can have. Specifically, we have computed the total value of the binding free energy as well as that of its various energetic components, and quantitatively inspected how these change when different selections are performed for the subgroup of amino acids, ranging from 3 to 10 in total, to be modelled at the fully atomistic level.

At first sight, one can appreciate that the binding free energy value rapidly converges to the atomistic reference when as few as 4 amino acids constituting the active site are described all-atom. This comforting result, however, unveils a greater complexity when the different terms constituting the free energy are looked at separately. These show an oscillating behaviour as the number of all-atom residues in the active site is increased, with a decreasing difference from the reference followed by a sudden jump to larger values, which dampens upon further addition of atomistic amino acids. The rationale in this behaviour is identified in the structure of the active site, which is constituted by a first shell of the six residues exposed to the solvent and closest to the ligand; when further amino acids beyond these are modelled with atomistic resolution, they interact with the substrate affecting the binding free energy components and shifting them away from the reference,

with a steadily lowering impact as the model's resolution is increased - as one can expect. Surprisingly, very little if no signal of this behaviour is observed in the value of the binding free energy as a whole, rather it becomes visible only upon inspection of its separate contributions.

The results of this work thus highlight the importance of mapping in the construction of multi-scale and multi-resolution models, as a higher degree of detail does not necessarily correlate with a higher accuracy of the quantities of interest. The implications of these observations should serve as a warning and a guide in the realisation of coarse-grained models concurrently employing various levels of detail for different regions of the same system, whose range of application spans from fundamental understating of a molecule's properties to real-life pharmaceutical applications.

3.5 Supporting Information

3.5.1 Thermodynamic Cycle for binding free energy

In order to compute the binding free energy ΔG_{bind} [154–156], we use a thermodynamic cycle which connects the protein-bound and protein-unbound ligand states through several intermediate ones as shown in the Fig. 3.1.

Starting from the top-right corner we have the complex, with the ligand and protein fully interacting as in a normal MD simulation. The first step is adding a set of restraints between ligand and protein (giving ΔG_{restr_on}) in order to avoid the problem of the ligand leaving the binding

pocket when interactions are being removed. The presence of restraints is indicated in the cycle scheme in the figure by a red circle, which represents the fact that the ligand is being confined to a certain volume. The set of restraints described by Boresch is used for this work [156]. These are quite useful as they restrain position and orientation of the compound relative to the protein, and they have an analytical solution for their removal.

The next step is decoupling the ligand from the system in order to get to the bottom-right corner of the cycle. This involves running a number of separate simulations at different λ values, first decoupling coulombic interactions ($\Delta G_{coul,c}$) and then Lennard-Jones ($\Delta G_{LJ,c}$).

Going up from the bottom-left corner of the cycles, the first step (ΔG_{restr_off}) is carried out analytically without need to run more simulations. At this point the ligand has come back to interact with the solvent, which means one needs to turn on charges ($\Delta G_{coul,\ell}$) and Van der Waals ($\Delta G_{LJ,\ell}$) parameters again, in order to obtain ΔG_{int_water} (or ΔG_{ligand}). Finally, at the top-left corner of the cycle, one sums up all the steps done so far to obtain the quantity ΔG_{bind}.

3.5.2 Annihilation and Binding Free Energy

The calculation of free energy can be done in two different ways: *decoupling* and *annihilation*. The difference between the two is the following: decoupling a molecular interaction refers to turning off that interaction between the molecule and its environment, whereas annihilation of a molecular interaction refers to turning off that interaction entirely.

We focus on the results of free energy in case of annihilation. This has two advantages: the first one is that it allows one to validate the implementation of protein free energy in ESPResSo++ [129, 130] doing a comparison with GROMACS [128]: this is feasible only in the case of annihilation in fully atomistic system because GROMACS cannot perform decoupling and dual resolution simulations.

The second advantage is that the simulation with annihilation allows us to give a further confirmation that the value of binding free energy in case of decoupling is correct, thereby proving the consistency between the two. By definition, in annihilation there are three components (ligand-ligand, ligand-water and ligand-protein) unlike decoupling which has two components (ligand-water and ligand-protein): hence, the values of complex and ligand free energy will be different each other, but the values of the resulting ΔG_{bind} in both cases agree each other within the error bar, as reported in the main text.

Without going into the simulation details (look in the apposite section of the article) we can see the results of binding free energy calculation by using Thermodynamic Integration (TI) [133].

A. Results of Binding free energy calculation

Recall that ΔG_{bind} consists in the algebraic sum of three terms: $\Delta G_{complex}$, ΔG_{ligand} and ΔG_{restr_off}. Let us focus, first of all, to the calculation of the latter because it is carried out analytically without needing to run simulations [156]:

$$-\frac{\Delta G_{restr\_off}}{kT} = \ln\left[\frac{8\pi^2 V^0}{r_0^2 sin\theta_{A,0} sin\theta_{B,0}} \frac{(K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C})^{\frac{1}{2}}}{(2\pi kT)^3}\right] \tag{3.6}$$

where: k is the ideal gas constant; T is the temperature in Kelvin; V^0 is the volume corresponding to the one molar standard state (1660 Å^3); r_0 is the reference distance for the restraints; θ_A,θ_B are the reference angles for the restraints; K_x is the force constant for the distance (r_0), two angles (θ_A,θ_B) and three dihedrals (ϕ_A, ϕ_B, ϕ_C) restraints we applied.

In our case we have that:

$$k = 8.31\frac{J}{mol \cdot K} = 1.987\frac{cal}{mol \cdot K}$$
$$T = 298K$$
$$V^0 = 1660\text{Å}^3$$
$$r_0 = 0.31nm = 3.1\text{Å}$$
$$k_x = 4184\frac{KJ}{mol \cdot nm^2} = 41.84\frac{KJ}{mol \cdot \text{Å}^2}$$
$$\theta_A = 120^o$$
$$k_{\theta_A} = 41.84\frac{KJ}{mol \cdot rad^2}$$
$$\theta_B = 90^o$$
$$k_{\theta_B} = 41.84\frac{KJ}{mol \cdot rad^2}$$
$$k_{\phi_A} = k_{\phi_B} = k_{\phi_C} = 41.84\frac{KJ}{mol \cdot rad^2}$$

Therefore the contribution to the binding free energy coming from the restraints amounts to:

$$\Delta G_{restr\_off} = -31.3 \ kJ \cdot mol^{-1} \tag{3.7}$$

The results of the $\Delta G_{complex}$ and ΔG_{ligand} terms and the alchemical changes, comparing ESPResSo++ and GRO-

Complex FE - Annihilation				
	$\Delta G_{coul,c}$	$\Delta G_{LJ,c}$	$\Delta G_{Restr\_on,c}$	$\Delta G_{complex}$
grom	1254.2 ± 8.0	57.3 ± 4.9	3.3 ± 0.3	1314.8 ± 13.2
espp	1250.7 ± 5.6	60.8 ± 10.4	3.6 ± 0.4	1315.1 ± 16.4

Table 3.4: Resulting values of free energy of Complex Free Energy (4th column) and its components (Coulomb, Lennard Jones and Restraints in the first three columns) in fully atomistic system in case of annihilation. All values are in $kJ \cdot mol^{-1}$ and performed with Thermodynamic Integration. All simulations are carried out in GROMACS and ESPResSo++ (*grom* and *espp* in the table). For each value of λ, the fully atomistic simulations lasts 1 ns by using both MD package program simulation. These results show that, within the error bars, both codes provide the same results.

Ligand FE - Annihilation			
	$\Delta G_{coul,\ell}$	$\Delta G_{LJ,\ell}$	ΔG_{ligand}
grom	1238.8 ± 2.3	20.2 ± 3.6	1259.0 ± 5.9
espp	1250.1 ± 6.2	25.2 ± 5.0	1275.3 ± 11.2

Table 3.5: Resulting values of Ligand Free Energy (3rd column) and its components (Coulomb, Lennard Jones in the first two columns) in fully atomistic system in case of annihilation. All the values are in $kJ \cdot mol^{-1}$ and performed with Thermodynamic Integration. All simulations are carried out in GROMACS and ESPResSo++ (*grom* and *espp* in the table). For each value of λ, the fully atomistic simulations lasts 1 ns by using both MD package program simulations.

MACS are shown respectively in Tab. 3.4 and Tab. 3.5 and illustrated in the Fig. 3.7.

The resulting Binding Free Energy is shown in the Tab. 3.6: we take up the final values of Complex and Ligand FE, in order to compute ΔG_{bind}. These results are illustrated in the Fig. 3.8.

These plots show how GROMACS and ESPResSo++ produce the same Binding FE results both in the components (Fig. 3.7) and in the total (Fig. 3.8).

As reported in the main article, comparing the results of the Binding FE values both in annihilation and decou-

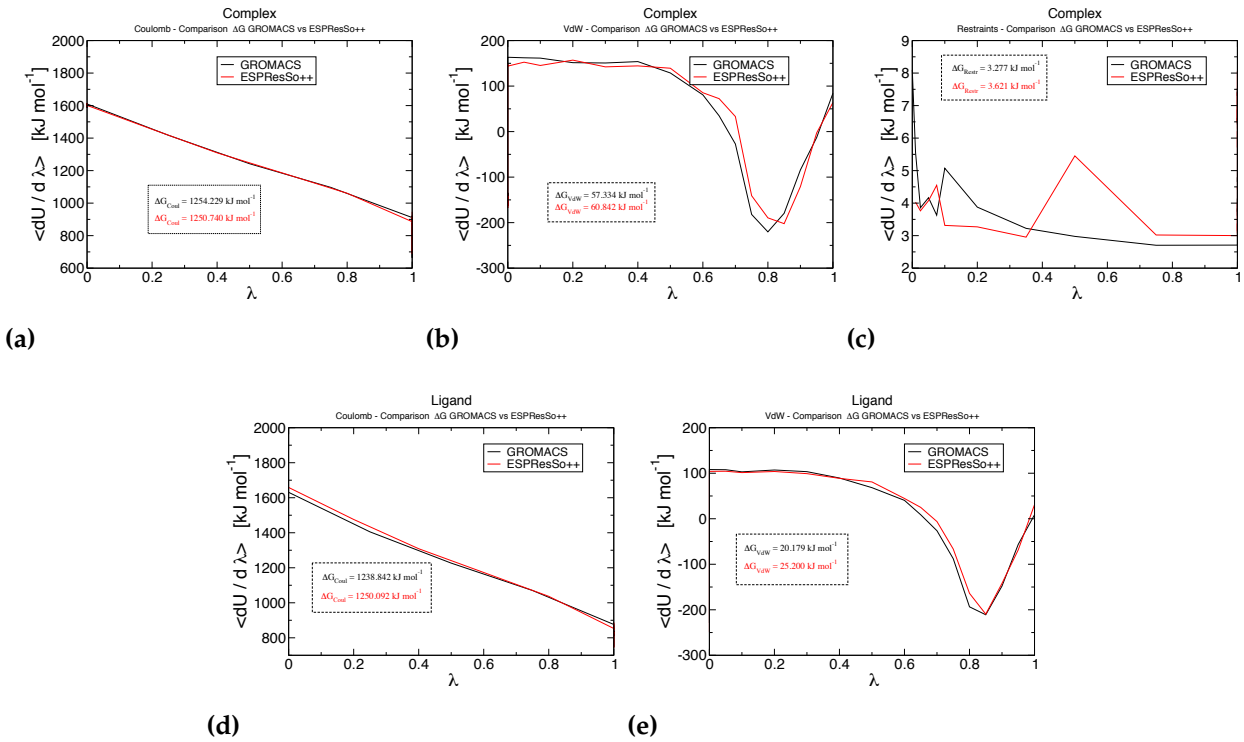


Figure 3.7: Comparison of the Thermodynamic Integration (TI) free energy derivative curves computed with ESPResSo++ and GROMACS for all atom protein. (a) Coulomb, (b) Lennard-Jones and (c) restraint free energies curves for the protein-ligand complex, and (d) Coulomb and (e) Lennard-Jones free energies curves for the ligand. These simulations use annihilation.

Table 3.6: Representation of Free Energies values computed in ESPResSo++ and GROMACS (*espp* and *grom* respectively in the table) in case of annihilation. The table is divided in three column: from left to right are represented the ligand, protein-ligand complex and binding FE. These results are in kJ · mol^{-1}.

Binding FE - Annihilation			
	ΔG_{ligand}	$\Delta G_{complex}$	$\Delta G_{binding}$
grom	-1259.0 ± 5.9	1314.8 ± 13.1	24.5 ± 19.1
espp	-1275.3 ± 11.2	1315.2 ± 16.3	8.6 ± 27.5

pling, we notice that there is a consistency between these two method of treating interactions (Fig. 3.6). Therefore we chose to work with decoupling instead of annihilation, but we choose the first one because this process is more intuitive with respect its annihilation turning off the interactions within it. Moreover, the ligand is always treated atomistically, therefore it is not involved in the change of free energy varying the protein resolution.

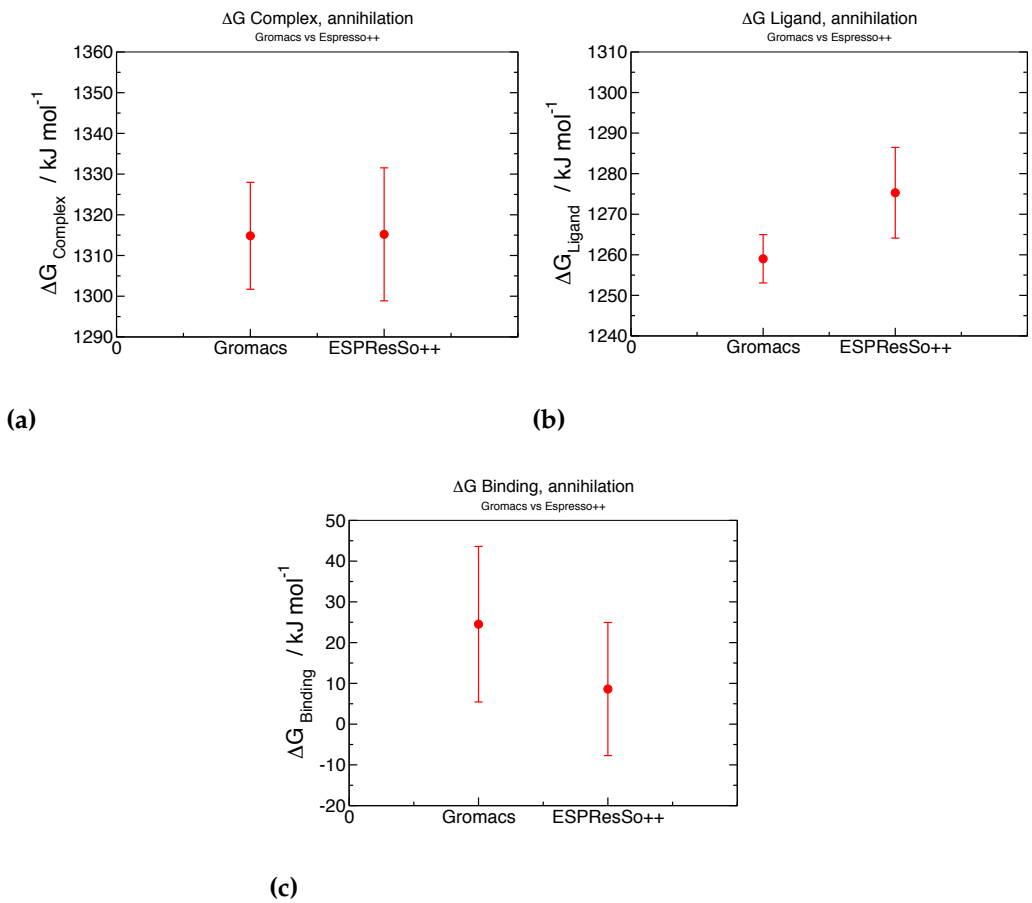


Figure 3.8: Comparison of the Thermodynamic Integration (TI) free energy values computed with ESPResSo++ and GROMACS for all atom protein. (a) protein-ligand complex, (b) Ligand and (c) Binding free energy values (with error bars). These simulations use annihilation. The plots show that the results obtained via ESPResSo++ and GROMACS are comparable within the error bars.

3.5.3 Parametrization of the dual-resolution model

Our protein is treated in dual-fixed-resolution, in particular the binding site of lysozyme is modelled in atomistic high level of resolution, whereas the rest of protein is treated in Coarse Grained and specifically in ENM [36].

In order to construct a good dual resolution model, the system needs a parametrization. The latter was already performed in [24]. Here we describe the key elements of the model parametrization, namely the elastic constant between consecutive ENM nodes and not consecutive ones,

and the parameters ϵ and σ of Week-Chandler-Anderson (WCA) [178]. In particular, the latter was found in the case of 8 atomistic residues, therefore in this section we must confirm that it is still good changing the protein resolution.

A. Determination of elastic constants between beads

First, we do a distinction between the value of elastic constants between consecutive C_α beads along the protein backbone (k_b) and not consecutive ones (k_{nb}) until the cut-off set to 1.2 nm. In particular, we take as k_b, the stiff value of $5 \cdot 10^4$ kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}. The global fluctuations are independent of this value. All other spring constants have a value k_{nb} = 160 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}, parametrised by minimising the average root mean square error in C_α rmsf and the S^2 order parameter calculated from ENM relative to fully atomistic simulations [24].

B. Determination of WCA parameters ε and σ

When the ENM is employed in multi-resolution simulations, an excluded volume interaction between ENM nodes and solvent molecules is required, in order to prevent from penetrating the protein and solvating the atomistic binding site from the interior. Thus, a WCA interaction is applied between C_α nodes and all the solvent molecules.

In its formulation, WCA needs two parameters: ϵ and σ. The former has a value of 0.34 kJ $\cdot$ mol^{-1}, arbitrarily chosen as the value for carbon in the atomistic forcefield, whilst $\sigma_i = R_{g,i} \cdot c$, where $R_{g,i}$ is the radius of gyration of a given residue i out of the twenty possible amino acids

aa-8	density / molecules nm^{-3}
fully at	100.2
0.59	99.5
0.61	99.8
0.63	99.9
0.65	100.1
0.67	100.4
0.69	100.5
0.71	100.7

Table 3.7: Density found in the case of 8 atomistic residues for different value of c and comparison with the atomistic reference. Each dual resolution simulation varying c lasts 1 ns.

and c is the same for all amino acids. The latter is not known a priori, because its value has to be tuned to give the correct bulk water density for a protein-water system (i.e. the water density far from the protein) from fully atomistic simulation.

In order to find the proper value of c we started with the 8 atomistic residues protein launching different dual resolution simulations of 1 ns, varying its value. After finding the correct c such that the density between atomistic and dual-res system are comparable, we checked that such a value is still good launching, this time, 1 ns simulations with different numbers of atomistic residues keeping c fixed.

Tab. 3.7 and Fig. 3.9 show the bulk water density in the fully atomistic reference system and in case of 8 amino acids modelled atomistically for different values of c. In particular Fig. 3.9 also shows a linear interpolation between points in order to get c as precise as possible. The resulting value of c is 0.658.

Tab. 3.8 shows that such a value is still valid when changing the number of amino acids of protein active site, because the relative error is no longer than 0.7%. We thus employed $c = 0.658$ in all considered cases.

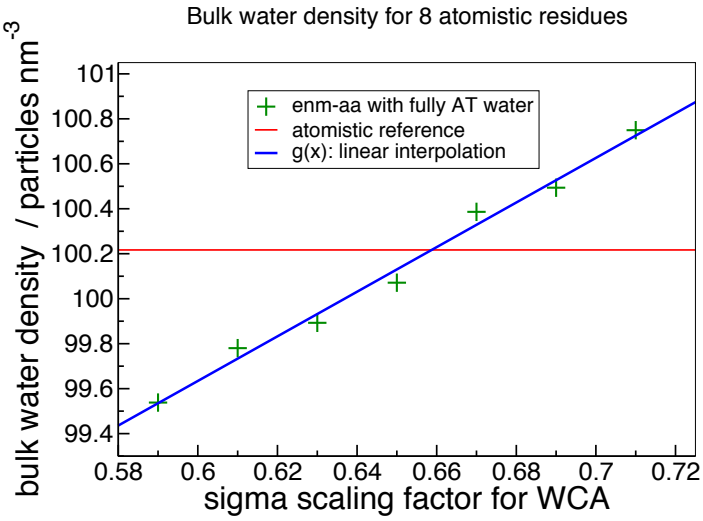


Figure 3.9: Bulk water density in the case of 8 atomistic residues for different value of c. The atomistic reference value is 100.2 and it represented with the red line, whereas the linear interpolation of points is called g(x) in the legend and it is shown with a blue line.

Table 3.8: Bulk water's average density (in molecules $\cdot$ nm^{-3}) and percentage relative error in dual-resolution simulation with different atomistic residues from 3 to 10, keeping c = 0.658. Each simulation lasts 1 ns.

# at residues	average density	relative error
3	100.3	0.1 %
4	100.1	0.1 %
5	100.1	0.1 %
6	100.1	0.1 %
7	100.0	0.2 %
8	100.2	0.0 %
9	100.0	0.2 %
10	100.9	0.7 %

In Fig. 3.10 we report the VMD [181] representation of all the considered cases changing the number of atomistic residues of active site from 3 to 10 (recall that in the article are reported only the most important cases namely three, six, eight and ten atomistic residues).

3.6 Acknowledgements

The authors are grateful to Robinson Cortes-Huerto and Thomas Tarenzi for a critical reading of the manuscript. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 758588).

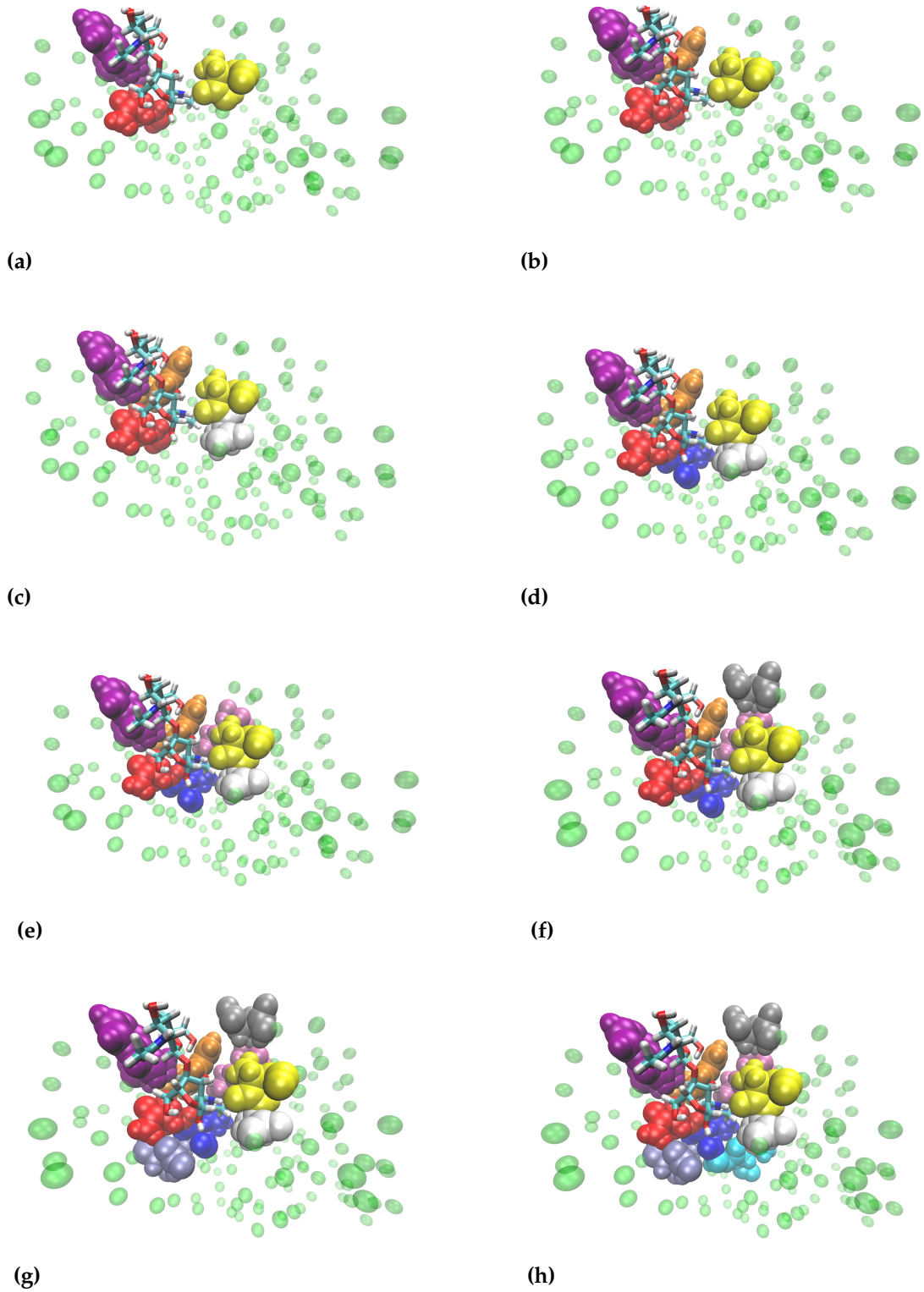


Figure 3.10: Representation of lysozyme and ligand in different resolution: (a) three, (b) four, (c) five, (d) six, (e) seven, (f) eight (g) nine, (h) ten atomistic residues. The ligand is always atomistic and it is represented in Licorice. In green are represented the ENM beads. With the other colors are represented, instead, the various atomistic residues which surround the ligand.

Free energy landscapes calculation in 1BBA investigated through an improvement version of the ENM.

4

In the previous chapter a dual resolution model has been employed to compute the binding free energy of lysozyme with a substrate. In this model the solvent is treated atomistically, as well as the binding site. To describe the coarse-graineded part an Elastic Network Model (ENM) has been employed, in which each residue is mapped onto a bead whose position corresponds at the C_α atom in the atomistic description. These beads are connected by identical harmonic springs.

In the first part of this chapter we employ the same dual-resolution model for the calculation of free energy landscapes in terms of collective variables appropriate to describe the reference system of a small protein *Bovine Pancreatic Polypeptide* (PDB code 1BBA). The choice of the atomistic and the coarse-grained part is carried out by using PiSQRD [182, 183], a tool that allows one to divide a protein into rigid subdomains according to the fraction of internal motion.

At difference with the previously discussed model, however, we employ different elastic constants between beads in the ENM. Specifically, the strength of the effective bonded

interactions between C_α atoms is based on their distance distribution in the all-atom representation, allowing a systematic way to establish a precise parametrization of each harmonic spring.

4.1 Introduction

From a simulation point of view, biomolecular systems are among the most challenging because of their heterogeneity and the wide range of length and time scales they encompass [184–186]. Simulating such systems therefore often lead to two important requirements:

- large systems and long simulation times;
- accurate and often computationally expensive models that contain sufficient physical and chemical detail to describe a given phenomenon, usually computationally expensive.

A promising way of mitigating the computational overhead is to employ a concurrent multi-resolution approach. This involves identifying those parts of the system where the physical and chemical details play a pivotal role in the phenomenon of interest and describing them using a sufficiently high-resolution model while using a less detailed, computationally more efficient model for the remainder of the system.

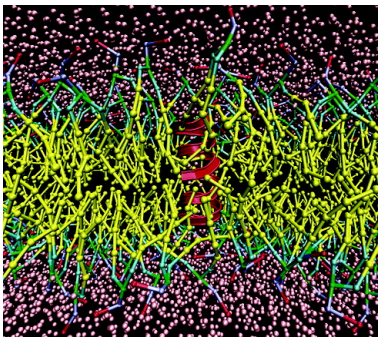


Figure 4.1: Configuration of the atomistic gA peptide in CG DMPC lipid and water, after 10ns AA-CG MD simulation. Adapted from Ref. [187]

In most concurrent multiple resolution simulation approaches employed, each system component is treated using only one level of resolution: for example an atomistic protein in a coarse-grained solvent or embedded in a coarse-grained membrane [187–189] as shown in Fig. 4.1.

However, when the goal is to construct a model which

includes only the minimum possible number degrees of freedom, one has to be able to place boundaries between resolutions at any arbitrary place within the system. Sometimes, this procedure is made simpler thanks to intrinsic properties of the system, e.g. in the case of enzymes.

In its simplest form, the latter can be seen as being composed of two parts: an active site, at which the ligand binding and catalytic reaction occur, and the remainder of the enzyme, which supports the active site and provides the thermal fluctuations necessary for its function. Moreover, the aqueous solvent plays an essential role in enzymatic function [190, 191]. An accurate model of ligand-binding, therefore, requires at minimum an atomistic level of detail in the description of the ligand, binding site, and neighbouring water molecules. The rest of the protein and the water sufficiently far away from the binding site can be modelled at a more coarse-grained level [36, 169].

The definition of a clear boundary between resolutions, in other cases, such as the test protein used here (*Bovine Pancreatic Polypeptide*), is not trivial. Therefore, there is no way to identify the block division intuitively. In this respect, the PiSQRD [182, 183] is a useful tool whose purpose is to find groups of amino acids that can be treated as rigid blocks and defining, consequently, the boundaries between these regions.

The coarse-grained part is treated as an Elastic Network Model (ENM) in which each residue is mapped onto its C_α in all-atom representation. In the original formulation, the beads are connected by a unique elastic constant. Here, a further refinement is introduced, in that elastic constants between beads are employed, according to their distance

distribution.

Here, we investigate this dual-resolution model of a biomolecule trying to identify the best possible strategy to face this issue. This validation paves the way to a new generalized ENM model.

4.2 Methodology

Pancreatic Polypeptides (PPs) are single-chain peptides of 36 amino acids. They were discovered as a contaminant in the purification of insulin and then isolated and purified from chicken pancreas (aPP) and bovine pancreas (bPP) in the mid '70s [192, 193]. Subsequently they have been purified from a variety of species [194–197]. All PPs except possibly anglerfish possess 36 amino acids with human, bovine, ovine, porcine, and canine species differing by 1-4 amino acids. They appear to play an important role in the physiological feedback inhibition that regulated pancreatic secretion after a protein meal [198].

The system under investigation in the present work is the *Bovine Pancreatic Polypeptide* [199] (bPP, PDB code 1BBA) in aqueous solution in two different cases: with salt in physiological concentration or pure water. The choice of studying this protein arises from two main reasons:

- its dynamics is not trivial providing a realistic, non-trivial test case of a system in which the protein division in all-atom (AA) and coarse-grained (CG) is complicated to identify;
- its size is tiny (only 582 atoms), allowing fast simulations.

In this model, the first challenge is to find the boundaries between the atomistic and coarse-grained part.

The model employed is not adaptive, that is, the resolution of a given residue is fixed (both atomistic and coarse-grained) and does not change during the simulation. Specifically, the coarse-grained model used to describe the low-resolution part of the protein is a classical Elastic Network Model. First, two values of elastic constants have been employed: one for consecutive C_α carbons along the backbone, and one for the other bonds. Afterwards, as further refinement, a specific value of elastic constant has been applied between beads according to their distance distribution.

Water molecules and ions are described in atomistic details inside the simulation box. The interaction with the high-resolution part of the protein takes place through the standard all-atom force field; on the other hand, an excluded volume interaction between ENM nodes, solvent molecules and ions is required in order to prevent the solvent from penetrating the protein. Hereafter, we provide a detailed description of the model. At first, we describe the PiSQRD tool, which is employed here with the purpose of dividing the protein in rigid domains and, consequently, finding the boundaries between atomistic and ENM part. Then, we outline the dual-resolution model and its coupling to the atomistic part. Afterwards, we focus on the strategy allowing us to find different elastic constants, based on the distance distribution between C_α beads. In the fourth part, we illustrate the methodology adopted to compare atomistic and dual-resolution simulation, and finally, we report information about the simulation setup.

4.2.1 Finding boundaries between atomistic and CG part

A possible strategy to find boundaries between the atomistic and the coarse-grained part consists in the identification of domains, in the protein structure, which move approximately as rigid bodies. For our purposes, we use the description of the molecule in terms of *quasi-rigid* domains as illustrated in Ref. [183]. In particular, the criterion to subdivide the protein into domains can be stated as follows: we are interested in assigning the amino acids of the protein to a given number Q of domains; the optimal partition is the one maximizing, over all possible assignments, the rigid roto-translation contribution of the domain (called $\mathrm{MSF}^{\parallel}$ in the following), or equivalently, the partition which minimizes the internal fluctuation inside the domain (called $\mathrm{MSF}^{\perp}$ in the following).

In order to prove the previous statement, let us start assuming that for each attempt of partitioning the amino acids, we consider the instantaneous displacement vector, $\mathbf{v}_q(t)$ of a putative domain. It turns out that the coordinates $\mathbf{r}_q(t)$ in a given trajectory frame are:

$$\mathbf{r}_q(t) = \mathbf{r}_q^0 + \mathbf{v}_q(t) \tag{4.1}$$

where $\mathbf{r}_q^0$ is the reference structure, and q is the label of a presumed domain.

The instantaneous displacement vector, $\mathbf{v}_q(t)$ can be separated in two contributions: $\mathbf{v}_q^{rb}(t)$, corresponding to a rigid roto-translation of the qth domain, and $\Delta\mathbf{v}_q$, which describes the fluctuations internal to the domain:

$$\mathbf{v}_q(t) = \mathbf{v}_q^{rb}(t) + \Delta\mathbf{v}_q \tag{4.2}$$

Furthermore, the rigid-body component $\mathbf{v}_q^{rb}(t)$ can be decomposed in a translation vector $\tau_q(t)$ and a rotation parametrized by the matrix $\mathcal{R}$ and the vector $\omega_q(t)$ as follows:

$$\mathbf{v}_q^{rb}(t) = \tau_q(t) + \mathcal{R}\left[\omega_q(t)\right](\mathbf{r}_q^0 - \mathbf{R}_q) \tag{4.3}$$

where $\mathbf{R}_q$ are the coordinates of the q-th domain's centre of mass.

The matrix $\mathcal{R}$ can be calculated by means of the Kabsch algorithm [200], which finds the optimal rotation of the sets of points minimizing the Root Mean Square Deviation (RMSD) between them.

The extremality conditions that are imposed to find $\tau_q(t)$ and $\omega_q(t)$ guarantee the orthogonality between the rigid-body displacement $\mathbf{v}_q^{rb}(t)$ and the internal fluctuation term, namely $\Delta\mathbf{v}_q$. The latter property allows one to decompose the total mean square fluctuation of the molecule in two contributions:

$$\begin{aligned} \mathrm{MSF} &\equiv \sum_{q=1}^{Q}\left\langle\left|\mathbf{R}_q - \mathbf{R}_q^0\right|^2\right\rangle = \sum_{q=1}^{Q}\left\langle\left|\mathbf{v}_q\right|^2\right\rangle = \\ &= \sum_{q=1}^{Q}\left\langle\left|\mathbf{v}_q^{rb}\right|^2 + \left|\Delta\mathbf{v}_q\right|^2\right\rangle = \mathrm{MSF}^{\parallel} + \mathrm{MSF}^{\perp} \end{aligned} \tag{4.4}$$

where:

$$\begin{aligned} \mathrm{MSF}^{\parallel} &= \sum_{q=1}^{Q}\left\langle\left|\mathbf{v}_q^{rb}\right|^2\right\rangle \\ \mathrm{MSF}^{\perp} &= \sum_{q=1}^{Q}\left\langle\left|\Delta v_q\right|^2\right\rangle \end{aligned} \tag{4.5}$$

This method is completely general since no assumption is made on the contiguity in space or sequence of the residues of the domains. In principle, any possible assignment of the residues to the domains is tried for a given Q, and the optimal choice is performed on the basis just mentioned. This algorithm is an open-source software called PiSQRD (*Protein Structure Quasi-Rigid Domain Decomposition*).

A block, or domain, is a sequence of residues along the protein with the property of being *quasi-rigid* according to the definition explained before. Therefore, after partitioning the structure, one can easily identify the interfaces between two different domains. Each boundary has an exact position along the backbone, placed intuitively, between two consecutive protein residues n and $n+1$. It can be denoted with $B_{n,n+1}$ with $n \in [1, \ N-1]$ where N is the total number of protein residues.

Increasing the number of imposed domains, one can visualize the position of all edges for each considered case. Then, we focus on the frequency $\mathcal{F}$ with which each one appears:

$$\mathcal{F}\left[B_{1,2}\right] = \mathsf{f}_1 \in \mathbb{N}$$

$$\mathcal{F}\left[B_{2,3}\right] = \mathsf{f}_2 \in \mathbb{N}$$

$$\vdots$$

$$\mathcal{F}\left[B_{N-1,N}\right] = \mathsf{f}_N \in \mathbb{N}$$

The higher its frequency, the higher the probability that the boundary under examination between two different blocks is physically meaningful - as opposed to those boundaries which strongly depend on the number of do-

mains Q and the local fluctuations of the system. It is possible to establish the most probable protein block division, considering only the high frequencies, assigning for each domain the high- or low-resolution, as shown in Section 4.3.

4.2.2 Dual-resolution model

In this work, we study the solvent treated with all-atom detail in two different cases: pure water and saline solution in physiological concentration. The protein, on the other hand, has a fixed (i.e., position- and time-independent) dual-resolution. The residues whose index is included in the range [11, 28] along the protein backbone are modelled in coarse-grained, while the remainder with atomistic resolution. This division has been obtained employing PiSQRD and the consequent boundaries frequency histogram, as shown in the section 4.3 (the theoretical part is illustrated in the Sec. 4.2.1).

To describe the lower-resolution part, first we employ the original ENM model [24, 36] in which each residue is mapped onto a bead whose position corresponds to the C_α atom. These beads are connected by two different harmonic springs, as shown in Fig. 4.2.

The potential energy is given by:

$$E = \sum_i \sum_j k_{ij} \left(r_{ij} - r_{ij}^0\right)^2 \theta(r_c - r_{ij}^0) \tag{4.6}$$

with spring constants k_{ij}, equilibrium distance r_{ij}^0, a cut-off distance r_c, i and j are the node index, and $\theta(r)$ is a Heaviside theta function taking value 1 if $r > 0$ and 0 oth-

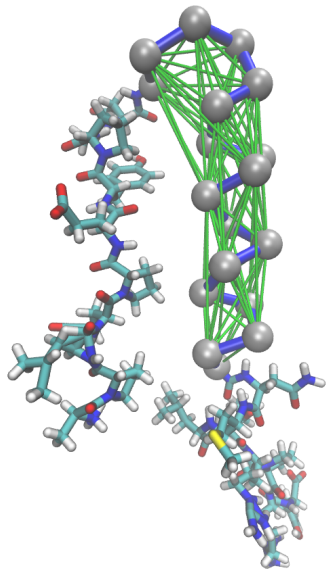


Figure 4.2: Visualization of the Bovine Pancreatic Polypeptide in dual-resolution. The residues included in atomistic detail are shown in red, blue, cyan and white (O, N, C and H atoms). The grey spheres are ENM nodes, the stiff backbone springs are shown as dark blue lines, and all others (weaker) springs are shown in green.

erwise. In this model we made use of two different elastic constants: a very stiff spring (k_b) for consecutive beads, represented in blue in Fig. 4.2, and a weaker spring k_{nb} for not consecutive beads whose distance in the reference (native) conformation lies below a fixed cutoff (in green).

The ENM used here, and specifically the value of elastic constant for non consecutive C_α beads, is parameterized to reproduce the conformational fluctuations of the reference all-atom model, these being quantified by the root mean square fluctuations (RMSF) of the all C_α carbons of the system [24].

The water-CG protein and ions-CG protein interactions consist in a simple excluded volume, modelled *via* a Weeks-Chandler-Anderson (WCA) potential [178]. Hereafter, we propose a refinement of the ENM in which a different elastic constant between beads is employed.

4.2.3 Different elastic constants in ENM

In the previous section, we have described the original formulation of the ENM [24, 36] and its application to a small protein test case. In the following, we propose an extension of this model parametrizing the spring constants on the basis of the distance distribution between the C_α beads involved in the coarse-grained part.

We are dealing with an elastic model whose beads are connected by harmonic springs; therefore in order to justify the harmonic behaviour, we study the C_α distances distribution in the fully-atomistic simulation: in particular, the only requirement is that the latter converges quickly at a Gaussian-like shape. This means that, taking into account

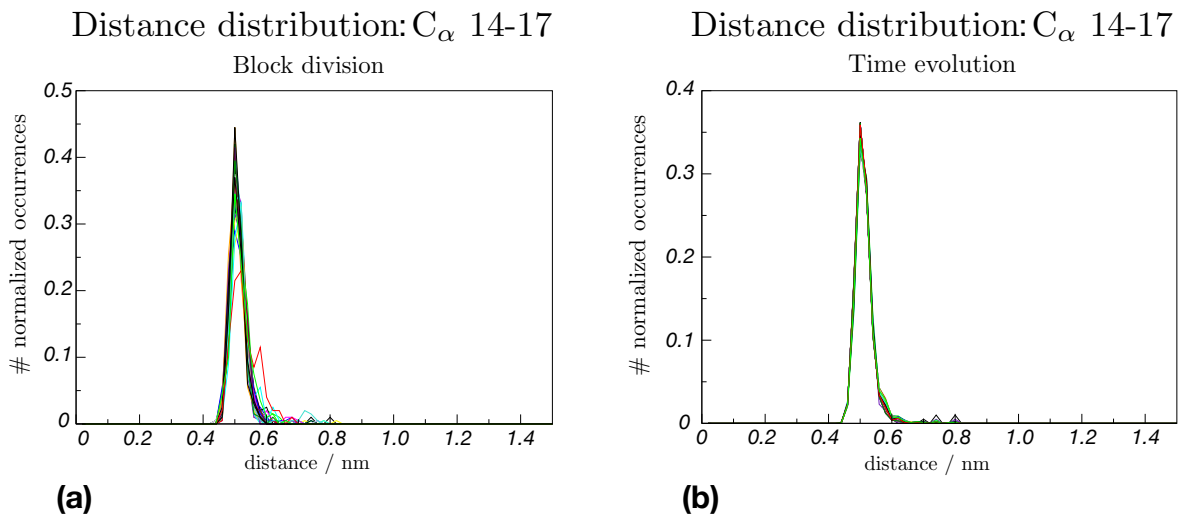


Figure 4.3: Distance distribution between the C_α carbons 14 and 17 of 1BBA in case of: (a) division of the entire atomistic trajectory in N shorter sub-trajectories of 10 ns each; (b) time evolution of the all-atom trajectory. In this particular example, we can notice that there is a perfect curves overlapping whose shape is a Gaussian-like.

two atoms A and B, two conditions have to be fulfilled:

- Divide the entire atomistic trajectory in N shorter sub-trajectories with the same length: for each of one, the distance distribution of A and B must present the same Gaussian curve (Fig. 4.3(a));
- consider the time evolution of the trajectory (increasing step by step its length until its entirety is reached): for each step, the distance distribution of A and B must show, once again, the same Gaussian curve (Fig. 4.3(b)).

Assuming that such distribution between the atoms A and B presents a harmonic behaviour, as shown in Fig. 4.3, it turns out that it can be fitted with a harmonic potential:

$$U = \frac{1}{2} K (d - d_0)^2 \tag{4.7}$$

where K is the elastic constant between A and B, while d_0

is their equilibrium distance.

Moreover, we can assume that the probability distribution calculated before is simply a Boltzmann distribution. Thus, the probability is given by:

$$\mathcal{P}(r) \propto e^{-\beta U} \tag{4.8}$$

where β is the reciprocal of the thermodynamic temperature of a system:

$$\beta = \frac{1}{K_B T} \tag{4.9}$$

with K_B corresponding at the Boltzmann constant and T is the temperature.

Substituting the Eq. 4.7 into the Eq. 4.8, it turns out that:

$$\mathcal{P}(r) \propto e^{-\frac{\beta}{2} K (d-d_0)^2} \tag{4.10}$$

The Eq. 4.10 shows that the probability distribution is a Gaussian, whose generic form is the following:

$$G(x) \propto e^{-\frac{1}{2\sigma^2}(x-x_0)^2} \tag{4.11}$$

where σ^2 and x_0 are, respectively, the variance and the mean value of the Gaussian.

Hence, comparing the Eqs. 4.10 and 4.11, it is possible to obtain the value of the elastic constant K:

$$K = \frac{1}{\beta \cdot \sigma^2} \tag{4.12}$$

In the Eq. 4.12 β is known, and it is constant since the temperature is given; whereas σ is extracted by the Gaus-

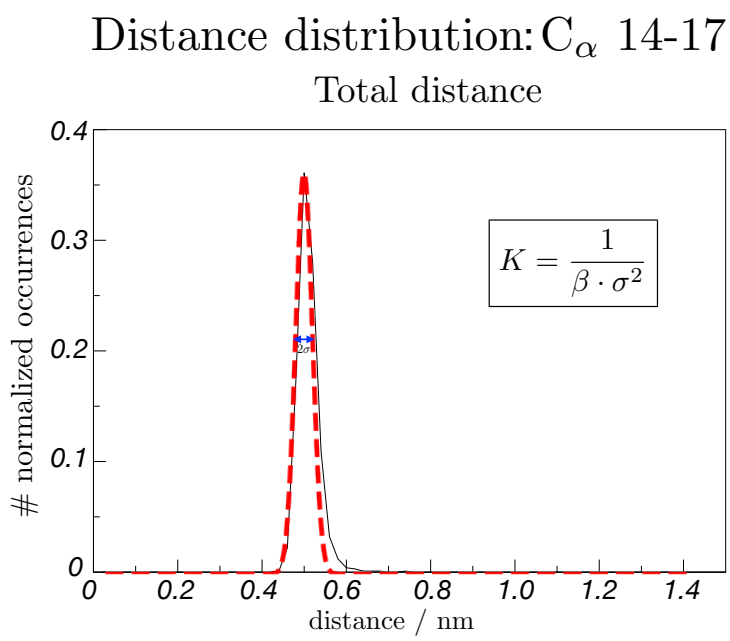


Figure 4.4: Distance distribution between the C_α carbons 14 and 17 of 1BBA (in black line) fitted (schematically) with a Gaussian (red line). The value of the elastic constant K is given by the Eq. 4.12. The β value is a constant since the temperature is known. On the other hand, σ is extracted by the Gaussian fit.

sian fit for each distance distribution that we take in account. Therefore, the value of the elastic constant in each case depends solely on the σ value obtained by the specific Gaussian fit.

Fig. 4.4, showing the distance distribution between C_α atoms 14 and 17 of 1BBA, illustrates schematically how to compute the value of elastic constant K.

4.2.4 Free energy landscapes

In order to compare the all-atom simulations and the dual-resolution ones, we choose two collective variables that describe the system under examination and, afterwards, analyze the result in terms of free energy landscapes. The procedure is shown in the following.

The first collective variable chosen for our purpose is the distance analysis of the two protein terminals, expressed in terms of distance between the first and last C_α carbons (red arrow of Fig. 4.5).

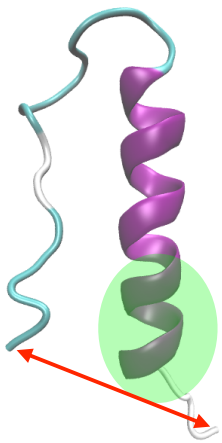


Figure 4.5: Visualization of the two collective variables chosen to describe the system. Specifically, the distance between the first and last C_α carbons is shown with a red arrow. On the other hand, the degree of unfolding of the protein in terms of RMSD of the C_α atoms $\in [29, 33]$ is schematically drawn with a green circle.

Moreover, another useful variable is the analysis of the unfolding degree of the α-helix analysis, described by means of the Root Mean Square Deviation (RMSD) of its last C_α carbons, calculated with respect to the native conformation which that presents a completely folded α-helix (green circle in Fig. 4.5). In particular, the C_α atoms taken in account correspond to the protein residue indices included between 29 and 33. The higher the RMSD, the higher the degree of unfolding the protein, namely the α-helix is the more unfolded.

After choosing the collective variables that describe our system, the analysis is undertaken by plotting the density of points (x and y) in terms of *2D-histogram*, which defines a probability $P(x, y)$. A further step can be done, calculating the free energy defined, here, as follows:

$$F(x, y) = -\ln(P(x, y)) \tag{4.13}$$

In general, a 2D histogram is represented as a heat map in which the colours of each surface is not random. In particular, the higher the probability, the more intense the colour: usually, it goes from red (P=0) to blue (P=1). Likewise, when constructing the heat-map of free energies, the latter range between 0 (P=1 and blue region) to infinity (P=0 and red colour) according to the Eq. 4.13.

The representation of the free energy in terms of a heat map has the advantage of emphasizing – applying the Eq. 4.13 directly – small probability variations. Indeed, it is possible to notice that two regions having a similar points frequency (namely normalized probabilities close each other) present, on the other hand, wide free energy

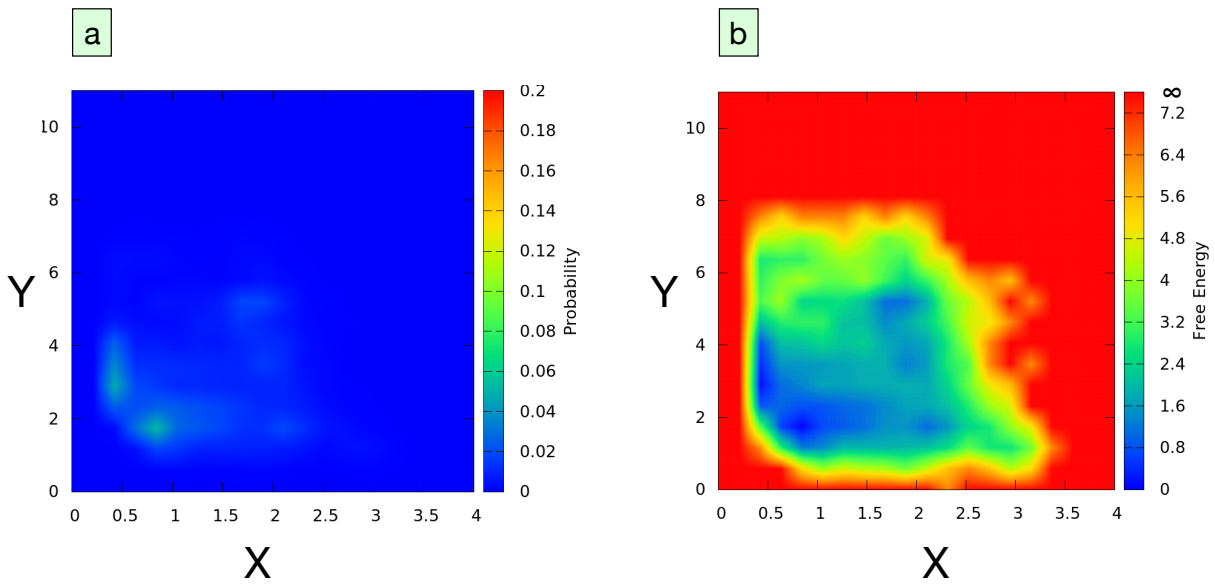


Figure 4.6: Comparison between the (a) heat-map treated in terms of point probability and (b) free energy, defined in the Eq. 4.13. The latter emphasizes small probability variation in terms of colour range scale. On the X- and Y-axis are reported the two collective variables chosen. Since, in this specific case, the free energy values range between 0 and 7.2, in red is shown, for simplicity, also the case in which it is ∞ corresponding thus at probability equals to 0. Moreover, figure (b) will be used later in the result section.

variations. In the Fig. 4.6(a) is reported an example of heat-map in terms of probability ranging between 0 and 0.1 The application of the Eq. 4.13, on the other hand, highlights much more details in terms of colour shades, as reported in Fig. 4.6(b).

4.2.5 Simulation details

The reference model is given by the 1 ns equilibrated PDB structure 1BBA in the NPT ensemble (the Parrinello-Rahman barostat [179] with a time constant of 2.0 ps and 1 bar was used) as shown Fig. 4.7. Both fully atomistic and dual-resolution models of 1BBA are solvated in water and placed in a cubic simulation box of 7.27 nm side. The force field employed is Amber99SB [77], whereas the water model is TIP3P [146]. The temperature is kept constant at 300 K by means of the *velocity-rescale* thermostat [83].

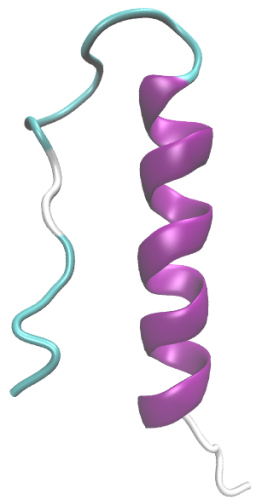


Figure 4.7: All-atom representation of Bovine Pancreatic Polypeptide (1BBA) after 1 ns equlibration in NPT ensamble, in terms of secondary structure. In particular, the α-helix is shown in purple color, while the β-turn are shown in cyan.

The integration step is 1 fs. The calculation of electrostatic interaction is performed using the reaction field method [201, 202] with a dielectric constant $\epsilon = 80$ and a cutoff of 1.2 nm. These parameters are a good compromise between speed and accuracy, as verified in Ref. [140]. The SETTLE [149] and RATTLE [150] algorithms for rigid water and rigid bonds to hydrogen have been used. Each system is prepared using fully atomistic minimization with the steepest descent and 1 ns of equilibration in NVT.

All-atom simulations, as well as fully atomistic equilibration ones, have been performed with the GROMACS simulation package[128]. On the other hand, dual-resolution-simulations have been carried out with the ESPResSo++ simulation package [129, 130]. The original tool has been employed for the first time in Fogarty et.al. work [24] and it is available upon request, included minor changes in the code for simulating a system that employs different elastic constants for each couple of bead in the ENM.

In the original formulation of the dual-resolution model the spring constant between consecutive C_α nodes along the backbone (k_b) has a stiff value of $5 \cdot 10^4$ kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}, whilst all the other ones (k_{nb}) have a value of 120 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}, until 1.2 nm as cutoff, parametrised by minimising the average root mean square error in the C_α RMSF.

On the other hand, considering the revised version of the dual-resolution model based on the distance distribution, the spring constant between consecutive C_α nodes along the backbone (k_b) has a stiff value of $5 \cdot 10^4$ kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}. All the other constants have a different value according to the resulting Gaussian fit: their wide range is between 100 and 5000 kJ $\cdot$ mol^{-1} $\cdot$ nm^{-2}, until 1.2 nm as cutoff.

In both dual-resolution models aforementioned, a WCA interaction is applied between C_α nodes and all solvent (including ions) molecules centre of mass. In the WCA potential, ϵ has a value of 0.34 kJ $\cdot$ mol^{-1} arbitrarily chosen as the value for carbon in the atomistic forcefield, and $\sigma_i = R_{g,i} \cdot c$ where $R_{g,i}$ is the radius of gyration of a given residue i, where c is the same for all amino acids. The value of c is tuned to give the correct bulk water density of reference for a protein-water system. The c value found is 1.15 for all cases. The procedure employed to compute such a value is the same as the one of the Supporting Information in Chapter 3.

We perform six different simulations because, on the one hand the solvent is pure water or with salt; on the other hand, the protein is treated in three different representations: all-atom, dual-resolution with a unique elastic constant in the CG-part, dual-resolution with different elastic constants between beads in the low-resolution part. The results are shown in the next section.

4.3 Results and discussion

As first, we performed the 1BBA division in blocks by means of the PiSQRD tool, increasing the number of imposed domains Q from 2 to 7. In this respect, the Fig. 4.8 shows with different colours the protein divided into blocks. On the other hand, the Fig. 4.9 illustrates the frequency within which each boundary appears increasing the number of Q; here we can also notice that some boundaries are present more often than other ones and therefore the probability that they exist is higher.

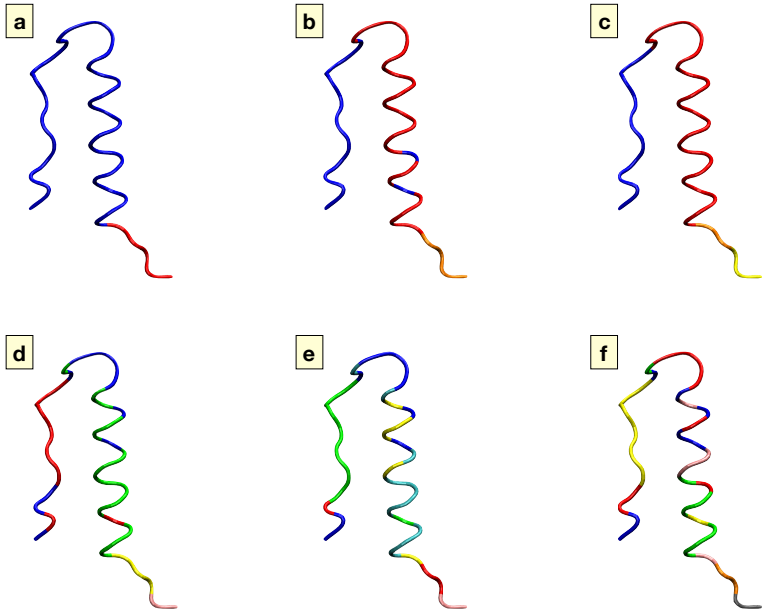


Figure 4.8: Representation of Bovine Pancreatic Polypeptide in (a) two, (b) three, (c) four, (d) five, (e) six, (f) seven blocks, by means of PiSQRD tool. In particular, each color represents a different domain.

The interfaces with higher frequencies are: $B_{10,11}$, $B_{11,12}$, $B_{27,28}$, $B_{28,29}$ and $B_{31,32}$, (indicated with the blue arrrow in Fig. 4.9), where B stands for "boundary", while the subscripts n and $n+1$ correspond at the residue index numbers. According with it, the protein under examination can be divided in four blocks:

- 1st block: 1-10
- 2nd block: 11-28
- 3rd block: 29-31
- 4th block: 32-36

Finally, each one requires the assignment: atomistic or coarse-grained. Since we expect that the α-helix fluctuates less than the remainder (and in particular the terminal part of the protein), as confirmed in the Ref. [199], a coarse-grained treatment of the 3rd block (which includes the most of α-helix residues) is preferable. One the other hand, 1st, 2nd and 4th block have been modelled with atomistic resolution.

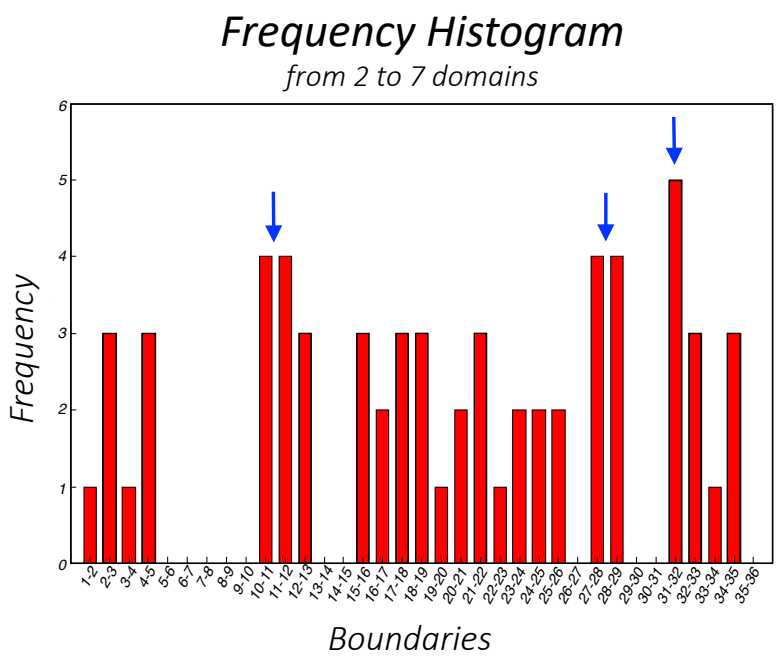


Figure 4.9: Frequency histogram for each boundary in 1BBA. The higher the bar is, the higher the probability is that the boundary under examination exists. On the x-axis are reported all each boundary $B_{n,n+1}$, while on the y-axis is shows the corresponding frequency. Moreover, with the blue arrows are reported the higher frequencies.

The pictorial representation of the block division is reported in Fig. 4.10, while the dual-resolution representation is reported on Fig. 4.2. The simulation analysis in terms of free energy landscapes is shown in Fig. 4.11.

The X- and Y- axes report the distance between the 1st and last C_α carbon and the RMSD of C_α atoms $\in [29, 33]$ residues indexes, respectively. The minima are coloured in blue, while in red we have the barriers corresponding to high value of free energies. All plots are constructed in such a way that the absolute minimum is zero in both cases allowing the best comparison among them.

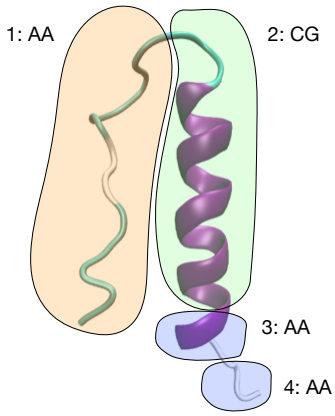


Figure 4.10: Schematic representation of 1BBA divided in 4 blocks. The 1st, 3rd and 4th blocks are labelled as all-atom (AA), while the 2nd one is labelled as coarse-grained (CG). The assignment is not random: preliminary considerations about the structure are needed.

First of all, starting from the all-atom simulation (Figs. **a** and **d**), it is possible to notice that the presence or absence of salt in water (with a 100 mM salt concentration) leads to completely different free-energy landscapes. In particular, in case of pure water (Fig. **a**), the presence of a tiny blue region means that the protein is trapped in one minimum corresponding at one unique configuration: stuck terminals (distance 1st-last C_α atoms between 0.5 and 1 Å) and

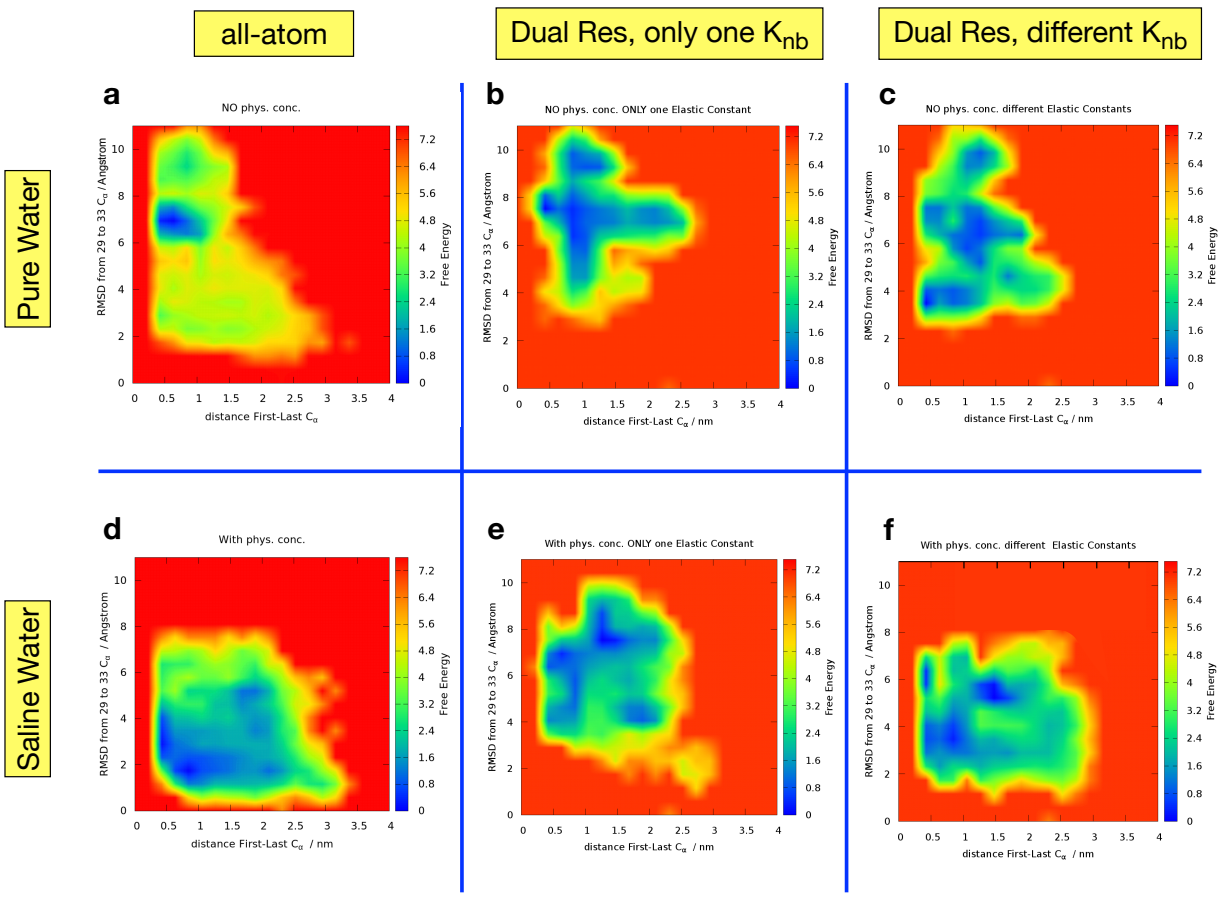


Figure 4.11: Free energy landscapes in six different cases after simulating for 500 ns, according to the presence or absence of salt in water, and the model used: all-atom or dual-resolution. On the x-axis is reported the distance 1st-last C_α, while on the y-axis is reported the RMSD from 29 to 33 C-alpha. Moreover, the FE value is shown with different colours in the unit of KT. All plots are shifted such that, for each one, the absolute minimum corresponds at 0 KT.

unfolded α-helix (RMSD of C_α indexes $\in [29, 33]$ about 7 Å). Contrarily, ion atoms presence in water (Fig. **d**), leads to a greater variability corresponding to the vast blue region. Indeed, the fluctuation of the two 1BBA terminals is very wide (distance 1st-last C_α between 0.5 and 3 Å); the α-helix, on the other hand, does not lose its structure for the most of simulation time (RMSD of C_α indexes $\in [29, 33]$ between 1 and 4 Å).

Looking at the analysis of free energy landscapes in case of the original formulation of the dual-resolution model (Figs. **b** and **e**), we can observe that, in case of pure wa-

ter (Fig. **b**), the blue region is wider with respect to the all-atom reference simulation, assuming diverse configurations not present in the reference all-atom simulation (Fig. **a**). However, the α-helix still retains an unfolded conformation (RMSD values between 6 and 10 Å), even though the two protein terminals are not stuck (the end-to-end distance is between 0.5 and 2.5 Å). In presence of salt (Fig. **e**), it is shown the same variability encountered in the reference all-atom simulation (Fig. **d**). However, the minima position is shifted towards higher RMSD values that lead to a higher probability of having an unfolded α-helix.

The analysis of the new version of the Dual-resolution model, by using different elastic constants between CG-beads (Figs. **c** and **f**) shows that in case of pure water (Fig. **c**) the FE landscape is analogous to Fig. **b**. Therefore the same considerations apply here: the α-helix keeps the unfolded conformation but the two terminals are not stuck, showing a variability not present in the all-atom reference (Fig. **a**). In presence of salt (Fig. **f**), it is possible to notice that the range of RMSD values explored by the two variables is more consistent with the reference all-atom simulation (Fig. **d**), although the minima positions are a little shifted.

Therefore, the results of the FE landscapes shows that the dual-resolution model works better in the case of water with salt, especially when using the new version with different elastic constants between CG-beads. In the case of pure water, the dual-resolution model explores more minima than the all-atom one. However, the overall "shape" of the free energy landscape is preserved: indeed it is possible to notice similar outlines in between all-atom simulation and dual-resolution one results.

4.4 Conclusions

In this work we have shown how the dual resolution model employed, constituted by an all-atom sub-region coupled to an elastic network remainder, can be used to catch the overall conformational variability of a small non-globular protein. However, it is not always easy to find a priori (as instead happens, for instance, for a ligand-enzyme system) the boundaries between the coarse-grained part and the full-atomistic one. In this respect, the block division in quasi-rigid domains is a useful tool capable of catching the fraction of internal motion after dividing the system in Q imposed domain.

Moreover, the coarse-grained part treated in Elastic Network Model has been refined by using different elastic constants between CG-beads, on the basis of the distance distribution between the corresponding C_α's in the atomistic simulation. This new strategy has led to an improvement in terms of FE-landscapes in comparison with the all-atom simulation. The FE profiles have been constructed after choosing two collective variables that well describe the system: the *end-to-end* distance and the degree of unfolding of the α-helix. Specifically, the all-atom simulation shows that the presence or the absence of salt in water leads to different dynamical properties: in the former case, the protein is fixed in one configuration consisting of unfolded α-helix and stuck terminals; in the latter case, on the contrary, there is more variability. However, the presence of salt stabilizes the α-helical structure in 1BBA.

At first sight, both dual-resolution models do not reproduce exactly the FE landscape in case of pure water showing a wider variability. On the other hand, the latter

is well reproduced in case of water with salt, especially when using the refined version of the model.

This is in line with what we expected since the protein used for our test is very small and very little *modular* from a dynamical point of view: it essentially fluctuates around its reference structure with wider terminals movements. Indeed, our model reproduces such fluctuations, as shown in the resulting free energy landscapes. However, the study of this system has had two prominent outcomes, that is:

- the validation of the usage of quasi-rigid domain division applied in the definition of the dual-resolution model;
- the validation of a dual resolution model with elastic constants between CG-beads.

The consequence of these observations provide useful practical and conceptual tools for the construction of more accurate dual-resolution models of larger proteins than the one studied here.

5 Simulating Adenylate Kinase through a Variable-Resolution model

This chapter is a draft of research paper that will be submitted in 2020.

The previous two chapters illustrate the dual-resolution model applied to the problem of ligand-binding in enzymes – specifically to the computation of binding free energy –, and to a small protein to investigate its dynamical properties in terms of FE landscapes. In this model, the more coarse-grained part is described as an ENM, in which each residue is mapped onto a bead whose position corresponds to the C_α atom in the atomistic description. Specifically, in the last chapter it was shown that a refinement of the ENM, considering different elastic constants for each couple of beads based on their distance distribution in the all-atom representation, can improve the accuracy of the model.

Here, we propose a novel multi-resolution scheme dubbed coarse-grained anisotropic network model for variable resolution simulations, or CANVAS. Its name is due to the fact that it allows to smoothly couple virtually any desired degrees of coarse-graining within the same model.

The model is here introduced, described in detail, and validated on a relatively small yet conformationally quite variable protein, adenylate kinase, a phosphotransferase enzyme that controls the energy balance in cells by catalyzing the interconversion of adenine nucleotides. The purpose of this work is to characterise the model's performance, advantages, and limits, and to identify possible modifications to improve its accuracy based on these preliminary results.

5.1 Introduction

Simulating bio-molecular systems is a particularly challenging task because of their structural and conformational heterogeneity, and the wide range of time and length scales they encompass. All-atom models provide the most accurate results compatibly with the limitations implicit in the parametrisation of the currently available force fields, starting from the fact that they provide a classical approximation to interaction of intrinsically quantum nature (e.g. Van der Waals dispersion interactions). However, the usage of a fully-atomistic representation usually containing all necessary physical and chemical detail characterising a given phenomenon is usually computationally expensive.

A promising way of mitigating the computational overhead is to employ a concurrent multi-resolution approach. This requires to identify the parts of the system where the physical and chemical details play a crucial role in the phenomenon of interest and describing them using a high-resolution model (all-atom) while using a less detailed, computationally more efficient model for the remainder of

the system.

In their simplest form, concurrent multi-resolution approaches employ, for each system component, a single level of resolution: one of the most known examples is simulating an atomistic protein in a coarse-grained solvent or embedded in a coarse-grained membrane [187–189]. On the other hand, when the purpose is to create a model apt to include the minimum number of degrees of freedom, one has to be able to place boundaries between resolutions within the system under examination.

One of the most known examples is the quantum mechanic/molecular mechanic (QM/MM) method [109–112]. It allows a connection between *ab initio* resolution and classical all-atom models. In particular, in a small domain forces acting on atoms are obtained through quantum calculations, while in the rest classical atomistic force fields are employed. Such a scheme is widely used in studying enzymatic chemical reactions [113, 114].

Another class of multi-resolution schemes focuses on the connection between atomistic and CG models simultaneously [115–118]. In practice, this idea lies in a smooth spatial interpolation on the atomistic and CG force field: a very popular technique is the Adaptive Resolution Scheme (AdResS) [115].

Here, we propose a novel multi-resolution approach which allows one to model at an atomistic resolution only the precise subset of degrees of freedom really necessary for the study of a given phenomenon, even when this leads to a boundary between resolutions which falls within a biomolecule. Furthermore, the model hereafter described allows one to set the level of resolution of the coarse-

grained subdomain(s) in a quasi-continuous range, spanning from the all-atom level to a degree of coarsening higher than one bead per amino acid. We emphasize that this property is the novelty of the method: in particular, one has an extremely high freedom in the choice of the level of coarse-graining. Furthemore, the interaction network on the basis of the model chosen has an automatic construction.

In the CANVAS model, the lower-resolution part is described with a new approach that of substantially differs from that Neri et.al. (who developed in 2005 a model in which an atomistically detailed active site was incorporated into a coarse-grained Gō model) [160] or from the ENM presented by Tirion in 1996 [36] and employed in Ref. [24] and in the chapters 3 and 4. In particular, two aspects assume particular prominence:

- Each bead has average properties depending on the atoms it represents: specifically, they correspond to the parameters used for the non-bonded interactions.
- Among coarse-grained beads, as well as between atomistic and coarse-grained beads, bonded interactions are placed which preserve the overall structure of the molecule; however, at odds with conventional ENM's, these bonds are not established based on the distance between particles in the reference structure, rather a different criterion, relying on a Voronoi-like subdivision of the system, is employed. Therefore, we also point out that the dependence on a cutoff is also missing is this model.

The model is then validated by comparison with reference atomistic simulations to the realistic case of a protein,

Adenylate Kinase (PBD code 4AKE) [203], which features a substantial degree of structural variability as well as conformational transitions. It is also known as **ADK** or **myokinase** and it is a phosphotransferase enzyme that catalyzes the interconversion of adenine nucleotides (ATP, ADP, and AMP)*. A network of adenylate kinase isoforms are distributed throughout intracellular compartments, interstitial space and body fluids to regulate energetic and metabolic signaling circuits, securing efficient cell energy economy, signal communication and stress response [204].

This novel multi-resolution approach, not only will lead to greater computational efficiency via a reduction in the number of degrees of freedom simulated; it will also allow the simulation of large biomolecular systems where the detailed atomistic structure is not known everywhere. Specifically, a lower accuracy in the structure is sufficient since the interactions are mediated on bigger regions of the protein independently on the precise position of the atoms within a Voronoi cell (with advantages for those proteins without experimental structure, for which one has to resort to homology modeling).

5.2 Methodology

In order to validate the novel multi-resolution scheme we have chosen the protein Adenylate Kinase in aqueous

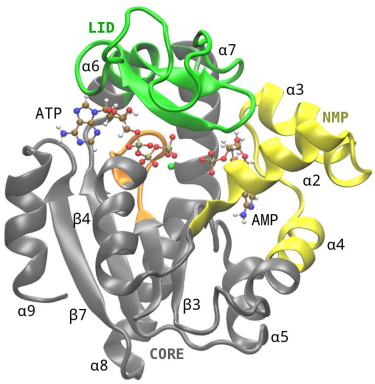


Figure 5.1: Structure of Adenylate Kinase: LID, NMP, and CORE. Adapted from Ref. [205]

* Phosphotransferases are a category of enzymes that catalyze phosphorylation reactions, that is the attachment of a phosphoryl group to a molecule. The general form of the reactions they catalyze is:

$$A\text{-}P + B \leftrightharpoons B\text{-}P + A$$

where P is a phosphate group and A and B are the donating and accepting molecules, respectively.

solution [203]. The latter has three domains, called CORE, LID and NMP and two distinct binding sites as shown in Fig. 5.1.

The LID and NMP domains are colored green and yellow, respectively. The CORE domain is colored in gray and the P-loop is colored in orange. Specifically, ATP, that is complexed with Mg^{2+}, is bound between the CORE and LID domains, in the so called ATP binding site, while AMP is sandwiched between CORE and NMP, in the AMP binding site. ATP and AMP ligands are represented as ball and stick.

The model employed in this work is not adaptive, that is, the resolution of a given amino acid is fixed (fully-atomistic or coarse-grained), and it does not change during the simulation. This means that the level of detail does not change during the simulation. Specifically, the coarse-grained model used to describe the low-resolution part of the protein consists of beads and springs. The selected sites (e.g. C_α carbons or other heavy atoms) have averaged properties depending on the group of atoms they represent and they are connected by harmonic bonds. Contrarily with the case of the ENM, in which each residue is usually mapped onto its C_α in the proposed model the mapping is not uniform, rather it depends on the neighbourhood of the atom selected as CG bead, as illustrated in Sec. 5.2.2.

Water molecules are modelled in atomistic details inside the simulation box. The interaction with the high-resolution part of the protein takes place through the standard all-atom force field. In the coarse-grained part we emphasise that bonded interaction connect CG sites, while standard non-bonded interactions (*Coulomb* and *Van der*

Waals) with appropriate values take place between solvent and CG beads. A more detailed description can be found in **"Properties of the Model"** in Sec. 5.2.2.

Hereafter, we provide a detailed description of the model illustrating, first, the Voronoi Tessellation on which the protein division in the low-resolution part is based, and second the actual model including the non-bonded and the bonded-interactions. Finally, we provide details about the simulation setup.

5.2.1 Voronoi Tessellation

One of the key aspects of this model is the block division in the coarse-grained part of the bio-molecule under examination. The method chosen is known as *Voronoi Tessellation* or *Voronoi Diagram*.

Generalities

In general, in mathematics, a *Voronoi diagram* is a partition of a plane into regions close to each of a given set of objects. In the simplest case, these objects are just finitely many points in the plane (called seeds, sites, or generators). Let us consider an example in two dimensions: after choosing a square region S, we define P_k special points (seeds) with $k \in \mathbb{N}$ as reported in Fig. 5.2(a). It turns out that, for each seed, there is a corresponding region consisting of all points ($x \in S$) closer to that seed than to any other. These regions are called *Voronoi cells* R_k (Fig. 5.2(b)).

The Voronoi diagram is named after the Russian mathematician Georgy Voronoy and is also called a Voronoi tessellation, a Voronoi decomposition, a Voronoi partition, or

a Dirichlet tessellation (after Peter Gustav Lejeune Dirichlet). Voronoi cells are also known as Thiessen polygons [206–208]. Voronoi diagrams have practical and theoretical applications in many fields, mainly in science and technology, but also in visual art [209, 210].

Formal definition

Mathematically, a Voronoi cell can be defined as follows:

$$R_k = \{x \in S \mid d(x, P_k) \leq d(x, P_j) \text{ for all } j \neq k\} \tag{5.1}$$

where $d(x, P_k)$ is the distance between the generic point x and the special point P_k.

Eq. 5.1 is always valid, because no conditions are given on the region S and on the distance d. In our example of Fig. 5.2(b), S is a square in 2D, whereas d is the *Euclidean distance*, employed in the CANVAS model described in Sec. 5.2.2, defined as:

$$\ell = d\,[(a_1, a_2), (b_1, b_2)] = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \tag{5.2}$$

For the sake of completeness, another metric that can be used (but not adopted in our model) is the **Manhattan distance** as reported in Fig. 5.2(c). The result is that the corresponding Voronoi diagram looks different:

$$L_2 = d\,[(a_1, a_2), (b_1, b_2)] = \mid a_1 - b_1 \mid + \mid a_2 - b_2 \mid \tag{5.3}$$

Moreover, according to Eq. 5.1, a Voronoi diagram can be intepreted as a list of Voronoi cells R_k.

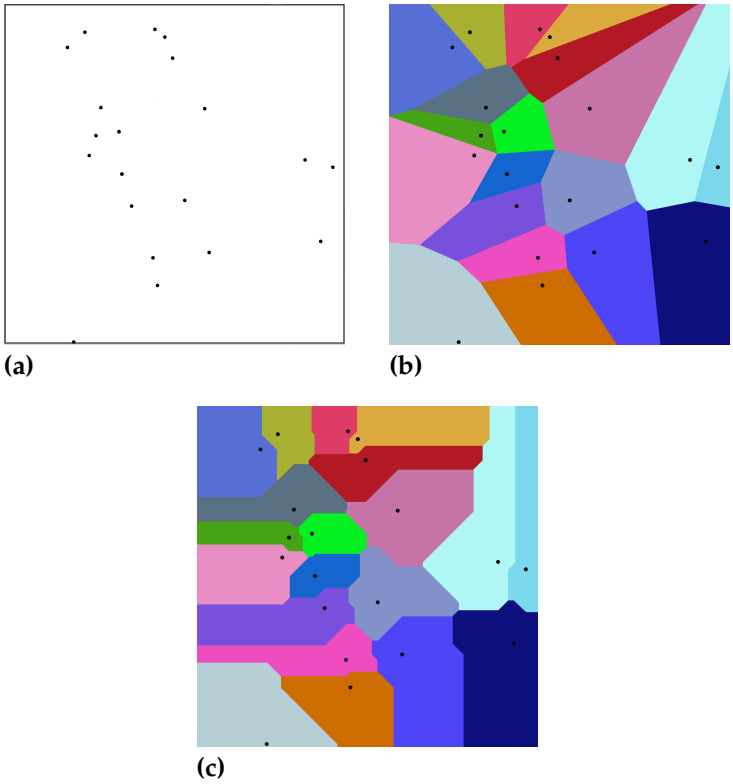


Figure 5.2: (a) Square region with 20 seeds (represented by black points) chosen randomly inside the figure. (b) Division of square region in 20 blocks with the euclidean distance as metric. Each of them is shown with a different colour according to the general rules of Voronoi Tessellation. Adapted from Wikipedia. (c) Division of Square region in 20 blocks with the Manhattan distance as metric. Adapted from Wikipedia.

5.2.2 The CANVAS model

In this work, the solvent is treated with all-atom detail, while the protein has a fixed (i.e. position- and time-independent) Variable-Resolution.

Model construction

Since we are dealing with a multiple/variable-resolution approach, we identify the region of the system where the chemical details play a crucial role, such that no simplification of the description is feasible: thus, it requires a high-resolution description; the remainder on the other hand, allows a lower resolution treatment.

In our approach, the high-resolution part is modelled fully atomistically, where the classical bonded and non-bonded interaction between atoms are employed.

To describe the lower resolution part, on the other hand, we identify the atoms that *survive* (red circles in Fig. 5.3(a)), that is, those atoms that will be treated as CG beads. According to Voronoi Tessellation rules, the decimated atoms are mapped onto the closest (in terms of euclidean distance ℓ defined in Eq. 5.2) survived atom (Fig. 5.3(b)), and the coarse-grained region of the protein is divided into blocks, each one represented by the reference atom chosen (Fig. 5.3(c)). Then, the CG beads are connected by harmonic springs as schematically reported in Fig. 5.3(d).

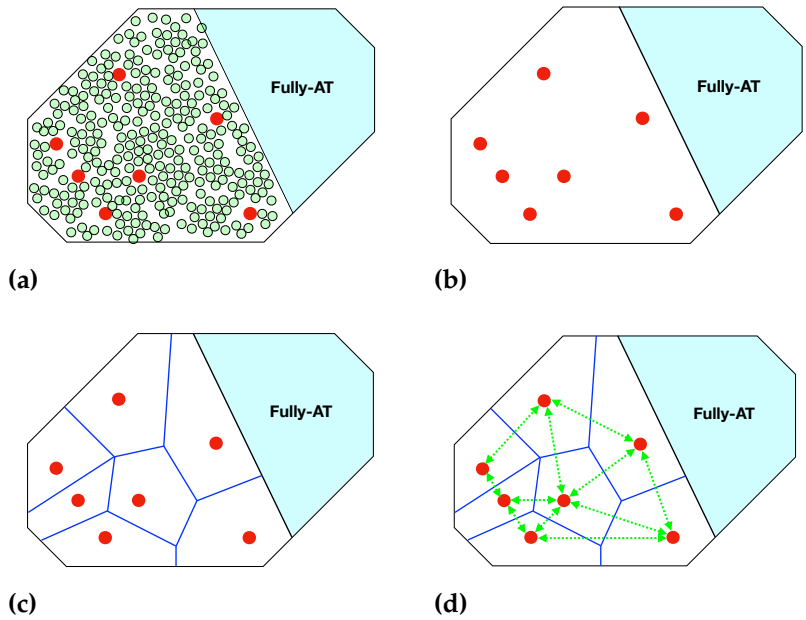


Figure 5.3: (a) Schematic all-atom representation of a generic protein. The green circles show the generic atoms, while the red ones are the atoms the survive in the white area. The cyan region, on the other hand, keeps the fully-atomistic representation; (b) schematic representation of the protein divided in fully-atomistic (cyan colour background) and coarse-grained with *only* the atoms that survive (red circles). (c) Division of the coarse-grained part of protein according with Voronoi Tessellation. Each survived atom is representative of a region that encloses the closest not-survived atoms mapped by it. (d) Addition of harmonic springs schematically represented with green arrows between survived atoms that belong to adjacent regions.

The potential energy is given by:

$$E = \sum_i \sum_j k_{ij} \left(r_{ij} - r_{ij}^0 \right)^2 \cdot \theta \left(\text{adj, tetrahedral} \right) \tag{5.4}$$

with spring constants k_{ij} equilibrium distance r_{ij}^0; i and j are the node index, and θ is a Heaviside theta function taking value 1 if i and j belong to adjacent region of

Voronoi Tessellation or they satisfy the *tetrahedral condition* described below, 0 otherwise.

Indeed, the insertion of harmonic springs between beads belonging to adjacent regions is not sufficient to keep the coarse-grained network stable; in fact, the system can collapse: the reason stems from the fact that the only adjacency criterion can lead the system to having null modes, that is movements with no energy cost, that make it unstable. For instance, in Fig. 5.4(a) the rotational movement around the axis A_3-A_4 allows the other atoms to move freely; in order to fix the network it is necessary to introduce a further bond between A_2 and A_5, as shown in Fig. 5.4(b).

Therefore, to avoid artifacts in the entire system with the consequent breakdown in the long run, further springs must be added to fix all CG beads: this can be carried out by requiring that each one occupies the vertex of a tetrahedron whose size is given by the harmonic bonds. Each bead has to be connected, at least, to three other particles, these in turn being connected to one another. Fig. 5.4(b) graphically illustrates the concept.

In this model we made use only one elastic constant k_b represented in green as shown in Fig. 5.5.

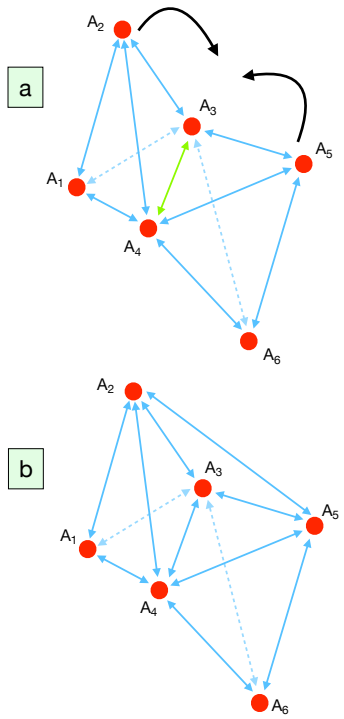


Figure 5.4: (a) shows a CG network: it is unstable as the rotational movement around the axis A_3-A_4 (in green) allows the other atoms (red circles labelled with the letter A_i) to move freely (black arrows); (b) the CG network is fixed adding a further spring between the atoms A_2 and A_3. The figure shows three tetrahedrons and each bead satisfies the tetrahedral condition.

How to choose the survived atoms

The method does not constrain in any way the choice of the survived atoms: indeed, it may be performed in several ways; two natural choices are the following:

- keep only the C_α and/or other heavy atoms of the residues modelled in a lower resolution.
- compute the internal deformation in each group of

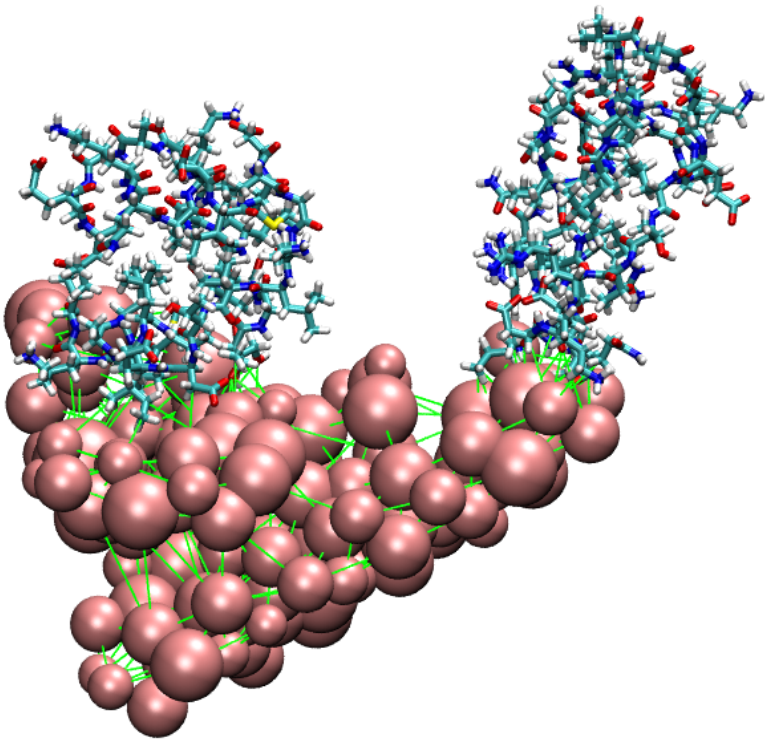


Figure 5.5: Visualization of Adenylate Kinase used in this work in Variable Resolution. The residues included in atomistic detail are shown in red, blue, cyan and white (O, N, C and H atoms). The pink spheres are the CG beads: the different radius size is due to the variable model used in coarse-grained part. Finally, the springs between beads belonging at adjacent Voronoi cells are shown in green. The radius of each bead is given by the *radius of gyration* of the entire block at which it belongs

atoms such that it is as low as possible. In other words, given a set of survived atoms, partition the system in Voronoi blocks according to these atoms, compute the fraction of internal motion (e.g. as in Ref. [106]), and perform a search for the subset of survived atoms which minimise the latter.

The mapping $C_\alpha \mapsto$ *entire residue* is the most common choice when coarse-graining proteins. Here we validate our approach in analogy with that mapping, by using the C_α atoms as representatives of the CG beads (that do not necessarily correspond to separate residues); however, we emphasise that the CANVAS model puts no restraints on the selection of survived atoms, neither by number nor by type. This strategy guarantees the highest level of freedom and system-specificity in the construction of a multiple-resolution model.

Properties of the Model

The atoms that survive in the coarse-grained part have average properties of the atoms they represent.

Specifically, for each Voronoi cell, as shown in Fig. 5.6, the reference CG bead has the following properties:

- The charge Q_{block} that it assumes is the algebraic sum of the charges q_i of the atoms it represents.
- The dimension of the block σ_{block} is twice the gyration radius R_g
- ϵ_{block} is the geometric average of the ϵ_i of the atoms it represents.

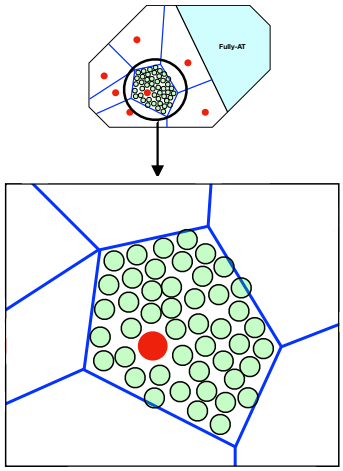


Figure 5.6: Visualization of a generic Voronoi cell in a protein: the red circle is the representative survived atom in the CG representation, while the green circles correspond at the not-survived atoms mapped, originally present in that domain in all-atom representation.

Mathematically, it turns out that:

$$Q_{\text{block}} = \sum_{i=1}^{N} q_i \tag{5.5}$$

$$\epsilon_{\text{block}} = \prod_{i=1}^{N} \epsilon_i^{\frac{1}{N}} \tag{5.6}$$

$$\sigma_{\text{block}} = 2 \cdot R_g \tag{5.7}$$

In the Eq. 5.7, R_g is gyration radius of the group of atoms under examination, and it is defined as:

$$R_g^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} |\mathbf{r}_i - \mathbf{r}_{\text{com}}|^2 \tag{5.8}$$

where, $\mathbf{r}_i$ are the coordinates of each atom inside the block, whereas $\mathbf{r}_{\text{com}}$ are the coordinates of the center of the mass of the atoms mapped in the all-atom representation, i.e. $\mathbf{r}_{\text{com}} = \frac{1}{N} \sum_i \mathbf{r}_i$.

Moreover, depending on the number and which atoms

survive, the size of each block is different: bigger and smaller domains are possible (Fig. 5.2(b) and Fig. 5.3(c)). The bigger the block is, the more coarse-grained the region is and, thus, more atoms are mapped onto a single bead. The extreme case occurs when the region consists of only one atom. In such a case, the survived atom is not a CG bead, but it conserves its own atomistic properties: specifically, the Eqs. 5.5 and 5.6 are still valid. On the other hand, The Eq. 5.7, is not fulfilled anymore: in this specific case, we take in account the real value of σ from non-bonded parameters of the force field used.

Furthermore, since the atoms representative of the CG beads are not necessarily C_α atoms, if two of them have a covalent bond in the all-atom representation, the latter replaces the aforementioned harmonic spring with constant k_b. In the same way, we also keep bending and torsion potential with their original values of angle and energy only if, respectively, the triplet and quadruplet of atoms in the all-atom representation are maintained in the coarse-grained model.

The parametrization of this model thus enables a quasi-continuous modulation of the resolution of a protein or part of it, in that the detail of the representation can be gradually reduced from the all-atom level to a very coarse one, possibly lower than a few amino acids per bead.

The water-CG protein interactions consist of a simple Coulomb and Van der Waals from the standard force-field parameters. WCA potential, usually used in other works [24] to avoid solvent penetrating in the CG network is unnecessary here, since each bead has a different σ_{block} representing twice the radius of gyration of the entire block;

thus the water molecules cannot go through the network.

5.2.3 Simulation details

The reference model is given by the 125 ns equilibrated PDB structure 4AKE in NPT ensemble (the Parrinello-Rahamn barostat[179] with a time constant of 2.0 ps a 1 bar was used). The conformation of the molecule after equilibration is reported in Fig. 5.7.

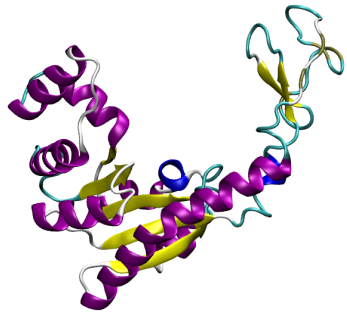


Figure 5.7: Fully atomistic representation of Adenylate Kinase (4AKE) after 125 ns equilibration in NPT ensemble in terms of secondary structure.

Both fully atomistic and CANVAS models of 4AKE, after the aforementioned equilibration, are solvated in water and placed in a cubic box of 9.41 nm side. The force field employed is Amber99SB-ildn [77], whereas the water model is TIP3P [146]. The temperature is kept constant at 300 K by means of the Velocity Rescale Thermostat [83]. The integration step is 1 fs. We performed two tests using different methods to calculate electrostatic interactions:

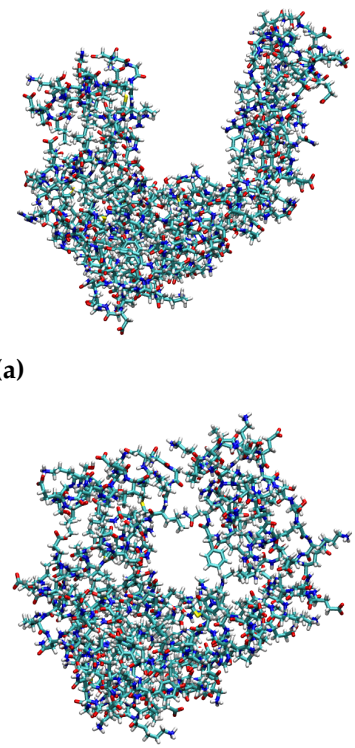
- the *reaction field method* [201, 202] with a dielectric constant $\epsilon = 80$ and a cutoff equals to $2.5 \cdot \sigma_{\text{max}} = 1.79$ nm both for fully-at and Variable Resolution. Since each block has a different value of σ_{block} (as reported in Sec. 5.2.2 and in particular in the Eq. 5.7), σ_{max} is the maximum of them;
- the *Particle Mesh Ewald method* [79] with fourth-order (cubic) interpolation and this Fourier spacing of 0.16. The cutoff is given by $2.5 \cdot \sigma_{\text{max}} = 1.79$ nm, as for the reaction-field method.

All simulations (atomistic and in CANVAS) are 500 ns long and were performed in GROMACS 2019 [211]. Moreover, the SETTLE [149] and RATTLE [150] algorithms for rigid water and rigid bonds to hydrogen have been used.

Each system is prepared, starting from the already equilibrated structure, using fully atomistic minimisation with the steepest descent and 50 ns of equilibration in NVT. In the Variable Resolution model, the spring constant between atoms in which at least a CG bead is involved has a stiff value of $5 \cdot 10^4$ kJ $\cdot$ mol$^{-1}\cdot$ nm^{-2}. Moreover, The interaction parameters for the CG part have been set through Eqs. 5.5-5.7.

5.3 Result and Discussion

The aim of this section is to compare results from the atomistic and multiscale simulations for the validation of the model.



(a)

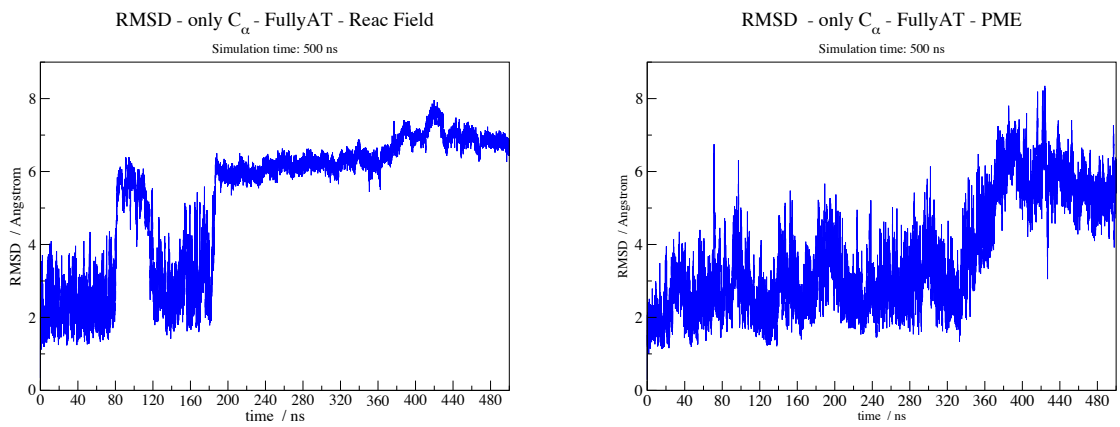
(b)

Figure 5.8: Fully atomistic representation of 4AKE in terms of primary structure. (a) shows the open conformation of the protein, while (b) the closed one.

5.3.1 All-atom simulation

First, we performed the all-atom simulation of 4AKE in explicit water employing Reaction-Field (RF) and Particle Mesh Ewald (PME) electrostatic. In both cases, the simulation time is 500 ns, and it shows clearly two main protein conformations: the open one, as reported in Fig. 5.8(a), and the closed one shown in Fig. 5.8(b)*. Intuitively, the former structure has the protein arms distant from other; on the other hand, in the closed conformation, the two arms adhere: specifically, the protein residue indexes 30-70 and 125-150 get close, while the remainder residues making up the protein residues maintain approximately their position in both conformations described above.

* For simplicity we call such conformation "closed"; however one has to be careful to name it that, since the real, crystallized closed conformation is a bit different (and we do not see it in our unbiased simulations)



(a) **(b)**

Figure 5.9: RMSD of all 4AKE C_α in case the electrostatic used is (a) reaction field and (b) Particle Mesh Ewald (PME). The simulation time is 500 ns. On the x-axis is reported the time evolution, while on the y-axis is represented the corresponding RMSD values. The presence of two different states, one corresponding at about 3 Å and the second one close to 6 Å are indicative of opened and closed structures, respectively.

The evolution of the two structures mentioned above can be visualised and quantified by calculating the Root Mean Square Deviation (RMSD) of the all protein C_α atoms with respect to the reference frame (reported in Fig. 5.7). Since the latter presents an open conformation, higher RMSD values are indicative of closer structures. The resulting plots are shown in Fig. 5.9; in particular, (a) corresponds to the simulation employing the reaction-field electrostatic, while (b) shows the same calculation when using the PME method.

As expected, two states are well visible in both plots: one corresponding to 3 Å, and the second one around 6 Å. Specifically, by looking at Fig. 5.9(a), corresponding to reaction-field method, the closed conformation (the state with higher RMSD values) appears for few nanoseconds after 80 ns, and subsequently after 200 ns, remaining as such until the end of the simulation. Contrarily, when using PME (Fig. 5.9(b)), the open conformation persists for 350 ns, until the closed one takes hold in the last part of the

simulation.

In support of the previous analysis, Fig. 5.10 shows the Root Mean Square Fluctuation (RMSF) for each C_α when using reaction-field (black line) and PME (red line) (Look the *Appendix* for further details about RMSD and RMSF).

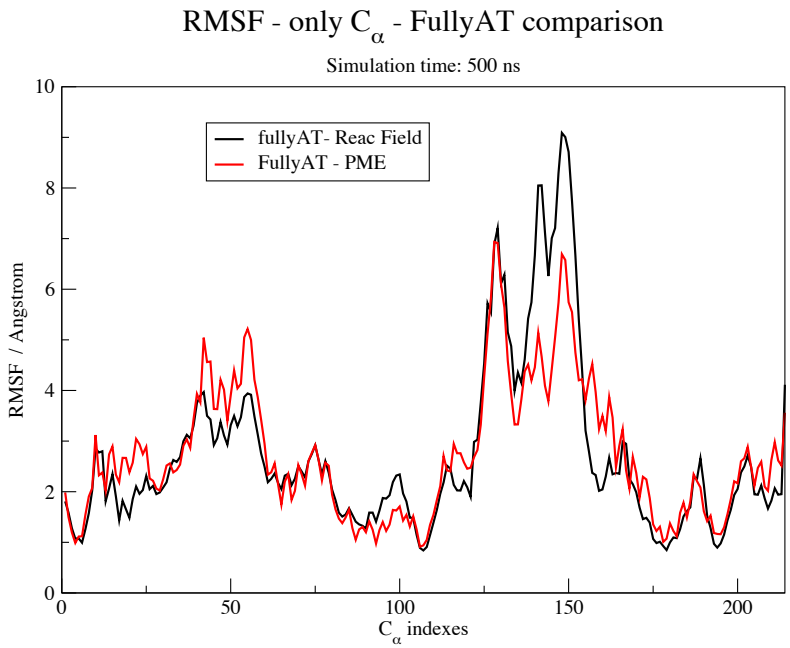
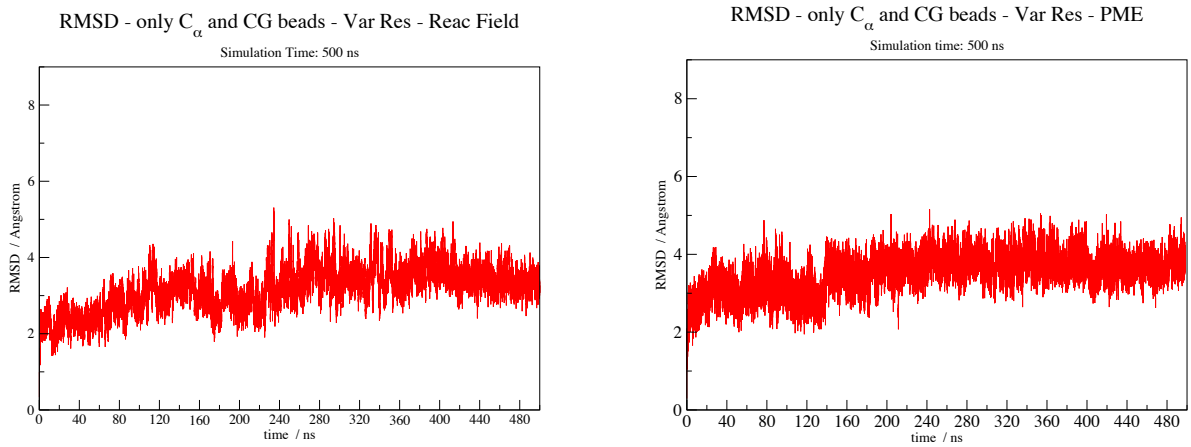


Figure 5.10: Root Mean Square Fluctuation (RMSF) for each C_α of 4AKE in case of Reaction-Field (black line) and PME (red line). The x-axis corresponds at the C_α indexes (from 1 to 214). The highest value of RMSF is in correspondence of the two protein arms (indexes 30-70 and 125-150), as expected, because they move more with respect the remainder of protein. The two plots are comparable since the fluctuation of all C_α are close each other.

Also the RMSF is computed with respect to the same reference frame: we can notice that the atoms constituting the protein arms (indexes 30-70 and 125-150, left and right protein arm, respectively) have wider fluctuations with respect to the remainder, namely the hinge. Hence, their relative orientation determines the *open-closed* conformation.

5.3.2 CANVAS simulations

As the all-atom simulations show that the two protein arms present more fluctuations with respect to the remainder of the protein, they are treated in high-resolution detail, while the hinge atoms are modelled in low-resolution.



(a) **(b)**

Figure 5.11: RMSD of multiscale resolution of 4AKE C_α (in the coarse-grained part, CG beads in the same position of C_α in all-atom simulation are taken in account) in case the electrostatic used is (a) reaction field and (b) PME. The simulation time is 500 ns. On the x-axis is reported the time evolution, while on the y-axis is represented the corresponding RMSD values. Only one state present corresponding at about 3-4 Å is indicative of the opened structure persistent for the entire simulation.

As explained in the Sec. 5.2.2, the coarse-grained part is described by keeping only the C_α atom of each residue. Moreover, the values of Q, σ and ϵ for each bead are different depending on the number and type of the atoms it maps. We remind, also, that each bead is connected to its neighbours with harmonic springs.

We performed the CANVAS simulation of 4AKE (as shown in shown in Fig. 5.5) in explicit water employing Reaction-Field (RF) and Particle Mesh Ewald (PME) electrostatics. In both cases it presents only one main protein conformation (corresponding at 3 Å), namely the opened one as illustrated in Fig. 5.11. This first result is in contrast with the fully-atomistic simulation that shows two main structures.

To gain further insight in this feature, we looked into the fluctuations of each protein C_α in the all-atom part and each bead in the coarse-grained one (whose position is the same of the corresponding C_α atoms in all-atom repre-

sentation), as displayed in Fig. 5.12. Specifically, the green area is useful to distinguish the fully-atomistic part and the coarse-grained one. Indeed, the RMSF values included in the green region correspond to the CG bead fluctuations, while the white ones show the fluctuations of C_α in the all-atom representation.

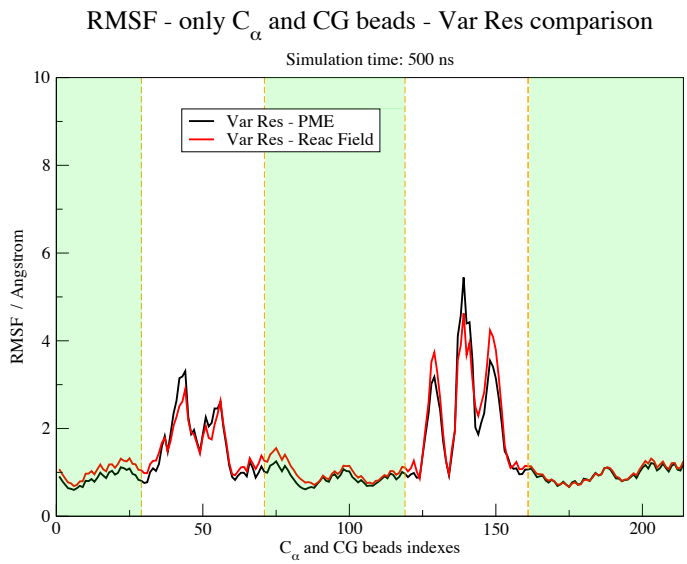


Figure 5.12: Root Mean Square Fluctuation (RMSF) for each C_α and the corresponding CG bead in the Coarse-Grained part, in case of Reaction-Field (black line) and PME (red line). The x-axis shows the C_α and CG beads indexes (from 1 to 214). In particular, the green area corresponds at CG beads indexes. The highest value of RMSF is in correspondence of the two protein arms (indexes 30-70 and 125-150), as expected because they move more in respect of the remainder of protein and they are modelled atomistically. The two plots are comparable since the fluctuation of all C_α (or CG beads) are close to each other.

We can notice that two aspects assume particular prominence:

- the C_α carbons that constitute the arms (white area) fluctuate more than the remainder in analogy with fully-atomistic simulation (Fig. 5.10), although only the open conformation is present;
- the RMSF from reaction-field and PME in CANVAS are much more similar among them than in the atomistic case (Fig. 5.10), probably as a consequence of the fact that here the system visits only one conformation (while the amount of time it was found in one or the other conformation was different in the two atomistic simulations).

Thus, in order to validate this model, we provide hereafter a comparison with the fully-atomistic simulation analysis.

5.3.3 Comparison

Since the Variable Resolution model has been constructed such that each CG bead is localised in the same position of the corresponding C_α in the all-atom representation, we can compare the RMSD of RMSF results for both all-atom and CANVAS model.

Starting with the analysis of RMSD plots, the fully atomistic simulation shows that the protein assumes the opened and the closed conformation; on the other hand, the CANVAS simulation reproduces only the opened structure. This is the first signal that the coarse-grained model is too rigid since it does not allow the protein to fluctuate sufficiently. This intuition is further confirmed by the analysis of the RMSF plots, overlapping the Figs. 5.10 and 5.12. The result is, therefore, provided in Fig. 5.13: first, In the fully-atomistic part (white area) the fluctuations of the 4AKE arms C_α carbons are lower in the CANVAS simulations, because the two arms never get stick. Second, by looking at the region highlighted in green it is possible to notice that also the CG bead fluctuations are lower than the corresponding C_α ones in all-atom simulation.

Since we have observed that in CANVAS simulations only the open conformation appears, we expected the above mentioned RMSF result showing, in general, lower value with respect to that from the atomistic simulation. Therefore, we have performed, in addition, a comparison

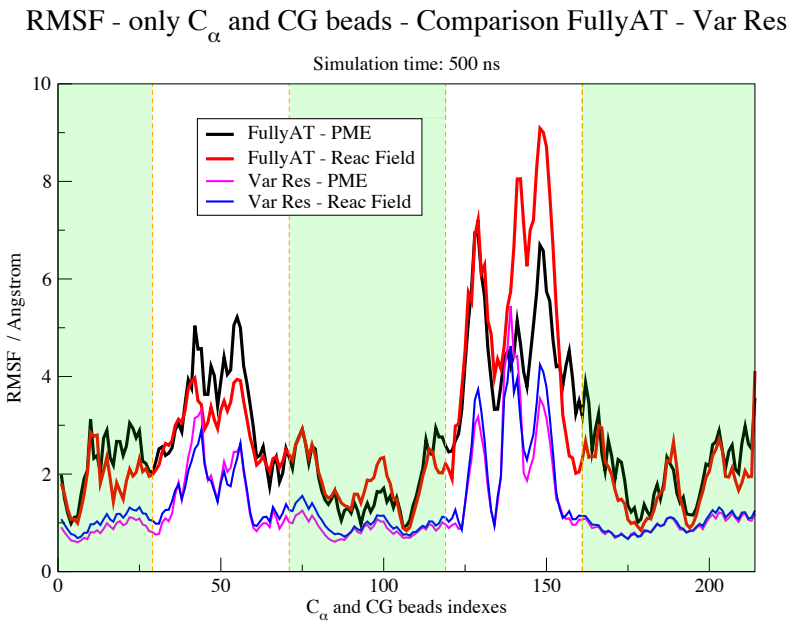


Figure 5.13: Root Mean Square Fluctuation (RMSF) for each C_α and the corresponding CG bead in the Coarse-Grained part, in case of Reaction-Field (red and blue lines) and PME (black and magenta lines). In particular, the plot displays a comparison between fully-atomistic and variable-resolution simulations. The x-axis shows the C_α and CG beads indexes (from 1 to 214). In particular, the green area corresponds at CG beads indexes. The highest value of RMSF is in correspondence of the two protein arms (indexes 30-70 and 125-150), as expected because they move more in respect of the remainder of protein and they are modelled atomistically. We can notice that the RMSF values in the variable-resolution simulation are much lower than the corresponding fully-atomistic one. This means that the CG network is too much rigid.

between the RMSF from CANVAS and the RMSF from the portions of the atomistic trajectories that are in the open state. The plot is shown in Fig. 5.14. It is possible to notice that the fluctuations in the CANVAS model are just slightly lower with respect to the fully-atomistic simulations, thus the open structure is well reproduced.

Therefore, summing up all results in terms of RMSD and RMSF calculations, the following properties are relevant:

- variable resolution simulation does not allow the closed conformation;
- the RMSD average value ($\approx 3-4$ Å) for the open conformation is the same both in the fully-atomistic simulation and in the variable resolution one, pointing out that the model reproduce well enough such structure (Fig. 5.14);

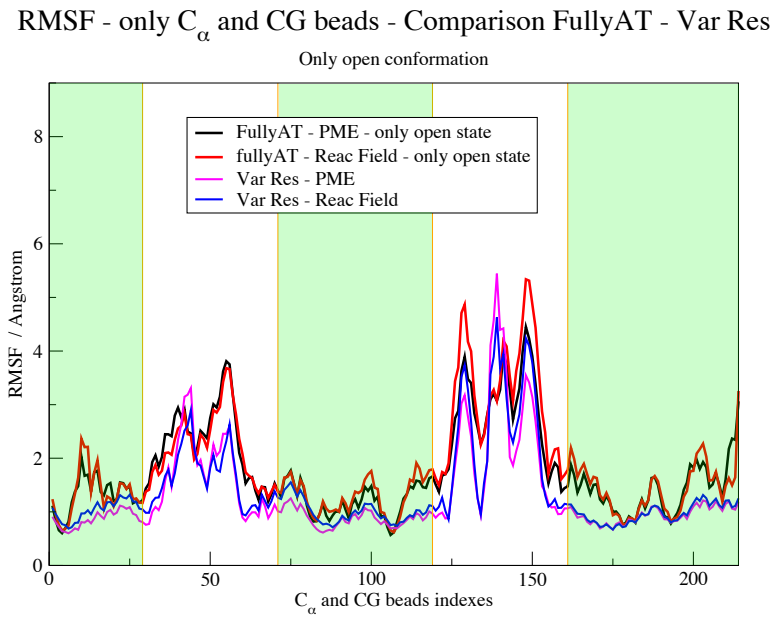


Figure 5.14: Root Mean Square Fluctuation (RMSF) for each C_α and the corresponding CG bead in the Coarse-Grained part, in case of Reaction-Field (red and blue lines) and PME (black and magenta lines). In particular, the plot displays a comparison between variable-resolution simulations and fully-atomistic ones in case the protein is only in the open state. The x-axis shows the C_α and CG beads indexes (from 1 to 214). In particular, the green area corresponds at CG beads indexes. The highest value of RMSF is in correspondence of the two protein arms (indexes 30-70 and 125-150), as expected because they move more in respect of the remainder of protein and they are modelled atomistically. We can notice that the RMSF values in the variable-resolution simulation are slightly lower than the corresponding fully-atomistic one. This means that, despite the CG network is rigid and requires refinements, the protein open structure is reproduced well enough by the CANVAS model.

- the RMSF plot shows that the C_α (and CG beads) fluctuation trend is the same both in CANVAS and all-atom simulations, but it shifted towards lower values (as shown in Fig. 5.13) in the former case, emphasising that the strength of the bonded interactions acting between CG particles to maintain the molecule's structure has proven to be too high.

The presented variable resolution model requires further refinements; a possible solution to the rigidity problem is provided in the conclusive part of the text.

5.4 Conclusions

In this work, we have illustrated a new multi-scale resolution scheme, dubbed CANVAS or coarse-grained anisotropic network model for variable resolution simulations. The term "variable" stems from the fact that such model allows complete freedom in coupling any desired level of coarse-graining, and to move smoothly between them.

In this scheme, each survived atom in the coarse-grained part has its own properties averaged on the atoms it represents, and these can differ even for the same kind of atoms representing the same kind of residues. In fact, the number and type of atoms mapping onto a given CG site depend on its local environment, which in turn depends on the molecule's structure and the distribution of retained atoms in it.

In this first application of this model, we have studied the conformational properties of the protein Adenylate Kinase [203] treating the two arms atomistically, while the remainder is coarse-grained. In particular, the latter has been obtained retaining only the C_α atoms of the corresponding residues. The fully-atomistic simulation and its analysis in terms of RMSD and RMSF shows that the protein assumes two main conformations: the first one has the two arms close to each other (closed structure). In contrast, the second one presents the two arms distant from each other, as in the reference conformation here employed (opened structure).

The variable resolution simulation, on the contrary, has shown that the model requires further refinements, since it explores only the open conformation; furthermore, even

though the RMSF profile of the CANVAS simulations follows rather accurately that of the reference, all-atom simulations, the absolute values are systematically too low. This points only towards an excessive rigidity of the structural bonds in the CG part, whose stiffness substantially dampens also the dynamics in the high-resolution domains.

However, the model defined here can be refined by applying the following rules:

- **The coarse-grained network requires flexibility**: it can be reached by using different elastic constant between CG beads according to their distance and/or treating both the arms and the hinge atomistically as shown schematically in Fig. 5.15. In particular, the atomistic hinge should allow the protein higher movements.
- **Retaining the C_α as representatives of the CG beads (that do not necessarily correspond to separate residues) could be suboptimal**: as proposed in Ref. [106] we can choose the mapping so as to minimises the fraction of internal motion: in this case, the survived atoms in the CG part are not necessarily C_α carbons.

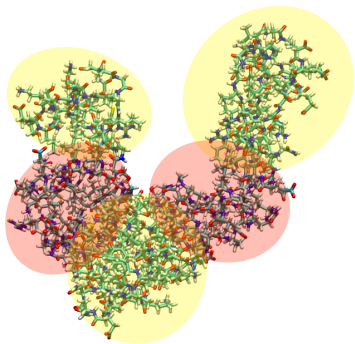


Figure 5.15: Schematic representation of 4AKE in Variable resolution: the two arms and the hinge are treated atomistically and coloured with the transparent yellow circles; the remainder is modelled in coarse-grained (red circle).

The results obtained in the first tests are pointing in the direction of the validity of this new approach. Nevertheless, only further simulations and analysis after refining the model will confirm it.

Appendix A: RMSD and RMSF

The RMSD (root mean square deviation) and RMSF (root mean square fluctuation) are commonly used to measure

the spatial variations of biomolecules in a molecular dynamics (MD) simulation. Their definition is similar, especially in terms of formula, but there is a substantial difference between the two.

The RMSD of certain atoms in a molecule in respect of a reference structure, r_0, is calculated as:

$$RMSD(t) = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (r_i - r_0)^2} \tag{5.9}$$

where N is the total number of chosen atoms, r_i is the coordinate of the atom i. Usually, the starting configuration is taken as reference.

The RMSF is a measure of the deviation between the position of particle i and some reference position r_0:

$$RMSF_i = \sqrt{\frac{1}{T} \cdot \sum_{i=1}^{N} (r_i - r_0)^2} \tag{5.10}$$

where T is the time over which one wants to average, whereas r_0 is the reference position of particle i. Usually, the starting configuration (or the time-averaged position one) is taken as reference.

The two previous equations have a similar expression, but they report two different analysis: the RMSF is averaged over the total time T giving the fluctuation value for each particle. On the other hand, for the RMSD, the average is taken over the particles, giving time-specific values.

6 Conclusions

The field of multi-scale modeling and simulation has enjoyed significant success in soft matter research within the past decade [113–118], also thanks to the boost impressed by the necessity to overcome the expensive cost of studying many phenomena in a single, highly detailed resolution.

Several methodologies have been developed in the last few years, and some of them have been the main object of this work. In section 6.1, we present a summary of the main results: particular prominence has the chapter 5 because it introduces a novel multi-scale approach: therefore, the latter can be considered as this thesis flagship.

In section 6.2 we provide an outlook and describe ongoing works and possible directions for future research about multiple-resolution simulations and their applications.

6.1 Summary

The introductory part of this thesis, chapter 1, provided an overview of relevant molecular simulation methodologies focusing on multiscale/multiple resolution simulation

schemes. On the other hand, the chapters 2-5 reported the results of the original scientific work.

2 – AdResS in amino acid sidechain analogues

In chapter 2, we have computed the solvation free energy of amino acid side chains analogoues solvated in water by using the combination of the force-based AdResS approach and the Thermodynamic Integration (TI). We have studied very small system, where the all-atom simulation is still feasible. This allowed us to validate the AdResS approach via comparison to fully-atomistic reference values. These calculations have highlighted three prominent results:

- **the strength of this approach**, in that it enables accurate control of the atomistic region density and a smooth transition between atomistic and coarse-grained regions, with no perturbation of the structural and thermodynamic properties of the solute and its solvation shell;
- **the speed-up** obtained via the AdResS approach: we have observed that the method provides a substantial reduction in simulation time with respect to fully atomistic.
- **accurate solvation free energies** by using the force-based adaptive resolution simulation scheme in combination with the Thermodynamic Integration.

3 – Ligand-protein interactions in Dual Resolution

In chapter 3, we have shown how the dual resolution model of a protein, constituted by an all-atom subdomain coupled to an elastic network model remainder, can be

used to calculate the binding free energy of an enzyme-substrate complex with accuracy comparable to that of a fully-atomistic setup. Particular attention has been paid to the impact of the mapping, i.e. the selection of atomistic and coarse-grained residues, on the binding free energy value: in fact, the active site is modeled with various numbers of amino acids treated atomistically.

Specifically, we have computed the total value of the binding free energy as well as that of its various energetic components and quantitatively inspected their dependence on the number of amino acids that are modeled at the fully atomistic level, ranging from 3 to 10, and on the location of these subgroups in the binding pocket.

It has been shown that, in spite of small variations of the total binding free energy with respect to the active site resolution, the separate contributions coming from different energetic terms (such as electrostatic and van der Waals interactions) manifest a stronger dependence on the mapping, thus suggesting the existence of an optimal level of intermediate resolution.

The results of this work thus have highlighted the importance of mapping in the construction of multi-scale and multi-resolution models, as a higher (but still intermediate) degree of detail does not necessarily correlate with a higher accuracy of the quantities of interest.

4 – 1BBA in Dual Resolution

In chapter 4 we have employed, once again, the dual-resolution scheme proposed in chapter 3 in which the lower-resolution part of the system is described by an Elastic Network Model. This model has been used in this

work focusing on the overall dynamic properties of a small non-globular protein known as Bovine Pancreatic Polypeptide (PDB code 1BBA). Particular attention has been paid on two aspects: first, the choice of the atomistic and the coarse-grained part, carried out by dividing the protein into quasi-rigid domains through the PiSQRD tool; second, and more importantly, the refinement of the Elastic Network Model employed for treating the lower-resolution part of the system. Specifically, we have used different elastic constants to connect the coarse-grained beads after a specific parametrization based on their distance distribution, with the purpose of improving the results obtained employing the original version of the ENM (only one elastic constant between CG beads), in terms of free energy landscapes, in comparison with all-atom simulation. Indeed, when using only one elastic constant, the value used is parametrized by minimazing the average root mean square error in C_α rmsf, therefore it is global parameter averaged on the entire system; on the other hand, when employing the refined version of ENM, each spring value is specific and it is based on the distance distribution of each couple C_α involved, thus leading to a greater accuracy.

The results of this work have emphasized, first, that the presence or the absence of salt in water leads to different dynamical properties: in the former case, the protein is fixed in one main conformation consisting of unfolded α-helix and stuck terminals; in the latter case, the protein presents more conformational variability. Furthermore, two important outcomes has been achieved:

- we have validated the usage of quasi-rigid domain subdivision in the context of dual-resolution models

with the purpose of dividing a biomolecule in CG and all-atom regions when the definition of a clear boundary between resolutions is not trivial;

- we have also validated the dual-resolution model with different elastic constant between CG beads. In this context, the free energy profiles are better reproduced with respect to the original formulation of ENM especially in case of water with salt.

5 – Introduction of a Variable Resolution model

In chapter 5 we have illustrated a new multi-scale resolution methodology dubbed CANVAS or coarse-grained anisotropic network model for variable resolution simulations. The term "variable" stems from the fact that CANVAS allows complete freedom in coupling any desired level of coarse-graining, and to move smoothly between them.

The lower resolution part is described with a new approach different from the Elastic Network Model or other works: each survived atom in the coarse-grained part has its own properties averaged on the group of atoms it represents –organised in domains defined by a Voronoi partition–, and these can differ even for the same kind of atoms representing the same kind of residues. In fact, the number and type of atoms mapping onto a given CG site depend on its local environment, which in turn depends on the molecule's structure and the distribution of retained atoms in it.

The common thread with the ENM is the employment of harmonic springs connecting the beads; however, at odds with the customary approach of ENM's, these bonds are

not placed among CG sites based on their distance, rather on the adjacency of the Voronoi domains.

In this work, we have performed fully-atomistic and variable-resolution simulations of Adenylate Kinase (PDB code 4AKE) comparing the results and, thus, the model performance. This first test has emphasised intriguing outcomes:

- The fully-atomistic simulation and its analysis in terms of RMSD and RMSF shows that the protein assumes two main conformations: the first one has the two protein arms close to each other (closed structure). In contrast, the second one presents the two arms distant from each other, as in the reference conformation here employed (opened structure) (see Fig. 5.8).
- The CANVAS simulation has shown that only the opened conformation is reproduced. Moreover, even though the RMSF profile follows rather accurately that of the reference, all-atom simulations, the absolute values are systematically too low. This points toward an excessive rigidity of the structural bonds in the CG part.

However, this novel multi-scale model, in spite of the necessity of further refinements displays encouraging results in the perspective of its application to larger systems than the ones examined here.

6.2 Outlook

The main object of this thesis are the multiple-resolution simulation of biomolecules. Each method is different in

some aspects with consequent pros and cons. However, the thread common to all multi-scale methods is that the region of the system playing a crucial role is treated at a high-resolution level, while the remainder is represented by a coarse-grained model.

In this thesis, we have presented different multiscale schemes. In particular, three models have been analyzed in detail: *Force-based Adaptive Resolution Scheme (AdResS)*, the *Dual-Resolution Model*, and the *Variable Resolution* one.

The AdResS methodology is a consolidated method proposed in 2005 by Praprotnik et al. [20, 115]. It allows to simulate a system where two different models (all-atom and coarse-grained, for instance) are concurrently employed in different sub-regions of the simulation domain. The particles are, moreover, allowed to diffuse from one region to the other freely. Specifically, between the atomistic domain (AT) and the coarse-grained one (CG), a hybrid (or transition) region (HY) is employed in which the coupling between different levels of resolution occurs. At the time, nearly 15 years after the first publication, all the physical bases of the method were fully established. However, most works have focused on the study of structural and dynamic properties, as well as basic thermodynamic quantities such as density, pressure, chemical potential, and compressibility. Only recently, the AdResS setup has been used to compute free energies, as proposed in Ref. [28] (this was discussed in Chapter 2 of this thesis). The long-term goal of this approach is the calculation of free energies in large, complex systems where fully-atomistic simulations are still largely unfeasible in practice. This includes, for example, ligand binding processes in high-molecular-weight

proteins, ligand intercalation in DNA, or small molecule-surface interactions. In such systems, the AdResS approach can be used to simulate at an atomistic level only the solvent molecules directly surrounding the region of interest, thus reducing the number of atomistic degrees of freedom in the system and consequently the associated computational cost.

In the Dual-Resolution methodology, at difference with AdResS, the resolution is fixed, i.e. position- and time-independent. The high-resolution level is treated with all-atom detail, while to describe the lower-resolution part the classical Elastic Network Model is employed. The latter has been introduced, for the first time, by Monique Tirion in 1996 [36] as a simplified approximation of the potential energy function a biomolecule near equilibrium. Its recent coupling with an atomistic level of detail in the framework of multi resolution schemes is providing encouraging results [24]: indeed, in chapter 3, we have proved that this strategy can be used for the calculation of binding free energy of an enzyme-substrate complex with atomistic accuracy. Moving away from ENMs, future refinements of this multiple-resolution description could include a coarse-grained model capable of of capturing anharmonic fluctuations.

This scheme is continuously evolving: in chapter 4 we have presented a refined version of the ENM, in which a different elastic constant connects the beads based on their distance distribution. The employment of such modification could be further improved, since the first test has shown promising results.

Finally, the Variable-resolution methodology is a new

multi-scale approach: the difference with the dual-resolution methodology lies in two new and pivotal aspects:

- The resolution can be modulated in a *quasi*-continuous way from atomistic to more coarse-grained according with the size of the Voronoi partition (itself dependent on the number and distribution of the retained CG sites).
- The potential parametrization does not require reference simulations, but only the all-atom structure of the protein.

To describe the lower resolution part, we have used a new approach (presented in chapter 5) based on beads and spring. While this method has shown some interesting results, it is still in the testing phase and requires further refinements. Specifically, we have noticed that in the variable resolution model the protein does not present the same conformational changes observed in fully-atomistic simulation. However, the fluctuation of each C_α both in all-atom reference simulation and in variable resolution one is the same albeit with lower values in the latter: this property is a signal of CG network rigidity. Therefore, the starting point is to focus on finding the optimal coarse-grained parametrization which involves the characterization of the elastic spring connecting CG beads (in analogy with the refined version of the ENM shown in chapter 4), and the best choice of the survived atom.

Der Unterschied zwischen Vergangenheit, Gegenwart und Zukunft ist nur eine Illusion, wenn auch eine hartnäckige.

The distinction between past, present and future is only a stubbornly persistent illusion.

A. EINSTEIN

We trust that time is linear. That it proceeds eternally, uniformly. Into infinity. But the distinction between past, present and future is nothing but an illusion. Yesterday, today and tomorrow are not consecutive, they are connected in a never-ending circle. Everything is connected.

Non ci abbracciamo oggi per abbracciarci piú forte domani.

Let's remain distant today so we can hug with more warmth and run faster tomorrow.

G. CONTE

Acknowledgements

These are uncertain times. Coronavirus is continuing its spread across the world, with more than three million confirmed cases in 185 countries and more than 200 thousand deaths. We will almost all (hopefully) have something to tell after this highly dramatic experience: we will remember to have suddenly lost interest in everything, in time, in sport, in our social life; we will remember to have seen empty streets and squares more than in August, to have queued outside the supermarket, to have been frightened by a sudden cough, to have not slept thinking about the risk of losing our jobs. And even when it is over, yes, because sooner or later it will be over, will we have the courage, at least at the beginning, to move again, to travel, to get together? This was a terrible punch in the stomach, from which we will psychologically struggle to recover, the psychosis will remain, because the enemy is invisible and you will never be sure that he has really been defeated. In the last two months it was difficult to reassure: those who wait for the heat, those who wait for a purifying rain, those who do not know what to wait for, but stay locked, inside their apartment, hoping it all goes away quickly. I just hope to return in my city to stay with everyone I love

after this pandemic. Four years are already passed. In May 2016, I left my homeland to undertake a new challenge. Therefore, it is time to retrace my steps and say thanks to the people who supported me during my studies.

As first, I would like to thank Prof. Dr. Kremer for giving me the opportunity to work in one of the most important and consolidated institutes on the European scene, and for allowing me to attend many coding courses, scientific collaborations and various conferences around Europe.

I am grateful to my supervisor, Dr. Raffaello Potestio, for his teaching, support, friendship during my studies and for allowing me to visit Trento in the last period of my Ph.D. to conclude my projects. Regrettably, he left the institute at the end of 2017; however, in these four years, I learned a lot, and I grew up from a scientific and personal point of view. All results presented in this thesis are the fruit of his passion for this research field and his constant presence.

My Master thesis supervisor, Prof. Gianluca Lattanzi, deserves a particular mention because he introduced me, for the first time, the Max Planck Institute for Polymer. He has never ceased showing me great affection and supporting me in my choices in these years.

At the beginning of my Ph.D., I was fortunate to collaborate with Aoife Fogarty, when she was a post-doc researcher. Regrettably, she also left the institute after I joined the Theory Group for less than one year. Nevertheless, I learned a lot, and, despite my difficulties, she was always kind and professional with me. I owe her so much.

I am grateful to all the people of the Theory Group of MPIP, with whom I had the luck to work during my Ph.D.

A big, big thank goes to Maziar, Roberto, Claudio, Cristina, Anirbal, Bin, Nancy, Hsiao-Ping, Yani, Kostas, Robin, and all the others, for helping me and for having intriguing scientific – and not – conversations. In particular, Robin deserves a special mention because he tried to compensate for Raffaello's absence in the last two years. I am also thankful to Burkhard, because he has been always available when I had some problems: I will never forget his help when I had problems in finding my first tutor job.

I also thank to Doris, Irene, Normen, and all the other people that helped me in administrative issues during these years.

I want to say thanks to all the people I encountered in my travels: stick in my memory are the guys of Bologna and Rome summer schools at CINECA. I had pleasant moments, and I hope to see them once again. Moreover, I found Trento surprising and plenty of cool people. See you soon!

I think that life is like a train ride. The passengers on the train are seemingly going to the same destination as you, but they will stay on the ride or they will get off somewhere during the trip. People can and will get off at any stop. I am very grateful to them because they, unconsciously, contributed to this result and in what I am now. Fortunately, some of the people (very few) are still facing this journey with me. Big, big thanks. I hug all of you.

A special thanks deserve Corrado and Domenico. The former has been the only one that traveled in Germany to visit me and living my everyday reality. Domenico, on the other hand, visited me in Trento in the last period of my Ph.D. Thanks a lot, guys.

Sometimes, thank a city is complicated. I had a strange relationship of love and hate towards Mainz. However, I will miss you, Mainz. I am sure. You were my hometown for four years. I met a lot of people in this city, and I feel to thank all of them.

Last, but maybe more important, I would like to thank my family for always loving and supporting me to whom this thesis is dedicated. Specifically, I want to apologize one thousand times to my mum. She knows why. Her suggestions, her presence always and everywhere, has been my strength. I wish her the same to fight against her evil for another hundred years. Grazie Mamma.

Probably, I owe the biggest thanks to me. My Ph.D. path was very tortuous, riddled with difficulties: Aoife and Raffaello left the MPIP after a few months. Their physical presence would have been of fundamental importance for everything. I leave the rest to the imagination.

I changed a lot in these four years; I am not the same as the guy I left my homeland. Now I am thirty, though. As the journalist Massimo Gramellini writes in his best book "Fai Bei Sogni":

«Thirty years. It is the age of the first evaluations. I know how you feel. You have the feeling of having lived along an inclined plane that has led you here. As if you were the product of choices that were not influenced by you, but by those around you».

I can only agree with him. Now, it is time to make choices. The right ones. But, despite everyone and everything, I can only say three words:

I DID IT.

Bibliography

Here are the references in citation order.

[1] J.M. Bijvoet, W.G. Burgers, and G. Hägg. *Early Papers on Diffraction of X-rays by Crystals*. Vol. 1. Springer US, 1969 (cited on pages 1, 9).

[2] J.M. Bijvoet, W.G. Burgers, and G. Hägg. *Early Papers on Diffraction of X-rays by Crystals*. Vol. 2. Springer US, 1972 (cited on pages 1, 9).

[3] Kurt Wüthrich. *NMR of proteins and nucleic acids*. New York: Wiley, 1986 (cited on pages 1, 9).

[4] Yifan Cheng et al. 'A Primer to Single-Particle Cryo-Electron Microscopy'. In: *Cell* 161.3 (Mar. 2015), pp. 438–449. DOI: doi:10.1016/j.cell.2015.03.050 (cited on page 1).

[5] Dieter W. Heermann. 'Computer-Simulation Methods'. In: *Computer Simulation Methods in Theoretical Physics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1990, pp. 8–12. DOI: 10.1007/978-3-642-75448-7_2 (cited on page 1).

[6] Jerome Sacks et al. 'Design and analysis of computer experiments. With comments and a rejoinder by the authors'. In: *Statistical Science* 4 (Jan. 1989). DOI: 10.1214/ss/1177012413 (cited on page 1).

[7] Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. 2nd ed. Springer-Verlag New York (cited on page 1).

[8] Gary S. Grest and Kurt Kremer. 'Molecular dynamics simulation for polymers in the presence of a heat bath'. In: *Phys. Rev. A* 33 (5 May 1986), pp. 3628–3631. DOI: 10.1103/PhysRevA.33.3628 (cited on page 2).

[9] Kurt Kremer, Gary S. Grest, and I. Carmesin. 'Crossover from Rouse to Reptation Dynamics: A Molecular-Dynamics Simulation'. In: *Phys. Rev. Lett.* 61 (5 Aug. 1988), pp. 566–569. DOI: 10.1103/PhysRevLett.61.566 (cited on page 2).

[10] J. A. McCammon and M. KARPLUS. 'Internal motions of antibody molecules'. In: *Nature* 268.5622 (1977), pp. 765–766 (cited on page 2).

[11] M. KARPLUS and J. MCCAMMON. 'Protein structural fluctuations during a period of 100 ps.' In: *Nature* 277.578 (1979). DOI: https://doi.org/10.1038/277578a0 (cited on page 2).

[12] Paolo Raiteri et al. 'Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics'. In: *The Journal of Physical Chemistry B* 110.8 (2006). PMID: 16494409, pp. 3533–3539. DOI: 10.1021/jp054359r (cited on page 2).

[13] Hongfeng Lou and Robert I. Cukier. 'Molecular Dynamics of Apo-Adenylate Kinase: A Distance Replica Exchange Method for the Free Energy of Conformational Fluctuations'. In: *The Journal of Physical Chemistry B* 110.47 (2006). PMID: 17125384, pp. 24121–24137. DOI: 10.1021/jp064303c (cited on page 2).

[14] Karunesh Arora and Charles L. Brooks. 'Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism'. In: *Proceedings of the National Academy of Sciences* 104.47 (2007), pp. 18496–18501. DOI: 10.1073/pnas.0706443104 (cited on page 2).

[15] F. Pontiggia, A. Zen, and C. Micheletti. 'Small and large scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics'. In: *Biophys J* 95.12 (Dec. 2008), pp. 5901–5912 (cited on pages 2, 71).

[16] K. Kremer and F. Müller-Plathe. 'Multiscale Problems in Polymer Science: Simulation Approaches'. In: 26 (2001) (cited on page 3).

[17] Nico F. A. van der Vegt, Christine Peter-Tittelbach, and Kurt Kremer. 'Structure-Based Coarse- and Fine-Graining in Soft Matter Simulations'. In: *Coarse-Graining of Condensed Phase and Biomolecular Systems*. Ed. by Gregory Voth. Boca Raton: CRC Press, 2009, pp. 379–395. DOI: 10.1201/9781420059564.ch25 (cited on page 3).

[18] W. G. Noid. 'Systematic Methods for Structurally Consistent Coarse-Grained Models'. In: *Biomolecular Simulations: Methods and Protocols*. Ed. by Luca Monticelli and Emppu Salonen. Totowa, NJ: Humana Press, 2013, pp. 487–531. DOI: 10.1007/978-1-62703-017-5_19 (cited on page 3).

[19] W. G. Noid. 'Perspective: Coarse-grained models for biomolecular systems'. In: *The Journal of Chemical Physics* 139.9 (2013), p. 090901. DOI: 10.1063/1.4818908 (cited on pages 3, 6, 28).

[20] Matej Praprotnik, Luigi Delle Site, and Kurt Kremer. 'Multiscale Simulation of Soft Matter: From Scale Bridging to Adaptive Resolution'. In: *Annual Review of Physical Chemistry* 59.1 (2008). PMID: 18062769, pp. 545–571. DOI: 10.1146/annurev.physchem.59.032607.093707 (cited on pages 3, 37, 38, 157).

[21] Gary S Ayton and Gregory A Voth. 'Systematic multiscale simulation of membrane protein systems'. In: *Current Opinion in Structural Biology* 19.2 (2009). Theory and simulation / Macromolecular assemblages, pp. 138–144. DOI: https://doi.org/10.1016/j.sbi.2009.03.001 (cited on pages 3, 15).

[22] M. Praprotnik et al. 'Comment on Adaptive Multiscale Molecular Dynamics of Macromolecular Fluids'. In: *Phys. Rev. Lett.* 107 (9 Aug. 2011), p. 099801. DOI: 10.1103/PhysRevLett.107.099801 (cited on pages 3, 41).

[23] Alexander B. Kuhn, Srinivasa M. Gopal, and Lars V. Schäfer. 'On Using Atomistic Solvent Layers in Hybrid All-Atom/Coarse-Grained Molecular Dynamics Simulations'. In: *J. Chem. Theory Comput.* 11.9 (2015), pp. 4460–4472. DOI: 10.1021/acs.jctc.5b00499 (cited on pages 3, 49).

[24] Aoife C. Fogarty, Raffaello Potestio, and Kurt Kremer. 'A multi-resolution model to capture both global fluctuations of an enzyme and molecular recognition in the ligand-binding site'. In: *Proteins: Struct., Func., and Bioinf.* 84.12 (2016), pp. 1902–1913. DOI: 10.1002/prot.25173 (cited on pages 3, 36, 48, 67, 70–72, 78, 95, 96, 109, 110, 116, 128, 138, 158).

[25] Richard J. Gowers and Paola Carbone. 'A multiscale approach to model hydrogen bonding: The case of polyamide'. In: *The Journal of Chemical Physics* 142.22 (2015), p. 224907. DOI: 10.1063/1.4922445 (cited on pages 3, 70, 72).

[26] Cameron F. Abrams, Luigi Delle Site, and Kurt Kremer. 'Dual-resolution coarse-grained simulation of the bisphenol-*A*-polycarbonate/nickel interface'. In: *Phys. Rev. E* 67 (2 Feb. 2003), p. 021807. DOI: 10.1103/PhysRevE.67.021807 (cited on pages 3, 70, 72).

[27] Matias R. Machado and Sergio Pantano. 'Exploring LacI–DNA Dynamics by Multiscale Simulations Using the SIRAH Force Field'. In: *Journal of Chemical Theory and Computation* 11.10 (2015). PMID: 26574286, pp. 5012–5023. DOI: 10.1021/acs.jctc.5b00575 (cited on pages 3, 70, 72).

[28] Raffaele Fiorentini et al. 'Using force-based adaptive resolution simulations to calculate solvation free energies of amino acid sidechain analogues'. In: *The Journal of Chemical Physics* 146.24 (2017), p. 244113. DOI: 10.1063/1.4989486 (cited on pages 3, 38, 40, 157).

[29] Raffaele Fiorentini et al. 'Using force-based adaptive resolution simulations to calculate solvation free energies of amino acid sidechain analogues'. In: *The Journal of Chemical Physics* 146 (June 2017), p. 244113. DOI: 10.1063/1.4989486 (cited on pages 3, 70, 73).

[30] Matej Praprotnik, Luigi Delle Site, and Kurt Kremer. 'A macromolecule in a solvent: Adaptive resolution molecular dynamics simulation'. In: *J. Chem. Phys.* 126.13 (2007), p. 134902 (cited on pages 3, 38, 65).

[31] R. Delgado-Buscalioni. 'Thermodynamics of adaptive molecular resolution'. In: *Phil. Trans. R. Soc. A* 374.2080 (2016). DOI: 10.1098/rsta.2016.0152 (cited on pages 3, 48).

[32] Han Wang, Christof Schütte, and Luigi Delle Site. 'Adaptive Resolution Simulation (AdResS): A Smooth Thermodynamic and Structural Transition from Atomistic to Coarse Grained Resolution and Vice Versa in a Grand Canonical Fashion'. In: *J. Chem. Theory Comput.* 8.8 (2012), pp. 2878–2887. DOI: 10.1021/ct3003354 (cited on pages 3, 52).

[33] Debashish Mukherji et al. 'Kirkwood–Buff Analysis of Liquid Mixtures in an Open Boundary Simulation'. In: *Journal of Chemical Theory and Computation* 8.2 (2012). PMID: 26596589, pp. 375–379. DOI: 10.1021/ct200709h (cited on page 3).

[34] Staš Bevc et al. 'Adaptive resolution simulation of salt solutions'. In: *New Journal of Physics* 15.10 (Oct. 2013), p. 105007. DOI: 10.1088/1367-2630/15/10/105007 (cited on page 3).

[35] Marilisa Neri et al. 'Coarse-Grained Model of Proteins Incorporating Atomistic Detail of the Active Site'. In: *Phys. Rev. Lett.* 95 (21 Nov. 2005), p. 218102. DOI: 10.1103/PhysRevLett.95.218102 (cited on page 3).

[36] Monique M. Tirion. 'Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis'. In: *Phys. Rev. Lett.* 77 (9 Aug. 1996), pp. 1905–1908. DOI: 10.1103/PhysRevLett.77.1905 (cited on pages 3, 29, 30, 35, 43, 71, 72, 78, 95, 103, 109, 110, 128, 158).

[37] Peter L. Freddolino et al. 'Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain'. In: *Biophysical Journal* 94.10 (2008), pp. L75–L77. DOI: doi:10.1529/biophysj.108.131565 (cited on page 5).

[38] Jan Kubelka, James Hofrichter, and William A Eaton. 'The protein folding speed limit'. In: *Current Opinion in Structural Biology* 14.1 (2004), pp. 76–88. DOI: https://doi.org/10.1016/j.sbi.2004.01.013 (cited on page 5).

[39] Christine Peter and Kurt Kremer. 'Multiscale simulation of soft matter systems: from the atomistic to the coarse-grained level and back'. In: *Soft Matter* 5 (22 2009), pp. 4357–4366. DOI: 10.1039/B912027K (cited on pages 6, 37).

[40] Shina C.L. Kamerlin et al. 'Coarse-Grained (Multiscale) Simulations in Studies of Biophysical and Chemical Systems'. In: *Annual Review of Physical Chemistry* 62.1 (2011). PMID: 21034218, pp. 41–64. DOI: 10.1146/annurev-physchem-032210-103335 (cited on page 6).

[41] Shoji Takada. 'Coarse-grained molecular simulations of large biomolecules'. In: *Current Opinion in Structural Biology* 22.2 (2012). Theory and simulation/Macromolecular assemblages, pp. 130–137. DOI: https://doi.org/10.1016/j.sbi.2012.01.010 (cited on page 6).

[42] Sereina Riniker, Jane R. Allison, and Wilfred F. van Gunsteren. 'On developing coarse-grained models for biomolecular simulation: a review'. In: *Phys. Chem. Chem. Phys.* 14 (36 2012), pp. 12423–12430. DOI: 10.1039/C2CP40934H (cited on page 6).

[43] Kurt Kremer and Gary S Grest. 'Monte Carlo and molecular dynamics simulations of polymers'. In: *Physica Scripta* T35 (Jan. 1991), pp. 61–65. DOI: 10.1088/0031-8949/1991/t35/013 (cited on page 6).

[44] Burkhard Dünweg, Gary S. Grest, and Kurt Kremer. 'Molecular Dynamics Simulations of Polymer Systems'. In: *Numerical Methods for Polymeric Systems*. Ed. by Stuart G. Whittington. New York, NY: Springer New York, 1998, pp. 159–195. DOI: 10.1007/978-1-4612-1704-6_10 (cited on page 6).

[45] J.R. Lustig and B.J.G. Strauss. 'BODY COMPOSITION'. In: *Encyclopedia of Food Sciences and Nutrition (Second Edition)*. Ed. by Benjamin Caballero. Second Edition. Oxford: Academic Press, 2003, pp. 550–557. DOI: https://doi.org/10.1016/B0-12-227055-X/00110-3 (cited on page 6).

[46] A.Keith Dunker et al. 'Intrinsically disordered protein'. In: *Journal of Molecular Graphics and Modelling* 19.1 (2001), pp. 26–59. DOI: https://doi.org/10.1016/S1093-3263(00)00138-8 (cited on page 7).

[47] H. Jane Dyson and Peter E. Wright. 'Intrinsically unstructured proteins and their functions'. In: *Nature Reviews Molecular Cell Biology* 6.3 (2005), pp. 197–208 (cited on page 7).

[48] A Keith Dunker et al. 'Function and structure of inherently disordered proteins'. In: *Current Opinion in Structural Biology* 18.6 (2008). Catalysis and regulation / Proteins, pp. 756–764. DOI: https://doi.org/10.1016/j.sbi.2008.10.002 (cited on page 7).

[49] F. Sanger and H. Tuppy. 'The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates'. In: *Biochemical Journal* 49.4 (Sept. 1951), pp. 463–481. DOI: 10.1042/bj0490463 (cited on page 7).

[50] A. V Finkelstein and O. Ptitsyn. *Protein Physics. A course of Lectures*. Academic Press, 2002 (cited on page 8).

[51] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Vol. 1. Computational Science Series. San Diego: Academic Press, 2002 (cited on pages 8, 13, 17, 19).

[52] M.F. Perutz, J.C. Kendrew, and H.C. Watson. 'Structure and function of haemoglobin: 2. Some relations between polypeptide chain configuration and amino acid sequence'. In: *Journal of Molecular Biology* 13.3 (1965), pp. 669–678. DOI: https://doi.org/10.1016/S0022-2836(65)80134-6 (cited on page 9).

[53] C. Branden and J. Tooze. *Introduction to Protein Structure*. 2nd. 1999 (cited on page 11).

[54] Linus Pauling, Robert B. Corey, and H. R. Branson. 'The Structure of Proteins. Two Hydrogen-Bonded Helical Configurations of the Polypeptide'. In: (1951) (cited on page 11).

[55] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Secondary structure: Polypeptide chains can fold into regular structures such as the alpha helix, the beta sheet, and turns and loops*. Ed. by NY: W. H. Freeman New York. 5th ed. 2002 (cited on pages 11, 12).

[56] Nicholas Metropolis et al. 'Equation of State Calculations by Fast Computing Machines'. In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114 (cited on page 13).

[57] E. Fermi et al. *STUDIES OF THE NONLINEAR PROBLEMS*. Tech. rep. Los Alamos Scientific Lab., N. Mex., May 1955. DOI: 10.2172/4376203 (cited on page 13).

[58] P. P. Ewald. 'Die Berechnung optischer und elektrostatischer Gitterpotentiale'. In: (Jan. 1921). DOI: 10.1002/andp.19213690304 (cited on page 13).

[59] 'Dimensional Analysis of Free Surface Flows'. In: *Rheology Bulletin* 74.2 (2005) (cited on page 14).

[60] Naresh B Kotadiya et al. 'Universal strategy for Ohmic hole injection into organic semiconductors with high ionization energies'. In: *Nature materials* 17.4 (Apr. 2018), pp. 329–334. DOI: 10.1038/s41563-018-0022-8 (cited on page 15).

[61] Pascal Kordt et al. 'Modeling of Organic Light Emitting Diodes: From Molecular to Device Properties'. In: *Advanced Functional Materials* 25.13 (2015), pp. 1955–1971. DOI: 10.1002/adfm.201403004 (cited on page 15).

[62] Roberto Menichetti et al. 'In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force'. In: *The Journal of Chemical Physics* 147.12 (2017), p. 125101. DOI: 10.1063/1.4987012 (cited on page 15).

[63] Stephen H. White and William C. Wimley. 'MEMBRANE PROTEIN FOLDING AND STABILITY: Physical Principles'. In: *Annual Review of Biophysics and Biomolecular Structure* 28.1 (1999). PMID: 10410805, pp. 319–365. DOI: 10.1146/annurev.biophys.28.1.319 (cited on page 15).

[64] LR Forrest and MS Sansom. 'Membrane simulations: bigger and better?' In: *Current opinion in structural biology* 10.2 (Apr. 2000), pp. 174–181. DOI: 10.1016/s0959-440x(00)00066-x (cited on page 15).

[65] Markus Deserno. 'Mesoscopic Membrane Physics: Concepts, Simulations, and Selected Applications'. In: *Macromolecular Rapid Communications* 30.9-10 (2009), pp. 752–771. DOI: 10.1002/marc.200900090 (cited on page 15).

[66] Thomas J Lane et al. 'To milliseconds and beyond: challenges in the simulation of protein folding'. In: *Current Opinion in Structural Biology* 23.1 (2013). Folding and binding / Protein-nucleic acid interactions, pp. 58–65. DOI: https://doi.org/10.1016/j.sbi.2012.11.002 (cited on page 15).

[67] Ron O. Dror et al. 'Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations'. In: *The Journal of General Physiology* 135.6 (May 2010), pp. 555–562. DOI: 10.1085/jgp.200910373 (cited on page 15).

[68] Ron O. Dror et al. 'Biomolecular Simulation: A Computational Microscope for Molecular Biology'. In: *Annual Review of Biophysics* 41.1 (2012). PMID: 22577825, pp. 429–452. DOI: 10.1146/annurev-biophys-042910-155245 (cited on page 15).

[69] M. F. Horstemeyer. *Multiscale Modeling: A Review*. 2009, pp. 87–135 (cited on page 15).

[70] D. Raabe et al. 'Multi-scale modeling in materials science and engineering'. In: (2009) (cited on page 15).

[71] Kieron Burke. 'Perspective on density functional theory'. In: *The Journal of Chemical Physics* 136.15 (2012), p. 150901. DOI: 10.1063/1.4704546 (cited on pages 16, 19).

[72] R. O. Jones. 'Density functional theory: Its origins, rise to prominence, and future'. In: *Rev. Mod. Phys.* 87 (3 Aug. 2015), pp. 897–923. DOI: 10.1103/RevModPhys.87.897 (cited on pages 16, 19).

[73] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, 2010, p. 712 (cited on pages 17, 19, 22, 23).

[74] Martin Karplus and Gregory A. Petsko. 'Molecular dynamics simulations in biology'. In: *Nature* 347.6294 (1990), pp. 631–639 (cited on page 19).

[75] Sandeep Patel and Charles L. Brooks III. 'CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations'. In: *Journal of Computational Chemistry* 25.1 (2004), pp. 1–16. DOI: 10.1002/jcc.10355 (cited on page 19).

[76] *GROMOS*. https://www.igc.ethz.ch/gromos.html (cited on page 19).

[77] Viktor Hornak et al. 'Comparison of multiple Amber force fields and development of improved protein backbone parameters'. In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 712–725. DOI: 10.1002/prot.21123 (cited on pages 19, 79, 115, 139).

[78] William L. Jorgensen and Julian Tirado-Rives. 'The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin'. In: *Journal of the American Chemical Society* 110.6 (1988). PMID: 27557051, pp. 1657–1666. DOI: 10.1021/ja00214a001 (cited on page 20).

[79] Tom Darden, Darrin York, and Lee Pedersen. 'Particle mesh Ewald: An N log(N) method for Ewald sums in large systems'. In: *The Journal of Chemical Physics* 98.12 (1993), pp. 10089–10092. DOI: 10.1063/1.464397 (cited on pages 23, 139).

[80] Loup Verlet. 'Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules'. In: *Phys. Rev.* 159 (1 July 1967), pp. 98–103. DOI: 10.1103/PhysRev.159.98 (cited on page 23).

[81] William C. Swope et al. 'A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters'. In: *The Journal of Chemical Physics* 76.1 (1982), pp. 637–649. DOI: 10.1063/1.442716 (cited on page 25).

[82] H. J. C. Berendsen et al. 'Molecular dynamics with coupling to an external bath'. In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3684–3690. DOI: 10.1063/1.448118 (cited on page 26).

[83] Giovanni Bussi, Davide Donadio, and Michele Parrinello. 'Canonical sampling through velocity rescaling'. In: *The Journal of Chemical Physics* 126.1 (2007), p. 014101. DOI: 10.1063/1.2408420 (cited on pages 27, 115, 139).

[84] Paul Langevin. 'Sur la théorie du mouvement brownien'. In: *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences* 146 (1908), pp. 530–533 (cited on page 27).

[85] T. Schneider and E. Stoll. 'Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions'. In: *Phys. Rev. B* 17 (3 Feb. 1978), pp. 1302–1322. DOI: 10.1103/PhysRevB.17.1302 (cited on page 27).

[86] Dominik Fritz et al. 'Multiscale modeling of soft matter: scaling of dynamics'. In: *Phys. Chem. Chem. Phys.* 13 (22 2011), pp. 10412–10420. DOI: 10.1039/C1CP20247B (cited on page 28).

[87] W. Tschöp et al. 'Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates'. In: *Acta Polymerica* 49 (2-3 Feb. 1998), pp. 61–74. DOI: 10.1002/(SICI)1521-4044(199802)49:2/3<61::AID-APOL61>3.0.CO;2-V (cited on pages 29, 50).

[88] Dirk Reith, Mathias Pütz, and Florian Müller-Plathe. 'Deriving effective mesoscale potentials from atomistic simulations'. In: *J Comput. Chem.* 24.13 (2003), pp. 1624–1636. DOI: 10.1002/jcc.10307 (cited on pages 29, 50, 58).

[89] R. L. McGreevy and L. Pusztai. 'Reverse Monte Carlo Simulation: A New Technique for the Determination of Disordered Structures'. In: *Molecular Simulation* 1.6 (1988), pp. 359–367. DOI: 10.1080/08927028808080958 (cited on page 29).

[90] Alexander P. Lyubartsev and Aatto Laaksonen. 'Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach'. In: *Phys. Rev. E* 52 (4 Oct. 1995), pp. 3730–3737. DOI: 10.1103/PhysRevE.52.3730 (cited on page 29).

[91] F Ercolessi and J. B Adams. 'Interatomic Potentials from First-Principles Calculations: The Force-Matching Method'. In: *Europhysics Letters (EPL)* 26.8 (June 1994), pp. 583–588. DOI: 10.1209/0295-5075/26/8/005 (cited on page 29).

[92] Sergei Izvekov and Gregory A. Voth. 'A Multiscale Coarse-Graining Method for Biomolecular Systems'. In: *The Journal of Physical Chemistry B* 109.7 (2005). PMID: 16851243, pp. 2469–2473. DOI: 10.1021/jp044629q (cited on page 29).

[93] Aviel Chaimovich and M Scott Shell. 'Coarse-graining errors and numerical optimization using a relative entropy framework'. In: *The Journal of chemical physics* 134.9 (Mar. 2011), p. 094112. DOI: 10.1063/1.3557038 (cited on page 29).

[94] M. S. Shell. *Thermodynamics and statistical mechanics: An integrated approach.* Cambridge University Press, 2015 (cited on page 29).

[95] Siewert J. Marrink et al. 'The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations'. In: *The Journal of Physical Chemistry B* 111.27 (2007). PMID: 17569554, pp. 7812–7824. DOI: 10.1021/jp071097f (cited on page 29).

[96] Siewert J. Marrink and D. Peter Tieleman. 'Perspective on the Martini model'. In: *Chem. Soc. Rev.* 42 (16 2013), pp. 6801–6822. DOI: 10.1039/C3CS60093A (cited on page 29).

[97] Guido Polles et al. 'Mechanical and Assembly Units of Viral Capsids Identified via Quasi-Rigid Domain Decomposition'. In: *PLOS Computational Biology* 9.11 (Nov. 2013), pp. 1–13. DOI: 10.1371/journal.pcbi.1003331 (cited on pages 29, 43).

[98] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. 'Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential'. In: *Folding and Design* 2.3 (1997), pp. 173–181. DOI: https://doi.org/10.1016/S1359-0278(97)00024-2 (cited on pages 29, 32, 43).

[99] Konrad Hinsen. 'Analysis of domain motions by approximate normal mode calculations'. In: *Proteins: Structure, Function, and Bioinformatics* 33.3 (1998), pp. 417–429. DOI: 10.1002/(SICI)1097-0134(19981115)33:3<417::AID-PROT10>3.0.CO;2-8 (cited on pages 29, 43).

[100] A.R. Atilgan et al. 'Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model'. In: *Biophysical Journal* 80.1 (2001), pp. 505–515. DOI: https://doi.org/10.1016/S0006-3495(01)76033-X (cited on pages 29, 43).

[101] M Delarue and Y.-H Sanejouand. 'Simplified Normal Mode Analysis of Conformational Transitions in DNA-dependent Polymerases: the Elastic Network Model'. In: *Journal of Molecular Biology* 320.5 (2002), pp. 1011–1024. DOI: https://doi.org/10.1016/S0022-2836(02)00562-4 (cited on pages 29, 43).

[102] Cristian Micheletti, Paolo Carloni, and Amos Maritan. 'Accurate and efficient description of protein vibrational dynamics: Comparing molecular dynamics and Gaussian models'. In: *Proteins: Structure, Function, and Bioinformatics* 55.3 (2004), pp. 635–645. DOI: 10.1002/prot.20049 (cited on pages 29, 43).

[103] Pemra Doruker and Ivet Atilgan Ali Rana andBahar. 'Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor'. In: *Proteins: Structure, Function, and Bioinformatics* 40.3 (2000), pp. 512–524. DOI: 10.1002/1097-0134(20000815)40:3<512::AID-PROT180>3.0.CO;2-M (cited on page 32).

[104] Q. Cui and I. Bahar. *Normal Mode Analysis*. Chapman and Hall/CRC, 2006 (cited on page 32).

[105] Timothy Lezon et al. 'Chapter 7 Elastic Network Models For Biomolecular Dynamics: Theory and Application to Membrane Proteins and Viruses'. In: (Dec. 2009). DOI: 10.1142/9789812838803_0007 (cited on page 35).

[106] Patrick Diggins et al. 'Optimal Coarse-Grained Site Selection in Elastic Network Models of Biomolecules'. In: *Journal of Chemical Theory and Computation* 0.0 (0), null. DOI: 10.1021/acs.jctc.8b00654 (cited on pages 35, 71, 72, 136, 149).

[107] L. Delle Site. 'What is a multiscale problem in molecular dynamics?' In: *Entropy* 16.1 (2014), pp. 23–40 (cited on page 37).

[108] Raffaello Potestio, Christine Peter, and Kurt Kremer. 'Computer Simulations of Soft Matter: Linking the Scales'. In: *Entropy* 16.8 (2014), pp. 4199–4245. DOI: 10.3390/e16084199 (cited on pages 37, 48, 70, 73).

[109] Paolo Carloni, Ursula Rothlisberger, and Michele Parrinello. 'The Role and Perspective of Ab Initio Molecular Dynamics in the Study of Biological Systems'. In: *Accounts of Chemical Research* 35.6 (2002). PMID: 12069631, pp. 455–464. DOI: 10.1021/ar010018u (cited on pages 37, 127).

[110] Hans Martin Senn and Walter Thiel. 'QM/MM Methods for Biomolecular Systems'. In: *Angewandte Chemie International Edition* 48.7 (2009), pp. 1198–1229. DOI: 10.1002/anie.200802019 (cited on pages 37, 127).

[111] Bo Wang and Donald G. Truhlar. 'Combined Quantum Mechanical and Molecular Mechanical Methods for Calculating Potential Energy Surfaces: Tuned and Balanced Redistributed-Charge Algorithm'. In: *Journal of Chemical Theory and Computation* 6.2 (2010). PMID: 26617295, pp. 359–369. DOI: 10.1021/ct900366m (cited on pages 37, 127).

[112] Frank H. Wallrapp and Victor Guallar. 'Mixed quantum mechanics and molecular mechanics methods: Looking inside proteins'. In: *WIREs Computational Molecular Science* 1.2 (2011), pp. 315–322. DOI: 10.1002/wcms.27 (cited on pages 37, 127).

[113] Wei Shi and Edward J. Maginn. 'Atomistic Simulation of the Absorption of Carbon Dioxide and Water in the Ionic Liquid 1-n-Hexyl-3-methylimidazolium Bis(trifluoromethylsulfonyl)imide ([hmim][Tf2N]'. In: *The Journal of Physical Chemistry B* 112.7 (2008). PMID: 18217747, pp. 2045–2055. DOI: 10.1021/jp077223x (cited on pages 37, 127, 151).

[114] Marc W. van der Kamp and Adrian J. Mulholland. 'Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology'. In: *Biochemistry* 52.16 (2013). PMID: 23557014, pp. 2708–2728. DOI: 10.1021/bi400215w (cited on pages 37, 127, 151).

[115] Matej Praprotnik, Luigi Delle Site, and Kurt Kremer. 'Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly'. In: *J. Chem. Phys.* 123.22 (2005), p. 224106 (cited on pages 37, 38, 40, 48, 52, 127, 151, 157).

[116] Raffaello Potestio et al. 'Hamiltonian Adaptive Resolution Simulation for Molecular Liquids'. In: *Phys. Rev. Lett.* 110 (10 Mar. 2013), p. 108301. DOI: 10.1103/PhysRevLett.110.108301 (cited on pages 37, 38, 40, 43, 48, 55, 127, 151).

[117] Raffaello Potestio et al. 'Monte Carlo Adaptive Resolution Simulation of Multicomponent Molecular Liquids'. In: *Phys. Rev. Lett.* 111 (6 Aug. 2013), p. 060601. DOI: 10.1103/PhysRevLett.111.060601 (cited on pages 37, 40, 127, 151).

[118] S. Fritsch et al. 'Adaptive resolution molecular dynamics simulation through coupling to an internal particle reservoir'. In: *Phys. Rev. Lett.* 108 (17 Apr. 2012), p. 170602. DOI: 10.1103/PhysRevLett.108.170602 (cited on pages 37, 41, 56, 127, 151).

[119] Aoife C. Fogarty, Raffaello Potestio, and Kurt Kremer. 'Adaptive resolution simulation of a biomolecule and its hydration shell: Structural and dynamical properties'. In: *J. Chem. Phys.* 142.19, 195101 (2015), p. 195101. DOI: http://dx.doi.org/10.1063/1.4921347 (cited on pages 38–40, 48, 63, 65).

[120] Karsten Kreis et al. 'Adaptive Resolution Simulations with Self-Adjusting High-Resolution Regions'. In: *Journal of Chemical Theory and Computation* 12.8 (2016). PMID: 27384753, pp. 4067–4081. DOI: 10.1021/acs.jctc.6b00440 (cited on page 38).

[121] S. Fritsch, C. Junghans, and K. Kremer. 'Structure Formation of Toluene around C60: Implementation of the Adaptive Resolution Scheme (AdResS) into GROMACS'. In: *J. Chem. Theory Comput.* 8.2 (Feb. 2012), pp. 398–403 (cited on pages 38, 48, 65).

[122] Matej Praprotnik, Luigi Delle Site, and Kurt Kremer. 'Adaptive resolution scheme for efficient hybrid atomistic-mesoscale molecular dynamics simulations of dense liquids'. In: *Phys. Rev. E* 73 (6 June 2006), p. 066701. DOI: 10.1103/PhysRevE.73.066701 (cited on pages 38, 40, 41).

[123] M. Heidari et al. 'Accurate and general treatment of electrostatic interaction in Hamiltonian adaptive resolution simulations'. In: *Eur. Phys. J. Spec. Top.* 225.8 (2016), pp. 1505–1526. DOI: 10.1140/epjst/e2016-60151-6 (cited on pages 39, 58).

[124] Silvina Matysiak et al. 'Modeling diffusive dynamics in adaptive resolution simulation of liquid water'. In: *J. Chem. Phys.* 128.2, 024503 (2008), p. 024503. DOI: http://dx.doi.org/10.1063/1.2819486 (cited on pages 40, 48).

[125] Margaret E. Johnson, Teresa Head-Gordon, and Ard A. Louis. 'Representability problems for coarse-grained water potentials'. In: *J. Chem. Phys.* 126.14, 144509 (2007), p. 144509. DOI: http://dx.doi.org/10.1063/1.2715953 (cited on pages 41, 56).

[126] Han Wang, Christoph Junghans, and Kurt Kremer. 'Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining?' In: *Eur. Phys. J. E* 28.2 (2009), pp. 221–229. DOI: 10.1140/epje/i2008-10413-5 (cited on pages 41, 56).

[127] A A Louis. 'Beware of density dependent pair potentials'. In: *J. Phys.: Condens. Matt.* 14.40 (2002), p. 9187 (cited on pages 41, 56).

[128] Berk Hess et al. 'GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation'. In: *Journal of Chemical Theory and Computation* 4.3 (2008). PMID: 26620784, pp. 435–447. DOI: 10.1021/ct700301q (cited on pages 42, 44, 61, 80, 91, 116).

[129] Jonathan D. Halverson et al. 'ESPResSo++: A modern multiscale simulation package for soft matter systems'. In: *Comput. Phys. Commun.* 184.4 (2013), pp. 1129–1149. DOI: http://dx.doi.org/10.1016/j.cpc.2012.12.004 (cited on pages 42, 43, 60, 80, 91, 116).

[130] Horacio V. Guzman et al. 'ESPResSo++ 2.0: Advanced methods for multiscale molecular simulation'. In: *Computer Physics Communications* 238 (2019), pp. 66–76. DOI: https://doi.org/10.1016/j.cpc.2018.12.017 (cited on pages 42, 43, 80, 91, 116).

[131] Karsten Kreis et al. 'A unified framework for force-based and energy-based adaptive resolution simulations'. In: *EPL (Europhysics Letters)* 108.3 (Nov. 2014), p. 30007. DOI: 10.1209/0295-5075/108/30007 (cited on page 43).

[132] K. Kreis et al. 'Advantages and challenges in coupling an ideal gas to atomistic models in adaptive resolution simulations'. In: *The European Physical Journal Special Topics* 224.12 (2015), pp. 2289–2304 (cited on page 43).

[133] John G. Kirkwood. 'Statistical Mechanics of Fluid Mixtures'. In: *J. Chem. Phys.* 3.5 (1935), pp. 300–313. DOI: 10.1063/1.1749657 (cited on pages 43, 48, 54, 74, 91).

[134] J. H. Peters, R. Klein, and L. Delle Site. 'Simulation of macromolecular liquids with the adaptive resolution molecular dynamics technique'. In: *Phys. Rev. E* 94 (Aug. 2016), p. 047701. DOI: 10.1103/PhysRevE.94.023309 (cited on page 48).

[135] Julija Zavadlav, Rudolf Podgornik, and Matej Praprotnik. 'Adaptive Resolution Simulation of a DNA Molecule in Salt Solution'. In: *J. Chem. Theory Comput.* 11.10 (2015). PMID: 26574288, pp. 5035–5044. DOI: 10.1021/acs.jctc.5b00596 (cited on page 48).

[136] Matej Praprotnik et al. 'Adaptive resolution simulation of liquid water'. In: *J. Phys. Condens. Matt.* 19.29 (2007), p. 292201 (cited on page 48).

[137] Francesca Stanzione and Arthi Jayaraman. 'Hybrid Atomistic and Coarse-Grained Molecular Dynamics Simulations of Polyethylene Glycol (PEG) in Explicit Water'. In: *J. Phys. Chem. B* 120.17 (2016). PMID: 27108869, pp. 4160–4173. DOI: 10.1021/acs.jpcb.6b02327 (cited on page 48).

[138] Jinglong Zhu, Rupert Klein, and Luigi Delle Site. 'Adaptive molecular resolution approach in Hamiltonian form: An asymptotic analysis'. In: *Phys. Rev. E* 94 (4 Oct. 2016), p. 043321. DOI: 10.1103/PhysRevE.94.043321 (cited on page 48).

[139] Animesh Agarwal et al. 'Chemical potential of liquids and mixtures via adaptive resolution simulation'. In: *J. Chem. Phys.* 141.3, 034102 (2014), p. 034102. DOI: http://dx.doi.org/10.1063/1.4886807 (cited on pages 48, 55).

[140] Michael R. Shirts et al. 'Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins'. In: *J. Chem. Phys.* 119.11 (2003), pp. 5740–5761. DOI: http://dx.doi.org/10.1063/1.1587119 (cited on pages 48, 59, 60, 64, 80, 116).

[141] A.K. Soper. 'Empirical potential Monte Carlo simulation of fluid structure'. In: *Chem. Phys.* 202.2–3 (1996), pp. 295–306. DOI: http://dx.doi.org/10.1016/0301-0104(95)00357-6 (cited on page 50).

[142] K. Kreis et al. 'Advantages and challenges in coupling an ideal gas to atomistic models in adaptive resolution simulations'. In: *Eur. Phys. J. Spec. Top.* 224 (12 2015), pp. 2289–2304 (cited on page 50).

[143] Luigi Delle Site. 'Some fundamental problems for an energy-conserving adaptive-resolution molecular dynamics scheme'. In: *Phys. Rev. E* 76 (4 Oct. 2007), p. 047701. DOI: 10.1103/PhysRevE.76.047701 (cited on pages 52, 54).

[144] Staš Bevc et al. 'Adaptive resolution simulation of salt solutions'. In: *New J. Phys.* 15.10 (2013), p. 105007 (cited on page 56).

[145] Wendy D. Cornell et al. 'A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules'. In: *J. Am. Chem. Soc.* 117.19 (1995), pp. 5179–5197. DOI: 10.1021/ja00124a002 (cited on page 59).

[146] William L. Jorgensen et al. 'Comparison of simple potential functions for simulating liquid water'. In: *The Journal of Chemical Physics* 79.2 (1983), pp. 926–935. DOI: 10.1063/1.445869 (cited on pages 59, 79, 115, 139).

[147] Victor Rühle et al. 'Versatile Object-Oriented Toolkit for Coarse-Graining Applications'. In: *J. Chem. Theory Comput.* 5.12 (2009), pp. 3211–3223. DOI: 10.1021/ct900369w (cited on page 59).

[148] M.J. Abraham et al. 'The GROMACS development team, GROMACS User Manual version 5.0.4'. In: *GROMACS User Manual version 5.0.4* (2014). www.gromacs.org (cited on pages 59, 80).

[149] Shuichi Miyamoto and Peter A. Kollman. 'Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models'. In: *J. Comput. Chem.* 13.8 (1992), pp. 952–962. DOI: 10.1002/jcc.540130805 (cited on pages 60, 80, 116, 139).

[150] Hans C Andersen. 'Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations'. In: *J. Comput. Phys.* 52.1 (1983), pp. 24–34. DOI: http://dx.doi.org/10.1016/0021-9991(83)90014-1 (cited on pages 60, 80, 116, 139).

[151] Student. 'The Probable Error of a Mean'. In: *Biometrika* 6.1 (1908), pp. 1–25. DOI: 10.2307/2331554 (cited on pages 61, 80).

[152] Sreeja Parameswaran and David L. Mobley. 'Box size effects are negligible for solvation free energies of neutral solutes'. In: *Journal of Computer-Aided Molecular Design* 28.8 (2014), pp. 825–829. DOI: 10.1007/s10822-014-9766-7 (cited on page 61).

[153] R. Wolfenden et al. 'Affinities of amino acid side chains for solvent water'. In: *Biochemistry* 20.4 (1981). PMID: 7213619, pp. 849–855. DOI: 10.1021/bi00507a030 (cited on page 64).

[154] Boyce SE et al. 'Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site.' In: *J Mol Bio.* (2009), pp. 747–763. DOI: 10.1016/j.jmb.2009.09.049 (cited on pages 70, 74, 89).

[155] Aldeghi M., Bluck J.P., and Biggin P.C. 'Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide'. In: *Computational Drug Discovery and Design* 1762 (2018), pp. 199–232. DOI: 10.1007/978-1-4939-7756-7_11 (cited on pages 70, 74, 89).

[156] Stefan Boresch et al. 'Absolute Binding Free Energies: A Quantitative Approach for Their Calculation'. In: *J. Phys. Chem. B* 107.35 (2003), pp. 9535–9551. DOI: 10.1021/jp0217839 (cited on pages 70, 74–76, 89–91).

[157] Zoe Cournia, Bryce Allen, and Woody Sherman. 'Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations'. In: *Journal of Chemical Information and Modeling* 57.12 (2017). PMID: 29243483, pp. 2911–2937. DOI: 10.1021/acs.jcim.7b00564 (cited on page 70).

[158] Robert Abel et al. 'Advancing Drug Discovery through Enhanced Free Energy Calculations'. In: *Accounts of Chemical Research* 50.7 (2017). PMID: 28677954, pp. 1625–1632. DOI: 10.1021/acs.accounts.7b00083 (cited on page 70).

[159] B. N. Dominy. 'Molecular Recognition and Binding Free Energy Calculations in Drug Development'. In: *Current Pharmaceutical Biotechnology* 9.2 (2008), pp. 87–95. DOI: 10.2174/138920108783955155 (cited on page 70).

[160] Marilisa Neri et al. 'Coarse-Grained Model of Proteins Incorporating Atomistic Detail of the Active Site'. In: *Phys. Rev. Lett.* 95 (21 Nov. 2005), p. 218102. DOI: 10.1103/PhysRevLett.95.218102 (cited on pages 70, 72, 73, 128).

[161] Marilisa Neri et al. 'Microseconds Dynamics Simulations of the Outer-Membrane Protease T'. In: *Biophysical Journal* 94.1 (2008), pp. 71–78. DOI: https://doi.org/10.1529/biophysj.107.116301 (cited on pages 70, 72).

[162] MatÃas Rodrigo Machado, Pablo Daniel Dans, and Sergio Pantano. 'A hybrid all-atom/coarse grain model for multiscale simulations of DNA'. In: *Phys. Chem. Chem. Phys.* 13 (40 2011), pp. 18134–18144. DOI: 10.1039/C1CP21248F (cited on pages 70, 72).

[163] Andrea Amadei, Antonius B. M. Linssen, and Herman J. C. Berendsen. 'Essential dynamics of proteins'. In: *Proteins: Structure, Function, and Bioinformatics* 17.4 (1993), pp. 412–425 (cited on page 71).

[164] Vincenzo Carnevale et al. 'Convergent Dynamics in the Protease Enzymatic Superfamily'. In: *J. Am. Chem. Soc.* 2 (2006), pp. 173–181 (cited on page 71).

[165] Andrea Zen et al. 'Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families'. In: *Protein Sci.* 17 (2008), pp. 918–929 (cited on page 71).

[166] K. Hinsen. 'Analysis of domain motions by approximate normal mode calculations'. In: *Proteins* 33 (1998), pp. 417–429 (cited on page 71).

[167] M Delarue and Y H Sanejouand. 'Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model'. In: *J Mol Biol* 320.5 (2002), pp. 1011–1024 (cited on page 71).

[168] C. Micheletti, P. Carloni, and A. Maritan. 'Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models'. In: *Proteins* 55.3 (May 2004), pp. 635–645 (cited on pages 71, 72).

[169] Tod D. Romo and Alan Grossfield. 'Validating and improving elastic network models with molecular dynamics simulations'. In: *Proteins: Structure, Function, and Bioinformatics* 79.1 (2011), pp. 23–34. DOI: 10.1002/prot.22855 (cited on pages 71, 103).

[170] R. Potestio, F. Pontiggia, and C. Micheletti. 'Coarse-grained description of proteins' internal dynamics: an optimal strategy for decomposing proteins in rigid subunits'. In: *Biophys J* 96 (2009) (cited on page 71).

[171] H. Golhlke and M. F. Thorpe. 'A natural coarse graining for simulating large biomolecular motion'. In: *Biophysical Journal* 91 (2006), pp. 2115–2120 (cited on page 71).

[172] Zhiyong Zhang et al. 'A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules'. In: *Biophysical Journal* 95.11 (2008), pp. 5073–5083 (cited on page 71).

[173] Zhiyong Zhang et al. 'Defining Coarse-Grained Representations of Large Biomolecules and Biomolecular Complexes from Elastic Network Models'. In: *Biophysical Journal* 97.8 (2009), pp. 2327–2337 (cited on page 71).

[174] Zhiyong Zhang and Gregory A. Voth. 'Coarse-Grained Representations of Large Biomolecular Complexes from Low-Resolution Structural Data'. In: *Journal of Chemical Theory and Computation* 6.9 (2010), pp. 2990–3002 (cited on page 71).

[175] Anton V. Sinitskiy, Marissa G. Saunders, and Gregory A. Voth. 'Optimal Number of Coarse-Grained Sites in Different Components of Large Biomolecular Complexes'. In: *The Journal of Physical Chemistry B* 116.29 (2012). PMID: 22276676, pp. 8363–8374 (cited on page 71).

[176] Guido Polles et al. 'Mechanical and Assembly Units of Viral Capsids Identified via Quasi-Rigid Domain Decomposition'. In: *PLOS Computational Biology* 9.11 (Nov. 2013), pp. 1–13 (cited on page 71).

[177] Thomas T. Foley, M. Scott Shell, and W. G. Noid. 'The impact of resolution upon entropy and information in coarse-grained models'. In: *The Journal of Chemical Physics* 143.24 (2015), p. 243104 (cited on pages 71, 72).

[178] John D. Weeks, David Chandler, and Hans C. Andersen. 'Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids'. In: *The Journal of Chemical Physics* 54.12 (1971), pp. 5237–5247. DOI: 10.1063/1.1674820 (cited on pages 79, 96, 110).

[179] M. Parrinello and A. Rahman. 'Polymorphic transitions in single crystals: A new molecular dynamics method'. In: *Journal of Applied Physics* 52.12 (1981), pp. 7182–7190. DOI: 10.1063/1.328693 (cited on pages 79, 115, 139).

[180] Karl N. Kirschner et al. 'GLYCAM06: A generalizable biomolecular force field. Carbohydrates'. In: *Journal of Computational Chemistry* 29.4 (2008), pp. 622–655. DOI: 10.1002/jcc.20820 (cited on page 79).

[181] William Humphrey, Andrew Dalke, and Klaus Schulten. 'VMD – Visual Molecular Dynamics'. In: *Journal of Molecular Graphics* 14 (1996), pp. 33–38 (cited on page 98).

[182] T. Aleksiev et al. 'PiSQRD: a web server for decomposing proteins into quasi-rigid dynamical domains'. In: *Bioinformatics* 25.20 (Aug. 2009), pp. 2743–2744. DOI: 10.1093/bioinformatics/btp512 (cited on pages 101, 103).

[183] Raffaello Potestio, Francesco Pontiggia, and C Micheletti. 'Coarse-Grained Description of Protein Internal Dynamics: An Optimal Strategy for Decomposing Proteins in Rigid Subunits'. In: *Biophysical journal* 96 (July 2009), pp. 4993–5002. DOI: 10.1016/j.bpj.2009.03.051 (cited on pages 101, 103, 106).

[184] Osman Burak Okan, Ali Rana Atilgan, and Canan Atilgan. 'Nanosecond Motions in Proteins Impose Bounds on the Timescale Distributions of Local Dynamics'. In: *Biophysical Journal* 97.7 (2009), pp. 2080–2088. DOI: 10.1016/j.bpj.2009.07.036 (cited on page 102).

[185] M. Kurplus and J. A. McCammon. 'DYNAMICS OF PROTEINS: ELEMENTS AND FUNCTION'. In: *Annual Review of Biochemistry* 52.1 (1983). PMID: 6351724, pp. 263–300. DOI: 10.1146/annurev.bi.52.070183.001403 (cited on page 102).

[186] Ken A. Dill and Justin L. MacCallum. 'The Protein-Folding Problem, 50 Years On'. In: *Science* 338.6110 (2012), pp. 1042–1046. DOI: 10.1126/science.1219021 (cited on page 102).

[187] Qiang Shi, Sergei Izvekov, and Gregory A. Voth. 'Mixed Atomistic and Coarse-Grained Molecular Dynamics: a Simulation of a Membrane-Bound Ion Channel'. In: *The Journal of Physical Chemistry B* 110.31 (2006). PMID: 16884212, pp. 15045–15048. DOI: 10.1021/jp062700h (cited on pages 102, 127).

[188] Tsjerk A. Wassenaar et al. 'Computational Lipidomics with insane: A Versatile Tool for Generating Custom Membranes for Molecular Simulations'. In: *Journal of Chemical Theory and Computation* 11.5 (2015). PMID: 26574417, pp. 2144–2155. DOI: 10.1021/acs.jctc.5b00209 (cited on pages 102, 127).

[189] Sereina Riniker, Andreas P. Eichenberger, and Wilfred F. van Gunsteren. 'Solvating atomic level fine-grained proteins in supra-molecular level coarse-grained water for molecular dynamics simulations'. In: *European Biophysics Journal* 41.8 (2012), pp. 647–661. DOI: 10.1007/s00249-012-0837-1 (cited on pages 102, 127).

[190] R Affleck et al. 'Enzymatic catalysis and dynamics in low-water environments.' In: *Proceedings of the National Academy of Sciences* 89.3 (1992), pp. 1100–1104. DOI: 10.1073/pnas.89.3.1100 (cited on page 103).

[191] Alexander M. Klibanov. 'Why are enzymes less active in organic solvents than in water?' In: *Trends in Biotechnology* 15.3 (Jan. 1997), pp. 97–101. DOI: 10.1016/S0167-7799(97)01013-5 (cited on page 103).

[192] Kyung Hwan Kim and R. Maynard Case. 'Effects of Pancreatic Polypeptide on the Secretion of Enzymes and Electrolytes by in Vitro Preparations of Rat and Cat Pancreas'. In: *Yonsei Med J* 21.2 (Dec. 1980), pp. 99–105 (cited on page 104).

[193] Lin T and Chance R. 'Bovine Pancreatic Polypeptide (BPP) and Avian Pancreatic Polypeptide (APP)'. In: *Gastroenterology* 67 (1974), pp. 737–738 (cited on page 104).

[194] Kimmel JR et al. 'Pancreatic polypeptide from rat pancreas'. In: *Endocrinology* 114 (1984), pp. 1725–1731 (cited on page 104).

[195] Janos Lonovics et al. 'Pancreatic Polypeptide: A Review'. In: *Archives of Surgery* 116.10 (Oct. 1981), pp. 1256–1264 (cited on page 104).

[196] Schwartz T. 'Pancreatic Polypeptide: A Hormone Under Vagal Control'. In: *Gastroenterology* 85 (1983), pp. 1411–1425 (cited on page 104).

[197] Taylor I. 'Pancreatic polypeptide family: pancreatic polypeptide, neuropeptide Y, and peptide YY.' In: *Handbook of Physiology: The Gastrointestinal System* 2 (1989) (cited on page 104).

[198] Thue W. Schwartz. 'Pancreatic Polypeptide: A Hormone Under Vagal Control'. In: *Gastroenterology* 85.6 (1983), pp. 1411–1425. DOI: doi:10.1016/S0016-5085(83)80027-4 (cited on page 104).

[199] Xiang Li et al. 'Sequence-specific proton NMR assignments and solution structure of bovine pancreatic polypeptide'. In: *Biochemistry* 31.4 (1992). PMID: 1734969, pp. 1245–1253. DOI: 10.1021/bi00119a038 (cited on pages 104, 118).

[200] W. Kabsch. 'A solution for the best rotation to relate two sets of vectors'. In: *Acta Crystallographica Section A* 32.5 (1976), pp. 922–923. DOI: 10.1107/S0567739476001873 (cited on page 107).

[201] J.A. Barker and R.O. Watts. 'Monte Carlo studies of the dielectric properties of water-like models'. In: *Molecular Physics* 26.3 (1973), pp. 789–792. DOI: 10.1080/00268977300102101 (cited on pages 116, 139).

[202] R.O. Watts. 'Monte Carlo studies of liquid water'. In: *Molecular Physics* 28.4 (1974), pp. 1069–1083. DOI: 10.1080/00268977400102381 (cited on pages 116, 139).

[203] CW Müller et al. 'Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding'. In: *Structure* 4.2 (1996), pp. 147–156. DOI: https://doi.org/10.1016/S0969-2126(96)00018-4 (cited on pages 129, 130, 148).

[204] Petras Dzeja and Andre Terzic. 'Adenylate Kinase and AMP Signaling Networks: Metabolic Monitoring, Signal Communication and Body Energy Sensing'. In: *International Journal of Molecular Sciences* 10.4 (2009), pp. 1729–1772 (cited on page 129).

[205] Elena Formoso, Vittorio Limongelli, and Michele Parrinello. 'Energetics and Structural Characterization of the large-scale Functional Motion of Adenylate Kinase'. In: *Scientific reports* 5.1 (2015) (cited on page 129).

[206] Peter A. Burrough, Rachael A. McDonnell, and Christopher D. Lloyd. *Principles of Geographical Information Systems*. 3rd. Oxford, 2015 (cited on page 132).

[207] Paul A. Longley et al. *Geographic Information Systems and Science*. 3rd. Wiley Publishing, 2010 (cited on page 132).

[208] Zekai Sen. *Spatial Modeling Principles in Earth Sciences*. 2nd. Springer, 2016 (cited on page 132).

[209] Franz Aurenhammer. 'Voronoi Diagrams&Mdash;a Survey of a Fundamental Geometric Data Structure'. In: *ACM Comput. Surv.* 23.3 (Sept. 1991), pp. 345–405. DOI: 10.1145/116873.116880 (cited on page 132).

[210] Atsuyuki Okabe et al. *Spatial Tessellations – Concepts and Applications of Voronoi Diagrams*. 2nd. Wiley Publishing, 2000 (cited on page 132).

[211] M.J. Abraham et al. 'GROMACS User Manual version 2019'. In: (2019). http://www.gromacs.org (cited on page 139).