University of Trento

Department of Physics

Thesis submitted for the Degree of
Doctor of Philosophy

The mapping problem in coarse-grained modelling of biomolecules

Author:
Marco Giulini

Supervisors:
Prof. Raffaello Potestio
Dr. Roberto Menichetti

Academic year 2021/2022

Contents

Introduction

Nucleic acids, lipids, proteins and sugars. These are the major biological macromolecules that interact together to form and sustain living organisms. Among all these players, proteins deserve a particular prominence because they are heavily involved in almost all biological processes, such as reaction catalysis, signalling and cell structure formation. Proteins are unbranched chains of amino acids, whose length ranges from few to tens of thousands of elementary units. They originate in the ribosome, where a sequence of messenger RNA is translated into a chain of amino acids that reaches its preferred three-dimensional conformation, or *native state*, through the process of protein folding. The folded protein should not be viewed as a static structure, but rather as a dynamical entity, as it continues to change shape in the cell.

Given their centrality in biological processes, protein structures and dynamics are extensively studied from both experimental and computational perspectives. Experimental techniques are capable of determining high-resolution native structures of proteins with a series of diverse methods, such as solution Nuclear Magnetic Resonance (NMR) [1], X-ray crystallography [2] and cryo-electron microscopy [3]. Since dynamical properties are more difficult and time-consuming to infer from wet-lab experiments, computational tools have been more and more successfully employed to reconstruct the behaviour of proteins in solution.

The most popular instrument to investigate the dynamical properties of biomolecules at the atomic scale is all-atom molecular dynamics (MD) [4, 5], a set of algorithms that treat the time evolution of a system with classical mechanics and numerically solve Newton's equations of motions of atomic nuclei. Data from MD experiments have proved to be of invaluable importance in a huge variety of scientific problems, such as protein folding, protein conformational changes, protein-ligand binding, protein-membrane interaction and many others [6, 7]. However, a major limitation of MD resides in the multiple time scales involved in the behaviour of a protein; for instance, the motion of hydrogen atoms is extremely fast, thus requiring a very small integration time step (1 or 2 femtoseconds) to avoid numerical

instabilities leading to unphysical sampling, whereas the characteristic time scales proper of slow biological processes can exceed the millisecond. As an example, a medium-size protein typically requires seconds or even minutes to fold properly [8, 9]. In order to sample interesting events one has to concatenate a huge number of short time steps. It is therefore evident how the full potential of MD is often hampered by the sampling problem: the biologically relevant events usually take place on a time scale that is hard to reach with the available computational tools.

The performances of MD engines improve on a yearly basis thanks to the tremendous technological advancements on software implementations and hardware, allowing one to sensibly reduce the physical time required to perform a single integration step, thus extending the time scales reachable by all-atom MD. Nevertheless, many interesting biological processes remain extremely challenging to tackle with unbiased MD simulations.

In this context, an approach to overcome the sampling challenge in biomolecular simulations relies on a set of methods aiming at accelerating the exploration of the conformational space by applying tailored modifications to the original, reference *all-atom, unbiased* MD protocol. In a first class of strategies, called enhanced sampling techniques, the resolution of the system is kept fixed at the atomistic (AT) level and the Hamiltonian of the system is modified so as to visit more frequently particularly interesting regions of the conformational space of a molecule. Examples include the observation of barrier crossing events (or transitions) between two metastable basins, whose probability is exponentially vanishing in the height of the free energy barrier. Among these methods let me cite metadynamics [10], umbrella sampling [11], temperature accelerated MD [12], replica exchange MD [13] and thermodynamic integration [14].

A second class of methods exist, which convert the highly detailed, atomistic description of the system into a simplified representation, or *mapping*, in terms of a lower number of degrees of freedom, called *sites*. Once accurate effective interactions among sites are introduced, the resulting reduced system aims at reproducing the properties of its high resolution counterpart. These *coarse-grained* (CG) models have proved to be an invaluable instrument to tackle a huge variety of biological problems ranging from protein folding [15, 16, 17] to the dynamics of large macromolecular complexes [18, 19, 20], which are extremely difficult to be simulated at the all-atom level.

The construction of a CG model consists of two intertwined but distinct steps, which are the definition of the CG mapping and the accurate parametrisation of the interactions among CG sites (CG force field). While the latter challenge has received much attention in the past, the former problem has been object of a limited number of works, and a unique strategy to

construct the CG mapping does not exist yet. Indeed, the mapping scheme has almost always been an ingredient, rather than an outcome of CG models of biomolecules, imposed *a priori* using chemical intuition.

This thesis proposes a series of approaches to investigate and characterise the representation problem in coarse-grained modelling of proteins. This is achieved by employing a collection of diverse methods, including statistical mechanics, machine learning algorithms and information-theoretical tools.

The manuscript begins with an introduction to the world of computational biophysics, highlighting the fundamental concepts of protein science as well as the impressive advancements in the application of fully atomistic simulations to the study of biological systems. The main families of coarse-grained models [21] are then introduced, focusing on the fully bottom-up approaches, that is, those strategies that exploit information retrieved from high-resolution, atomistic simulations to construct accurate low-resolution models. The four major algorithmic procedures that have been proposed in the literature to determine accurate coarse-grained effective potentials are discussed. Among these, the *relative entropy* protocol [22] is described with a high level of detail.

The second chapter of the thesis is devoted to a comprehensive review of the coarse-grained representations that have been employed in the low-resolution modelling of proteins. In the first part of the the chapter the several levels of resolution that can be used to describe a biomolecule are discussed; then, the focus is shifted to the enumeration and discussion of models that treat a biomolecular system with a non-uniform level of detail. I conclude with a comprehensive analysis of the methods that have been proposed to optimize the choice of the coarse-grained mapping of a protein in an automated manner. The chapter should be considered as a personal summary of Ref. [23].

The reader that is not interested in this quite broad and comprehensive introduction to the mapping problem in coarse-graining can immediately jump from this introduction to Chapter 3, which is entirely devoted to a mathematical object called *mapping entropy*, that can be used to measure the quality of a reduced representation of a biomolecule. More specifically, the mapping entropy quantifies the loss of information arising from the removal of degrees of freedom from a fully atomistic structure operated by the coarse-grained mapping. The theoretical calculations connecting the relative to the mapping entropy are shown in detail, together with the approximations allowing the computation of the latter. Subsequently, I describe the numerical implementation of the mapping entropy minimisation scheme, an unsupervised procedure whose aim is to identify optimal, reduced representations of the molecule of interest. The last part of the chapter presents the

application of this protocol to three model proteins, heterogeneous in terms of size, shape and biological role. The chapter is entirely based on Ref. [24].

The computation and, consequently, the minimisation of the mapping entropy are particularly onerous from a computational point of view. The acceleration of these tasks lies at the core of the fourth chapter, which presents a protein-specific, graph-based machine learning algorithm able to speed-up the computation of the mapping entropy by several orders of magnitude. The novel combination of the trained neural network with the Wang-Landau enhanced sampling scheme allows the quasi-exhaustive exploration of the space of available reduced representations of a protein, providing the correct, unbiased, mapping entropy density of states. Ref. [25] is employed as reference for this chapter.

Given the possibility of exploring the huge space of coarse-grained mappings, it is instructive to investigate its structure and metric properties. This analysis is the subject of the fifth chapter of this thesis, which is based on Ref. [26]. In this context, a purely structural notion of scalar product, norm and distance between coarse-grained representations is introduced, and these tools are employed to explore and characterise the immense mapping space; such exploration leads to the emergence of qualitatively different mappings. The notion of distance developed here allows one to assess that representations with low mapping entropy are close to each other in the mapping space, thus proving that a geometrical similarity exists among them.

The sixth chapter of this manuscript is devoted to the analysis of the relationships between the mapping entropy and two other information-theoretical quantities, namely the resolution and the relevance [27]. Their properties are exploited to provide an unsupervised strategy to extract microstates out of molecular dynamics trajectories, each one weighted with its frequentist, non-Boltzmann, probability. This algorithmic procedure is then used to provide an alternative method to compute the mapping entropy of a coarse-grained representation. In the last sections of the chapter, resolution, relevance and mapping entropy are calculated for CG representations of two discrete systems, namely a set of non-interacting spins and a simple model of the Nasdaq stock market. The mapping entropy proves to be an extremely precise and useful tool to pinpoint the features providing an optimal coarse-graining of the system of interest.

The calculation of the mapping entropy, norm, and distance is implemented and freely available in the EXtensible COarse-GraIning TOol (EXCOGITO), a fast and flexible software suite, whose algorithmic and numerical details are explained and reviewed in the last chapter of the thesis.

The manuscript ends with a brief, critical discussion about the relevance of the presented methodologies, together with some personal perspectives.

Chapter 1

Modelling of proteins

In this chapter I first summarize few fundamental concepts about the structural and chemical properties of proteins. Then, I go through a concise overview of the basic strategies underlying MD simulations, with a focus on the semi-empirical potentials employed in the MD Hamiltonian. The major advancements achieved and challenges faced by the field of plain, all-atom MD simulations of protein systems are discussed. The chapter proceeds with a long introduction to low-resolution models, focusing on bottom-up coarse-graining strategies.

Proteins are unbranched, heterogeneous polymers constituted by a long sequence of elementary units, namely the twenty-one proteinogenic amino acids. These are chemical entities that share the same backbone structure, with an amino (NH_2) and a carboxyl (COOH) group on the two sides, both connected to a central C_α atom via single covalent bonds (C-C and C-N). The C_α carbon is involved in other two covalent interactions, namely one with a hydrogen atom and one with the R group, or side chain, which is the variable region of amino acids. Side chains differ immensely in size and chemical composition; for instance, glycine ($m \sim 75$ Daltons) is the only achiral amino acid, having an R group with only one hydrogen atom, while tryptophan ($m \sim 204$ Daltons) contains two aromatic rings in the side chain, which amounts at more than a half of its molecular weight. Two consecutive amino acids are patched together thanks to the formation of a peptide bond, with carboxyl and amino groups that lose an oxygen and two hydrogen atoms, respectively, thus resulting in the expulsion of a water molecule as a reaction byproduct. The properties of amino acids in solution are determined by the chemical features of their side chains, as schematically illustrated in Fig. 1.1.

The sequence of amino acids, also called primary structure of the protein, is univocally determined by the mRNA chain translated inside the ribosome. Once the amino acid sequence is translated by the ribosomal complex, local

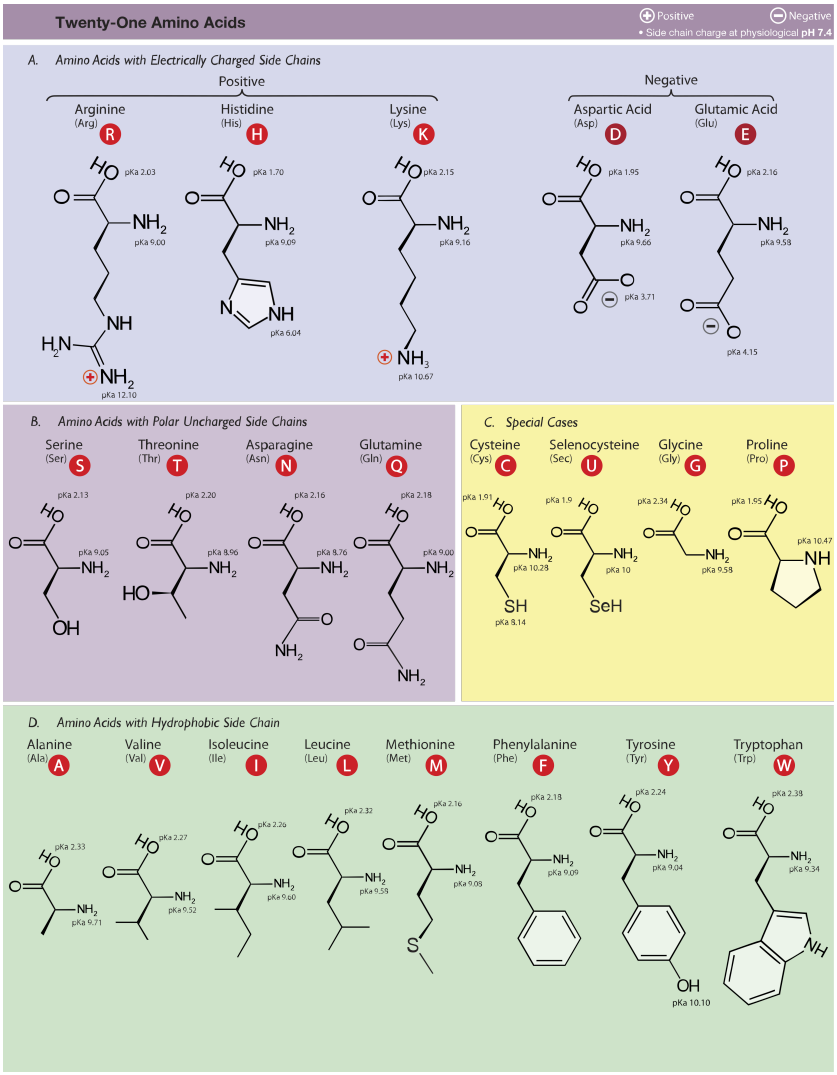


Figure 1.1: The proteinogenic amino acids, divided in groups according to the properties of their side chains. Charged residues tend to be located in regions that are exposed to the solvent, while hydrophobic amino acids cluster together and are usually buried in the core of the structure. Source: Wikimedia Commons

segments of the protein start to assume a specific three-dimensional arrangement, dictated by the surrounding, intramolecular hydrogen bond network. The set of local structural arrangements defines the secondary structure of the protein. There exist eight different secondary structure elements according to the DSSP method [28], which can be ultimately reduced to four major categories:

- *α helices*: rod-like objects where the side chains of the amino acids point towards the outside and the internal structure is stabilized by hydrogen bonds between CO and NH groups of the backbone chain;

- *β sheets*: sequences of *β strands*, bonded to each other by hydrogen bonds between NH groups on one strand and C=O on the other. β strands are subregions of the protein where the backbone shows a regular, zigzagging configuration. In a β sheet, the side chains lie orthogonal to the plane formed by the sheet, alternating between up and down orientations;

- *turns*: small portions of the protein where the direction of the polypeptide chain is reversed. They occur when there is a hydrogen bond between two non-consecutive amino acids separated by few peptide bonds;

- *loops*: patternless regions linking two secondary structure elements.

The overall three-dimensional organisation of the polypeptide chain is called *tertiary structure*, which is responsible for the basic function of the protein. In this arrangement, hydrophobic side chains of neutral amino acids are mainly localised in the internal region of the protein, while charged, hydrophilic side chains tend to be highly exposed to the solvent. Furthermore, disulfide bonds and salt bridges are essential interactions to stabilize this native state. Multiple tertiary structures can combine and fold into an assembly called *quaternary structure*. Fig. 1.2 schematically shows this hierarchy of structures for an example system.

Proteins are the key elements of an immense variety of biological processes, such as enzymatic reactions, signalling cascades and the formation of macromolecular complexes. Understanding the non-trivial structural, thermodynamic and energetic properties of proteins and protein assemblies is fundamental to understand and rationalise their role in the huge zoo of cellular processes.

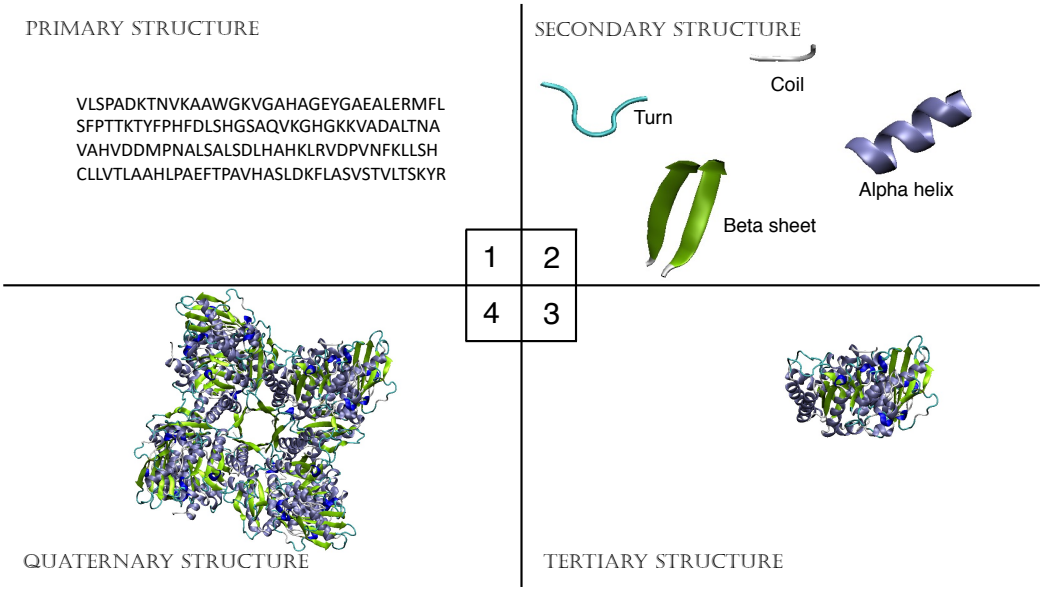


Figure 1.2: The four levels of structures for the Rubisco protein (PDB code 1BXN), namely [1]: the sequence of amino-acids, [2]: the secondary structure elements, [3]: the tertiary, three-dimensional arrangement, [4]: the assembly of several tertiary structures in the overall complex.

1.1 All-atom simulations

In the previous section the biochemical and structural properties of proteins have been discussed from an empirical point of view. Let me now introduce the key physical concepts that lie behind the field of theoretical modelling of these complex systems.

With the notable exception of metalloproteins, proteins can be viewed as sets of atoms with only five different atomic species (carbon, hydrogen, nitrogen, oxygen and sulphur) and with a very peculiar pattern of interactions. In physiological conditions, these objects are immersed in a solution of water and ions.

Just like any other molecular systems, proteins are composed by atoms that follow the laws of quantum mechanics, and in principle should be modelled with the corresponding tools and methods such as Density Functional Theory [29] and Quantum Monte Carlo [30]. Nevertheless, the relatively huge size of biomolecules and their relevant time scales cannot be approached by such computationally heavy strategies. Therefore, it is common practice to perform the Born-Oppenheimer approximation and to separate the motions of the electrons from those of the nuclei. In classical all-atom molecular dy-

namics [4, 5] calculations, the electronic degrees of freedom are not explicitly considered, but rather implicitly integrated in the model, and the atomic nuclei are viewed as point-like particles, whose positions and momenta are evolved in time using the laws of classical mechanics.

The effective, interatomic potential, or *force field* that governs the interaction between the atoms is derived both from experiments and *ab-initio* calculations. It can be approximated as

$$V_{MD} = V_{prot-prot} + V_{prot-solv} + V_{solv-solv}, \tag{1.1}$$

where $V_{prot-prot}$ is the intramolecular potential describing the interactions among the protein constituent atoms, which is often expressed in the following, approximated form:

$$\begin{aligned} V_{prot-prot} &= \sum_{bonds} k_b \left(b - b_0\right)^2 + \sum_{angles} k_\theta \left(\theta - \theta_0\right)^2 + \\ &+ \sum_{dihedrals} k_\theta \left[1 + cos\left(n\phi - \delta\right)\right] + \sum_{impr} k_\omega \left(\omega - \omega_0\right)^2 + \\ &+ \sum_{ij} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{ij} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}. \end{aligned} \tag{1.2}$$

In Eq. 1.2, the terms associated to bonds, angles and improper dihedrals are described by harmonic potentials, each one with a proper equilibrium value $(b_0, \theta_0, \omega_0)$ and spring constant $(k_b, k_\theta, k_\omega)$. Dihedral terms are parametrised with a Fourier series usually truncated at the fifth or sixth term, in order to take into account for multiple, acceptable values of the dihedral angle. Classical mechanics cannot incorporate reactive chemistry, which is intimately quantum in nature, and therefore bonds cannot be created nor broken during a MD simulation. The topology, defined as the set of covalent interactions present in the system, remains constant in MD.

The last line of Eq. 1.2 contains non-bonded terms, which account for the presence of electrostatic and van der Waals interactions between pairs of atoms (i and j) that are not covalently bonded; these interactions are usually treated with a sum of Coulomb and Lennard-Jones potentials, where r_{ij} is the distance between the atoms, ϵ_0 is the vacuum electrostatic constant, and ϵ_{ij} and σ_{ij} are pair-specific energy and distance parameters for the Lennard-Jones terms.

In addition to the intramolecular potential acting among the protein's constituent atoms, the last two terms of Eq. 1.1 account for the presence of the solvent in the simulation. $V_{prot-solv}$ is constituted by the set of non-covalent interactions regulating the interplay between the protein structure

and the solvent, and is therefore a sum of non-bonded terms similar to those in the last line of Eq. 1.2. $V_{solv-solv}$ contains instead non-covalent and covalent terms, where the latter are introduced to maintain the correct topology of solvent molecules. As an example, almost all the existing classical, explicit models of water [31] impose harmonic constraints on the oxygen-hydrogen bond length and on the hydrogen-oxygen-hydrogen angle.

The accuracy of atomistic, semi-empirical force fields such as CHARMM [32] and AMBER [33] is improving on a daily basis thanks to the huge quantity of data obtained from experimental observations and quantum mechanical, *ab initio* calculations. For instance, the accurate treatment of intrinsically disordered proteins [34, 35] and amyloid assemblies [36] has been incorporated in several force fields.

In parallel with the improvements on the atomistic force fields, recent technological advancements such as GPU computing, special purpose architectures and distributed computing are steadily pushing the application of plain, all-atom MD to the investigation of previously unconceivable timescales. Graphic processing units find a fertile ground in the realm of biomolecular simulations, and the GPU implementation of popular MD softwares represents a turning point for the whole field of computational biophysics [37, 38]. The supercomputer ANTON [39] is specifically tailored and optimised to run MD simulations: in 2010 Lindorff-Larsen et al. [40] employed this architecture in the first computational experiment on protein folding that used all-atom MD on the scale of the microseconds. The newly released version of ANTON, ANTON 3 [41], promises to reach a performance peak such that millisecond-long simulations are expected to be run in a workweek. With respect to distributed computing, it is crucial to highlight the tremendous effort carried out by the *Folding@home* consortium [42] in combining millions of personal computers around the globe to study protein folding and misfolding, with a recent focus on Sars-ncov-2 viral proteins [43].

Overall, the growing accuracy of atomistic force fields combined with the increasing computing power allows to approach relevant biological problems that were previously thought to be impossible to tackle with plain MD simulations. As an example, in a recent paper by Singharoy et al. [44], an atomistic model of a whole cell organelle, composed by ~ 100 million atoms, has been simulated in full detail. Nevertheless, the intrinsic complexity of the majority of biological processes makes them out of reach for atomistic simulations both now and in the foreseeable future. It is in this context that coarse-grained models, which describe biological systems with a level of detail lower than atomistic, can play an extremely relevant role.

1.2 Coarse-grained modelling

This section focusses on the conceptual and theoretical basis of equilibrium coarse-grained modelling of biomolecules. It begins with a brief introduction of the main concepts of this field and then proceeds to a formal distinction between the available methodologies developed to derive CG potentials, with particular attention dedicated to bottom-up coarse-graining.

Molecular coarse-graining is defined as the effective reduction of the number of degrees of freedom of an atomistic system. This is a process that requires two elements, namely a reduced representation of the system and effective interactions. The former, that is, the CG mapping, fixes the level of detail of the resulting low-resolution model and, therefore, its range of validity: for example, protein mappings that only consider the backbone atoms cannot provide any information about the side chains. The latter, namely the CG interactions, are introduced so as to reproduce the behaviour of the atomistic system once observed at a CG level.

CG models span several levels of resolution, ranging from few atoms per constituent unit (or CG *site*) up to few atoms per molecule and to continuous models, where molecules are represented with continuous objects such as density fields.

When the discussion is restricted to discrete models, the CG mapping $\mathbf{M}$ specifies the position of the CG site I, $\mathbf{R}_I$, with a linear combination of atomistic coordinates $\mathbf{r}_i$:

$$\mathbf{M}_I(\mathbf{r}) = \mathbf{R}_I = \sum_i c_{Ii}\mathbf{r}_i. \tag{1.3}$$

The linear coefficients c_{Ii} in Eq. 1.3 are constant, positive and subject to the normalisation condition $\sum_i c_{Ii} = 1$ to preserve translational invariance. Furthermore, in the vast majority of the cases [23], coefficients are generally taken to be *specific* to each site [21], that is, an atom i taking part to the definition of CG site I cannot be involved in the construction of another site J ($c_{Ji} = 0 \; \forall \; J \neq I$).

The vast majority of models proposed in the literature [21] consider a given, fixed choice of the CG mapping, focusing on the problem of correctly parametrizing effective interactions. In this respect, three main families of methods exist [1], which differ mainly in the *source of information* employed to construct the low-resolution potential.

Interactions in *top-down* CG models are built with the help of general principles with no prior assumptions on the existence of a more detailed,

[1] I here follow the distinction between CG models operated by W. Noid in Ref. [21].

atomistic model for the considered system. The functional form of the coarse-grained force field usually relies on a very simple basis set, such as the harmonic potentials, whose free parameters are determined by matching a set of selected macroscopic properties of the system, obtained mainly from experiments. These strategies always "work", although their predictive power is limited and dependent on the property used to build the model: there is no guarantee that a top-down model parametrised so as to reproduce a given structural or thermodynamic feature of the system will succeed in estimating another quantity that is not explicitly employed in the process of model construction. The power of this approach shows up in the fact that the implementation and simulation of top-down models requires limited system knowledge and computational resources: from the analysis of the behaviour of the CG system, one can qualitatively estimate which are the most important structural and chemical features that should be present in a more sophisticated model.

Knowledge-based models build CG potentials using a different philosophy. Often confused with top-down approaches, these methods rely on a dataset-wide approach to compute the parameters of the effective CG force-field, which are extracted from statistical analyses of local structural information present in the literature. Like all dataset-wide methods, knowledge-based potentials are prone to provide unphysical results when the features of the system of interest are not well represented in the data set.

The third class of methodologies, *bottom-up* CG, is based on the idea that the fundamental properties of a molecule should not be employed to parametrise the model, but rather they are expected to emerge systematically from an accurate coarse-graining of the fully detailed system. A bottom-up protocol is based on the exploitation of information obtained from a high-resolution model, that, in the case of biological molecules, is usually an atomistic MD simulation. If such detailed model does not exist or is not precise enough, bottom-up strategies can possibly succeed in reproducing the low-resolution observables of the high-resolution model, but are bound to fail in predicting experimentally measurable, emergent properties: as an example one can consider the case of DNA systems, which were coarse-grained using mainly top-down [45] or hybrid approaches until accurate atomistic force fields were developed [46, 47] and the fully bottom-up approach was viable [48].

Most of the models developed in the current days cannot be classified in only one category, but rather they exploit the strengths of all of them in an integrative fashion, for example by incorporating experimental information and data set statistics into constraints for bottom-up CG models.

A prominent example of an integrative CG model that has been success-

fully employed over the last two decades is the MARTINI CG model [49, 50], in which the non-bonded force field (electrostatic and Lennard-Jones terms) is determined from experimental data, while the bonded terms are constructed using reference atomistic simulations.

The next section is devoted to a comprehensive discussion of bottom-up strategies for coarse-grained modelling, while the interested readers are referred to the excellent review of W. Noid [21] for an exhaustive discussion of top-down and knowledge-based CG models.

1.2.1 Bottom-up coarse-graining

I here provide a general introduction to the main quantities that play a relevant role in bottom-up CG, together with a brief description of some approaches that enable to derive accurate CG potentials starting from a fine-grained model.

Bottom-up CG of proteins aims at constructing a reduced model with N sites exploiting information retrieved from an atomistic description of a molecule with $n \geq N$ atoms. Each configuration $\mathbf{r}$ of the high-resolution system is associated to its equilibrium probability that, in the case of the canonical ensemble, takes the form of the Boltzmann distribution:

$$p_r(\mathbf{r}) = \frac{1}{z} e^{-\beta u(\mathbf{r})}, \tag{1.4}$$

where $\beta = \frac{1}{k_B T}$ and $u(\mathbf{r})$ is the potential energy of the system (such as the ones in Eqs. 1.2 and 1.1). In Eq. 1.4 z is the standard atomistic configurational partition function [51]:

$$z = \int d\mathbf{r} e^{-\beta u(\mathbf{r})}, \tag{1.5}$$

where the integrals implicitly depend on the volume V, which is constant in the canonical ensemble.

The same procedure can be employed to derive the CG equilibrium distribution: upon fixing the mapping and the interactions, or CG potential $U(\mathbf{R})$, between different sites, the probability of sampling the CG configuration $\mathbf{R}$ can be written as

$$P_R(\mathbf{R}|U) = \frac{1}{Z_U} e^{-\beta U(\mathbf{R})}, \tag{1.6}$$

where Z is the CG partition function $Z_U = \int d\mathbf{R}\ e^{-\beta U(\mathbf{R})}$. Intuitively, one expects that the most accurate candidate CG model approximates perfectly the atomistic equilibrium probability distribution p_r. Attention must be paid to the fact that AT and CG models live in two different configurational

spaces, respectively with $3n$ and $3N$ dimensions: in order to compare AT and CG probabilities it is crucial to first convert p_r into its CG configurational space analogue, p_R, defined as

$$\begin{aligned} p_R(\mathbf{R}) &= \int d\mathbf{r}\ p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \\ &= \frac{1}{z}\int d\mathbf{r}\ e^{-\beta u(\mathbf{r})}\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \end{aligned} \tag{1.7}$$

where the functional delta is employed to restrict the integral to those configurations such that $\mathbf{M}(\mathbf{r}) = \mathbf{R}$. This quantity assigns to each CG configuration $\mathbf{R}$ a statistical weight equal to the sum of the atomistic probabilities of the microstates $\mathbf{r}$ that map onto it. This definition allows to introduce a quantity of crucial importance in bottom-up CG, that is, the *multi-body potential of mean force* (MB-PMF) U^0:

$$U^0(\mathbf{R}, T) = -k_B T \ln\left(\int d\mathbf{r}\ e^{-\beta u(\mathbf{r})}\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})\right) \tag{1.8}$$

$$= -k_B T \ln(p_R(\mathbf{R})) + const, \tag{1.9}$$

a state-dependent free energy that samples the CG configuration space *as if it was sampled by the high-resolution model*. In order words, if an atomistic simulation of the system is observed through the degrees of freedom specified by a CG mapping, the resulting free energy profile is identical to the one generated by a CG model simulated employing the MB-PMF. U^0 provides the exact potential that should be utilized to describe interactions at a resolution lower than the atomistic one. A CG model that samples the CG configuration space with an equilibrium probability distribution equal to $p_R(\mathbf{R})$ is said to be *consistent* with the atomistic model [52], and all the thermodynamical properties of the original system can be recovered [53].

As for any conventional free energy, the MB-PMF can be decomposed in an energetic and entropic contribution [54, 55]:

$$U^0(\mathbf{R}, T) = E^0(\mathbf{R}, T) - TS^0(\mathbf{R}, T), \tag{1.10}$$

where the energetic component is simply a canonical, temperature-dependent average of the atomistic potential energy function over the coarse-grained macrostate $\mathbf{R}$:

$$E^0(\mathbf{R}, T) = \frac{1}{p_R(\mathbf{R})}\int d\mathbf{r}\ p_r(\mathbf{r})u(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) = \langle u(\mathbf{r})\rangle_{\mathbf{R}}\,, \tag{1.11}$$

where $\langle ...\rangle_{\mathbf{R}}$ denotes a canonical average restricted to the CG configuration $\mathbf{R}$. The entropic component S^0 can be expressed as the temperature variation

of the MB-PMF [54, 56]:

$$\begin{aligned} S^0(\mathbf{R},T) &= -\left(\frac{\partial U^0(\mathbf{R},T)}{\partial T}\right)_{\mathbf{R}} \\ &= \frac{k_B}{p_R(\mathbf{R})}\int d\mathbf{r}\ p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r})-\mathbf{R})\ln\left(\frac{V^N p_R(\mathbf{R})}{V^n p_r(\mathbf{r})}\right) \end{aligned} \quad (1.12)$$

It is important to underline that this equation for S^0 does not assume that the energetic component E^0 is temperature-independent [54]. As discussed in Chapter 3, $S^0(\mathbf{R},T)$ is related to the mapping entropy, which is the crucial quantity investigated in this thesis.

Let me conclude this brief introduction to the properties of the MB-PMF by defining the coarse-grained force acting on site I, which can be computed by taking the first derivative of the potential with respect to the CG coordinate of site I:

$$\begin{aligned} \mathbf{F}_I^0(\mathbf{R}_I) &= -\frac{\partial U^0}{\partial \mathbf{R}_I} \\ &= \frac{-k_B T}{p_R(\mathbf{R})}\frac{\partial p_R(\mathbf{R})}{\partial \mathbf{R}_I}. \end{aligned} \quad (1.13)$$

It is possible to show [52, 57] that, assuming the specificity of at least one atom per CG site, the resulting force is given by:

$$\mathbf{F}_I^0(\mathbf{R}_I) = \frac{\int d\mathbf{r}\ p_r(\mathbf{r})\mathbf{f}_I(\mathbf{r})\delta(\mathbf{M}(\mathbf{r})-\mathbf{R})}{p_R(\mathbf{R})}, \quad (1.14)$$

where $\mathbf{f}_I(\mathbf{r})$ is the AT force acting on the CG site I. Eq. 1.14 is an average (with the usual canonical weight) of all possibles values of $\mathbf{f}_I(\mathbf{r})$ over all the atomistic configurations $\mathbf{r}$ that map onto $\mathbf{R}$. Finally, it is possible to appreciate how U^0 is capable of generating *mean forces* thanks to this averaging procedure over all the neglected degrees of freedom.

Unfortunately, U^0 is impossible to calculate except for extremely trivial cases because of the insurgence of many-body terms [58, 59], caused by the effective incorporation of the eliminated degrees of freedom into CG interactions operated by Eq. 1.8 [59]:

$$U^0(\mathbf{R}) = U_0^0(N,V) + U_1^0(\mathbf{R}) + U_2^0(\mathbf{R}) + ... + U_N^0(\mathbf{R}), \quad (1.15)$$

where the subscript denotes the order of the interaction. In biomolecular coarse-graining the zero-body potential (*volume term*) is simply equal to a constant [60, 59]. The one-body term is always zero in absence of external fields, in order to preserve the translational invariance of the system.

In practical applications, the sum in Eq. 1.15 is usually truncated to the second or third term, therefore losing, in principle, the possibility of being thermodynamically consistent with the underlying atomistic model [52, 53].

A natural question arises now: once the series in Eq. 1.15 is reduced to the truncated (to the T-order term) potential of mean force $U_T^0(\mathbf{R})$, can this term be effectively calculated? In order to answer it is necessary to introduce the approximated T-body coarse-grained potential $U_T(\mathbf{R})$ neglecting the first two terms [57]:

$$U_T(\mathbf{R}) = \sum_{t=2}^{T} \sum_{k} \sum_{\lambda} V_k^t(\Psi_k(\mathbf{R}_\lambda), \{\alpha_k\}), \tag{1.16}$$

where k represents a t-body interaction term with functional form V_k^t (such as the harmonic potential for CG bonds), which depends on a scalar variable Ψ_k, function of a set of CG coordinates λ (such as the site-site distance) and on some set of hyperparameters $\{\alpha_k\}$ (such as the bond equilibrium length and the coupling strength). The set of potentials V_k^t is bound to be finite and therefore constitutes an incomplete basis set in the space of CG force fields [52]: reproducing the MB-PMF is practically impossible not only because the effective potential is intrinsically multi-body in nature, but also because the truncated T-body PMF is approximated by a necessarily inadequate CG potential U_T, built using a finite basis set.

Given the impossibility of recovering the correct MB-PMF, several strategies exist whose aim is to construct simple but reliable approximations to it. Two main families of approaches can be distinguished in this context: in a group of methods, one first determines a set of low-resolution, structural distribution functions; then, consistency between atomistic and CG distributions is enforced by means of an iterative procedure. Notable examples of such *structure-based* procedures are the Boltzmann Inversion methods [61] and Inverse Monte Carlo [62]. Other strategies rely on the definition of a functional to quantify the distance between fine-grained and coarse-grained models: such functional vanishes if and only if the fully atomistic model is considered. Upon minimisation of this object, a unique CG potential is obtained, which provides an optimal approximation to the true MB-PMF U^0. The Force Matching [63, 64, 52] and Relative Entropy [22, 65, 66] frameworks are two most prominent examples of this philosophy, which is commonly defined as the *variational* approach to bottom-up CG.

1.2.1.1 Direct and Iterative Boltzmann Inversion

Boltzmann inversion-based methods are bottom-up strategies that aim at determining the CG interactions by reproducing some low-resolution, struc-

tural properties of the high-resolution system. In this context, the objective of approximating the MB-PMF is replaced by a more modest goal, namely the reproduction of a set of correlation functions. In the Direct Boltzmann Inversion (DBI) scheme, each term k of the CG potential (Eq. 1.16) is calculated by Boltzmann-inverting the corresponding probability distribution $p_k(x)$:

$$V_k(x) = -\frac{1}{\beta} \ln \left(\frac{p_k(x)}{J_k(x)} \right), \tag{1.17}$$

where $J_k(x)$ is a Jacobian factor. In protein systems, k usually refer to a bonded or non-bonded pair potential, to an angular term (three body), or to a dihedral potential (four body). CG potentials obtained through DBI can possibly succeed in reproducing the atomistic distributions *if* the probability distributions p_k are statistically independent (uncorrelated) [67], that is:

$$\begin{aligned} p_{k^1,...,k^L} &= p_{k^1,...,k^L}(x_1, ..., x_L) \\ &= \prod_k p_k(x_k), \end{aligned} \tag{1.18}$$

where $p_{k^1,...,k^L}$ is the joint probability associated to all L terms of the CG potential and $x_1, ..., x_L$ are the corresponding scalars (such as the angle between three CG beads). Probabilities factorise as in Eq. 1.18 when CG interaction terms are decoupled, that is, when the system is extremely dilute [68] and for bonded terms [21], which usually do not display significant statistical dependencies. In all the other cases, however, CG interactions are strongly coupled, giving rise to non-negligible correlations: it is in this context that potentials obtained through DBI fail to reproduce the target atomistic distributions.

The generalized Yvon-Born-Green theory developed by Noid and Mullinax [69, 70] offers the mathematical tools to quantify the statistical correlations that are missing in DBI potentials.

In order to reproduce the target structural property in presence of statistically dependent CG potentials, it is possible to resort to the Iterative Boltzmann Inversion (IBI) [61, 71], which extends DBI by guaranteeing that the target structural properties are replicated up to a pre-defined level of accuracy. IBI iteratively corrects *each* CG potential V_k via the following equation [68]:

$$V_{k,i+1}(x) = V_{k,i}(x) + \frac{\gamma}{\beta} \ln \left(\frac{P_{k,i}(x)}{p_k(x)} \right), \tag{1.19}$$

where i is the iteration index and γ is a prefactor $\in (0, 1]$, introduced to avoid numerical instabilities in the first stages of the optimisation. $P_{k,i}(x)$ is the CG distribution function obtained through a simulation of the CG system with potential $V_{k,i}$.

In this framework the correlations among CG interactions are treated in an implicit way by the iterative procedure, with the advantage of having a simple protocol to optimize each term V_k separately from the others.

1.2.1.2 Inverse Monte Carlo

A second important example of structure-based approach is the Inverse Monte Carlo (IMC) method, introduced by Lyubartsev and Laaksonen in two pioneering works [62, 72]. As in IBI, the potentials V_k are determined from atomistic ensemble averages of the set of target structural properties, here denoted with $\{\langle A_s \rangle\}$. In IMC, the problem of finding an optimal value for the parameters of the CG force field ($\{\alpha_k\}$, see Eq. 1.16) starting from averages is treated as a nonlinear, multidimensional equation, assuming that each element of $\{\langle A_s \rangle\}$ can be influenced by each CG force field parameter α_k. In the first stage of the algorithm, the system is simulated with a reasonable approximation to the CG potential; from this simulation one obtains a set of averages $\{A_s^0\}$, arbitrarily far from the reference values but such that the differences $\Delta\langle A_s \rangle$ are finite. The Newton-Raphson method is employed to iteratively solve this system of equations [73, 74]:

$$\Delta\langle A_s \rangle = \sum_k \frac{\partial \langle A_s \rangle}{\partial \alpha_k} \Delta\alpha_k + O(\Delta\alpha^2) \tag{1.20}$$

where $O(\Delta\alpha^2)$ denotes a higher order term in $\Delta\alpha$ and $\frac{\partial\langle A_s\rangle}{\partial\alpha_k}$ is a Jacobian matrix that quantifies how the target average properties depend on each parameter of the CG potential. Given that $\langle A_s \rangle = \frac{\int d\mathbf{r}\ e^{-\beta\mathcal{H}(\mathbf{r})} A_s}{Z}$, one can find a close-form for all the elements of this matrix by making use of basic statistical mechanics. The derivative in Eq. 1.20 reads [74]:

$$\begin{aligned} \frac{\partial \langle A_s \rangle}{\partial \alpha_k} &= \frac{Z \frac{\partial}{\partial \alpha_k} \left(\int d\mathbf{r}\ e^{-\beta\mathcal{H}(\mathbf{r})} A_s \right) - \frac{\partial Z}{\partial \alpha_k} \left(\int d\mathbf{r}\ e^{-\beta\mathcal{H}(\mathbf{r})} A_s \right)}{Z^2} \\ &= \left\langle \frac{\partial A_s}{\partial \alpha_k} \right\rangle - \beta \left(\left\langle \frac{\partial \mathcal{H}}{\partial \alpha_k} A_s \right\rangle - \left\langle \frac{\partial \mathcal{H}}{\partial \alpha_k} \right\rangle \langle A_s \rangle \right). \end{aligned} \tag{1.21}$$

The elements of the Jacobian matrix ($\frac{\partial\langle A_s\rangle}{\partial\alpha_k}$) can be readily calculated once the system is simulated with the Hamiltonian $\mathcal{H}$, which contains the potential parameters α_k. This allows one to solve the system of equations in Eq. 1.20 and to obtain a new set of parameters $\{\alpha_k\}$, which is then employed in a new simulation, until convergence is reached.

The major difference between IMC and IBI lies in the fact that in Eq. 1.20 the updates of different parameters of the CG potential are interdependent,

while in Eq. 1.19 different potentials are completely decoupled, thus causing numerical problems in the case of multicomponent systems [75, 68]. A (positive) consequence of this decoupling is that IBI calculations are faster than IMC ones [76].

Practical applications of the IMC method range from the first, pioneering studies on electrolite solutions to the approximate description of lipid membranes [77] and DNA systems [78]. IMC and IBI are compared in a number of different works [75, 79] and both methods are currently implemented in several software packages for CG, such as VOTCA [75] and MagiC [80].

1.2.1.3 Force Matching

The first example of the application of a variational approach to the bottom-up construction of a CG potential is the Multiscale Coarse-Graining method (MS-CG), or force matching, introduced by Voth and Izvekov in 2005 [63, 64]. The central concept of MS-CG is that, for each CG site I, an accurate CG potential U should reproduce the atomistic force acting on I ($\mathbf{f}_I(\mathbf{r})$). As it is always the case for variational approaches, such requirement is enforced by defining an appropriate functional:

$$\chi^2[U] = \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I(\mathbf{M}(\mathbf{r})|U)|^2 \right\rangle, \tag{1.22}$$

where $\mathbf{F}(\mathbf{R}_I|U)$ is the force generated by the approximated CG potential U and the angular brackets denote a canonical ensemble average for the high-resolution, atomistic model. It is instructive to introduce the force generated by the MB-PMF (Eq. 1.13) into Eq. 1.22 [52]:

$$\begin{aligned}
\chi^2[U] &= \frac{1}{3N}\left\langle \sum_{I=1}^{N} \left| \big(\mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I^0(\mathbf{M}(\mathbf{r}))\big) - \big(\mathbf{F}_I(\mathbf{M}(\mathbf{r})|U) - \mathbf{F}_I^0(\mathbf{M}(\mathbf{r}))\big) \right|^2 \right\rangle \\
&= \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I^0(\mathbf{M}(\mathbf{r}))|^2 \right\rangle + \\
&+ \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\mathbf{F}_I(\mathbf{M}(\mathbf{r})|U) - \mathbf{F}_I^0(\mathbf{M}(\mathbf{r}))|^2 \right\rangle - \\
&- \frac{2}{3N}\left\langle \sum_{I=1}^{N} \big(\mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I^0(\mathbf{M}(\mathbf{r}))\big) \times \big(\mathbf{F}_I(\mathbf{M}(\mathbf{r})|U) - \mathbf{F}_I^0(\mathbf{M}(\mathbf{r}))\big) \right\rangle \\
&= \chi^2[U^0] + \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\mathbf{F}_I(\mathbf{M}(\mathbf{r})|U) - \mathbf{F}_I^0(\mathbf{M}(\mathbf{r}))|^2 \right\rangle.
\end{aligned} \tag{1.23}$$

The third term in the second equality is a canonical average of a product, in which the first factor has zero average over a CG configuration $\mathbf{R}$ and the second is constant on that domain [52]. Hence, this term is always zero.

In the last line of Eq. 1.23, $\chi^2[U^0]$ is a constant, positive-definite term measuring the distance between the MB-PMF and the atomistic system. Importantly, this quantity does not show any dependency upon the approximated CG potential U, which appears only in the second term of the equality. This latter factor vanishes if and only if $U = U^0$, thus proving that the MB-PMF is the unique minimum of $\chi^2[U]$ [81, 52, 70, 57]. CG potentials of the form of Eq. 1.16 can be variationally optimized to make them as close as possible to the projection of the atomistic force field onto the subspace of CG force fields, that is, the MB-PMF.

As the MB-PMF intrinsically depends on the choice of the CG sites (Eq. 1.8), the term $\chi^2[U^0]$ changes when the mapping is modified. In this context it is evident that a quite accurate CG force field $U_{\mathbf{M}} \sim U^0_{\mathbf{M}}$ may be more distant from the atomistic force field with respect to a less sophisticated potential $U_{\mathbf{M}'} \neq U^0_{\mathbf{M}'}$ constructed upon a different mapping function $\mathbf{M}'$.

Recently, Noé, Clementi and coworkers translated the MS-CG protocol into a machine learning problem [82, 83], demonstrating how the χ^2 of Eq. 1.22 can be employed as the loss function of a graph neural network. Such intuition is applied to simulate the dynamics of short, coarse-grained peptides and, more recently, to tackle the problem of implicit solvation from a CG perspective [84]. This last contribution can be utterly beneficial for the molecular simulation community, as a correct, molecule-independent implementation of an implicit solvent CG force field allows to considerably increase the sampling time of MD simulations.

1.2.1.4 Relative Entropy

The functional developed by Voth and colleagues in the MS-CG strategy (Eq. 1.22) is entirely based on the reproduction of the average atomistic forces. Alternative variational approaches to bottom-up coarse-graining exist, such as the notable one developed by Shell and co-workers [22, 65, 66], who combine the power of probability distribution functions and information theory to define a new functional, called *relative entropy*.

In the original paper on this topic [22], Shell first introduces two ensemble probabilities, p_T and p_M, which are related to a detailed (*target*) and approximate (*model*) characterisation of a system, respectively. Then, assuming that these two descriptions possess the same number of degrees of freedom, one can compute a discrete likelihood that n samples extracted with

probability p_M reproduce p_T:

$$L(T|M) = n! \prod_i^{N_c} \frac{p_M(i)^{np_T(i)}}{(np_T(i))!}, \tag{1.24}$$

where the product runs over the available configurations N_c and $n\ p_T(i)$ is the expected number of times configuration i is observed in the target ensemble. When one considers the limit of infinite sampling ($n \to \infty$), the (logarithm of the) likelihood can be simplified by means of the Stirling approximation:

$$\ln L(T|M) = -n \sum_i^{N_c} p_T(i)\ \ln\left(\frac{p_T(i)}{p_M(i)}\right). \tag{1.25}$$

Apart from a proportionality constant ($-n$), this quantity is equivalent to a discrete Kullback-Leibler divergence [85] (or relative entropy) between the two probability distributions:

$$S_{rel} = \sum_i^{N_c} p_T(i)\ \ln\left(\frac{p_T(i)}{p_M(i)}\right). \tag{1.26}$$

Such formula is of limited use in the context of bottom-up CG without an extension to the case in which the model system contains a lower number of degrees of freedom than the target. If the probability distribution of this model ensemble, now dubbed P_M, is defined over a space constituted by a set of $N_C < N_c$ configurations, it is necessary to back-map P_M to p_M, that is, to relate P_M to the configurational space of the target system; it is crucial to keep in mind that multiple configurations of the target ensemble, i, can map to a single one in the model ensemble, $M(i)$. One thus introduces a *degeneracy* $\Omega(k)$ associated to each model configuration k:

$$\Omega(k) = \sum_i \delta(k, M(i)). \tag{1.27}$$

The delta inside the sum filters all the target configurations i that map onto k. In the original formulation of the method, Shell defines the back-mapped probability p_M as

$$p_M(i) = \frac{1}{\Omega(M(i))} P_M(M(i)). \tag{1.28}$$

The back-mapped probability in the target ensemble is given by the probability of the mapped configuration in the model ensemble divided by the corresponding degeneracy $\Omega(M(i))$.

At this point, the parallelism with bottom-up coarse-graining can be rendered more explicit by defining the relative entropy in the canonical ensemble. In order to do so, it is necessary to translate Eq. 1.26 into its continuous configurational space analogue and to define the target and the model probabilities; the former is the properly normalized atomistic Boltzmann weight of Eq. 1.4, while the latter is replaced with $P_R(\mathbf{R}|U)$ (Eq. 1.6), namely the coarse-grained probability distribution defined by the approximate model of the high-resolution system. As in the discrete case, it is necessary to translate $P_R(\mathbf{R}|U)$ into the higher-dimensional, atomistic configurational space, assigning a probability weight to each microscopic configuration $\mathbf{r}$:

$$P_{1r}(\mathbf{r}|U) = \frac{P_R(\mathbf{M}(\mathbf{r})|U)}{\Omega_1(\mathbf{M}(\mathbf{r}))} \tag{1.29}$$

$$\Omega_1(\mathbf{M}(\mathbf{r})) = \int d\mathbf{r}\ \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}). \tag{1.30}$$

The reason for the subscript 1 will become clear in the following pages. Putting all these ingredients together the relative entropy of Eq. 1.26 becomes

$$S_{rel} = \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left(\frac{p_r(\mathbf{r})\ \Omega_1(\mathbf{M}(\mathbf{r}))}{P_R(\mathbf{M}(\mathbf{r})|U)}\right). \tag{1.31}$$

By the Gibbs inequality the relative entropy is always non-negative, with the value of zero that can be reached if the CG model *coincides*, from a probabilistic perspective, with the atomistic one. Now it is possible to explicitly write the distributions inside the logarithm of Eq. 1.31 and to decompose S_{rel} in several canonical averages [22]:

$$\begin{aligned} S_{rel} &= \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left(\frac{Z_U}{z}\frac{e^{-\beta u(\mathbf{r})}}{e^{-\beta U(M(\mathbf{r}))}}\Omega_1(\mathbf{M}(\mathbf{r}))\right) \\ &= \left\langle \ln\left(\frac{Z_U}{z}\right)\right\rangle + \left\langle \ln\left(\frac{e^{-\beta u(\mathbf{r})}}{e^{-\beta U(M(\mathbf{r}))}}\right)\right\rangle + \langle \ln\left(\Omega_1(\mathbf{M}(\mathbf{r}))\right)\rangle \\ &= \beta(A_r - A_U) + \beta\left\langle (U(\mathbf{M}(\mathbf{r})) - u(\mathbf{r}))\right\rangle + \langle \ln\left(\Omega_1(\mathbf{M}(\mathbf{r}))\right)\rangle, \end{aligned} \tag{1.32}$$

where A_r and A_U are the Helmholtz free energies of the atomistic and CG model, respectively:

$$A_r = -\frac{1}{\beta}\ln(z), \tag{1.33}$$

$$A_R = -\frac{1}{\beta}\ln(Z_U). \tag{1.34}$$

Being uniquely determined from the partition functions, these quantities are independent of the specific atomistic configuration $\mathbf{r}$.

The second term in Eq. 1.32 is a canonical average of a potential energy difference between a fine-grained configuration and a CG one [65].

The third term in Eq. 1.32 corresponds to a Boltzmann-weighted average of the degeneracy associated to each atomistic configuration $\mathbf{r}$; this mathematical object, called *mapping entropy*, is of crucial importance for this thesis and is discussed more than extensively in the following chapters.

In 2011, Rudzinski and Noid [57] proposed a substantial modification to the definition of the relative entropy. In their approach, the mapped coarse-grained probability distribution (see Eq. 1.29) does not necessarily assign equal probability weight to all the atomistic configurations that map onto the same one:

$$P_r(\mathbf{r}|U) = \frac{g(\mathbf{r})}{\Omega(\mathbf{M}(\mathbf{r}))} P_R(\mathbf{M}(\mathbf{r})|U), \tag{1.35}$$

where $g(\mathbf{r})$ is a weighting factor dependent on the configuration of the atomistic ensemble. $\Omega(\mathbf{R})$ now becomes a weighted sum over all the atomistic states $\mathbf{r}$ that map onto the same CG configuration $\mathbf{R}$:

$$\Omega(\mathbf{R}) = \int d\mathbf{r}\ g(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}). \tag{1.36}$$

In Shell's formulation $g(\mathbf{r}) = 1\ \forall\ \mathbf{r}$ and all the microscopic configurations mapping onto the same coarse-grained one receive an equal *a priori* weight. Rudzinski and Noid [57] propose to differentiate between configurations and to employ the usual Boltzmann factor $p_r(\mathbf{r})$ as $g(\mathbf{r})$ in Eqs. 1.35 and 1.36:

$$\Omega(\mathbf{R}) = \int d\mathbf{r}\ p(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \tag{1.37}$$

which is exactly the probabilistic weight $p_R(\mathbf{R})$ (Eq. 1.7) assigned to the CG configuration $\mathbf{R}$ by the MB-PMF. Additionally, the Boltzmann constant k_B is introduced in order to enforce consistency between the relative entropy and the standard Gibbs formula for the entropy in statistical mechanics [86]:

$$S = -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left(p_r(\mathbf{r})\right), \tag{1.38}$$

thus leading to the following equation:

$$S_{rel} = k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left(\frac{p_R(\mathbf{M}(\mathbf{r}))}{P_R(\mathbf{M}(\mathbf{r})|U)}\right). \tag{1.39}$$

Since the two quantities inside the logarithm depend only on the low-resolution configuration, S_{rel} can be straightforwardly expressed as a Kullback-Leibler

divergence over the CG configurational space [87, 57] by introducing a functional delta $1 = \int d\mathbf{R}\ \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$:

$$S_{rel} = k_B \int d\mathbf{R}\ p_R(\mathbf{R}) \ln\left(\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)}\right). \tag{1.40}$$

Rudzinski and Noid [57] employ the argument of the logarithm, $\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)} = \Phi(\mathbf{R}|U)$, to bridge the Relative Entropy and the Force Matching variational approaches. More specifically, they realise that the first derivative of this mathematical quantity, namely

$$\frac{\partial \Phi(\mathbf{R}|U)}{\partial \mathbf{R}_I} = \beta\left(\mathbf{F}_I^0(\mathbf{M}(\mathbf{r})) - \mathbf{F}_I(\mathbf{M}(\mathbf{r})|U)\right) \tag{1.41}$$

can be directly plugged into the MS-CG equation (Eq. 1.23) to obtain the following expression:

$$\chi^2[U] = \chi^2[U^0] + \frac{1}{3N\beta^2}\left\langle |\nabla\Phi(\mathbf{R}|U)|^2 \right\rangle, \tag{1.42}$$

where $\nabla\Phi = \sum_{I=1}^{N} \frac{\Phi(\mathbf{R}|U)}{\partial \mathbf{R}_I}$. While a CG force field with zero relative entropy minimises the average of $\Phi(\mathbf{R}|U)$ (Eq. 1.40), the force matching strategy induces the minimisation of the average of $|\nabla\Phi(\mathbf{R}|U)|^2$.

As in the case of MS-CG, the relative entropy attains a (unique) global minimum at constant mapping if the MB-PMF is employed as the potential in the model ensemble. In the context of CG force field development, this functional is minimized with respect to each interaction parameter α_k of the CG potential U so as to maximize the consistency between the two probability distribution functions:

$$\frac{\partial S_{rel}}{\partial \alpha_k} = 0. \tag{1.43}$$

It is important to notice how the MB-PMF at the numerator of the logarithm in Eq. 1.40 is independent of α_k. Therefore, Eq. 1.43 is equivalent to the derivative of a coarse-grained ensemble average of the (logarithm) of the approximate $P_R(\mathbf{R}|U)$:

$$\begin{aligned}
\frac{\partial S_{rel}}{\partial \alpha_k} &= -k_B \int d\mathbf{R}\ p_R(\mathbf{R}) \frac{\partial \ln P_R(\mathbf{R}|U)}{\partial \alpha_k} \\
&= -k_B \int d\mathbf{R}\ p_R(\mathbf{R}) \beta \left[-\frac{\partial U}{\partial \alpha_k} + \int d\mathbf{R}'\ p_R(\mathbf{R}'|U) \frac{\partial U}{\partial \alpha_k} \right] \\
&= \frac{1}{T}\left[\left\langle \frac{\partial U}{\partial \alpha_k} \right\rangle_{U^0} - \left\langle \frac{\partial U}{\partial \alpha_k} \right\rangle_{U} \right]
\end{aligned} \tag{1.44}$$

A modification of a parameter in the CG force field has two effects on the coarse-grained probability and, therefore, on the relative entropy: it changes the potential energy function and the Helmholtz free energy (the partition function). The first term of Eq. 1.44 describes the first change: the derivative of the CG potential is averaged over the MB-PMF ensemble. The second term of Eq. 1.44 accounts for the modification of the coarse-grained partition function Z_U (Eq. 1.6), and is therefore weighted with the approximate probability $P_R(\mathbf{R}|U)$. At the relative entropy minimum these two averages are exactly equivalent.

The relative entropy minimisation protocol is described in detail in some excellent papers [65, 88] and is successfully applied to the coarse-graining of water and proteins [89, 90]. Applications of this strategy in biomolecular modelling extend beyond bottom-up CG methods, ranging from native structure prediction in atomistic protein folding [91] to the coarsening of Markov State Models [92].

Chapter 2

The representation problem

In the previous chapter I introduced the theoretical basis of bottom-up coarse-graining, describing the four major methods that are routinely employed to derive CG potentials. For the sake of brevity, the discussion was limited to classical, particle-based CG models, in which each CG site is representative of a group of atoms of the molecular structure. The definition of these sites is performed via the mapping operator, see Eq. 1.3. Importantly, the mapping was considered simply an ingredient of the CG model, and no reference was made to how its selection should be performed.

In principle, the choice of a mapping with a specific level of resolution allows one to *observe* all phenomena in the system that occur at a length scale equal to or larger than the characteristic size of the elemental CG units; in the construction of a CG model, though, it is the choice of the interactions that limits its ability to *reproduce* such phenomena. If the CG potential accurately reproduces the MB-PMF (Eq. 1.8), all thermodynamical properties and observables of the system can be obtained, even if they originate from processes that take place at a scale below the resolution level of the model [53, 93, 94]. However, Sec. 1.2.1 shows how, in practical applications, it is not possible to calculate *all* many-body contributions that appear in the PMF, let alone embodying them into computationally manageable functional forms. With a limited expansion of the MB-PMF, the modeller must expect that a reduction in the resolution level will correspond to a decrease in the spectrum of properties and phenomena that the model is able to predict.

The interplay between resolution level and range of observable phenomena lies at the core of the first two sections of this chapter, which describe the most common choices to represent the atomistic protein system at lower resolution by fixing *a priori* the CG mapping. In Sec. 2.3 the few existing methods that aim at optimising the choice of the CG representation are listed, focusing on the lack of a unique consensus among these strategies,

which rely on different criteria.

Ref. [23] should be considered the reference article for this chapter. Once again, the discussion is restricted to classical, particle-based coarse-grained models of proteins, without considering the huge number of excellent works on continuous CG models [95, 96] and hybrid quantum-classical protocols [97].

2.1 On the choice of the resolution level

This section has been entirely written by my colleague and supervisor Roberto Menichetti, whom I here acknowledge for his crucial contribution to this chapter.

The mapping can be interpreted as the mathematical prescription that connects the high-resolution description of the system to the coarse-grained one. It is evident that a first feature that is immediately determined by a particular choice of the mapping is the resolution level of the CG model, that is, the minimum degree of detail one can have access to in describing the system's properties. Once a certain level of resolution is fixed, it is impossible to observe phenomena occurring at lower length scales, but only at or above such level. As an example, consider the fully atomistic description of a protein: to a certain extent, this is an already coarse-grained model of an inherently quantum system, and processes that directly involve electrons cannot be inspected from this perspective; nevertheless, one can hope to *reproduce* the emergent properties of the system by taking into account the implicit presence of the electronic degrees of freedom in an effective force field. From these considerations, it follows that the choice of the appropriate resolution level is not trivial in general, as it impacts the lower limit of the observable length scales.

Particle-based CG models can be subdivided in three main groups according to their resolution level. Going from the highest to the lowest detail, one meets a first category of nearly chemically-accurate protocols, which include an explicit description of the solvent, though treated with a detail lower than atomistic; the second group of models, instead, rely on an implicit description of the solvent degrees of freedom, retaining a medium-to-high level of chemical accuracy on the solute elements. The third class is constituted by ultra coarse-grained (UCG) models, in which the solvent is treated implicitly and a molecule is reduced to a set of few sites. In the following, models belonging to these categories are outlined.

2.1.1 Explicit solvent CG models

Within explicit solvent CG models, a quite conspicuous number of effective interaction sites is employed to represent a single amino acid composing the protein structure, thus resulting in a mapping rule that projects a small chemical moiety onto a CG bead. The protein is then surrounded by a relatively "granular" solvent, resulting in a rather moderate level of CG. Particular attention is further paid to approximately capturing the "local" chemical features and flexibility of amino acid side chains, so that several beads can be employed in their description. Overall, this fairly high level of detail can limit the computational speed-up generated by these models, especially due to the presence of the solvent; at the same time, it often allows an almost one-to-one reconstruction, or backmapping, of the microscopic structure starting from the CG one [98].

Notable examples in this class of models include the popular SIRAH [98, 99] and MARTINI [49, 50] force fields. In both cases, interactions among the CG sites are parametrised to account for the average properties of the atoms they enclose, and include bonded as well as non-bonded contributions. The former are tailored so as to reproduce (a subset of) structural features, such as bond distances and the bending and dihedral angles between consecutive units. Different philosophies lie instead at the core of the determination of the non-bonded potentials: while SIRAH aims at providing an accurate description of the system's electrostatics and sterics [98, 99], MARTINI mainly targets experimental free energies of partitioning of small chemical fragments between a polar and an apolar phase [49, 50]. In both cases, the result of this overall parametrisation protocol is a "dictionary" of CG building blocks, one per amino acid, that can be combined together to model the protein structure of interest and investigate its behaviour.

The resolution level and chemical specificity characterising the fundamental units of SIRAH and MARTINI enables their application to the investigation of large-scale conformational and/or thermodynamic properties of a system, as well as to problems in which the local detail, down to a sub-residue level, can play a crucial role on the system's emergent phenomena: among these it is important to mention the rearrangement of side chains, hydrogen bonding, and protein-solvent, protein-protein, or protein-substrate interactions. Despite the similar length scales characterising the elemental units composing the two models, however, already at this limited degree of CG the delicate interplay between resolution level and effective interactions has a considerable impact on the spectrum of observable phenomena. As an example, SIRAH has shown to be able to preserve the stability of proteins comprising α-helix as well as β-sheet elements in absence of explicit

topological biases [98]. On the contrary, MARTINI requires secondary structure motifs to be enforced *a priori*, thus preventing its application in studies involving folding or general conformational rearrangements [50, 100, 101]. While this limitation is commonly associated to the relatively low resolution at which the protein backbone is treated in MARTINI (one bead per peptide), it should rather be considered a direct consequence of the particular choice in the parametrisation of the interactions: in fact, effective models exist that rely on MARTINI-like CG representations and are capable of stabilising secondary structure elements without introducing ad hoc constraints [102, 103].

2.1.2 Implicit solvent CG models

The level of resolution proper of explicit solvent CG models can be considered excessive when dealing with phenomena that take place at larger length scales, such as protein folding, conformational rearrangements, or self-assembly. Consider for example the case in which a net attraction/repulsion between pairs of amino acids constitutes the driving force of the macroscopic process; for this to emerge from the CG model, a much lower resolution than that of SIRAH or MARTINI might be sufficient, e.g. removing the solvent and describing each amino acid as an effective interaction unit.

In implicit solvent CG models the solvent degrees of freedom are integrated out from the description, and one only accounts for the effect they *on average* exert on the proteins under investigation. Such proteins, on the other hand, are still decomposed in terms of their constituent residues, albeit in an increasingly simpler form as the structural coarsening progresses. It is in this context that the correlation between resolution level, CG interactions, and range of observable phenomena becomes particularly strong: a decrease in the first is usually not balanced by an increase in the second, which in turn can result in a reduction of the third.

Among implicit solvent CG models, the more detailed ones aim at preserving the "chemical identity" of each amino acid. Since such information is inherently contained in the side chain, this directly translates into the usage of one or more explicit CG beads representing it and accounting for its chemical features, in addition to the effective sites that are employed to describe the peptide backbone: in analogy with the case of explicit solvent models (Sec. 2.1.1), the desired outcome is again a protocol in which the fundamental units embodying each amino acid type can be joined together to assemble the specific system under investigation. Examples of such *intermediate resolution* CG force fields are OPEP [104, 105, 106], the one by Bereau and Deserno (BD) [107], PRIME [108, 109], AWSEM [110, 111], and

UNRES [112, 113, 114].

The first model, OPEP, is characterised by a high degree of structural detail [104, 105, 106]. All the heavy atoms composing the protein backbone as well as the amide hydrogens are retained as CG sites, while a single bead describes the side chain of each amino acid—except for proline, which is represented by all its heavy atoms. Interactions among these fundamental units are then parametrised via a combination of structural, thermodynamic and knowledge-based approaches. Interestingly, while the original version of the model neglected the solvent degrees of freedom, hydrodynamic interactions were later incorporated in OPEP by coupling it with a Lattice Boltzmann representation of the solvent [115]. As for BD and PRIME, they lean on a similar CG mapping prescription to describe each amino acid, namely three beads for the backbone and one for the associated side chain. Notable differences exist, however, in the derivation of their constitutive interactions. In particular, in analogy with OPEP, BD is again defined in terms of a conventional basis set for the bonded and non-bonded interactions, whose fundamental parameters are tuned by combining structural and knowledge-based protocols [107]. PRIME, on the other hand, resorts to a very crude interaction network, in which extremely simplified potentials such as hard-sphere and square-well functions describe steric repulsion and bonding/attractive interactions among the effective sites, respectively [108]. This choice enables the usage of discontinuous molecular dynamics [116, 117], further speeding up simulations. Originally blind to the side chain chemical detail, PRIME was later generalised via a knowledge-based approach so as to capture their specificity [109]. In AWSEM, three CG sites, respectively located on the C_α, C_β, and backbone oxygen atoms, are employed to represent a single protein amino acid [110, 118, 111]. Bonded potentials among the AWSEM CG units are then complemented with a complex network of non-bonded interactions: these include hydrogen-bonding terms, bioinformatic terms biasing the formation of local structures, nonlocal terms describing contacts—either direct or water/protein-mediated—among distal residues along the sequence, and burial terms that aim at accommodating an amino acid into its preferential environment—e.g. the protein bulk or surface. The corresponding parameters are tuned *via* a combination of structural and knowledge-based approaches. Finally, UNRES maps each amino acid onto three CG sites, namely the C_α atom, the center of the peptide bond, and the side chain, the latter being described as an ellipsoid of revolution [112]. Only the last two elements, however, are explicit effective interaction sites, while C_α sites only serve the purpose of tracing the protein's geometry. Interactions among the UNRES building blocks are then parametrised through a rigorous bottom-up procedure: the potential of mean force of the system is expanded in a trun-

cated series of Kubo-cluster cumulants, which enable the derivation of the multi-body interactions acting among the CG sites in a systematic manner [112, 113, 114].

The power of the intermediate resolution CG models lies in their transferability, that is, the possibility of employing them to provide insight on the behaviour of systems that are not directly involved in the models' parametrisation. It follows that particular care must be taken as far as meso- to macroscopic properties are concerned; while these can be explicitly included in the construction of the effective potential, for the latter to be transferable the introduced restraints should be flexible enough so as not to bias the model's predictions towards very specific outcomes, associated to particular systems. It is thus possible, and indeed often advantageous, to design transferable implicit solvent CG models tackling well-defined large-scale problems; at the same time, one should make their constitutive ingredients as general as possible, so as to enable the characteristic phenomenon of the system of interest to arise from the model, without the need of imposing it a priori. On the other hand, one might need implicit solvent CG models that are more severely bound to a subset of known macroscopic properties associated to a specific biomolecule. In this case, the model could be asked, e.g. to reproduce the experimentally resolved tertiary structure of a particular system. The emergent property now directly represents an input of the CG protocol.

One could clearly resort to standard CG strategies and develop a dedicated effective model in which these conditions are satisfied [63, 57, 66]; this often lengthy parametrisation procedure, however, should at least in principle be repeated from the ground up every time a new system is investigated, for which the same kind of external piece of information is available. It is therefore highly desirable to construct CG models that rely on more "intuitive" interaction potentials and are easily generalisable to arbitrary systems through a minimal fine-tuning. The particular choice of the phenomenological potential will play a pivotal role in defining the class of phenomena the model can *additionally* provide insight on. The simplification of the interaction network typically goes on par with an additional reduction in the resolution level and chemical detail, with every amino acid composing the molecule being now described as a single interaction site.

A notable example of this second class of implicit solvent CG models is represented by *structure-based* ones, such as Gō-like models (GLM) [119, 120] or elastic network models (ENM) [121, 122]. Here, the external macroscopic input involved in the construction of the effective CG potential is the static, either stable or metastable, three-dimensional spatial conformation assumed by the protein of interest. Both GLM and ENM describe the interaction among the elemental CG units in terms of very general functional forms,

tailored to *reproduce* the target structure but easily applicable to arbitrary ones; the complexity and richness of the basis set, however, significantly decreases while moving from GLMs to ENMs, generating a crucial impact on the spectra of phenomena these two classes of models can respectively capture.

2.1.3 Ultra CG models

The class of models presented in the previous sections, although characterised by a gradual decrease in the level of detail, always rely on a residue-based decomposition of a protein, in which only one or few effective interaction centroids describe *each* amino acid composing the biomolecule. To push the applicability of particle-based CG models to the investigation of phenomena occurring at even larger time and length scales, one possibility is that of resorting to ultra coarse-graining (UCG) methods. Here, each CG site becomes representative of larger chemical entities, be that few residues, whole proteins or even entire molecular complexes [123, 124, 125]. Several examples of UCG models, ranging from more "chemically accurate" to more heuristic ones, have been presented in the literature. While more traditional applications typically focus on single proteins [125, 126], UCG methods have provided impressive insights into the behaviour of overwhelmingly complicated macromolecular structures [127, 128], including actin filaments [123], bacterial flagella [129], and viral capsids [130, 131, 132].

As pointed out in Ref. [133], from a conceptual point of view UCG models pose notable additional challenges compared to their more detailed counterparts, which are, as it is the case for the previously discussed studies, often overlooked in the construction of the UCG effective interaction potential of a system. Specifically, as the structural coarsening progresses, several internal states of the system can end up being mapped onto the same CG configuration. For instance, let me consider the case of a macromolecular complex, a whole protein of which is represented as a single UCG site. If the protein undergoes a conformational rearrangement between two states that leave the CG site coordinates unaltered, both states contribute to the energetic landscape of a single CG macrostate and, as far as the model is concerned, they are indistinguishable. At the same time, the rearrangement could play a key role in the generation of the macroscopic phenomenon of interest, and it would thus be desirable to construct a UCG model able to discriminate the two conformational basins. To tackle the problem of constructing CG models for systems possessing internal states, Voth and coworkers have recently developed an extremely elegant Theory of Ultra Coarse-Graining (UCGT) in a series of works [133, 134, 135]. While applications of this theory have

been, to my knowledge, so far limited to relatively high-resolution CG representations of liquids, UCGT represents an extremely promising framework for the development of accurate UCG models of biologically relevant macromolecules.

2.2 On the choice of the resolution distribution

This section has been entirely written by my colleague Thomas Tarenzi, whom I here acknowledge for his crucial contribution to this chapter.

The application of finer or coarser mappings to the atomistic system of interest determines the "average" level of detail of its reduced representation. The mapping, though, can characterise different regions of the system with different levels of resolution, that is, with a variable *density* of CG beads. In the trivial, limit case of implicit solvent CG models (Sec. 2.1.2), one can appreciate how they imply by definition a non-uniform distribution of CG sites, assigning significant and zero resolution to solute and solvent degrees of freedom, respectively. Except for this peculiar case, why would one employ a variable resolution representation of a biomolecular system?

In the context of computational simulations, the biological phenomena of interest can be confined to a very specific area of the simulation box. This is the case, for example, of protein-protein and protein-ligand interactions, where the contact region usually involves a tiny fraction of the overall atoms composing the system. If such domains involved in the interaction are known in advance, e.g. from experimental evidence or previous computational analysis, one can additionally exploit the inherently multiscale nature of the problem to build a hybrid atomistic/coarse-grained (AA/CG) set-up, where the atomistic detail is retained only in the region of interest (in the above example, the binding site of a receptor). The rest of the macromolecule is instead treated at a lower, CG resolution, bringing the immediate advantage of a reduced computational cost.

This general idea gave rise to a variety of approaches, where the details of each method (namely, the resolution distribution and the parametrisation of interactions) are specifically designed to tackle the system under investigation. Examples range from the multi-resolution model of a polyamide melt [136], where only the amide groups involved in the formation of the hydrogen bonds are maintained at atomistic resolution, to multimeric complexes including both proteins and nucleic acids, as in [137, 138, 139].

In most of AA/CG applications the size of the atomistic region is larger

than a single chemical moiety, but substantially smaller than the protein itself. This is the case of ligand-binding multiscale studies, where an atomistic resolution is required for only a few protein residues [140, 141].

A hybrid method specifically designed for the study of ligand-protein interactions is the so-called Molecular Mechanics/Coarse-Grained approach (MM/CG) [142, 143, 144]. In the latest implementation of the method (Open Boundary MM/CG) [145], the multi-resolution model of the protein is coupled to an adaptive resolution description of the solvent through the Hamiltonian adaptive resolution (H-AdResS) scheme [146]. Water is modelled with atomistic accuracy in the two hemispheres capping the intracellular and extracellular parts of the receptor, and free diffusion is ensured with a surrounding reservoir of CG water molecules. The improved hydration model leads to the simulation of a rigorous statistical ensemble and enables accurate binding free energy calculations for a drug design purpose [147].

Moving away from the AA/CG framework, it is important to mention the existence of multi-resolution models, where the two or more resolutions concurrently employed are coarse-grained, that is, lower than atomistic. These approaches aim at reproducing the large-scale conformational dynamics of large biomolecules in a particularly efficient manner. In this context, proteins have been modelled as networks of a small number of CG sites, fewer than the total number of residues [148, 149, 150], that are further unevenly distributed along the primary structure. The partitioning among resolution levels can be performed on the basis of previous knowledge of the system functions: this is the case of the multiscale network model [151]: here, the fine-grained region is constituted by specific functional sites represented at the residue level as an ENM; the remaining regions are described at a lower resolution, including only a subset of the C_α atoms as interaction sites.

In other approaches, the choice of the level of resolution and its distribution along the protein structure is not so obvious. This is the case of the essential dynamics coarse-graining (ED-CG [152, 153]), where residues undergoing collective dynamics are represented by pseudo-nodal points.

It is important to emphasise that all the hybrid multiscale models considered in this section require the definition of the resolution domains during the phase of simulation set-up, *on the basis of some previous knowledge of the system.*

2.3 Strategies for mapping optimisation

The previous sections have showed how coarse-graining techniques model soft matter systems, proteins in particular, using a plethora of simplified repre-

sentations, each one characterised by its level of detail. Approaches exist that displace the same level of detail throughout the whole system; additionally, I discussed how it is possible to concurrently employ, in the same simulation setup, models at different resolution, so as to provide a small subregion with an accurate description and the remainder of the system with a computationally efficient one. In both cases, the level of detail and its distribution is usually determined *a priori* on the basis of various characteristics (chemical identity, biological function, intuition), depending on the usage one does of the model. Recently, however, interest has grown around the idea of allowing the system itself to decide its "best" coarse-grained description. Clearly, the notion of "best" is relative, and it necessarily has to answer to the question *best for what?*

In this final section let me report on the recent attempts to find the optimal resolution of a biomolecule, namely the "most appropriate" number and selection of degrees of freedom to describe it, together with their spatial distribution. These two concepts are deeply intertwined and several studies suggest the existence of a link among the optimal resolution, the distribution of detail assigned in the coarse-grained model, and the relevant properties of the system of interest. This connection has its roots in the philosophy behind bottom-up CG modelling, which assumes that the properties of a system should emerge from the behaviour of a statistical mechanics-based, simplified model obtained through the (exact) integration of a subset of its degrees of freedom (see Sec. 1.2.1). Usually, this concept of "behaviour" refers to the conformational space sampling of the simulated CG model, which is a combination of mapping and interactions. Here, I argue that the process of simplification (mapping) itself, acting as a *filter*, can provide hints to non-trivial features of the high-resolution model. This hypothesis has immediate consequences, such as the conversion of coarse-graining methods into analysis tools, a change of paradigm that could constitute a valuable instrument for the analysis of high-resolution, fully atomistic representations of biomolecules.

In bottom-up CG modelling, the choice of the CG mapping has proved to be critical for the properties of interest to emerge systematically [154, 57]. This idea is pushed forward by Rudzinski and Noid [155], who quantitatively rationalise how the quality of the modelling is influenced by the quality of the mapping. Specifically, the authors group the configurations sampled in a MD simulation into n (resp. m) distinct molecular states of the high-resolution (resp. low-resolution) system; as the low-resolution macrostates clearly depend on the choice of the mapping scheme, Rudzinski and Noid posit that the most informative CG representation should generate a bijective correspondence between atomistic and CG molecular states. This approach

allows, in principle, to estimate the optimal level of resolution as well as its distribution. It is thus the system itself that informs the modeller about its low-resolution description that maximises the consistency with the high-resolution behaviour.

This promising paradigm is at the heart of a recent work by Fiorentini and coworkers [141], in which a protein-ligand system is considered and the relationship between the binding free energy and the chosen level of resolution is quantified. The authors consider several hybrid AT-CG representations of the protein by treating a variable number of amino acids around the binding site at the all-atom level. The resulting values of binding free energy are compared with the atomistic reference, showing that the accuracy of the dual-resolution model does not necessarily increase with the spatial extension of the atomistic region. This result suggests the existence of a system-specific, optimal number of amino acids that should be modelled with high detail in such hybrid schemes.

In general, then, the idea has started to emerge that a macromolecular system admits one or more *optimal* reduced models, that is, simplified representations in terms of which it can be *observed* with a marginal loss of information in spite of a loss of detail. Furthermore, it appears more and more evident that such an optimal representation cannot, in general, be uniform: the degree of fidelity with which the original, high-resolution structure is reproduced in the simplified model can vary from point to point, in parallel with the system's chemical, mechanical, dynamical, and functional properties.

Foley and coworkers [54, 156] have pioneered the analysis of the CG model spectrum in a formal and systematic way. In [54] they consider a one-bead-per-residue Gaussian network model (GNM) of proteins as the reference, high-resolution representation; then, taking advantage of the exact integrability of GNMs, they perform a systematic *decimation* of the system's beads to investigate how reduced models at varying degrees of resolution manage to reproduce fluctuations and correlations of the original model. In so doing, they show that the information loss that is inherent in the process of coarse-graining is not a monotonic function of the resolution, as an optimal value of the latter was found for which the information content per CG bead (quantified by an appropriate measure) exhibits a maximum. These works thus highlight the relation between the informativeness of a representation and its resolution *level*, that is, the *number* of CG sites.

The impact of resolution *distribution* was later studied by Koehl and coworkers, also in this case making use of ENMs: the *Decimate* [157] algorithm progressively reduces the resolution of a biomolecule by creating a hierarchy of increasingly simplified models, in the spirit of the renormalisation

group theory. As expected, such CG mappings show an uneven distribution of detail: as an example, optimal models of globular proteins tend to concentrate atoms on the surface of the molecule, thus heavily coarse-graining the inner region–whose mechanical properties require fewer degrees of freedom to be aptly reproduced. A related approach is employed in a work by Diggins et al. [158]: here, the authors identify the CG beads that produce a coarse-grained ENM whose Hamiltonian interaction matrix is as close as possible, measured according to an appropriate distance, to the high-resolution, atomistic ENM. The proposed selection of atoms proves to outperform a random assignment in terms of several observables, such as the intra-block dynamics fraction.

Most of the mentioned approaches can be grouped under the umbrella of methods to optimise the representation of a biomolecule in order to improve the capability of the reduced models to faithfully reproduce the atomistic properties of interest. Let me now summarise the existing methods that, acting as pure filters, focus only on the choice of the representation *itself* without considering the parametrisation of the effective interactions.

The first prominent attempts at finding the most informative reduced description of a biomolecule can be ascribed to Voth and coworkers, who employed the χ^2 residual of essential dynamics to estimate the optimal number and partitioning of coarse-grained sites for large protein complexes (ED-CG) [153, 159, 160]. In particular, in Ref. [160] this χ^2 is subject to a constrained minimisation, in which the addition of a CG site to a simplified description of a molecule is accepted only if there is a substantial gain in information about the system. Related works [161, 162, 163] by Xia and colleagues take the moves from the ED-CG method to develop several protocols for the determination of the optimal representations of biomolecules. In Ref. [161] the authors introduce the stepwise optimisation with boundary constraint (SOBC) algorithm to enhance the numerical performances of ED-CG [153, 159] on large proteins. Subsequently they propose to maximise the ENM pairwise fluctuations between atoms that are mapped to different CG sites (fluctuation maximisation) [162]. The resulting reduced representations, once equipped with simple, harmonic interactions, are capable of matching the large-scale fluctuations of the corresponding fine-grained counterparts. More recently, Wu *et al.* [163] adopt a combination of ED-CG and internal clustering validation indices to estimate the proper number of sites to coarse-grain proteins. Their results suggest that the appropriate number of C_α atoms to be preserved in a simplified model should lie between one half and one fourth of the total.

Always in the context of employing the CG mapping as a filter, multiple examples of the application of CG methods to analyse simulation data

of biomolecules rely on quasi-rigid domain decomposition [164, 165, 166]. Specifically, Polles *et al.* [167] employ a quasi-rigid domain decomposition of several viral capsids to single out their fundamental mechanical blocks; once validated on a dataset of known viruses, this method is used to formulate predictions about structures whose mechanical subunits had not been characterised yet. Following a similar approach [168], Morra *et al.* study MD trajectories of three representatives of the heat shock protein 90 (Hsp90) family, simulated with and without substrates. They observe that, when the protein is partitioned in as few as three quasi-rigid domains, the relative rigid-like movements of the latter can account for a significant fraction of the system's fluctuations, thus allowing to pinpoint two *optimal axes* for rigid rotations of the domains. In turn, the position of these hinges was shown to correspond to two interfaces: while the biological importance of one of them had already been assessed, the other one was hitherto unknown, thus highlighting a potentially druggable functional site.

These remarkable results prove that it is possible to exploit CG methodologies to perform a detailed analysis of the fundamental aspects of an atomistic system. Nevertheless, it is important to notice how these approaches rely on the examination of *mechanical* properties of the system of interest; although they certainly represent simple, intuitive variables to look at, such features do not seem to be as fundamental as the underlying problem they are applied to. Examples of more profound approaches exist that aim at optimising the coarse-grained representation of biomolecules in a systematic way [169, 170, 171, 172]. Delvenne *et al.* [169] rank CG mappings according to the quality of the corresponding partitioning induced on the protein graph. Chen and Habeck [170] propose a Bayesian procedure that extracts the optimal representation from a single macromolecule or cryo-EM map. Boninsegna et al. [171] combine time-averaged diffusion maps [173] and Markov State Models [174] to select groups of atoms that are mutually close (coherent) over a conformational basin. Wang and Gómez-Bombarelli [172] employ a variational autoencoder to learn a set of latent CG variables (that is, a CG representation) from the atomistic configuration: in the decoding process the former aims at reconstructing the latter in a deterministic procedure.

All the works showcased here reflect the emergence of a profound need in the computational biophysics community: that of a strategy to build a faithful simplified representation of a molecular system in an entirely unsupervised manner. In standard coarse-graining recipes, such reduced descriptions must be equipped with proper effective interactions in order to *generate data*. However, the impressive development of techniques to enhance the performances of atomistic simulations is making this necessity less and less pressing. In

contrast, the huge amount of high-resolution data produced at each MD run these days might benefit from the capacity of CG models to serve as powerful instruments to *make sense of the data*.

The next chapter focusses on the mapping entropy, a quantity that aims at answering to these needs by considering *both* structural and energetic properties of a coarse-grained biomolecule at the same time. Rigorous information-theoretical calculations allow one to compare these properties to those of the high-resolution system, and to search for the CG reduced representation that adheres as much as possible to the atomistic one.

Chapter 3

Mapping Entropy

In the past section I have discussed a number of different coarse-grained representations employed in the last decades to generate low-resolution models of proteins, highlighting the few existing methods proposed to optimize their choice. Let me now focus on a quantity, the mapping entropy, capable of measuring the amount of information contained in a simplified description of a biomolecule. In the first part of the chapter I illustrate how a series of controllable approximations allows one to evaluate this information-theoretical function from a fully atomistic simulation of a biomolecule. The mapping entropy is then used as the driving observable of an optimisation process that leads to the identification of the most informative representations of a protein. This chapter must be considered a personal re-elaboration of Ref. [24], that should be employed as the main reference.

In the first chapter of this thesis (see Sec. 1.2.1.4), the relative entropy and its properties were extensively discussed. Let me recapitulate the crucial concepts about this complex mathematical object:

$$\begin{aligned} S_{rel} &= k_B \times D_{KL}(p_r(\mathbf{r})||P_r(\mathbf{r}|U)) \\ &= k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{p_r(\mathbf{r})}{P_r(\mathbf{r}|U)} \right]. \end{aligned} \tag{3.1}$$

Here the probability $P_r(\mathbf{r}|U)$ is considered to be equal to that defined by Rudzinski and Noid in Ref. [57], namely

$$P_r(\mathbf{r}|U) = \frac{p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))} P_R(\mathbf{M}(\mathbf{r})|U), \tag{3.2}$$

where $p_R(\mathbf{M}(\mathbf{r})) = \frac{e^{-\beta U^0(\mathbf{M}(\mathbf{r}))}}{Z_{U^0}}$ is the probability assigned by the MB-PMF to the configuration $\mathbf{M}(\mathbf{r})$. Putting this quantity into Eq. 3.1 one can obtain Eq. 1.40, that is, the expression of the relative entropy over the CG

conformational space:

$$S_{rel} = k_B \int d\mathbf{R}\; p_R(\mathbf{R}) \ln\left(\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)}\right). \tag{3.3}$$

Importantly, the atomistic weight at the numerator of the logarithm in Eq. 3.1 cancels out and Eq. 3.3 quantifies the distance, in the Kullback-Leibler sense, between the MB-PMF and the approximate CG potential U. In other words, this functional assesses the quality of the effective CG interactions introduced among the CG sites in place of the exact MB-PMF.

The objective of a genuine bottom-up coarse-graining procedure should be the minimisation of the distance between the effective model and a first principles theory, in our case a fully atomistic, fine-grained description of the system. Given that a CG model is the result of a combination of a mapping $\mathbf{M}$ and some effective interactions, the quality of the former should be somehow measured. In this context Eq. 3.3 is of limited help, as the relative entropy vanishes when U coincides with the MB-PMF, irrespectively of the selected mapping. In other words, different CG mappings can give rise to different many-body potentials of mean force (Eq. 1.8) and S_{rel} can only measure the similarity between the chosen, mapping-dependent MB-PMF and the approximate CG model.

Starting from Eq. 3.1, it is useful to keep track of the atomistic weight by decomposing the relative entropy in two, distinct Kullback-Leibler distances:

$$\begin{aligned} S_{rel} &= k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln\left[\frac{p_r(\mathbf{r})}{P_r(\mathbf{r}|U)}\right] \\ &= k_B \int d\mathbf{r}\, p_r(\mathbf{r}) \ln\left[\frac{V^n}{V^N}\frac{p_r(\mathbf{r})}{P_R(\mathbf{M}(\mathbf{r})|U)}\right] - k_B \int d\mathbf{r}\, p_r(\mathbf{r}) \ln\left[\frac{V^n}{V^N}\frac{p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))}\right] \\ &= S_{tot} - S_{map} \end{aligned} \tag{3.4}$$

where n and N denote the number of atomistic and CG sites, respectively. Here S_{tot} represents a KL divergence between the atomistic probability and the CG one, that is, the most comprehensive measure of distance one can think of in bottom-up CG. Indeed, it depends both on the mapping and on the approximate CG potential U. The second term in Eq. 3.4, S_{map}, is the mapping entropy of the CG model, which measures the distance between the MB-PMF and the AT reference probability density, thus being completely independent of the choice of the CG potential U.

The sign of S_{map} differs from the one employed in the works of Noid and coworkers [57, 54, 56], being consistent with the convention introduced by Shell [22] (Eq. 1.32). On one hand, this enables the mapping entropy to be

directly related to a loss of information in the KL sense—a *positive* KL divergence implies a *loss* of information. On the other hand, it allows the relative entropy in Refs. [57, 54] to be considered a difference of information losses—those of U and U^0, see Eq. 3.4—calculated with respect to the atomistic system, so that the vanishing of S_{rel} for $U = U^0$ in Refs. [57, 54] effectively amounts at recalibrating the zero of the relative entropy as originally defined in Ref. [22].

It is important to notice that the mapping entropy introduced in Eq. 3.4 can be related to the entropic component of the MB-PMF, S^0 (Eq. 1.12)[56]:

$$\begin{aligned} S_{map} &= k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{V^n}{V^N} \frac{p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))} \right] \\ &= -\int d\mathbf{R}\; p_R(\mathbf{R})\; S^0(\mathbf{R}) \end{aligned} \tag{3.5}$$

A mapping that possesses zero mapping entropy would thus imply, see Eq. 1.12, that the entropic component of the MB-PMF is bound to vanish for each CG configuration $\mathbf{R}$, therefore eliminating the temperature-dependence of U^0 and making the latter transferable in temperature.

Starting from the third line of Eq. 3.4, the mapping entropy can be further decomposed in two terms:

$$S_{map} = -k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{V^N}{V^n} \Omega_1(\mathbf{M}(\mathbf{r})) \right] + k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})} \right]. \tag{3.6}$$

Here $\Omega_1(\mathbf{M}(\mathbf{r}))$, see Eq. 1.29, is the number of atomistic configurations associated to the CG configuration $\mathbf{M}(\mathbf{r})$.

The mapping entropy introduced here in Eqs. 3.4 and 3.6 differs substantially from the object introduced by Shell in Ref. [22] (Eq. 1.32); in fact, when considering $P_r(\mathbf{r}|U) = P_{1r}(\mathbf{r}|U) = \frac{P_R(\mathbf{M}(\mathbf{r})|U)}{\Omega_1(\mathbf{M}(\mathbf{r}))}$ as in Eq. 1.29, the mapping entropy reduces to a canonical average of the number of atomistic microstates mapping onto the same CG macrostate.

In the decomposition of the mapping entropy presented in Eq. 3.6, the first term possesses a purely geometrical origin, representing how well the chosen mapping function *partitions* the atomistic states inside CG configurations. The second, more interesting term accounts for the smearing of probabilities that is inherent to the process of coarse-graining. It is still a KL divergence, in which the atomistic reference density is compared to its smeared counterpart $\bar{p}_r(\mathbf{r})$:

$$\bar{p}_r(\mathbf{r}) = \frac{p_R(\mathbf{M}(\mathbf{r}))}{\Omega_1(\mathbf{M}(\mathbf{r}))}. \tag{3.7}$$

Mathematically speaking, Eqs. 3.6 and 3.7 do not provide any substantial simplification to the original definition of the mapping entropy, since I only divided and multiplied by the factor $\Omega_1(\mathbf{M}(\mathbf{r}))$. Nevertheless, this operation is necessary to understand the nature of this quantity; assume that there exists an observer living in the CG configurational space, located on the CG macrostate $\mathbf{R}$. It is known that the probability weight of sampling $\mathbf{R}$ is given by the MB-PMF. Now imagine that the observer is asked to estimate the probability weight associated to the AT microstates that map onto $\mathbf{R}$, which live in the reference configurational space. The observer does not possess any detailed information about this high-resolution space, but only a cumulative property ($p_R(\mathbf{R})$): he will conclude that all the atomistic microstates mapping on $\mathbf{R}$ display the same weight, namely the CG probability divided by the number of states $\Omega_1(\mathbf{M}(\mathbf{r}))$, which is exactly $\bar{p}_r(\mathbf{r})$. In doing so, however, the observer is neglecting the fact that microscopic probabilities are indeed distinct, and even if two microstates map onto the same CG macrostate, the associated Boltzmann weights can be substantially different.

The first, geometric term in Eq. 3.6 does not vanish in general [57]. However, a simple prescription on the functional form of the mapping allows to set it to zero; indeed, the first logarithm in Eq. 3.6 is identically zero if the mapping takes the form of a *decimation* [175, 176]:

$$\begin{aligned} &\mathbf{M}_I(\mathbf{r}) = \sigma_i \mathbf{r}_i, \ \sigma_i = 1 \text{ for one } I, \text{ 0 otherwise}, \\ &\sum_{i=1}^{n} \sigma_i = N. \end{aligned} \tag{3.8}$$

In this context, only a subset of the atoms of the system is considered, and the remaining ones are integrated out. A simple example of this rule is provided by atomically detailed implicit solvent models [31, 84]. According to this rule, the number of configurations mapping to the same CG one is:

$$\Omega_1(\mathbf{M}(\mathbf{r})) = V^{n-N}, \tag{3.9}$$

which immediately leads to a simplified expression for the mapping entropy, where only the smearing of probabilities is present:

$$S_{map} = k_B \int d\mathbf{r} \; p_r(\mathbf{r}) \ln \left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})} \right]. \tag{3.10}$$

This quantity is always non-negative and vanishes only if, for each CG macrostate $\mathbf{R}$, the atomistic configurations $\mathbf{r}$ that map onto it have the same probability. In the canonical ensemble this is equivalent to require their *isoenergeticity*, i.e., they must possess the same value of the potential energy.

A different way of retrieving the mapping entropy involves the non-ideal configurational entropies of the high and low-resolution systems. It is well-known that a system of n free, non-interacting particles possesses an entropy given by the Boltzmann formula:

$$\begin{aligned} S_b &= -nk_B \int_V d\mathbf{r_1} \frac{1}{V} \ln\left(\frac{1}{V}\right) \\ &= k_B \ln\left(V^n\right), \end{aligned} \tag{3.11}$$

which is the sum of the ideal configurational entropies of the n particles. Switching on some interactions between the particles causes a decrease in entropy, as statistical dependencies among the particles are introduced and more information about the system is readily available. In other words, the probability of sampling a configuration $\mathbf{r}$ is not uniform, i.e., maximally entropic, and equal to $\frac{1}{V^n}$, but rather becomes the usual Boltzmann weight of Eq. 1.4. The associated entropy S, also called Gibbs entropy (Eq. 1.38), is always lower than its uniform counterpart [177]:

$$S = -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln(p_r(\mathbf{r})) \tag{3.12}$$

The integral extends over the volume V^n as usual. The non-ideal configurational entropy measures the gain in information guaranteed by introducing interactions with respect to the free-particles system:

$$s_r = -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln(V^n p_r(\mathbf{r})) \tag{3.13}$$

When, given a mapping, interactions in a CG model are described by the MB-PMF, one can define the non-ideal configurational entropy of the atomistic system over the CG coordinates:

$$s_R = -k_B \int d\mathbf{R}\ p_R(\mathbf{R}) \ln(V^N p_R(\mathbf{R})), \tag{3.14}$$

where $p_R(\mathbf{R})$ is compared to the uniform distribution in the CG space ($\frac{1}{V^N}$). By inserting the known expression for $p_R(\mathbf{R})$ (Eq. 1.7) into the previous equation:

$$\begin{aligned} s_R &= -k_B \int d\mathbf{R}\ \left[\int d\mathbf{r}\ p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}))\right] \ln(V^N p_R(\mathbf{R})) \\ &= -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln(V^N p_R(\mathbf{M}(\mathbf{r}))), \end{aligned} \tag{3.15}$$

it is possible to show that the difference between the two information gains in Eqs. 3.15 and 3.13 is:

$$s_R - s_r = k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left(\frac{V^{n-N} p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))}\right), \tag{3.16}$$

and by virtue of Eq. 3.7 and 3.9, one finally obtains

$$s_R - s_r = S_{map}, \tag{3.17}$$

further highlighting that the mapping entropy represents the difference in information content between the distribution obtained by reducing the level of resolution at which the system is observed, $p_R(\mathbf{R})$, and the original, microscopic reference, $p_r(\mathbf{r})$.

The mapping entropy provides an important link between a fully detailed model and a reduced description of it. More specifically, this mathematical object provides a natural tool to measure the inherent loss of information that arises due to the process of dimensionality reduction. Still, the mapping entropy presented in Eq. 3.10 cannot be explicitly calculated from a computational perspective. Indeed, the Kullback-Leibler divergence is a canonical average of the logarithm of two high-dimensional probability distributions, which is extremely hard to calculate except for very simple systems. In the next section I investigate the nature of this canonical average, showing that it can be approximated by a much simpler expression.

3.1 Explicit calculation of the mapping entropy

In the last section I discussed how the mapping entropy of Eq. 3.10 vanishes if and only if the atomistic configurations mapping onto each CG macrostate possess the same potential energy. Let me make this statement more evident by explicitly writing the probability distributions inside the logarithm:

$$\begin{aligned} S_{map} &= -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left[\frac{\bar{p}_r(\mathbf{r})}{p_r(\mathbf{r})}\right] \\ &= -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left[\frac{z \times e^{\beta u(\mathbf{r})} \times \int d\mathbf{r}' e^{-\beta(u(\mathbf{r}'))} \delta(\mathbf{M}(\mathbf{r}') - \mathbf{M}(\mathbf{r}))}{z \times \Omega_1(\mathbf{M}(\mathbf{r}))}\right] \\ &= -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left[\frac{\int d\mathbf{r}' e^{-\beta(u(\mathbf{r}')-u(\mathbf{r}))} \delta(\mathbf{M}(\mathbf{r}') - \mathbf{M}(\mathbf{r}))}{\int d\mathbf{r}' \delta(\mathbf{M}(\mathbf{r}') - \mathbf{M}(\mathbf{r}))}\right] \end{aligned} \tag{3.18}$$

so that if $u(\mathbf{r}') = u(\mathbf{r})\ \forall\ \mathbf{r}'$ s.t. $\mathbf{M}(\mathbf{r}') = \mathbf{M}(\mathbf{r})$, the argument of the logarithm is unity and the right-hand side of Eq. 3.18 vanishes. Here, for the sake

of clarity, the mapping entropy is expressed with a minus sign due to the inversion of numerator and denominator inside the logarithm.

It is possible and convenient to express S_{map} as an integral over the CG configuration space by introducing a delta $1 = \int d\mathbf{R}\, \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$ in Eq. 3.18:

$$S_{map} = -k_B \int d\mathbf{R} \int d\mathbf{r}\; p_r(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \times \tag{3.19}$$
$$\ln \left[\frac{\int d\mathbf{r}' e^{-\beta(u(\mathbf{r}')-u(\mathbf{r}))} \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}{\int d\mathbf{r}' \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})} \right]$$
$$= \int d\mathbf{R}\; p_R(\mathbf{R}) S_{map}(\mathbf{R}), \tag{3.20}$$

The total mapping entropy becomes an integral over the CG configurational space, where each term is weighted with the probability determined by the MB-PMF. The mapping entropy of a CG macrostate is then given by

$$S_{map}(\mathbf{R}) = -\frac{k_B}{p_R(\mathbf{R})} \int d\mathbf{r}\; p_r(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \times \tag{3.21}$$
$$\ln \left[\frac{\int d\mathbf{r}' e^{-\beta(u(\mathbf{r}')-u(\mathbf{r}))} \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}{\int d\mathbf{r}' \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})} \right].$$

The integral in Eq. 3.21 runs over the microscopic configurations $\mathbf{r}$ and, inside the logarithm, their energy is compared to that of all the other configurations $\mathbf{r}'$, provided that they map onto the (same) coarse-grained macrostate $\mathbf{R}$. In this way, all pairs of energies are considered.

It is useful to introduce another identity on the energies $1 = \int dU' \delta(u(\mathbf{r}') - U')$, that fixes the energy of configuration $\mathbf{r}'$. This quantity can be inserted in the logarithm of Eq. 3.21 to switch from a configurational to an energetic integral:

$$\ln \left[\frac{\int d\mathbf{r}' e^{-\beta(u(\mathbf{r}')-u(\mathbf{r}))} \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}{\int d\mathbf{r}' \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})} \right] = \ln \int dU' P(U'|\mathbf{R}) e^{-\beta(U'-u(\mathbf{r}))}, \tag{3.22}$$

where

$$P(U'|\mathbf{R}) = \frac{\int d\mathbf{r}' \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R}) \delta(u(\mathbf{r}') - U')}{\int d\mathbf{r}' \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})} \tag{3.23}$$

is the microcanonical (unweighted) conditional probability of possessing energy U' given that the CG macrostate is $\mathbf{R}$. It is possible to write it as $\Omega_1(U', \mathbf{R})/\Omega_1(\mathbf{R})$, that is, the multiplicity of atomistic configurations such that $\mathbf{M}(\mathbf{r}) = \mathbf{R}$ and $u(\mathbf{r}') = U'$ normalised by the multiplicity of configurations that map to $\mathbf{R}$. A second identity $1 = \int dU \delta(u(\mathbf{r}) - U)$ on the energies

provides the following expression for $S_{map}(\mathbf{R})$:

$$\begin{aligned} S_{map}(\mathbf{R}) &= -k_B \int d\mathbf{r} \frac{p_r(\mathbf{r})}{p_R(\mathbf{R})} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \times \\ &\quad \ln \left[\int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - u(\mathbf{r}))] \right] \\ &= -k_B \int dU \ln \left[\int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - U)] \right] \times \\ &\quad \int d\mathbf{r} \frac{p_r(\mathbf{r})}{p_R(\mathbf{R})} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \delta(u(\mathbf{r}) - U). \end{aligned} \tag{3.24}$$

The last integral in Eq 3.24, here dubbed $P_\beta(U|\mathbf{R})$,

$$P_\beta(U|\mathbf{R}) = \int d\mathbf{r} \frac{p_r(\mathbf{r})}{p_R(\mathbf{R})} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \delta(u(\mathbf{r}) - U) \tag{3.25}$$

is now the canonical—i.e., Boltzmann-weighted—conditional probability of possessing energy U provided that $\mathbf{M}(\mathbf{r}) = \mathbf{R}$, namely $p_R(U, \mathbf{R})/p_R(\mathbf{R})$. One thus obtains:

$$\begin{aligned} S_{map}(\mathbf{R}) &= -k_B \int dU P_\beta(U|\mathbf{R}) \times \ln \left[\int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - U)] \right] \\ &= -k_B \ln \left[\int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - \langle U \rangle_{\beta|\mathbf{R}})] \right], \end{aligned}$$

where

$$\langle U \rangle_{\beta|\mathbf{R}} = \int dU P_\beta(U|\mathbf{R}) U \tag{3.26}$$

is the canonical average of the microscopic potential energy over the CG macrostate $\mathbf{R}$. A direct calculation of $S_{map}(\mathbf{R})$ starting from the last line of Eq. 3.26 requires to perform an average over the microcanonical distribution $P(U'|\mathbf{R})$, which is not straightforwardly accessible in NVT simulations. However, there is a connection between $P(U|\mathbf{R})$ in Eq. 3.23 and $P_\beta(U|\mathbf{R})$ in Eq. 3.25: if one writes $p_R(\mathbf{R})$ as $\int dU' \exp[-\beta(U')] \Omega_1(U', \mathbf{R})$ and $p_R(U, \mathbf{R})$ as $\exp[-\beta(U)] \Omega_1(U, \mathbf{R})$, standard reweighing provides

$$P(U|\mathbf{R}) = \frac{P_\beta(U|\mathbf{R}) \exp[\beta U]}{\int dU' P_\beta(U'|\mathbf{R}) \exp[\beta U']}. \tag{3.27}$$

Eq. 3.27 enables one to convert the microcanonical average in Eq. 3.26 to a canonical one, so that

$$S_{map}(\mathbf{R}) = k_B \ln \left[\int dU' P_\beta(U'|\mathbf{R}) \; e^{\beta(U' - \langle U \rangle_{\beta|\mathbf{R}})} \right]. \tag{3.28}$$

In principle, this equation can be employed to calculate the mapping entropy of a coarse-grained macrostate $\mathbf{R}$. Nevertheless, the average of the exponential function is usually difficult to compute from a numerical perspective. It is indeed common practice to resort to a cumulant expansion of the logarithm of the exponential average of an observable y [178]:

$$\begin{aligned} \ln(E[\alpha y]) &= \sum_i \frac{k_i \alpha^i}{i!} \\ &= \alpha E[y] + \frac{\alpha^2}{2}(E[y^2] - E[y]^2) + ... \end{aligned} \tag{3.29}$$

where α is a constant and k_i is the i-th cumulant of the expansion, which is equal to $E[y]$ and $E[y^2] - E[y]^2$ for $i = 1$ and $i = 2$, respectively. Truncating the expansion of Eq. 3.28 to the second order one obtains

$$S_{map}(\mathbf{R}) \simeq k_B \frac{\beta^2}{2} \langle (U - \langle U \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}. \tag{3.30}$$

Again, the mapping entropy of a CG macrostate is zero if all the microscopic configurations mapping onto it possess the same potential energy. In the real cases this is not true, and the information loss due to the choice of the mapping is proportional to the variance of the energies of these microstates.

Inserting Eq. 3.30 into Eq. 3.19, one obtains the following expression for the *total* mapping entropy:

$$S_{map} \simeq k_B \frac{\beta^2}{2} \int d\mathbf{R}\; p_R(\mathbf{R}) \langle (U - \langle U \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}, \tag{3.31}$$

where each configuration-dependent component is properly weighted according to the MB-PMF.

It is important to highlight that the expression obtained in Eq. 3.30 resonates with a number of articles by Noid and coworkers [53, 93, 56]. In particular, they show how it is possible to put the energy fluctuation internal to each CG macrostate (Eq. 3.30) in relation with the atomic specific heat at constant volume, c_V, defined as:

$$c_V = \frac{\partial}{\partial T} \langle u(\mathbf{r}) \rangle = \frac{\sigma_u^2}{k_B T^2}, \tag{3.32}$$

where σ_u^2 is the variance of the atomistic potential over the high-resolution configurational space:

$$\sigma_u^2 = \int_{V^n} d\mathbf{r}\; p_r(\mathbf{r}) \left(u(\mathbf{r}) - \bar{u}\right)^2, \tag{3.33}$$

assuming the explicit temperature-independence of the atomistic potential u. In Ref. [93] Lebold and Noid show how this variance can be decomposed in two terms once the energetic component of the MB-PMF of the coarse-grained system is explicitly introduced:

$$\sigma_u^2 = \int_{V^N} d\mathbf{R}\; p_R(\mathbf{R}) \left(E^0(\mathbf{R}) - \bar{u}\right)^2 + \int_{V^N} d\mathbf{R}\; p_R(\mathbf{R}) \sigma_{u|\mathbf{R}}^2, \tag{3.34}$$

where

$$\sigma_{u|\mathbf{R}}^2 = \int_{V^n} d\mathbf{r}\; \frac{p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})}{p_R(\mathbf{R})} \left(u(\mathbf{r}) - E^0(\mathbf{M}(\mathbf{r}))\right)^2. \tag{3.35}$$

It is important to note that the mixed terms in Eq. 3.33 containing $(u(\mathbf{r}) - E^0(\mathbf{M}(\mathbf{r})))$ can be safely neglected since $E^0(\mathbf{M}(\mathbf{r})) = \langle u(\mathbf{r})\rangle_{\mathbf{R}}$ (Eq. 1.11). The overall fluctuations of the atomistic potential are divided in the fluctuations of the energetic component of the MB-PMF, E^0, *plus* a term that accounts for the internal deviations of u with respect to E^0 within each CG macrostate. The latter quantity measures the average energetic discrepancy between $E^0(\mathbf{R})$ and $u(\mathbf{r})$ for all atomistic microstates $\mathbf{r}$ such that $\mathbf{M}(\mathbf{r}) = \mathbf{R}$.

$\sigma_{u|\mathbf{R}}^2$ is related to the coarse-grained configuration-dependent specific heat:

$$C_{U^0}(\mathbf{R}) = \left(\frac{\partial E^0}{\partial T}\right)_{\mathbf{R},V} = T\left(\frac{\partial S^0}{\partial T}\right)_{\mathbf{R},V} \tag{3.36}$$

$$= \frac{\sigma_{u|\mathbf{R}}^2}{k_B T^2} \tag{3.37}$$

Realising that $S_{map}(\mathbf{R}) \simeq \frac{\sigma_{u|\mathbf{R}}^2}{k_B T^2}$, it is possible to obtain the following approximate relation between S_{map} and $C_{U^0}(\mathbf{R})$:

$$C_{U^0}(\mathbf{R}) \simeq 2S_{map}(\mathbf{R}). \tag{3.38}$$

The atomic specific heat at constant volume c_V is constant at a given state point, and it can be decomposed in two terms, as showed in Eq. 3.34; an optimal mapping such that $S_{map}(\mathbf{R}) = 0\;\forall\,\mathbf{R}$ is able to *concentrate* the overall contribution to c_V into the inter-macrostate energetic fluctuations (first term of Eq. 3.34), minimizing the intra-macrostate ones (second term of 3.34) and eliminating the temperature dependence of the MB-PMF, (Eqs. 1.12 and 3.36).

Noid and coworkers employed those concepts from classical thermodynamics to develop a dual potential approach [53] that introduces a novel ingredient in a bottom-up CG procedure to enforce the accurate modelling

of the atomistic energetics, and to predict the temperature-dependence of the low-resolution model. Given a CG approximate potential U, built so as to approximate the MB-PMF U^0 at a fixed state point, it is not guaranteed that U reproduces the underlying atomistic energetic properties [56]. Therefore, Lebold and Noid propose an energy-matching functional to construct an energetic operator E, explicitly parametrised to variationally approximate E^0 through the minimisation of an appropriate functional:

$$\chi^2_E[E] = \langle |E(\mathbf{M}(\mathbf{r})) - u(\mathbf{r})|^2 \rangle \tag{3.39}$$

This term can be further decomposed in two separated averages:

$$\chi^2_E[E] = \langle |\langle u(\mathbf{r}) \rangle_{\mathbf{R}} - u(\mathbf{r})|^2 \rangle + \langle |E(\mathbf{M}(\mathbf{r})) - \langle u(\mathbf{r}) \rangle_{\mathbf{R}}|^2 \rangle \tag{3.40}$$

Minimising $\chi^2_E[E]$ on $E(\mathbf{R})$ for a given, *fixed* mapping as in Refs. [53, 93] is tantamount to minimising the second term of Eq. 3.40, with the objective of reducing the error introduced by approximating $\langle E^0 \rangle_{\beta|\mathbf{R}}$ through $E(\mathbf{R})$. However, a comparison of Eqs. 3.31 and 3.40 displays that S_{map} coincides, up to a multiplicative factor, with the first term of Eq. 3.40. Critically, the latter depends only on the mapping $\mathbf{M}$ and would be nonzero also in the case of an *exact* parametrisation of E, that is, if $E(\mathbf{R}) \equiv \langle E^0 \rangle_{\beta|\mathbf{R}}$. The approach illustrated here goes in a direction complementary to that of Refs. [53, 93], as the ultimate objective of this chapter consists of identifying those mappings that minimise the one contribution to $\chi^2[E]$ that is due to, and depends only on, the CG representation $\mathbf{M}$.

3.2 Numerical implementation

With the exception of the cumulant expansion in Eq. 3.30, the calculations performed in the previous section are exact. I now show how the mapping entropy can be practically computed, provided that a certain amount of information about the reference system is available. The latter ultimately consists in a finite set of L fully atomistic equilibrium configurations sampled from the canonical ensemble; these microstates must be *mapped* onto a certain number of CG macrostates K. In this context, the discretized mapping entropy $\tilde{S}_{map}$ reads:

$$\tilde{S}_{map} = k_B \frac{\beta^2}{2} \sum_{i=1}^{K} p_R(\mathbf{R}_i) \langle (U - \langle U \rangle_{\beta|\mathbf{R}_i})^2 \rangle_{\beta|\mathbf{R}_i}, \tag{3.41}$$

where the sum runs over each CG configuration and each factor $\langle (U - \langle U \rangle_{\beta|\mathbf{R}_i})^2 \rangle_{\beta|\mathbf{R}_i}$ is weighted with a discretized probability $p_R(\mathbf{R}_i)$:

$$p_R(\mathbf{R}_i) = \int d\mathbf{r}\ p_r(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}_i) \simeq \frac{1}{L} \sum_{j=1}^{L} \delta(\mathbf{M}(\mathbf{r}_j) - \mathbf{R}_i), \qquad (3.42)$$

which is exactly the fraction of atomistic configurations j mapping onto $\mathbf{R}_i$. It is important to emphasise that this is valid as long as all the atomistic microstates are sampled according to the Boltzmann statistics, which amounts at requiring that the simulations are carried out in the canonical ensemble.

In the numerical implementation of the method the MD trajectories of three candidate proteins are considered. They are first equilibrated in the NVT ensemble making use of a stochastic velocity rescaling thermostat [179] with a coupling constant $\tau_T = 0.1$ ps. Then, a Parrinello-Rahman barostat ($\tau_P = 2$ ps) [180] is added to the thermostat to set the pressure of the system in a NPT equilibration. Finally, the proteins are simulated in the NVT ensemble for 200 nanoseconds at $300K$ with the GROMACS 2018 package [181, 37] and the AMBER99SB-ILDN force field [182], employing a time step of 2 femtoseconds. The LINCS algorithm is used to constrain all the covalent bonds involving hydrogen atoms and long-range electrostatics are treated by means of the *Particle Mesh Ewald* method [183].

Here I proceed to a description of the three candidate proteins, focusing on their biological role and on the qualitative description of their simulation.
[TAM] A recently released [184] 31-residue *tamapin* mutant (PDB code 6D93). Tamapin is the toxin produced by the Indian red scorpion. It features a remarkable selectivity towards a peculiar calcium-activated potassium channel (SK2), whose potential use in the pharmaceutical context has made it a preferred object of study during the past decade [185, 186]. Throughout the simulation almost every residue is highly solvent-exposed. Side chains fluctuate substantially, thus giving rise to a notable structural variability.
[AKE] *Adenylate Kinase* (PDB code 4AKE). It is a 214 residue-long phosphotransferase enzyme that catalyses the interconversion between adenine diphosphate (ADP) and monophosphate (AMP) and their energetically rich complex, adenine triphosphate (ATP) [187]. It can be subdivided in three structural domains, CORE, LID, and NMP [188]. The CORE domain is stable, while the other two undergo large conformational changes. Its central biochemical role in the regulation of the energetic balance of the cell and relatively small size, combined with the possibility to observe conformational transitions over timescales easily accessible by plain MD [189], make it the ideal candidate to test and validate novel computational methods [190, 191, 158, 192]. During the MD simulation the protein displays many

rearrangements in the two motile domains, which occur to be quite close at many points. Nevertheless, the protein does not undergo a full *open* $\leftrightarrow$ *closed* conformational transition.
[AAT] $\alpha-1$ *antitrypsin* (PDB code 1QLP). With 5934 atoms (372 residues), this protein is almost two times bigger than adenylate kinase. $\alpha - 1$ antytripsin is a globular biomolecule and it is well known to exhibit a conformational rearrangement over the timescales of the minutes [193, 194, 195]. In the course of the simulation the molecule experiences fluctuations particularly localised in correspondence of the most solvent-exposed residues. The protein bulk appears to be very rigid, and there is no sign of major conformational rearrangements.

The energies contained in Eq. 3.41 correspond to the atomistic intramolecular potential energy of the protein, thus neglecting the solvent-solvent and solvent-protein interaction terms. This simplification is formally wrong, as one should consider the full system energy in the calculation, the vast majority of which is due to the solvent. Nevertheless, I deem it appropriate to neglect such contributions since the *relative* fluctuations of the overall energy in a MD simulation of a protein at room temperature are negligible with respect to those of the intramolecular energy.

3.2.1 Definition of coarse-grained macrostates

In order to construct K CG macrostates from L microstates *observed with coarse-grained glasses*, it is necessary to introduce a notion of similarity between the latter. In the context of protein structures, the most intuitive measure of pairwise similarity is the Root Mean Squared Deviation (RMSD): given two structures $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same molecule with n atoms, the RMSD between them is defined as:

$$\mathrm{RMSD}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \mathcal{R}\mathcal{T}\boldsymbol{y}_i)^2}, \tag{3.43}$$

where $\mathcal{R}\mathcal{T}$ is the roto-translation that superimposes $\boldsymbol{y}$ to $\boldsymbol{x}$ according to some optimality criterion, thus minimizing the overall displacement. Among the several methods to align protein structures, I employ the one from Kabsch [196, 197]. The CG RMSD between two atomistic structures filtered by a decimation mapping $\mathbf{M}$ (Eq. 3.8) with N sites is immediately given by

$$\mathrm{RMSD}^{\mathrm{CG}}(\mathbf{M}(\boldsymbol{x}), \mathbf{M}(\boldsymbol{y})) = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(\mathbf{M}_I(\boldsymbol{x}) - \mathcal{R}\mathcal{T}^{\mathrm{CG}}\mathbf{M}_I(\boldsymbol{y}))^2}. \tag{3.44}$$

Here the sum runs over the retained CG sites and $\mathcal{RT}^{CG}$ is a rigid roto-translation that superimposes the two coarse-grained structures, without taking into account the presence of the removed degrees of freedom.

Once a notion of similarity between coarse-grained configurations of an atomistic trajectory is defined, it is necessary to specify a prescription to lump them together to form CG states. This calls for the introduction of a clustering algorithm that, upon being applied on the pairwise $\mathrm{RMSD}^{\mathrm{CG}}$ matrix, outputs the set of desired K CG configurations, each one associated to a certain subset of the L microstates.

I here employ the UPGMA algorithm with average linkage [198], a hierarchical clustering procedure that iteratively aggregates configurations in a dendrogram according to the following strategy:

1. At the first step, the minimum of the $\mathrm{RMSD}^{\mathrm{CG}}$ similarity matrix is retrieved and the two corresponding entries x, y (the *leaves*) are merged together in a new cluster k;

2. k is placed in the middle of its two constituent, meaning that the similarity matrix is updated in order to account for the presence of the newly formed cluster $d(k, z) = (d(x, z) + d(y, z))/2$, where z is another structure;

3. Steps 1. and 2. are iterated until one *root* is found. The distance among clusters k and w is generalised as follows:

$$d(k, w) = \sum_{i \in k} \sum_{j \in w} \frac{d(k[i], w[j])}{|k| \times |w|}, \tag{3.45}$$

   where $|k|$ and $|w|$ are the populations of the clusters and $k[i]$ and $w[j]$ their elements;

The choice falls on this clustering algorithm because it is robust, completely deterministic and widely applied. It is important to emphasize that different versions of the method exist, which differ notably in the definition of the inter-cluster distance. As an example, in the *complete linkage* hierarchical clustering the distance between two clusters (Eq. 3.45) is defined as

$$d(k, w) = \max_{i \in k, j \in w} d(k[i], w[j]), \tag{3.46}$$

namely the maximum value of the distance among their constituent elements. In the context of this calculation, the average linkage criterion seems more appropriate, since the complete linkage prescription induces a non-trivial

distribution in the space of frequencies, which could introduce an additional bias in the computation of the mapping entropy.

The scipy [199, 200] implementation [201, 202] of the method is employed, which scales with $\mathcal{O}(L^2 logL)$ with the number of atomistic configurations L.

Once the dendrogram is obtained, the actual division in clusters can be retrieved in two different ways; in the first approach, a real number is employed as a threshold on the inter-cluster distance: when the minimum of the reduced distance matrix exceeds this value, the agglomerative procedure stops and the current clusters are saved. In the second strategy, the number K of clusters is specified *a priori*, and the dendrogram is cut when there are exactly K leaves. The latter criterion is selected for a practical reason connected to the minimisation of the mapping entropy, as the first prescription pushes the optimisation procedure to create as many clusters as possible, in order to minimize the energy variance inside them. This is due to the fact that a CG cluster containing only one microstate possesses an energy variance equal to zero and, consequently, zero mapping entropy. Since the aim is to minimise the impact of such flaw, which is entirely due to the finiteness of the sampling, let me fix the number of CG macrostates.

The choice of K has an impact on the overall value of the mapping entropy. When $K << L$, few CG configurations are identified, and the energy variance of the microstates mapping onto them is non-negligible. Instead, if K has the same order of magnitude of L, few microstates map onto the same CG configuration, and their energetic displacement is expected to be small. In the first scenario, a low value of mapping entropy is reached by CG mappings that best account for large energetic fluctuations. In the second case, $\tilde{S}_{map}$ will be in general very small, favouring those reduced representations able to discriminate structures possessing tiny differences in energy.

This aspect is not investigated in detail here, but rather I choose to employ five, evenly spaced values of K in the definition of an "average" mapping entropy:

$$\Sigma = \frac{1}{5} \sum_{K \in \mathcal{K}} \tilde{S}_{map}(K) \tag{3.47}$$

The set of values $\mathcal{K}$ is reported in Table 3.1. In general, the number of CG clusters ranges between $\frac{1}{500}$ and $\frac{1}{100}$ of the original number of atomistic configurations.

3.3 Mapping optimisation

In the previous subsection I have introduced the necessary ingredients to calculate the mapping entropy provided a set of atomistic configurations and

Protein	$\mathcal{K}$
Tamapin	$\{34, 48, 62, 76, 91\}$
Adenylate Kinase	$\{29, 58, 87, 116, 147\}$
$\alpha - 1$ antytripsin	$\{7, 29, 51, 73, 96\}$

Table 3.1: number of clusters employed to average the mapping entropy (Eq. 3.47).

Protein	N_α	$N_{\alpha\beta}$	N_{bkb}	N_{heavy}
Tamapin (TAM)	31	59	124	230
Adenylate Kinase (AKE)	214	408	856	1656
$\alpha - 1$ antytripsin (AAT)	372	723	1488	2956

Table 3.2: Values of N_α, $N_{\alpha\beta}$, N_{bkb} and N_{heavy} (see text) for each analysed protein.

the associated intramolecular potential energies. It is interesting to employ this observable to explore and rank the set of coarse-grained mappings of a protein. Unfortunately, the overall number of decimation mappings that can be applied to biomolecules is astronomical even for the smallest peptides:

$$\Theta = \sum_{N=1}^{n} \frac{n!}{N!\ (n-N)!} = 2^n - 1. \tag{3.48}$$

The sum is carried on over the number of CG sites N and the case with N equal to 0 is excluded. Two approximations can be introduced to reduce the number of mappings that are the subject of the analysis; first, only the heavy atoms of the protein are employed as putative CG sites ($n = N_{heavy}$), thus completely neglecting the presence of the hydrogen atoms, which are always integrated out; second, the exploration is restricted to three fixed, chemically-relevant values of N for each protein: (*i*) N_α, i.e., the number of C_α atoms of the structure (equal to the number of amino acids); (*ii*) $N_{\alpha\beta}$, the number of C_α and C_β atoms; and (*iii*) N_{bkb}, which results from counting all the heavy atoms belonging to the main chain of the protein. The values of N for mappings (*i*)-(*iii*) in the case of TAM, AKE and AAT are listed in Tab. 3.2, together with the corresponding N_{heavy}.

Even limiting the investigation to the aforementioned subset of CG mappings, an exhaustive exploration of it is not feasible. As an example, the number of reduced representations of AAT that contain half of the heavy atoms ($N_{heavy} = 2956$) is of the order of 10^{890}. Hence, it is necessary to resort to an approximate exploration of the space of reduced representations,

which is composed by two subsequent steps; at first, a set of 500 CG mappings is generated by random sampling, and the corresponding values of Σ are calculated, thus constructing a baseline, unbiased distribution to which one can compare "peculiar" values of mapping entropy. Then, a Monte Carlo Simulated Annealing [203, 204] minimisation of Σ is conducted to retrieve the reduced representations that decrease as much as possible the loss of information with respect to the fully detailed atomistic system. Specifically, for each protein of interest and value of N, 48 independent optimisation runs are performed, i.e., minimisations of the mapping entropy with respect to the CG site selection; during these optimisation processes, the CG representation characterised by the lowest value of Σ in each run is saved, thus creating a pool of *optimised* representations.

The protocol is initiated with the generation of a mapping such that the overall number of retained sites is equal to N. Then, at each SA step, the following operations are performed:

1. swap a retained site ($\sigma_i = 1$) and a removed site ($\sigma_j = 0$) in the mapping;

2. compute a distance matrix among CG configurations using the $\mathrm{RMSD}^{\mathrm{CG}}$ (see Eq. 3.44);

3. apply a clustering algorithm on the $\mathrm{RMSD}^{\mathrm{CG}}$ matrix in order to identify the CG macrostates $\mathbf{R}$;

4. compute Σ using Eqs. 3.47 and 3.41.

Once the new value of Σ is obtained, the move is accepted/rejected using a Metropolis-like rule. The overall workflow of the algorithm is schematically illustrated in Fig. 3.1. It is important to underline that, in general, Monte Carlo Simulated Annealing is not the most efficient algorithm for an optimisation; this technique is selected since the variable of interest, i.e., the mapping, is a discrete object, which renders gradient-based approaches complicated to implement. Moreover, the non-negligible computational cost of a single computation of Σ makes more exhaustive approaches, such as Wang-Landau sampling [205, 206], too expensive in terms of CPU time.

For the sake of the accuracy of the optimisation, the more exhaustive the sampling the better, hence the number of sampled atomistic configurations should be at least of the order of the tens of thousands. However, in that case step 2 requires to align a huge number of structure pairs for each proposed CG mapping, which in turn would dramatically slow down the entire process. This problem is circumvented performing a reasonable approximation in the

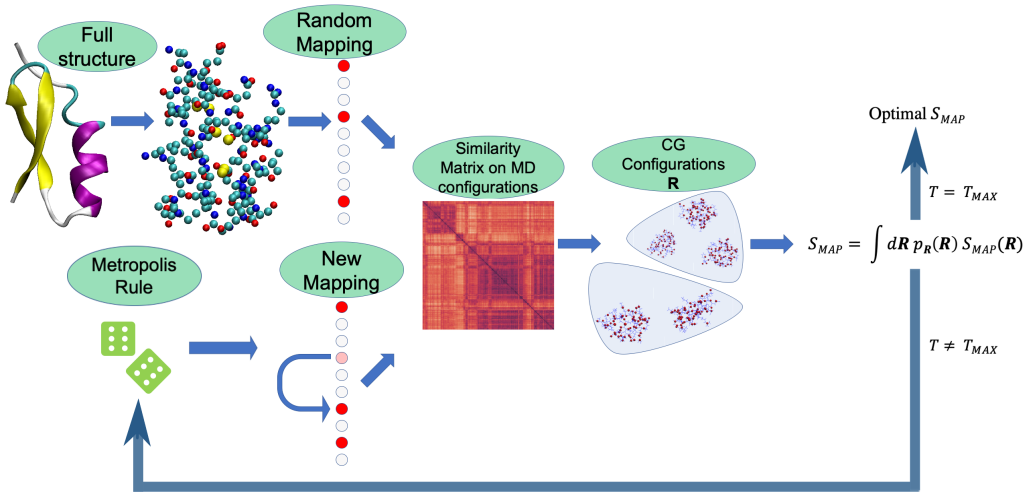


Figure 3.1: Schematic representation of the algorithmic procedure employed to minimise the mapping entropy, the latter being calculated by means of Eq. 3.47. The full similarity matrix is computed once every T_K steps, while in the intermediate steps the approximation of Eq. 3.49 is applied. T_K depends both on the protein and on N. T_{MAX} is the number of simulated annealing steps, $T_{MAX} = 2 \times 10^4$. Image taken from Ref. [24].

calculation of the $\mathrm{RMSD}^{\mathrm{CG}}$ matrix, that is to consider the Kabsch alignment between two structures constant for a certain number of Simulated Annealing steps T_K. It is indeed intuitive to expect negligible variations in the Kabsch alignment between two CG structures differing by a pair of swapped atoms. This assumption is particularly appealing from the point of view of speed and memory, since the expensive and relatively slow alignment procedure produces a result (a rotation matrix) that can be stored with negligible use of resources. After T_K steps, the full Kabsch alignment is applied again.

This approximation results in a substantial reduction of the number of operations that must be executed at each Monte Carlo step. At first, given the initial random mapping operator $\mathbf{M}^0$, the overall $\mathrm{RMSD}^{\mathrm{CG}}$ matrix (Eq. 3.44) is computed between every pair of aligned *mapped* structures, $\mathrm{RMSD}^{\mathrm{CG}}(\mathbf{M}^0(\boldsymbol{x}), \mathbf{M}^0(\boldsymbol{y}))$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ run over the MD configurations and $\mathcal{RT}_0^{\mathrm{CG}}$ is the corresponding optimal roto-translation (see Eq. 3.44). Then, for all moves $\mathbf{M} \to \mathbf{M}'$ within a block of T_K Monte Carlo steps, $\mathbf{M}$ and $\mathbf{M}'$ only differing in a pair of swapped atoms, this quantity is updated with the simple rule

$$\mathrm{MSD}^{\mathrm{CG}}(\mathbf{M}'(\boldsymbol{x}), \mathbf{M}'(\boldsymbol{y})) = \mathrm{MSD}^{\mathrm{CG}}(\mathbf{M}(\boldsymbol{x}), \mathbf{M}(\boldsymbol{y})) - \frac{1}{N}\mathrm{MSD}(\boldsymbol{x}_s, \boldsymbol{y}_s) + \frac{1}{N}\mathrm{MSD}(\boldsymbol{x}_a, \boldsymbol{y}_a), \tag{3.49}$$

where s and a are the removed (substituted) and added atom, respectively, and MSD is the Mean Squared Deviation. Importantly, all the MSD calculations employ $\mathcal{RT}_0^{\mathrm{CG}}$ to superimpose the structures.

This approach clearly represents an approximation to the correct procedure; it has to be emphasised, however, that the impact of such approximation is increasingly perturbative as the size of the system grows. Furthermore, the computational gain that the described procedure enables is sufficient to counterbalance the fact that the exact protocol would be so inefficient to make the optimisation impossible when the number of sampled configurations exceeds the few thousands. For example, choosing $T_K = 1000$ for AAT with $N = N_{bkb}$ the proposed approximation gives a speed-up factor of the order of 10^3.

The optimisation runs for 2×10^3 MC epochs, each of which is composed by 10 steps. This amounts at keeping the temperature constant for 10 steps and then decreasing it according to an exponential law. For the i-th epoch

$$T(i) = T_0 \; e^{-i/\nu}. \tag{3.50}$$

The choice of the hyperparameters T_0 and ν are crucial for a well-behaved MC optimisation. Let me choose $\nu = 300$ so that the temperature at $i = 2000$ is approximately $T_0/1000$. In order to feed the algorithm with reasonable values of T_0, for each of 100 random mappings 10 stochastic moves are performed, measuring $|\Delta\Sigma|$, namely the absolute value of difference between the observables computed at two consecutive steps. Then T_0 is estimated so that a move that leads to an increment of the observable equal to the average of $|\Delta\Sigma|$ would possess an acceptance probability of 0.75 at the first step, that is, when $i = 0$.

3.3.1 Results

Fig. 3.2 displays, for each value of N considered, the distribution of mapping entropies obtained from a random choice of the CG representation of TAM, AKE, and AAT together with each protein's optimised counterpart. For $N = N_{bkb}$ and $N = N_\alpha$, Fig. 3.2 also reports the values of Σ associated to physically-intuitive choices of the CG mapping that are commonly employed in the literature: the backbone mapping ($N = N_{bkb}$), which neglects all atoms belonging to the side chains; and the C_α mapping ($N = N_\alpha$), in which the C_α atoms of the structures are retained. The first is representative of united-atom CG models, while the second is a ubiquitous and rather intuitive choice to represent a protein in terms of a single bead per amino acid [207].

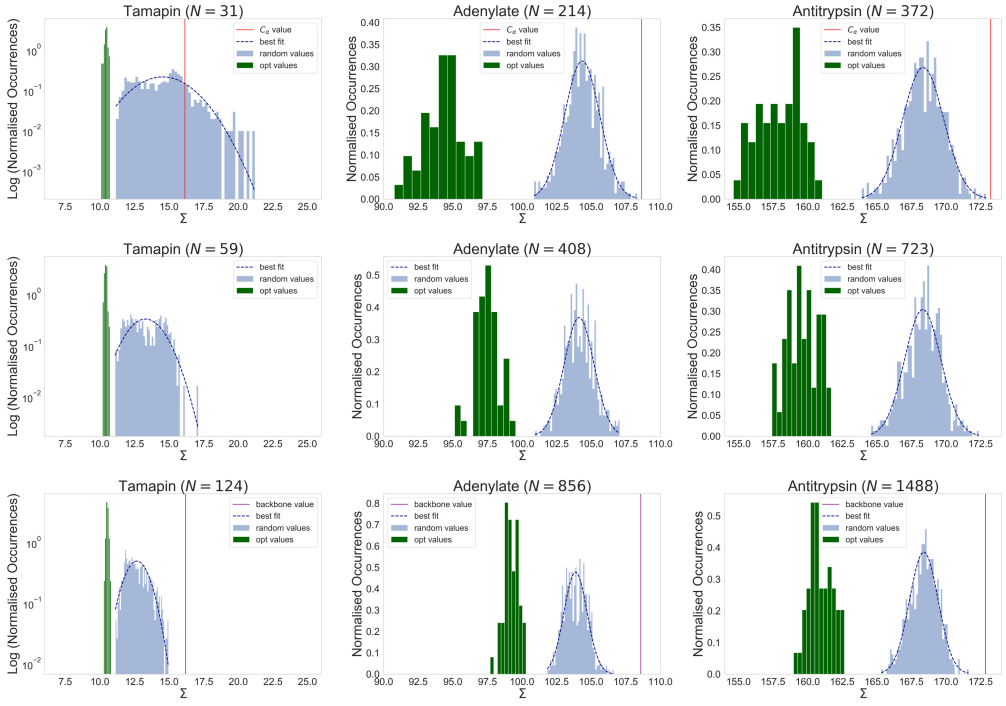


Figure 3.2: Distributions of the values of mapping entropy Σ $[kJ/\mathrm{mol}/K]$ in Eq. 3.47 for random mappings (light blue histograms) and optimised solutions (green histograms). Dark blue dashed lines show the best fit with normal distributions over the random cases. Each column corresponds to an analysed protein, each row to a given number N of retained atoms. In the first and last rows, corresponding to numbers of CG sites equal to the number of C_α atoms and of backbone atoms, N_α and N_{bkb} respectively, the values of the mapping entropy associated to the physically-intuitive choice of the CG sites (see text) is indicated by vertical lines (red for $N = N_\alpha$, purple for $N = N_{bkb}$). Note that the Σ ranges have the same width in all plots. Image taken from Ref. [24].

The optimality of a given mapping with respect to a random choice of the CG sites can be quantified in terms of the Z-score

$$Z = \frac{\Sigma_{opt} - \mu}{\sigma}, \tag{3.51}$$

where μ and σ represent mean and standard deviation of the distribution of Σ over randomly sampled mappings, respectively. Table 3.3 summarises the values of Z found for each N for the proteins under examination, including $Z[backbone]$ and $Z[\mathrm{C}_\alpha]$, which are computed with respect to the random distributions generated with $N = N_{bkb}$ and $N = N_\alpha$ respectively.

N	TAM	AKE	AAT
$\overline{Z}[N_\alpha]$	-2.22 ± 0.06	-7.85 ± 1.14	-6.96 ± 1.03
$\overline{Z}[N_{\alpha\beta}]$	-2.38 ± 0.08	-6.09 ± 0.79	-6.64 ± 0.84
$\overline{Z}[N_{bkb}]$	-2.65 ± 0.09	-5.55 ± 0.62	-7.24 ± 0.85
$Z[backbone]$	4.37	5.65	4.31
$Z[\mathrm{C}_\alpha]$	0.87	3.36	3.28

Table 3.3: Table of Z scores of each analysed protein, reporting mean and standard deviation of the distribution of Z values of the optimised solutions, $\overline{Z}$, for all values of N investigated. Results for the standard mappings—$Z[backbone]$ for backbone atoms only and $Z[\mathrm{C}_\alpha]$ for C_α atoms only—are also included.

As for the physically intuitive CG representations, Fig. 3.2 shows that the value of Σ associated to the backbone mapping is very high for all structures. For TAM in particular, the amount of information retained is so low that the mapping entropy falls 4.37 standard deviations higher than the reference distribution of random mappings, see Table 3.3. This suggests that neglecting the side chains in a CG representation of a protein is detrimental, at least as far as the structural resolution is concerned. In fact, the backbone of the protein undergoes relatively minor structural rearrangements when exploring the neighbourhood of the native conformation, thereby inducing negligible energetic fluctuations; for side chains, on the other hand, the opposite is true, with comparatively larger structural variability and a similarly broader energy range associated to it. Removing side chains from the mapping induces the clustering of atomistically different structures with different energies onto the same CG configuration, the latter being solely determined by the backbone. The corresponding mapping entropy is thus large—worse than a random choice of the retained atoms—since it is related to the variance of the energy in the macrostate (Eq. 3.31).

Calculations employing the C_α mapping for the three structures show that this provides Σ values that are very close to the ones obtained with the backbone mapping, thus suggesting that C_α atoms retain about the same amount of information that is encoded in the backbone. This is reasonable, given the rather limited conformational variability of the atoms along the peptide chain. However, a comparison of the random case distributions for a number N_α and N_{bkb} of retained atoms in Fig. 3.2 reveals that the former generally has a broader spread than the latter, due to the lower number of CG sites; consequently, the Σ of the C_α atoms mapping is closer to the bulk of the distribution of the random case than that of the backbone mapping.

Let me now discuss the case of optimised mappings, that is, CG representations retaining the maximum amount of information about the atomistic reference. Each of the 48 minimisation runs, which have been carried out for each protein in the set and value of N considered, provided an optimal solution—a deep local minimum in the space of CG mappings; the corresponding Σ's spread over a compact range of values that are systematically lower than, and do not overlap with, those of the random case distributions (Fig. 3.2).

Optimal solutions for AKE and AAT span a wide interval of values of Σ; when $N = N_\alpha$ in particular, the support of this set and of the corresponding random reference have comparable sizes. A quantitative measure of this broadness is displayed in the distributions of Z scores of optimal solutions presented in Table 3.3. In both proteins, the Σ's associated to optimal mappings increase with the degree of CG'ing N; this is a consequence of keeping the number of CG configurations of each system (conformational clusters, see Sec. 3.2.1 and Tab. 3.1) constant across different resolutions. As N increases, the available CG conformational clusters are populated by more energetically diverse conformations, thereby incrementing the associated energy fluctuations. On the other hand, TAM shows narrowly peaked distributions of optimal values of Σ, whose position does not vary with the amount of retained sites. Both effects can be ascribed to the fact that most of the energy fluctuations in TAM—and consequently the mapping entropy—are due to a subset of atoms that are almost always maintained in each optimal mapping (see Sec. 3.3.3) in contrast to a random choice of the CG representation. At the same time, the associated Z scores are lower than the ones of the bigger proteins for all values of N under examination, as TAM conformations generally feature a lower variability in energy than the other molecules.

To conclude, let me observe how the values of Σ reported in Fig. 3.2 do not display the expected dependence on the number of CG sites N, that is, mappings with high (resp. low) N corresponding to low (resp. high) Σ. Once again, this is a result of fixing a priori the number of CG macrostates (see

Sec. 3.2.1 and Tab. 3.1) for all values of N, thereby inhibiting any possible scaling. In Sec. 3.4 this limitation is relaxed and the predicted behaviour is finally observed.

3.3.2 Transitions between optimal mappings

For all the investigated proteins, the absence of an overlap between the distributions of Σ associated to random and optimised mappings raises some relevant questions. First, one might wonder what kind of structure the *solution space* has, that is, if the identified solutions lie at the bottom of a rather flat vessel or, on the contrary, each of them is located in a narrow well, neatly separated one form the other.

In order to answer this question, for each structure the four pairs of mapping operators $\mathbf{M}^{opt}$ resulting in the lowest values of Σ are selected. Subsequently, 100 independent transitions between these solutions are performed, constructing intermediate mappings by randomly swapping two non-overlapping atoms from the two solutions at each step and calculating the associated mapping entropy. Fig. 3.3(a-c) shows the results of this analysis for the pair of mappings with the lowest Σ, all the other transitions being reported together in Fig. 3.3(d-f). It is interesting to notice that the endpoints (that is, the optimised mappings) correspond to the lowest values of Σ along each transition path; by increasing the size of the proteins, the values of Σ for intermediate mappings get closer to the average of Σ_{random}. The absence of lower minima over all the possible paths cannot be ruled out, although it seems quite unlikely given the available sampling.

This analysis shows that the deepest solutions of the optimisation procedure are distinct from each other. Hence, it is impossible to (quasi) continuously transform an optimal mapping into another through a series of steps keeping the value of the mapping entropy low.

3.3.3 Properties of optimal mappings

The previous subsection shows how the optimal solutions of the mapping entropy optimisation protocol seem to be very separated from each other *in terms of the same observable* Σ. Now, the second question that arises is if there exists some similarity among these disconnected solutions.

The degree of similarity between the optimal mappings can be assessed by a simple average, returning the frequency P_{cons} with which a given atom is retained in the 48 solutions of the optimisation problem.

Fig. 3.4 shows the value of P_{cons} separately for each analysed protein and degree of coarse-graining N investigated, computed as the fraction of times

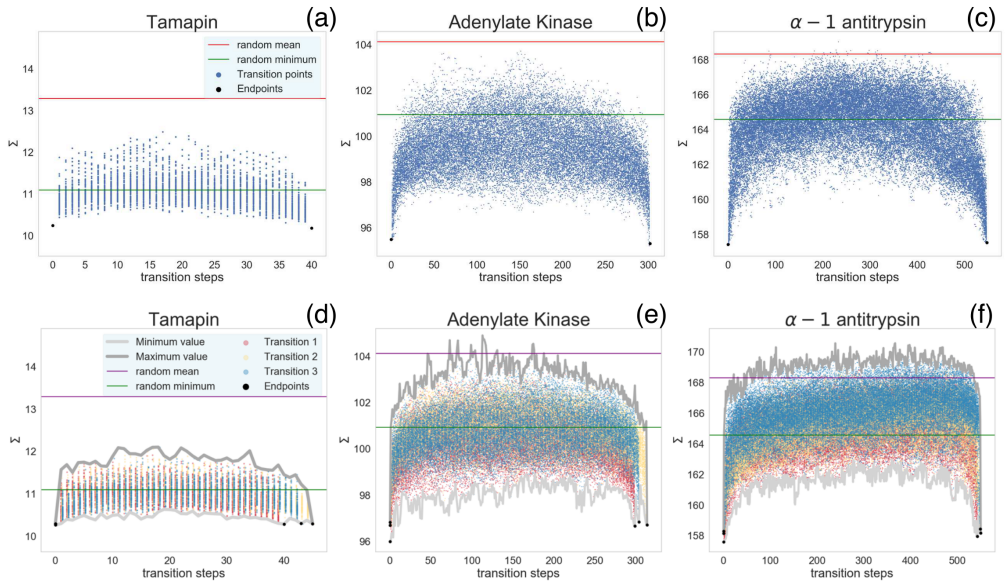


Figure 3.3: (a-c)Values of the mapping entropy Σ $[kJ/\mathrm{mol}/K]$ of mappings connecting two optimal solutions. In each plot, one per protein under examination, the two lowest-Σ mappings are taken as initial and final endpoints (black dots) for paths constructed by swapping pairs of atoms between them (blue dots). For each protein, 100 independent paths at given $N = N_{\alpha\beta}$ are constructed and the mapping entropy of each intermediate point is computed. In each plot, horizontal lines represent the mean (red) and minimum (green) S_{map} obtained from the corresponding distribution of random mappings presented in Fig. 3.2. (d-f) the same analysis performed for the three next-to-lowest-Σ pairs of optimal mappings at $N = N_{\alpha\beta}$. Image adapted from Ref. [24].

it appears in the corresponding pool of optimised solutions. One can notice the presence of regions that appear to be more or less conserved. Quantitative differences can be observed between the three cases under examination: while the heat map of TAM shows narrow and pronounced peaks of conservation probability, optimal solutions for AKE feature a more uniform distribution, where the maxima and minima of P_{cons} extend over secondary structure fragments rather than small sets of atoms. The distribution gets even more blurred for AAT.

As index proximity does not imply spatial proximity in a protein structure, the aforementioned probabilities are mapped on the three-dimensional configurations. Results for TAM are shown in Fig. 3.5, while the corresponding ones for AKE and AAT are displayed in Fig. 3.6. From the distribution

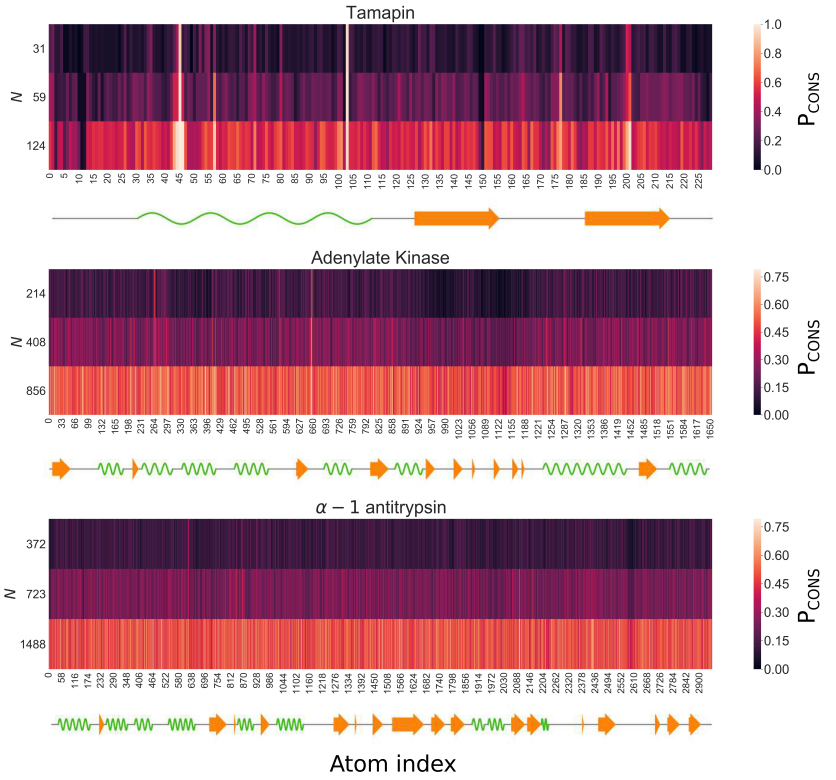


Figure 3.4: Probability P_{cons} that a given atom is retained in the optimal mapping at various numbers N of CG sites and for each analysed protein, expressed as a function of the atom index. Atoms are ordered according to their number in the PDB file. The secondary structure of the proteins is depicted using Biotite [208]: green waves represent alpha helices and orange arrows correspond to beta strands. Image taken from Ref. [24].

of P_{cons} at different number of retained sites N it is possible to infer some relevant properties of optimal mappings.

For what concerns TAM (Fig. 3.5), it seems that, at the highest degree of CG ($N = N_\alpha$), only two sites are always conserved, namely two nitrogen atoms belonging to ARG6 and ARG13 residues ($P_{\text{cons}}(\text{NH1}, \text{ARG6}) = 0.92$, $P_{\text{cons}}(\text{NH2}, \text{ARG13}) = 0.96$). The atoms that constitute the only other arginine residue, ARG7, are well conserved but with lower probability. By increasing the resolution ($N = N_{\alpha\beta}$), i.e., employing more CG sites, it is possible to see that the atoms in the side chain of LYS27 appear to be retained more than average together with atoms of GLU24 ($P_{\text{cons}}(\text{NZ}, \text{LYS27}) = 0.65$, $P_{\text{cons}}(\text{OE2}, \text{GLU24}) = 0.75$). At $N = 124$ the distribution becomes more uniform, but still sharply peaked around terminal atoms of ARG6 and

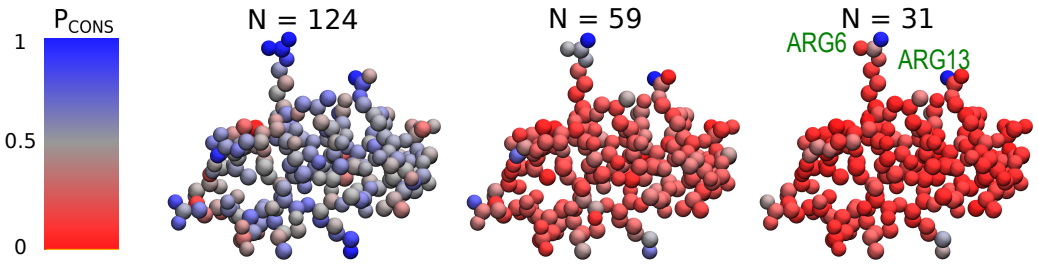


Figure 3.5: Structure of tamapin (one bead per atom) coloured according to the probability P_{cons} for each atom to be retained in the pool of optimal mappings. Each structure corresponds to a different number N of retained CG sites. Residues presenting the highest retainment probability across N (ARG6 and ARG13) are highlighted. Image taken from Ref. [24].

ARG13.

Interestingly, ARG6 and ARG13 have been identified to be the main actors involved in the TAM-SK2 channel interaction [209, 210, 211]: Andreotti *et al.* [209] suggest that these two residues strongly interact with the channel through electrostatics and hydrogen bonding. Furthermore, Ramírez-Cordero *et al.* [211] show that mutating one of the three arginines of TAM dramatically decreases its selectivity towards the SK2 channel.

It thus appears that the mapping entropy minimisation protocol is capable of singling out the two residues that are crucial for a complex biological process. The rationale for this can be found in the fact that such atoms strongly interact with the remainder of the protein, so that small variations of their relative coordinates have a large impact on the value of the overall system's energy. Retaining these atoms, and fixing their position in the coarse-grained conformation, thus enables the model to discriminate effectively a macrostate from another.

Notably, this result is achieved solely relying on data obtained in standard MD simulations. This aspect is particularly relevant as the simulation is performed in absence of the channel, whose size is substantially larger than that of TAM. Consequently, valuable biological information, otherwise obtained *via* large-scale, multi-complex simulations, bioinformatic approaches, or experiments, is here retrieved by means of straightforward simulations of the molecule of interest in absence of its substrate.

As for AKE (Fig. 3.6(a-c)), when $N = N_\alpha$ the external, solvent-exposed part of the LID domain is heavily coarse-grained, while its internal region is more conserved. The CORE region of the protein is always largely retained, without noteworthy peaks in probability. Such peaks, on the contrary, appear in correspondence of some terminal nitrogens of ARG36, LYS57 and ARG88

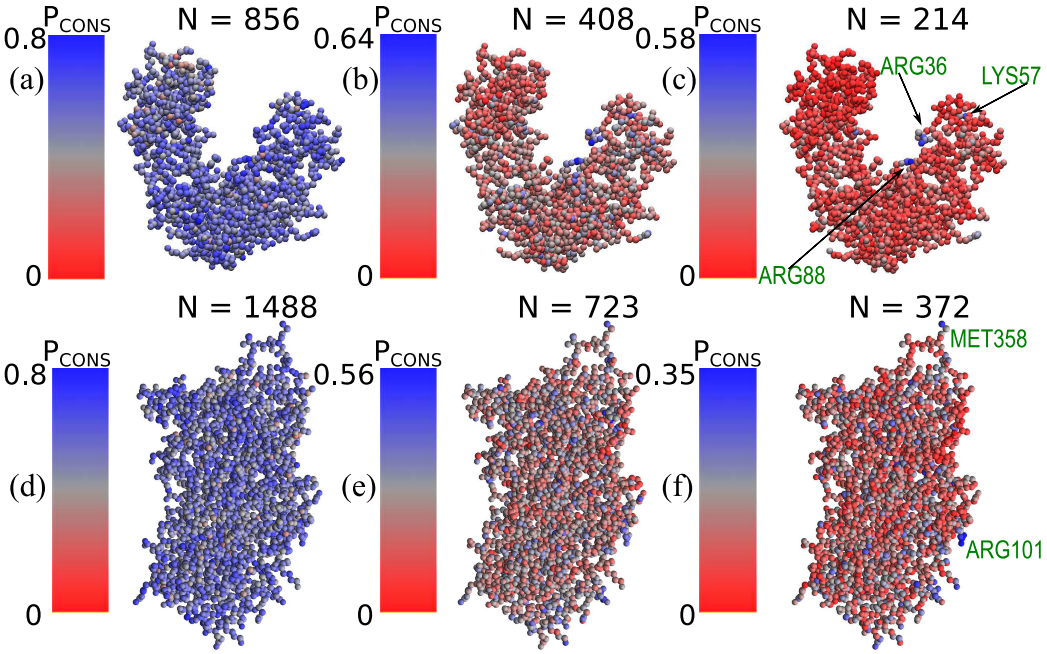


Figure 3.6: Adenylate Kinase [(a),(b) and (c)], and $\alpha - 1$ antitrypsin [(d), (e) and (f)]: probability of conserving sites over the optimised solutions as a function of the number N of retained sites. The residues containing those atoms that are conserved with the highest probability (see text) have been explicitly indicated in figure. Image taken from Ref. [24].

($P_{\rm cons}(\mathrm{NH2, ARG36}) = 0.52$, $P_{\rm cons}(\mathrm{NZ, LYS57}) = 0.48$, $P_{\rm cons}(\mathrm{NH2, ARG88}) = 0.58$). The two arginine amino acids are located in the internal region of the NMP arm, at the interface with the LID domain. ARG88 is known to be the most important residue for catalytic activity [212, 213], being central in the process of phosphoryl transfer [214]. Phenylglyoxal [215], a drug that mutates ARG88 to a glycine, has been shown to substantially hamper the catalytic capacity of the enzyme [214]. ARG36 is also bound to phosphate atoms [213]. Finally, LYS57 lies on the external part of NMP and has been identified to play a pivotal role in collaboration with ARG88 to block the release of adenine from the hydrophobic pocket of the protein [216]. More generally, this amino acid is crucial for stabilising the closed conformation of the kinase [217, 218], which is never observed throughout the simulation. The overall probability pattern persists as N increases, even though less pronounced.

Finally, in the case of AAT, Fig. 3.6(d-f) shows that the associated optimisations heavily coarse-grain the reactive center loop of the protein. On the

other hand, two of the most conserved residues in the pool of optimised mappings, MET358 and ARG101, are central to the biological role of this serpin. MET358 ($P_{\text{cons}}(\text{CE}, \text{MET358}) = 0.31$) constitutes the reactive site of the protein [219]. Being extremely inhibitor-specific, mutations or oxidation of this amino acid lead to severe diseases. In particular, heavy oxidation of MET358 is one of the main causes of emphysema [220]. The AAT *Pittsburgh* variant shows MET358–ARG mutation, which leads to diminished anti-elastase activity but markedly increased antithrombin activity [221, 219, 193]. In turn, ARG101 ($P_{\text{cons}}(\text{CZ}, \text{ARG101}) = P_{\text{cons}}(\text{NH1}, \text{ARG101}) = P_{\text{cons}}(\text{NH2}, \text{ARG101}) = 0.35$) has a crucial role due to its connection to mutations that lead to severe AAT deficiency [195, 194].

In summary, the presented approach identifies biologically relevant amino acids in all the proteins investigated. Most notably, these residues, which are known to be biologically active in presence of other compounds, are singled out *from substrate-free MD simulations*. With the exception of MET358 of AAT, the most probably retained atoms belong to amino acids that are charged and highly solvent-exposed. To quantify the statistical significance of the selection operated by the algorithm, let me note that the latter detects those fragments out of a pool of 8, 69 and 100 charged residues for TAM, AKE and AAT, respectively. If solvent exposition is accounted for, these numbers reduce to 7, 32 and 40 considering amino acids with solvent accessible surface area (SASA) higher than 1 nm^2.

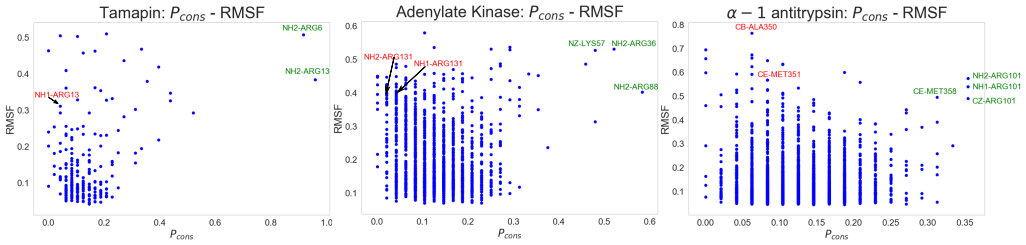


Figure 3.7: Atom-wise comparison between RMSF and P_{cons} (calculated at $N = N_\alpha$) for the three proteins of interest. The most conserved atoms, for which the RMSF is always non negligible, are highlighted in green. Highly mobile atoms that are almost never included in the optimised solution are pinpointed in red. Image taken from Ref. [24].

Another aspect worth mentioning is the fact that several atoms pinpointed as highly conserved in optimal mappings are located in the side chains of relatively large residues, such as arginine, lysine and methionine. It is thus legitimate to wonder whether a correlation might exist between an amino acid size and the probability of one or more of its atoms to be

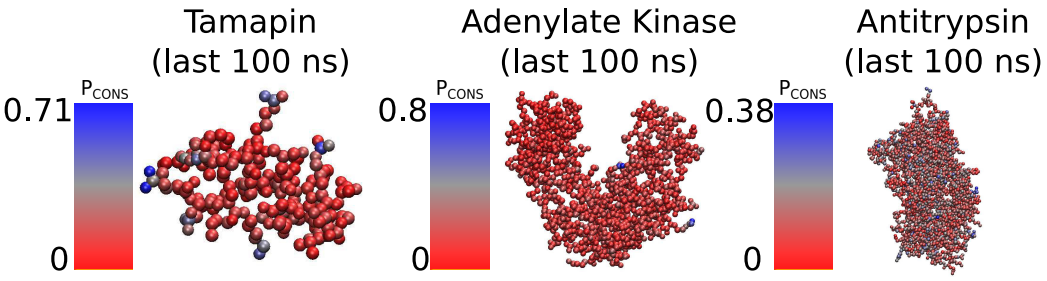


Figure 3.8: P_{cons} of conserving atoms calculated taking into account only the last 100 ns of the MD simulations. A visual comparison with Fig. 3.5 (for TAM) and Fig. 3.6 (for AKE and AAT) can show the differences between the two cases. Image taken from Ref. [24].

present in a low S_{map} reduced representation. An inspection of the RMSF values of the three proteins' atoms *vs.* their conservation probability (see Fig. 3.7) shows no significant correlation for low or intermediate values of P_{cons}; highly conserved atoms, on the other hand, tend to be located on highly mobile residues because a relatively large conformational variability is a prerequisite for an atom to be determinant in the mapping. In conclusion, highly mobile residues are not necessarily highly conserved, while the opposite is more likely.

3.3.4 On the sampling dependence of the protocol

How do the presented results depend on the sampling of the atomistic system? In this subsection I briefly summarise the output of the minimisation of the mapping entropy in all three proteins under examination, taking as sampled structures the configurations extracted from the last 100 ns of MD trajectories. Frames are separated by 10 ps in order to consider 10^4 configurations, as it is done in the previous sections. 5×10^3 of them are already included in the 200 ns of sampling considered before, while the other half of them consists of new, "intermediate" snapshots. The following analysis represents a first assessment of how the results of the protocol illustrated in this chapter depend on the extent of the sampling.

Fig. 3.8 shows the results of the optimisations carried out over the shorter trajectories, restricted to the case $N = N_\alpha$. Regarding the subset of atoms that are more conserved by the optimisation procedure, let me highlight the changes protein by protein:

- [**TAM**]: the terminal atoms in the arginine residues ARG6 and ARG13 are conserved with medium-to-high values of Σ ($P_{\text{cons}}(\text{CZ}, \text{ARG13}) =$

0.58, $P_{\text{cons}}(\text{NH2}, \text{ARG6}) = 0.46$, $P_{\text{cons}}(\text{CZ}, \text{ARG6}) = 0.42$). These values are lower than the ones observed with 200 ns of conformational sampling. Interestingly, the atoms retained with higher probabilities in the terminal regions of these arginine residues are not the ones identified in the main text. Overall, the atoms with highest P_{cons} are the terminal oxygens of GLU24 ($P_{\text{cons}}(\text{OE1}, \text{GLU24}) = 0.71$, $P_{\text{cons}}(\text{OE2}, \text{GLU24}) = 0.69$);

- [**AKE**]: a visual comparison between Fig. 3.6(c) and Fig. 3.8(b) highlights how the external portion of the LID domain is heavily coarse-grained. Looking at specific atoms, ARG88 is retained with values of P_{cons} ($P_{\text{cons}}(\text{NH2}, \text{ARG88}) = 0.79$, $P_{\text{cons}}(\text{CZ}, \text{ARG88}) = 0.65$) even higher than those obtained with 200 ns of sampling. Instead, ARG36 and LYS57 are well conserved but without the peaks in probability described in the previous section. This result may suggest that, in the full simulation, ARG88 is always involved in highly energetic medium-to-large scale rearrangements, while the other two residues play a less prominent role in the last 100 ns;

- [**AAT**]: while the residue ARG101 does not possess atoms with P_{cons} higher than 0.21, MET358 terminal atoms are well conserved throughout the optimised solutions ($P_{\text{cons}}(\text{SG}, \text{MET358}) = P_{\text{cons}}(\text{CE}, \text{MET358}) = 0.25$). In the case of AAT 100 ns of MD sampling seem to be too few to extract relevant information from the trajectory, giving rise to a uniform conservation probability distribution.

The evaluation of the dependence of mapping entropy values on the duration and other features of the employed MD trajectories is a fundamental step to critically assess advantages and limitations of the method. It is reasonable to expect that, as it is the case with any approach that relies on MD simulations as input data, a variation of the latter induces a variation of the results. This variation can be employed to investigate the features of the input, e.g. in the present case different mappings can emerge from trajectories sampling different structural basins.

The results described above suggest that several features of the optimised mappings are retained even when a different set of configurations is employed. The usage of the last 100 ns of the trajectories has showed small variations, coherent with the different duration of the input and the stochastic nature of the optimisation procedure, as well as an overall consistent pattern of results, which demonstrates the solidity of the approach.

3.4 Scaling with the number of coarse-grained sites

From Fig. 3.2 it is possible to observe that the values of Σ do not seem to show a clear dependence on the number of coarse-grained sites N. This is a consequence of fixing the number of macrostates (Table 3.1) for all the degrees of coarse-graining, thus ultimately constraining the mapping entropy to a certain range of values. In Sec. 3.2.1 it is discussed how this choice is forced by the fact that a *distance-based* criterion for clustering is inappropriate for the mapping optimisation procedure, since the algorithm would be strongly biased towards the mapping resulting in the maximum number of macrostates at the chosen distance.

Nevertheless, such distance-based criterion can be combined with a metric that is slightly different from the $\mathrm{RMSD}^{\mathrm{CG}}$ in order to enforce a scaling of the values of $\tilde{S}_{map}$(Eq. 3.41) with the number of coarse-grained sites N. In this context, the unweighted CG Root-Square-Deviation ($\mathrm{RSD}^{\mathrm{CG}}$) is used to quantify the distance between configurations $\mathbf{x}, \mathbf{y}$ in the mapped trajectory:

$$\mathrm{RSD}^{\mathrm{CG}}(\mathbf{M}(\boldsymbol{x}), \mathbf{M}(\boldsymbol{y})) = \min_{\mathcal{RT}} \left[\sqrt{\sum_{I=1}^{N} \left(\mathbf{M}_I(\boldsymbol{x}) - \mathcal{RT}^{\mathrm{CG}} \mathbf{M}_I(\boldsymbol{y}) \right)^2} \right] \tag{3.52}$$

where $\mathcal{RT}^{CG}$ (see Eq. 3.43) is a rigid rototranslation that superimposes $\mathbf{M}(\boldsymbol{x})$ and $\mathbf{M}(\boldsymbol{y})$, and the sum is performed over all retained atoms.

Now it is possible to apply average linkage bottom-up hierarchical clustering (UPGMA [198], see Sec. 3.2.1) with a distance threshold d such that d is the *minimal* number such that all the atomistic configurations can be distinguished. If the optimal rototranslation $\hat{\mathcal{RT}}^{CG}$ solution to Eq. 3.52 does not change when removing atoms, the $\mathrm{RSD}^{\mathrm{CG}}$ becomes an additive property *on the number of retained sites*, therefore giving rise to smaller distances between the sampled points as N decreases. Hence, in a distance-based clustering, separated configurations start to be lumped together as the degree of CG increases, thus producing the scaling observed in Fig. 3.9, where this strategy is applied to a subset of CG mappings of TAM. Notably, Fig. 3.9 highlights how the mapping entropy calculated with the proposed prescription is always low when $N > \frac{N}{2}$. When N decreases, the average value of mapping entropy increases, together with its variance at fixed N: low-resolution CG mappings are more "diverse" than high-resolution ones, giving rise to strong fluctuations in the observable.

As already discussed, employing the aforementioned clustering prescription is not feasible during the optimisation process, as the procedure would

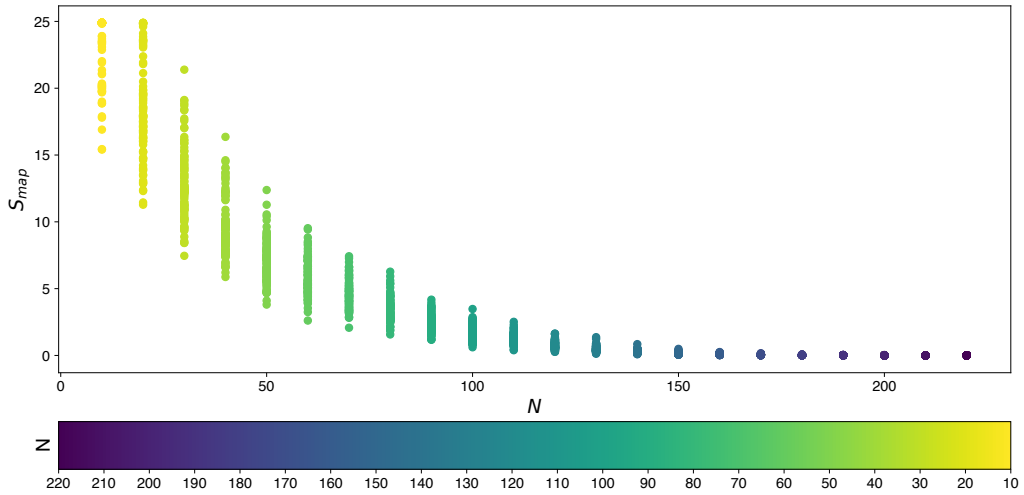


Figure 3.9: Values of mapping entropy calculated using RSD^{CG} (Eq. 3.52) and distance threshold as distance between configurations and clustering criterion, respectively. 100 mappings are extracted randomly for each value of $N \in \{10, 20, ..., 220\}$. It is possible to see that the mapping entropy remains quite low until $N \sim 150$. Then, CG macrostates start to contain more and more energetically diverse configurations, leading to an important increase of S_{map}, which attains its maximal values when $N < N_\alpha$, where the trajectory is divided in few CG clusters. Such scaling is not observed in Fig. 3.2, where RMSD^{CG} and fixed number of clusters are selected as distance metric and clustering criterion, respectively.

tend to generate an increasing number of clusters, as $\tilde{S}_{map}$ would (trivially) vanish when there exists one macrostate for each atomistic configuration.

3.5 Limitations of the method

Let me conclude this long chapter with a brief critical discussion of the current major limitations inherent to the calculation of the mapping entropy:

1. the first thing that must be mentioned is the well-known difficulty to obtain equilibrium configurations from MD trajectories. All the calculations carried out in this chapter are rooted on the idea of possessing a sample of uncorrelated configurations extracted at the equilibrium, which is clearly not the situation one encounters when sampling protein structures every 20 picoseconds, as performed in the current implementation of the method; this is the well-known sampling problem, intrinsic

to bottom-up CG of biomolecules;

2. an important ambiguity revolves around the choice of the energy to be employed in the calculations. In this chapter, the intramolecular potential energy of the protein is employed, completely neglecting solvent degrees of freedom in the calculation. Alternative strategies might involve the incorporation of (a fraction of) of solvent molecules in the calculation of the energies, as well as the usage of an accurate implicit solvent force field, such as Rosetta [222], to properly account for solvent effects. The first section of Chapter 6 presents a different protocol to compute the mapping entropy that eliminates this ambiguity, being solely focused on the probability and without requiring any knowledge on the energy;

3. the single calculation of the mapping entropy is computationally expensive when the number of available configurations exceeds the few thousands. Chapters 4, 6 and 7 present different strategies to reduce the computational burden associated to these computations.

3.6 Conclusions

This chapter presents an information theoretical measure, the mapping entropy, to rank coarse-grained representations of biomolecules. The power of this quantity resides in the fact that both the conformational and the energetic variability of the high-resolution system are accurately taken into account in the calculations.

Lowest-S_{map} CG mappings can be used as the starting ingredient for CG force field development, employing one of the methods outlined in Chapter 1, Sec. 1.2.1. In this context, given a mapping associated to a negligible value of S_{map}, zeroing the relative entropy is tantamount at minimizing the overall Kullback-Leibler distance (S_{tot}, Eq. 3.4) between atomistic and coarse-grained models.

Importantly, optimal CG representations resulting from the mapping entropy minimisation share common features, related to the function of the selected molecule. Specifically, a high level of detail is assigned to those regions that are crucial for the biological role of the protein. Notably, such hotspots are identified by means of substrate-free MD simulations, thus implying that a substantial amount of information about functional residues is encoded in the protein structure and energetics.

This consideration paves the way for a new significance of optimal CG mappings, that can be employed to *filter* the huge amount of data provided

by atomistic simulations into few, informative messages.

Chapter 4

A Deep Graph Network–Enhanced Sampling Approach to Efficiently Explore the Space of Reduced Representations of Proteins

In the previous chapter I described in detail a strategy that aims at finding the maximally informative reduced representations of a biomolecule through the minimisation of the associated mapping entropy S_{map} (Eqs. 3.10, 3.31 and 3.41). In this chapter I illustrate how a combination of a machine learning algorithm and an enhanced sampling method can allow an extremely fast and accurate exploration of the space of CG mappings of a biomolecule. This chapter has to be considered a personal re-elaboration of Ref. [25], which is the main reference.

The method outlined in the last chapter suffers from three bottlenecks (see Sec. 3.5): first, the protocol requires in input a set of configurations of the high-resolution system that are sampled through an MD simulation, a task that is well-known to be expensive; second, the determination of the mapping entropy is *per se* computationally intensive: even though smart workarounds (see Eq. 3.49 for an example) can be conceived and implemented to speed up the calculation, its relative complexity introduces a nontrivial slowdown in the minimisation process. Third, the sheer size of the space of possible CG mappings of a biomolecule is so ridiculously large that it makes a random search practically useless and an exhaustive enumeration simply impossible (see Eq. 3.48). Hence, an optimisation procedure is required to identify the simplified descriptions that entail the largest amount of information about

the system. Unfortunately, this procedure nonetheless implies the calculation of S_{map} over a very large number of tentative mappings, making the optimisation, albeit possible, computationally intensive and time consuming.

Throughout this chapter I present a protocol that suppresses the computing time of the optimisation procedure by several orders of magnitude, while at the same time boosting the sampling accuracy. This strategy relies on the fruitful, and to the best of my knowledge unprecedented combination of two very different techniques: graph-based machine learning models [223, 224, 225] and the Wang-Landau enhanced sampling algorithm [205, 206, 226, 227]. Based on a graph representation of a protein, the first serves the purpose of reducing the computational cost associated with the estimation of the mapping entropy; the second enables the efficient and thorough exploration of the mapping space of a biomolecule.

4.1 Data sets

The machine learning-based mapping entropy prediction model developed in this study is applied to two proteins extracted from the set investigated in Chapter 3, namely the tamapin mutant (PDB code *6d93*) and adenylate kinase (PDB code *4ake*).

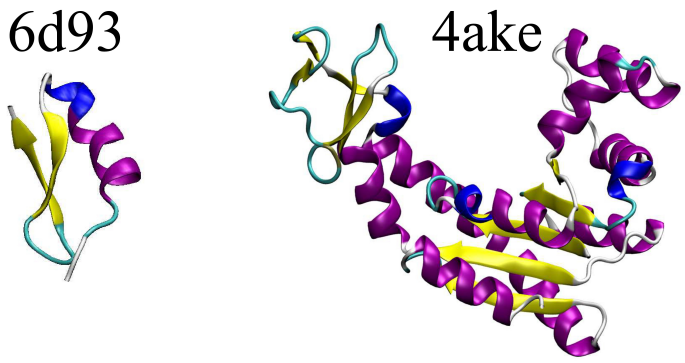


Figure 4.1: Protein structures employed in this work: the tamapin mutant (PDB code: 6d93) and the open conformation of adenylate kinase (PDB code: 4ake). The former, although small, possesses all the elements of proteins' secondary structures, while the latter is bigger in size and has a much wider structural variability. Image taken from Ref. [25].

The machine learning model of each protein is trained on a data set containing the molecular structure—the first snapshot of the MD trajectory—and many CG representations, the latter being selected with the constraint

of having a number of retained sites equal to the number of amino acids composing the molecule ($N = N_\alpha$, see Tab. 3.2). The data sets combine together randomly selected CG mappings (respectively 4200 for *6d93* and 1200 for *4ake*) and optimised ones (768 for both systems). The mapping entropy values are calculated using the approximate expression of Eq. 3.47[1], and the optimised representations are obtained through the Monte Carlo Simulated Annealing optimisation protocol described in Sec. 3.3. More specifically, the 768 SA runs of each protein are divided in four groups of 192 elements depending on their length, respectively 2×10^4 (full optimisation, as in Sec. 3.3 and in Ref. [24]), 1×10^4, 5×10^3 and 2.5×10^3 steps.

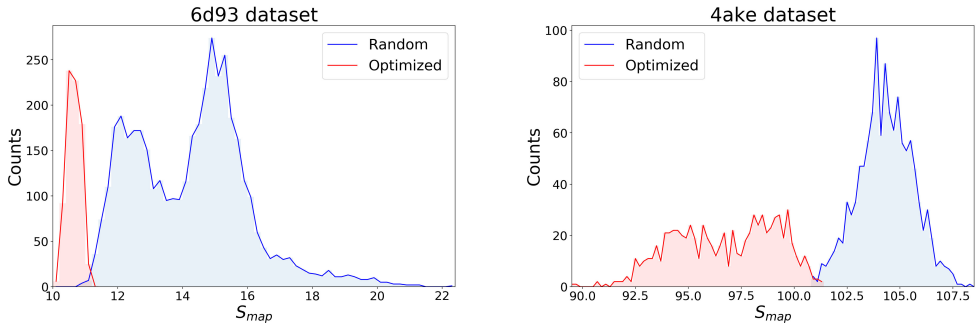


Figure 4.2: Distributions of target values for both data sets, *6d93* (left) and *4ake* (right). For each protein, S_{map} data are displayed in two distinct, non-overlapping histograms depending on their origin: blue curves are filled with random instances, while red histograms represent optimised CG mappings. All values of S_{map} are in $[kJ/\text{mol}/K]$. Image taken from Ref. [25].

Fig. 4.2 displays the distribution of S_{map} values in the data sets separately for the two systems, discriminating between random (blue) and optimised (red) CG mappings. In both structures the two curves have a negligible overlap, meaning that the set of values spanned by the optimised CG representations cannot be reached by a random exploration of the mapping space, i.e, this region possesses a very low statistical weight. A comparison of the S_{map} distribution of the two proteins, on the other hand, highlights that the mapping entropy increases with the system's size: while the range of values covered has similar width in the two cases, the lower bound in mapping entropy of *4ake* differs of roughly one order magnitude from that of *6d93*.

For each analysed protein, the computational time required to perform

[1]The numerical values reported in this chapter refer to the average mapping entropy Σ (Eq. 3.47), as explained in Sec. 3.2.1. For the sake of clarity, this observable is here denoted with S_{map}.

Protein	*MD CPU time*	*MD walltime*	*Single measure*
6d93	40.7 days	2.55 days	$\simeq$ 2.1 mins
4ake	153.9 days	3.20 days	$\simeq$ 8.0 mins

Table 4.1: Computational cost of all-atom MD simulations and mapping entropy calculations for the two investigated proteins. Specifically, *MD CPU time* (resp. *MD walltime*) represents the core time (resp. user time) necessary to simulate the system for 200 ns on the GROMACS 2018 package [181]. Both *6d93* and *4ake* runs were performed on Intel Xeon-Gold 5118 processors, respectively using 16 and 48 cores. *Single measure* is the amount of time that is required to compute, on a single core of the same architecture, the S_{map} of a given CG mapping by relying on the algorithm introduced in Sec. 3.2 (Fig. 3.1).

the MD simulation and a single S_{map} estimate is reported in Table 4.1. The time associated with the calculation of S_{map} for a single CG mapping through the algorithm discussed in Sec. 3.2 grows from 2 to 8 minutes while moving from *6d93* to *4ake*. It is worth stressing that the proteins studied here are small, so that this value would dramatically increase in the case of bigger biomolecules.

4.2 Data Representation and Machine Learning model

Part of this section has been written by my collaborators Federico Errica, Davide Bacciu, and Alessio Micheli, whom I here acknowledge for their crucial contribution to this chapter.

With their long and successful story both in the field of coarse-graining [228, 229, 230] and in the prediction of protein properties [231, 232, 233], graph-based learning models represent a rather natural and common choice to encode the (static) features of a molecular structure.

A graph g can be formally defined as a tuple $(\mathcal{V}_g, \mathcal{E}_g)$, where $\mathcal{V}_g$ is the set of vertices (i.e., the entities of interest) and $\mathcal{E}_g = \{\{u, v\} \mid u, v \in \mathcal{V}_g\}$ is the set of undirected edges (i.e., how entities are related). The neighborhood of a vertex v is defined as the set of vertices connected to v by an edge, that is, $\mathcal{N}_v = \{u \in \mathcal{V}_g \mid \{u, v\} \in \mathcal{E}_g\}$. For the purpose of this chapter, each heavy atom composing the molecule corresponds to a vertex, and edges connect pairs of atoms that in the reference structure are closer than a selected threshold—in this case, 1 nm. Information about the decimation mapping can be directly encoded in the vertices of the protein's graph by using a binary feature, with

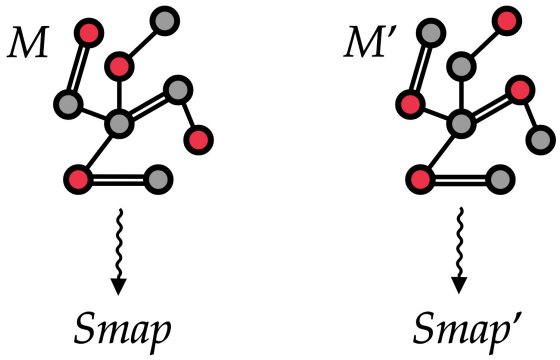


Figure 4.3: Two different mappings M and M' associated with the same (schematic) protein structure. Each protein is treated as a graph where vertices are atoms, and edges are placed among atoms closer than a given threshold. The selected CG sites in each of the two mappings are marked in red and encoded as a vertex feature. The goal of this chapter is to automatically learn to associate both mappings to the correct values S_{map} and $S_{map}{}'$ of mapping entropy. Image taken from Ref. [25].

different selections of CG sites—an example being provided in Fig. 4.3—corresponding to different values of S_{map}. In addition, each vertex is enriched with 10 features, summarised in Tab. 4.2, describing the physico-chemical properties of the underlying atom; similarly, the inverse atomic distance e_{uv} between vertices u and v is employed as an edge feature.

Once the protein structure and the CG mapping data sets (see Sec. 4.1) are converted into this graph-like format (statistics in Table 4.3), Deep Graph Networks (DGNs) are employed [225] with the aim of learning the desired property, namely the mapping entropy S_{map}.

The main advantages of DGNs are their efficiency and the ability to learn from graphs of different size and shape. This is possible for two reasons: first, DGNs focus on a local processing of vertex neighbors, so that calculations can be easily distributed; secondly, in a way that is similar to Convolutional Neural Networks for images [234], DGNs stack multiple layers of graph convolutions to guarantee an efficient exchange of information between vertices. The output of a DGN is a vector for each vertex of the graph, as sketched in Fig. 4.4, and these can be aggregated to make predictions about a graph class or property. The computational efficiency of the DGN is especially important in this context, where the goal is to approximate the complex calculation of S_{map} in a fraction of the time originally required.

Feature name	Description
C	Carbon atom
N	Nitrogen atom
O	Oxygen atom
S	Sulphur atom
HPhob	Part of a hydrophobic residue
Amph	Part of a amphipathic residue
Pol	Part of a polar residue
Ch	Part of a charged residue
Bkb	Part of the protein backbone
Site	Atom selected as a CG site

Table 4.2: Binary features (0/1) used to describe the physico-chemical properties of an atom in the protein, i.e. a vertex in the graph representation of the latter. In this simple model, the DGN is only provided with the chemical nature of the atom and of its residue, together with the flag *Bkb* that specifies if the atom is part of the backbone of the polypeptide chain.

Protein	Vertices	Edges	Avg. Degree	Samples
6d93	230	21474	93	4968
4ake	1656	207618	125	1968

Table 4.3: Basic statistics of the data sets fed to the machine learning model. For each protein, the table reports the number of vertices (i.e., heavy atoms) in its graph representation, the total number of edges connecting them, and the average number of edges per vertex (Avg. Degree). The total number of CG representations of known mapping entropy provided in input to the protocol (Samples) is included, including random and optimised ones.

The main building block of a DGN is the "graph convolution" mechanism. At each layer ℓ, the DGN calculates the new state of each vertex v, i.e., a vector $\mathbf{h}_v^{\ell+1} \in \mathbb{R}^K$, as a function of v's neighboring states $\mathbf{h}_{\mathcal{N}_v}^{\ell} = \{\mathbf{h}_u^{\ell} \in \mathbb{R}^K \mid u \in \mathcal{N}_v\}$, where $K \in \mathbb{N}$ is an hyper-parameter of the model.

In general, a graph convolutional layer first applies a permutation invariant function to the neighbors of each vertex, such as the sum or mean. The resulting aggregated vector is then passed to a multi-layer perceptron (MLP) that performs a non-linear transformation of the input, thus producing the new vertex state $\mathbf{h}_v^{\ell+1}$. The graph convolutional layer can be formalised as

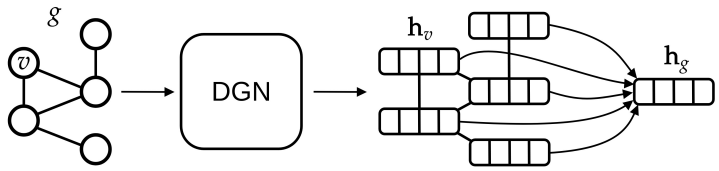


Figure 4.4: High-level overview of typical deep learning methodologies for graphs. A graph g is given as input to a Deep Graph Network, which outputs one vector, also called embedding or state, for each vertex v of the graph. All vertex states are aggregated via a differentiable permutation-invariant operator, i.e., the mean, to obtain a single value that encodes the whole graph structure. Then, the graph embedding is fed into a machine learning regression model (a linear model in this case), to output the S_{map} value associated with g. Image taken from Ref. [25].

follows:

$$\mathbf{h}_v^{\ell+1} = \mathrm{MLP}^{\ell}\Big(\big(1+\epsilon^{\ell}\big) * \mathbf{h}_v^{\ell} + \sum_{u \in \mathcal{N}_v} \mathbf{h}_u^{\ell} * e_{uv},\Big) \tag{4.1}$$

where $*$ denotes element-wise scalar multiplication, $\epsilon^{\ell} \in \mathbb{R}$ is an adaptive weight of the model, and e_{uv} is the scalar edge feature holding the inverse atomic distance between two atoms u and v. A pictorial representation of the transition between layer ℓ and layer $\ell+1$ is presented in Fig. 4.5.

A few remarks about Eq. 4.1 are in order. First, the initial layer is implemented with a simple non-linear transformation of the vertex features, i.e., $\mathbf{h}_v^1 = \mathrm{MLP}^1(\mathbf{x}_v)$, where $\mathbf{x}_v$ is the vector of 10 features associated to each node (see Tab. 4.2); secondly, at each layer ℓ, the *same* non-linear transformation MLP^{ℓ} is applied to all nodes (i.e., a graph traversal), thus allowing to handle graphs with variable size. Finally, the MLP weights are not shared across different layers, meaning that a different MLP is trained for each layer. It is worth noting that this weight sharing scheme at each layer resembles the one employed in Convolutional Neural Networks, where the same adaptive filter is applied to all the pixels in an image.

When building a deep graph network, $L \in \mathbb{N}$ graph convolutional layers are stacked, until the model produces a final state for each vertex, $\mathbf{h}_v$; in addition, a global graph state $\mathbf{h}_g$ is computed by aggregating all vertex states (see Fig. 4.4). Being in vectorial form, $\mathbf{h}_g$ can then be fed to standard machine learning models to solve graph regression or classification tasks.

To produce a prediction $\hat{S}_{map}$ one first needs to process and aggregate all node states into a single graph representation. Here, the importance of

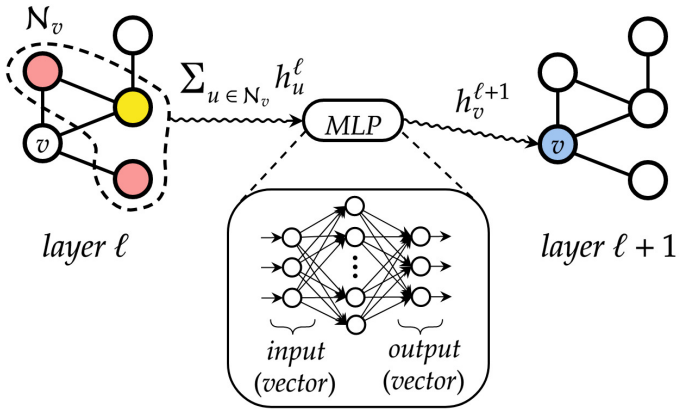


Figure 4.5: A simplified representation of how a graph convolutional layer works. First, neighboring states of each vertex v are aggregated by means of a permutation invariant function, to abstract from the ordering of the nodes and to deal with variable-sized graphs. Then, the resulting vector is fed into a multi-layer perceptron that outputs the new state for node v. Image taken from Ref. [25].

selected (resp. unselected) CG sites $\mathcal{V}_g^s \subset \mathcal{V}_g$ (resp. $\mathcal{V}_g^n$) is taken into account by means of a scalar adaptive weight w_s (resp. w_n). The resulting formula is:

$$\hat{S}_{map} = \mathbf{w}_{out}^T \Big(\sum_{u \in \mathcal{V}_g^s} \big([\mathbf{h}_u^1, \ldots, \mathbf{h}_u^L] * w_s \big) + \sum_{u \in \mathcal{V}_g^n} \big([\mathbf{h}_u^1, \ldots, \mathbf{h}_u^L] * w_n \big) \Big), \quad (4.2)$$

where $\mathbf{w}_{out} \in \mathbb{R}^{K*L}$ is a set of parameters to be learned, while square brackets denote concatenation of the different vertex states computed at different layers. More specifically, the models possess $L = 5$ layers and each MLP^ℓ is implemented as a one-layer feed-forward network with $K = 64$ hidden units followed by an element-wise Rectifier linear unit (ReLU) activation function [235]. As the number of weights, without considering the bias, of MLP^ℓ is K^2 ($10 * K$ for MLP^1), the total number of weights in the architecture is $10 * K + K^2 * (L-1) + (L * K) + (L-1) + 2 = 17350$.

The loss objective used to train the DGN is the Mean Absolute Error (MAE). The optimisation algorithm is Adam [236] with a learning rate of 0.001 and no regularisation. The models are trained for a maximum of 10000 epochs with an early stopping patience of 1000 epochs and a mini-batch size equal to 8, accelerating the training using a Tesla V100 GPU with 16 GB of memory.

To assess the performance of the model on a single protein, the corresponding data set is first divided into training, validation and test realisa-

tions following an 80%/10%/10% hold-out strategy. Early stopping [237] is applied to pinpoint the training epoch with the best validation score, and the resulting, selected model is evaluated on the unseen test set. The evaluation metric for the regression problem is the coefficient of determination (or R^2-score), given by:

$$R^2 = 1 - \frac{\sum_M \left(S_{map}(M) - \hat{S}_{map}(M) \right)^2}{\sum_M \left(S_{map}(M) - \overline{S_{map}} \right)^2}, \tag{4.3}$$

where the sums run over the whole data set of mappings and $\overline{S_{map}}$ is the average mapping entropy.

4.2.1 Results

Table 4.4 reports the R^2 score (Eq. 4.3) and MAE in training, validation and test for *6d93* and *4ake*. It is possible to observe that the machine learning model can fit the training set and has excellent performances on the test set. More quantitatively, extremely low values of MAE are obtained for *6d93*, with an R^2 score higher than 0.95 in all cases. The model performs slightly worse in the case of *4ake*: the result of $R^2 = 0.84$ on the test set is still acceptable, although the gap with the training set ($R^2 = 0.92$) is non-negligible.

Protein	TR MAE	TR R^2	VL MAE	VL R^2	TE MAE	TE R^2
6d93	0.13	0.99	0.33	0.95	0.33	0.96
4ake	0.91	0.92	1.2	0.85	1.35	0.84

Table 4.4: Results of the machine learning model in predicting the mapping entropy on the training (TR), validation (VL) and test (TE) sets for the two analysed proteins. The models are evaluated using both the R^2 score (Eq. 4.3) and the mean average error (MAE, $[kJ/\text{mol}/K]$). The R^2 scores range from $-\infty$ (worst predictor) to 1 (best predictor).

Fig. 4.6 shows how predicted values for training and test samples differ from the ground truth. Ideally, a perfect result corresponds to the points being on the diagonal dotted line, and both training and test predictions for *6d93* are extremely close to the true target. The deviation from the ideal case becomes wider for *4ake*, but no significant outlier is present. A more detailed inspection of the *4ake* scatter plot in Fig. 4.6, on the other hand, reveals that the network tends to slightly overestimate the value of S_{map} of optimized CG mappings for $S_{map} \lesssim 100\ kJ/\text{mol}/K$, whereas the opposite

is true for $S_{map} \gtrsim 100\ kJ/\text{mol}/K$, where random CG mappings' values are mildly underestimated.

The dissimilarity in performance between the two data sets is not surprising if one takes a closer look at their nature. In fact, as highlighted in Fig. 4.1, adenylate kinase is both larger and more complex than the tamapin mutant, and the CG mapping data sets' sizes are very different due to the heavy computational requirements associated with the collection of annotated samples for *4ake*. As a consequence, training a model for *4ake* with excellent generalisation performance becomes a harder tasks. What is remarkable, though, is the ability of a completely adaptive machine learning methodology to well approximate, in both structures, the long and computationally intensive algorithm for estimating S_{map} of Chapter 3. Critically, this is achieved only relying on a combination of static structural information and few vertex attributes, that is, in absence of a direct knowledge for the DGNs of the complex dynamical behaviour of the two systems as obtained by onerous MD simulations.

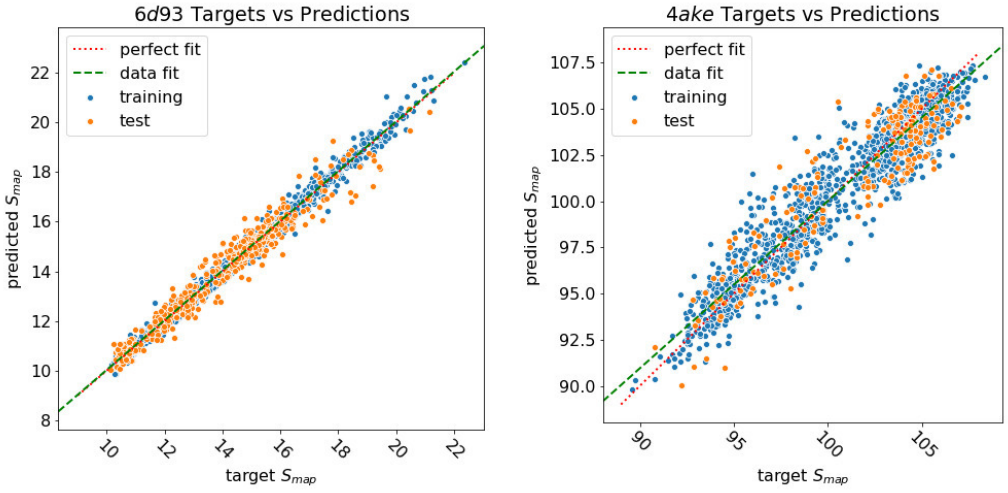


Figure 4.6: Plot of S_{map} target values against predictions of all samples for *6d93* (left) and *4ake* (right). Training samples are in blue, while test samples are in orange. A perfect prediction is represented by points lying on the red dotted diagonal line (perfect fit). A green dashed line, obtained by fitting a linear model on the data, is introduced to show that in the case of *4ake* the model slightly overestimates the S_{map} of optimized mappings and underestimates the rest. All values of S_{map} are in $[kJ/\text{mol}/K]$. Image taken from Ref. [25].

In Tab. 4.5 the computational time required by the machine learning

model to perform a single S_{map} calculation is compared to the one of the algorithm presented in Chapter 3. Overall, employing the machine learning model for inference guarantees a substantial speed-up of mapping entropy calculations, reaching the factor of 5 orders of magnitude when using a GPU machine. Noteworthy, these improvements do not come at the cost of a significantly worse performance of the machine learning model. In addition, this methodology is easily applicable to other kinds of molecular structures, as long as a sufficiently large training set is provided as input.

Protein	Single measure	Inference GPU (CPU)	Time Ratio GPU (CPU)
6d93	$\simeq$ 2.1 mins	$\simeq$ 0.9(98.7) ms	$\simeq 140000 \times (1276\times)$
4ake	$\simeq$ 8.0 mins	$\simeq$ 4.8(1103.2) ms	$\simeq 100000 \times (435\times)$

Table 4.5: Comparison between the time required to compute the S_{map} of a single CG mapping through the algorithm presented in Chapter 3 (CPU only) and the inference time of the model (CPU as well as GPU). For both proteins, CPU calculations where performed on a single core of a Intel Xeon-Gold 5118 processor, while GPU ones were run a Tesla P100 with 16 GB of memory. The machine learning model generates a drastic speed-up, enabling a wider exploration of the S_{map} landscape of each system.

4.3 Wang-Landau Sampling

The computational speed-up guaranteed by the DGN in the calculation of the mapping entropy (see Tab. 4.5) allows one to employ the trained network as a tool to identify optimal representations of the proteins, as in Chapter 3, Sec. 3.3, and, more generally, to quickly explore the immense space of reduced CG representations.

However, Fig. 4.2 highlights how an attempt to detect the most informative CG representations of a protein—i.e., those minimising S_{map}—through a completely unbiased exploration of its mapping space would prove extremely inefficient, if not practically pointless. Indeed, such optimised CG representations live relatively far away in the left tails of the S_{map} distributions obtained from random sampling, thus constituting a region of exponentially vanishing size within the broad mapping space. It would then be desirable to embed the trained network in a sampling strategy in which no specific value of S_{map} is preferred, but rather a *uniform coverage* of the spectra of possible mapping entropies—or at least of a subset of it, vide infra—is achieved.

To obtain this "flattening" of the S_{map} landscape the algorithm proposed by Wang and Landau (WL) [205, 206, 226, 227] is exploited. In WL sampling, a Markov chain Monte Carlo (MC) simulation is constructed in which a transition between two states M and M'—in this case, two mappings containing N sites but differing in the retainment of one atom—is accepted with probability

$$W(M \to M') = \min\left[1, \frac{\Omega_N(S_{map}(M))}{\Omega_N(S_{map}(M'))}\right]. \tag{4.4}$$

In Eq. 4.4, $\Omega_N(S_{map})$ is the number of CG representations with N retained sites exhibiting a mapping entropy equal to S_{map}, that is, the mapping entropy density of states,

$$\Omega_N(S_{map}) = \sum_M \delta(N(M), N)\delta(S_{map}(M), S_{map}), \tag{4.5}$$

where the sum is performed over all possible CG representations with N sites of the system.

When compounded with a symmetric proposal probability T for the attempted move, $T(M \to M') = T(M' \to M)$, the Markov chain defined in Eq. 4.4 generates, at convergence, CG representations distributed according to $P(M) \propto 1/\Omega_N(S_{map}(M))$ [205, 206]. As the equilibrium probability of visiting a mapping is proportional to the inverse of the S_{map}'s density of states, the WL simulation results in a flat histogram of sampled mapping entropies *over the whole range of possible ones*.

Critically, the density of states $\Omega_N(S_{map})$ is a priori unknown and is itself a byproduct of the WL scheme. $\Omega_N(S_{map})$ is self-consistently constructed by means of a sequence $k = 0, ..., K$ of nonequilibrium simulations that provide increasingly accurate approximations to the exact result, iterations being stopped when a predefined precision is achieved.

Having divided the range of possible values of the mapping entropy in bins of width δS_{map}, the WL self-consistent protocol is based on three quantities: the overall density of states $\Omega_N(S_{map})$, the histogram of sampled mapping entropies at iteration k, $H_k(S_{map})$, and the modification factor f_k governing convergence—for $k = 0$, one typically initialises $\Omega_N(S_{map}) = 1$ for each value of S_{map} and $f_0 = e$.

At the beginning of WL iteration k, the histogram $H_k(S_{map})$ is reset. Subsequently, a sequence of MC moves among CG mappings driven by the acceptance probability presented in Eq. 4.4 is performed. If a transition between two CG representations M and M'—respectively with mapping entropies S_{map} and S'_{map} predicted by the trained DGNs—is accepted, the entries of

Parameter	*6d93*	*4ake*
$\ln(f_{end})$	10^{-6}	10^{-6}
p_{flat}	0.8	0.8
range	$[10 - 22.4]$	$[89.4 - 108.6]$
δS_{map}	0.2	0.2

Table 4.6: Set of parameters employed for the WL exploration of the mapping entropy space for both analysed proteins. $\ln(f_0)$ and $\ln(f_{end})$ respectively represent the modification factor at the beginning and at the end of the self-consistent scheme in a logarithmic setup, see Sec. 4.3. p_{flat} is the minimal histogram flatness required to halve the modification factor: with $p_{flat} = 0.8$ all bins in the histogram $H(S_{map})$ must have a population between 0.8 and 1.2 times its average $\langle H \rangle$. *range* is the interval of permitted values of the mapping entropy in the WL scheme, while δS_{map} is the bin size employed for its discretisation. Both *range* and δS_{map} are expressed in $[kJ/\mathrm{mol}/K]$.

the histogram and density of states are updated according to

$$H_k(S'_{map}) = H_k(S'_{map}) + 1, \tag{4.6}$$

$$\Omega_N(S'_{map}) = f_k \times \Omega_N(S'_{map}). \tag{4.7}$$

In case the move $M \to M'$ is rejected, one has to replace S'_{map} with S_{map} in Eqs. 4.6 and 4.7.

The sequence of MC moves is stopped—that is, iteration k ends—when $H_k(S_{map})$ is "flat", meaning that each of its entries does not exceed a threshold distance from the average histogram $\langle H_k \rangle$: a typical requirement is $p_{flat} \times \langle H_k \rangle < H_k(S_{map}) < (2 - p_{flat}) \times \langle H_k \rangle$ for every value of S_{map}, p_{flat} being the selected flatness parameter. At this stage, WL iteration $k+1$ begins with a reduced modification factor, namely $f_{k+1} = \sqrt{f_k}$.

Convergence of the self-consistent scheme is achieved when $f_k \approx 1$—more precisely, when $\ln(f_k)$ becomes smaller than a predefined value $\ln(f_{end})$. Up to a global multiplicative factor, the resulting density of states $\Omega_N(S_{map})$ reproduces the exact result with an accuracy of order f_{end} [238].

In order to avoid numeric overflow of $\Omega_N(S_{map})$ along the WL simulation it is useful to consider its logarithm $\Xi_N(S_{map}) = \ln \Omega_N(S_{map})$. Starting from Eq. 4.4, the acceptance probability $W(M \to M')$ expressed in terms of Ξ reads

$$W(M \to M') = \min\left[1, \exp(\Xi_N(M) - \Xi_N(M'))\right], \tag{4.8}$$

while within iteration k of the self-consistent scheme, the update prescription of Ξ after an (accepted) MC move—see Eq. 4.7—becomes

$$\Xi_N(S'_{map}) = \Xi_N(S'_{map}) + \ln(f_k). \tag{4.9}$$

Finally, in a logarithmic setup the modification factor $\ln(f_k)$ follows the simple reduction rule $\ln(f_{k+1}) = \ln(f_k)/2$, with $\ln(f_0) = 1$.

The WL algorithm in principle enables the reconstruction of the density of states of an observable over the whole range of possible values of the latter; at the same time, knowledge of the sampling boundaries proves extremely beneficial to the accuracy and rate of convergence of the self-consistent scheme [239, 240]. In this case, for each analysed protein, such boundaries would correspond to the minimum and maximum achievable mapping entropies S_{map}^{min} and S_{map}^{max} in the space of all CG representations of the system obtained by retaining N of its constituent atoms. As this information is *a priori* unknown, in the implementation of the WL algorithm the range of explorable values of S_{map} is limited by rejecting all MC moves $M \to M'$ for which $S'_{map} < S_{map}^{min}$ or $S'_{map} > S_{map}^{max}$, in each system setting S_{map}^{min} and S_{map}^{max} as respectively the minimum and maximum values of the mapping entropy in the corresponding data set. Note that for each protein S_{map}^{min} is the outcome of a thorough optimisation procedure (see Sec. 3.3), and can thus be considered a reasonable approximation of the system's *absolute* minimum of the mapping entropy. Imposing an upper bound on S_{map} through S_{map}^{max}, on the other hand, simply amounts at requiring the WL sampling algorithm not to visit uninteresting regions of the mapping space of each biomolecule, that is, CG representations characterised by a huge amount of information loss with respect to the all-atom reference. The values of S_{map}^{min} and S_{map}^{max} employed for the two proteins investigated in this work is presented in Table 4.6, together with the input parameters required by the WL protocol—the bin size δS_{map}, the convergence modification factor $\ln(f_{end})$ and the flatness parameter p_{flat}.

4.3.1 Results

By embedding the trained networks in a Wang-Landau sampling scheme it is possible to retrieve the density of states $\Omega_N(S_{map})$ defined in Eq. 4.5 for *6d93* and *4ake*, that is, one can estimate the number of CG representations throughout the mapping space of each protein that exhibit a specific amount of information loss with respect to the all-atom reference. Let me stress that reaching convergence of the self-consistent WL protocol required to probe approximately 4.8×10^6 and 3×10^7 CG representations for *6d93* and *4ake*, respectively: such an extensive sampling is only made feasible by the computational gain provided by the trained machine learning model (see Tab. 4.5).

WL predictions for the logarithm of the density of states $\Xi_N(S_{map}) = \ln \Omega_N(S_{map})$ of the two proteins are presented in Fig. 4.7. As for *6d93*, I observe the presence of a steep increase of Ξ starting from low values of the mapping entropy, followed by two main peaks respectively located at

$S_{map} \approx 12.5$ and 15 $kJ/\text{mol}/K$. After the second peak Ξ decreases exhibiting a shoulder for high mapping entropies. On the other hand, the Ξ of *4ake* displays a relatively gradual growth towards its unique maximum, the latter being located at $S_{map} \approx 105$ $kJ/\text{mol}/K$, before starting to decrease.

Given the WL $\Omega_N(S_{map})$—or equivalently $\Xi_N(S_{map})$—it is possible to calculate the probability $P(S_{map})$ of observing a particular mapping entropy by performing a completely random exploration of the space of CG representations of a system,

$$P(S_{map}) = \frac{\Omega_N(S_{map})}{\sum_{S_{map}} \Omega_N(S_{map})}. \tag{4.10}$$

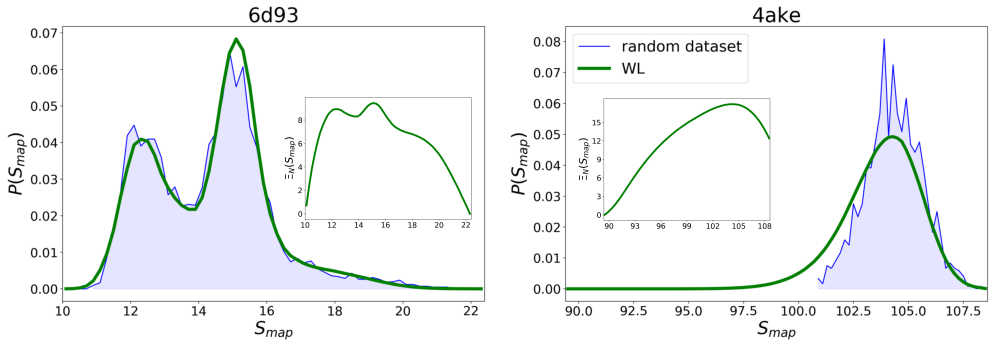


Figure 4.7: Comparison between the probability densities $P(S_{map})$ for the two systems estimated via the Wang-Landau algorithm enhanced by the DGNs (green lines) and the distributions generated by a random sampling of mappings (blue areas). In inset the logarithm of the WL density of states, $\Xi(S_{map})$, is reported, after a scaling that assigns to the Ξ of the most scarcely populated bin the value of zero. All values of S_{map} are in $[kJ/\text{mol}/K]$. Image adapted from Ref. [25].

Results for the $P(S_{map})$ of *6d93* and *4ake* are shown in Fig. 4.7. In the case of *6d93*, let me note that the WL sampling scheme produces a probability density that is fully compatible with the (normalised) histograms of Fig. 4.2. In particular, the WL graph resembles the histograms in Fig. 4.2 if the non-random, optimized instances are removed, whose statistical weight is negligible. This result is highly non-trivial, as it proves that the trained DGN of *6d93* does not overfit the training set and is able to predict the correct population of the true mapping entropy landscape.

As for *4ake*, the agreement between the two curves presented in Fig. 4.7 is still remarkable, though not as precise as in the case of *6d93*. More quantitatively, the left tail of the probability density predicted by the WL scheme

is shifted of roughly 1 $kJ/$ mol$/K$ towards lower values of S_{map} with respect to the distribution obtained from random sampling. This mismatch can be ascribed to the mild overfitting problem observed in Fig. 4.6: the network has the tendency to underestimate (resp. overestimate) the value of S_{map} associated with random (resp. optimised) CG representations, resulting in an increase in the predicted population of mappings at the intersection of the two sets.

The data sets employed for this study and the code that performs the Wang-Landau-based exploration of the mapping space are freely available at https://github.com/CIML-VARIAMOLS/GRAWL.

4.4 Conclusions

The main outcome of this chapter is a strategy that enables an extensive exploration of the space of possible CG mappings of a biomolecule. The combination of trained networks and Wang-Landau sampling allows one to characterise the mapping entropy landscape of a system with impressive accuracy.

The natural following step would be to apply the knowledge acquired by the model on different protein structures, so that the network can predict values of S_{map} even in the absence of an MD simulation. As of now, however, it is difficult to assess if the information extracted from the training over a given protein's trajectory can be fruitfully employed to determine the mapping entropy of another, just feeding the structure of the latter as input. More likely one would have to resort to database-wide investigations, training the network over a large variety of different molecular structures before attempting predictions over new data points. In other words, obtaining a transfer effect among different structures by the learning model may not be straightforward, and additional information could be needed to achieve it.

In conclusion, it is important to emphasise the complete generality of the proposed approach: first, the specific nature and properties of the mapping entropy played no special role in the construction of the deep learning scheme; second, the DGN formalism enables one to input graphs of variable size and shape, allowing the method to be transferred to other problems where different selections of a subset of the molecule's atoms give rise to different values of a given observable, such as the example of Ref. [158]. In principle, the combination of WL sampling and DGNs can be leveraged to recover the correct density of states of any observable: provided that the network does not overfit the possibly biased training set, this is a highly efficient strategy to reconstruct the original, unbiased distribution.

Chapter 5

A journey through mapping space

The previous chapters of this manuscript showed how different coarse-grained prescriptions can be employed to describe a biological system with reduced resolution, with a strong focus on decimation mappings and their information content, as quantified by the mapping entropy. It is important to underline that all the comparisons between coarse-grained representations presented so far are either qualitative (see Sec. 2.3) or based on an objective function, such as the mapping entropy (see Sec. 3.3.2), which contains more information than the purely structural detail provided by the mapped coarse-grained structure. Furthermore, the choice of this objective function is not unique, as explained in Sec. 2.3, proving that the introduction of a quantitative definition of similarity between coarse-grained mappings is a crucial step towards the understanding of the nature of reduced representations of biomolecules.

Following Ref. [26], I here introduce a general, purely geometrical notion of scalar product, and consequently of norm and distance, between reduced representations of a biomolecular system. As in the previous chapters, the analysis is restricted to the *discrete* subspace of CG representations that can be obtained for a system through a *decimation* (see Eq. 3.8) of its microscopic degrees of freedom: a subset of N constituent atoms is retained while the remaining ones are neglected.

Consider a protein composed by n constituent atoms; the number of representations Ω_N that can be constructed by selecting N of them as effective CG sites is given by the binomial coefficient, so that the *total* number of possible decimation mappings Ω is equal to $2^n - 1$ (Eq. 3.48); this number becomes prohibitively large as the size of the system increases even if one focusses only on the heavy atoms of the molecule, as it is done throughout the whole chapter. This set of coarse-grained representations constitutes the mapping space $\mathcal{M}$ (Fig. 5.1) of the molecule.

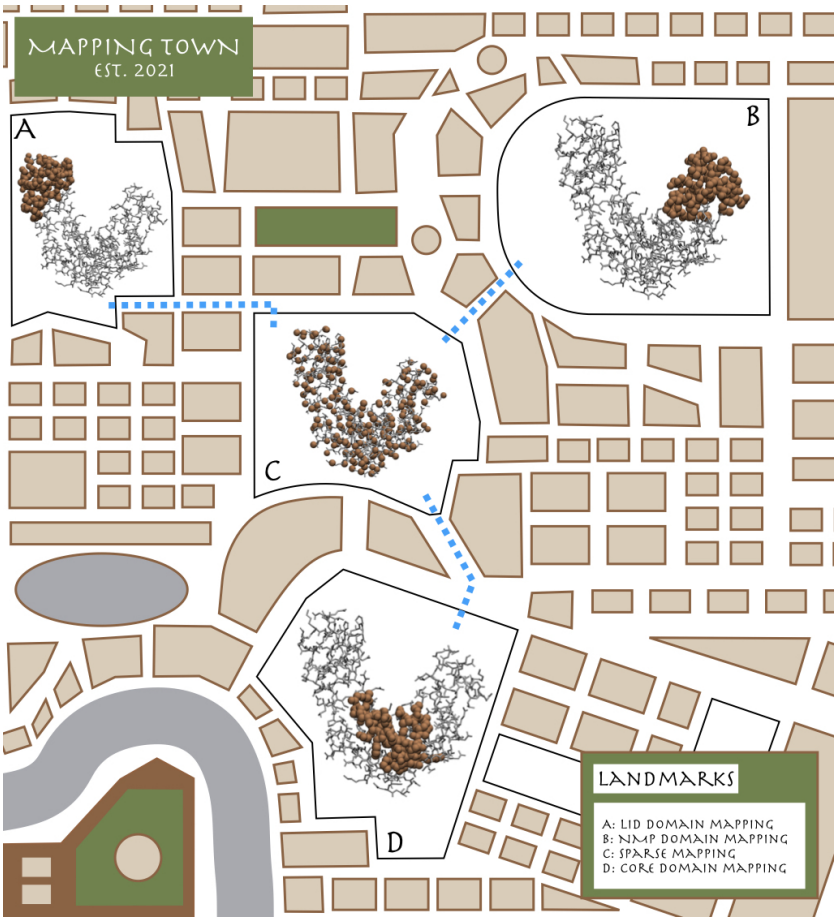


Figure 5.1: CG mappings populate an incredibly large space, whose extension grows exponentially with the number of atoms in the molecule. A notion of distance is introduced that allows one to explore the mapping space and establish relationships among them.

The investigation of the topological structure of $\mathcal{M}$ calls for the introduction of a distance $\mathcal{D}(M, M')$, $M, M' \in \mathcal{M}$, able to quantify the "separation" between pairs of points M and M' belonging to the space of decimation mappings, that is, pairs of CG representations employed to represent the system that differ in the choice of the retained atoms. Such distance must be equipped with all the associated metric properties, namely identity, symmetry, and triangle inequality.

To construct $\mathcal{D}(M, M')$, a *static* configuration of the molecule is considered, namely the crystallographic one, with (heavy) atoms located in positions $\mathbf{r}_i$, $i = 1, ..., n$ and a set of selection operators $\chi_{M,i}$, $i = 1, .., n$ defining

mapping M,

$$\chi_{M,i} = \begin{cases} 1 & \text{if atom } i \text{ is retained,} \\ 0 & \text{if atom } i \text{ is not retained,} \end{cases} \tag{5.1}$$

$$\sum_{i=1}^{n} \chi_{M,i} = N(M), \tag{5.2}$$

where $N(M)$ is the number of retained atoms in the mapping. Taking inspiration from the Smooth Overlap of Atomic Positions method (SOAP) developed by Csány *et al.* [241, 242], each $M \in \mathcal{M}$ is associated to an element $\phi_M(\mathbf{r})$ of the Hilbert space of square-integrable real functions $L_2(\mathbb{R}^3)$ as

$$\phi_M(\mathbf{r}) = \sum_{i=1}^{n} \phi_{M,i}(\mathbf{r}) = \sum_{i=1}^{n} C e^{-(\mathbf{r}-\mathbf{r}_i)^2/2\sigma^2} \chi_{M,i}, \tag{5.3}$$

obtained by centering a three-dimensional Gaussian—whose normalisation factor C will be fixed in the following—on the position of each atom of the macromolecule retained in the mapping.

The inner product $\langle \phi_M, \phi_{M'} \rangle$ of $L_2(\mathbb{R}^3)$ between two mappings M and M',

$$\langle \phi_M, \phi_{M'} \rangle = \int d\mathbf{r}\ \phi_M(\mathbf{r}) \phi_{M'}(\mathbf{r}), \tag{5.4}$$

induces a norm $||\phi_M||$ for mapping M, with

$$\mathcal{E}(M) = ||\phi_M||^2 = \langle \phi_M, \phi_M \rangle, \tag{5.5}$$

starting from which the distance $\mathcal{D}(M, M')$ can be defined as

$$\mathcal{D}(M, M') = ||\phi_M - \phi_{M'}|| = \langle \phi_M - \phi_{M'}, \phi_M - \phi_{M'} \rangle^{\frac{1}{2}}, \tag{5.6}$$

satisfying all the aforementioned metric properties. By inserting Eq. 5.3 in Eq. 5.4, the inner product $\langle \phi_M, \phi_{M'} \rangle$ between mappings generated by two distinct selection operators χ_M and $\chi_{M'}$ becomes

$$\langle \phi_M, \phi_{M'} \rangle = \sum_{i,j=1}^{n} J_{ij} \chi_{M,i} \chi_{M',j}, \tag{5.7}$$

while the associated distance $\mathcal{D}(M, M')$ in Eq. 5.6 reads

$$\begin{aligned} \mathcal{D}(M, M') &= (\mathcal{E}(M) + \mathcal{E}(M') - 2\langle \phi_M, \phi_{M'} \rangle)^{\frac{1}{2}} \\ &= \left(\sum_{i,j=1}^{n} J_{ij} \chi_{M,i} \chi_{M,j} + \sum_{i,j=1}^{n} J_{ij} \chi_{M',i} \chi_{M',j} - 2 \sum_{i,j=1}^{n} J_{ij} \chi_{M,i} \chi_{M',j} \right)^{\frac{1}{2}} \end{aligned} \tag{5.8}$$

where the coupling constant $J_{ij} = J_{ij}(\mathbf{r}_i, \mathbf{r}_j)$ between two atoms i and j is given by

$$J_{ij}(\mathbf{r}_i, \mathbf{r}_j) = C^2 \int d\mathbf{r}\ e^{-[(\mathbf{r}-\mathbf{r}_i)^2+(\mathbf{r}-\mathbf{r}_j)^2]/2\sigma^2}, \tag{5.9}$$

with

$$J_{ij}(\mathbf{r}_i, \mathbf{r}_j) = J_{ij}(|\mathbf{r}_i - \mathbf{r}_j|) = J_{ij}(r_{ij}). \tag{5.10}$$

due to translational and rotational invariance. By introducing polar coordinates in Eq. 5.9, one has

$$\begin{aligned} J_{ij}(r_{ij}) &= 2\pi C^2 \int dr d\theta\ r^2 \sin\theta e^{-\frac{1}{2\sigma^2}(2r^2+r_{ij}^2-2rr_{ij}\cos\theta)} \\ &= \frac{4\pi\sigma^2}{r_{ij}} C^2 e^{-r_{ij}^2/2\sigma^2} \int dr\ r e^{-r^2/\sigma^2} \sinh\left(\frac{r r_{ij}}{\sigma^2}\right), \end{aligned} \tag{5.11}$$

and a chain of Gaussian integrals provides

$$J_{ij}(r_{ij}) = \pi^{3/2} C^2 \sigma^3 e^{-r_{ij}^2/4\sigma^2} = e^{-r_{ij}^2/4\sigma^2}, \tag{5.12}$$

where the last equality is obtained by setting, without loss of generality, $C^2 = \frac{1}{\pi^{3/2}\sigma^3}$. Finally, by combining Eq. 5.7 and 5.12 the inner product $\langle \phi_M, \phi_{M'} \rangle$ reads

$$\langle \phi_M, \phi_{M'} \rangle = \sum_{i,j=1}^{n} e^{-r_{ij}^2/4\sigma^2} \chi_{M,i} \chi_{M',j}, \tag{5.13}$$

i.e. a sum of Gaussian factors over the positions of all pairs of atoms retained in the two mappings. Notably, the factorisation with respect to the operators χ_M and $\chi_{M'}$ in Eq. 5.7 and 5.13 enables the inner product (and therefore the distance $\mathcal{D}$ and the squared norm $\mathcal{E}$) to be determined starting from a coupling matrix J_{ij} that can be calculated *a priori* over the static structure of the molecule.

One might ask what kind of information the previously defined quantities provide about the possible CG representations of a system. To answer this question, let me first focus on the squared norm of a mapping $\mathcal{E}(M)$, see Eq. 5.5 and 5.13,

$$\mathcal{E}(M) = \langle \phi_M, \phi_M \rangle = \sum_{i,j=1}^{n} e^{-r_{ij}^2/4\sigma^2} \chi_{M,i} \chi_{M,j}. \tag{5.14}$$

Consider now two limiting cases: *(i)* extremely sparse and homogeneous CG representations, in which each retained atom does not have any retained neighbour within a radius of order $\sqrt{2}\sigma$—this condition can only be fulfilled

provided that N is not too large, or σ is much smaller than the typical interatomic distance. In this case, one has $e^{-r_{ij}^2/4\sigma^2} \approx \delta_{ij}$ and consequently $\mathcal{E}(M) \approx N$; *(ii)* globular mappings characterised by densely populated (i.e. almost atomistic) regions of retained sites surrounded by "empty" ones. In this case, the average number of retained atoms in the mapping that are located within a sphere of radius $\sqrt{2}\sigma$ from a CG site will roughly resemble its atomistic value:

$$\bar{z} = \frac{1}{n} \sum_{i,j=1}^{n} e^{-r_{ij}^2/4\sigma^2}, \tag{5.15}$$

and thus $\mathcal{E}(M) \approx N\bar{z}$. It follows that the squared norm $\mathcal{E}(M)$ captures the average homogeneity of a CG representation, that is, whether the associated retained atoms are uniformly distributed across the macromolecule or are mainly localised in well-defined regions of it. Fig. 5.2(a-c) reports examples of CG mappings extracted for these two extreme categories in the case of *4ake* together with a CG representation in which the retained atoms are randomly selected.

In general, the existence of an inner product enables the definition of an angle $\theta_{M,M'}$ between mappings, whose cosine reads

$$\cos\theta_{M,M'} = \frac{\langle \phi_M, \phi_{M'} \rangle}{(\mathcal{E}(M)\mathcal{E}(M'))^{\frac{1}{2}}}. \tag{5.16}$$

The orthogonality of mappings ($\cos\theta_{M,M'} \approx 0$) has a relatively straightforward interpretation in terms of their spatial complementarity, as $\langle \phi_M, \phi_{M'} \rangle \approx 0$ implies that it is sufficient that each atom in mapping M does not have any neighbour in M' (and vice-versa). The condition of parallelism, $\cos\theta_{M,M'} \approx 1$, is a bit less intuitive. If the mappings M and M' possess the same number of atoms N and the same norm ($\mathcal{E}(M) = \mathcal{E}(M') = \mathcal{E}$), one can see how Eq. 5.16 becomes:

$$\cos\theta_{M,M'} = \frac{\langle \phi_M, \phi_{M'} \rangle}{\mathcal{E}} = \frac{\sum_{i=1}^{n} \chi_{M,i} \left[\sum_{j=1}^{n} e^{-r_{ij}^2/4\sigma^2} \chi_{M',j} \right]}{\sum_{i=1}^{n} \chi_{M,i} \left[\sum_{j=1}^{n} e^{-r_{ij}^2/4\sigma^2} \chi_{M,j} \right]} \tag{5.17}$$

The parallelism requires that the average number of neighbors one atom of M has from mapping M' has to be equal to the average number of neighbors the atom has from itself. This means that the two mappings must place retained atoms across the macromolecule in a similar fashion. Examples of approximately parallel and orthogonal CG representations for *4ake* are presented in Fig 5.2(d-e).

While $\mathcal{E}(M)$ quantifies the average sparseness of a CG representation, $\langle \phi_M, \phi_{M'} \rangle$—or equivalently $\cos\theta_{M,M'}$—characterises the average degree of

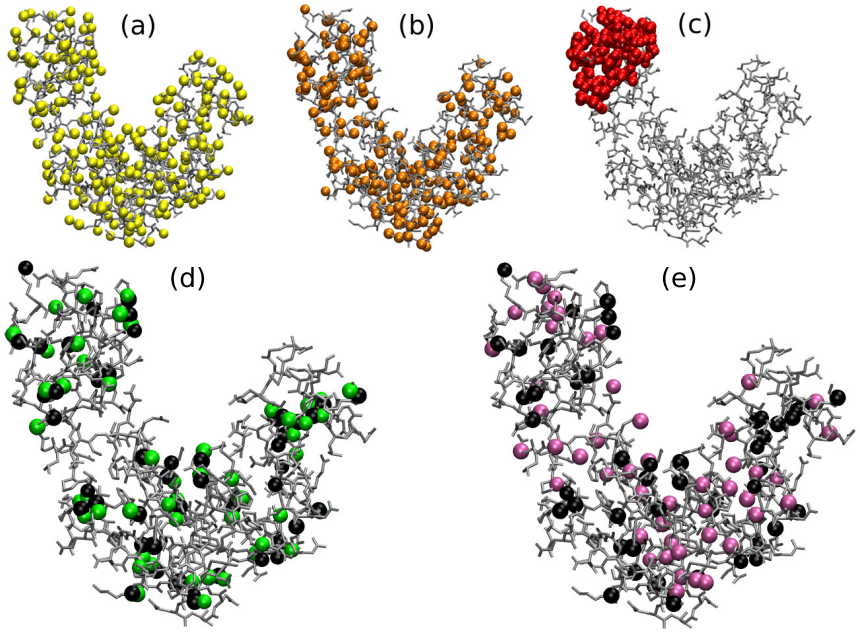


Figure 5.2: *Top row*: Example of possible CG representations for adenylate kinase (PDB code *4ake*) with $N = 214$ sites (represented as beads) characterised by a low (a), intermediate (b) and high (c) mapping squared norm $\mathcal{E}$. By increasing $\mathcal{E}$ one moves from maximally homogeneous to extremely globular CG representations. *Bottom row*: Examples of CG mappings with $N = 53$ sites that are approximately parallel (d) and orthogonal (e) to a given one. The atoms composing the reference CG representation are represented as black beads. Parallel (resp. orthogonal) mappings tend to displace CG sites on similar (resp. complementary) regions of the system. Image taken from Ref. [26].

spatial similarity between two different decimations of the microscopic degrees of freedom of the system. The distance $\mathcal{D}(M, M')$ in Eq. 5.8 combines these two notions to extract how "far" a pair of CG representations is in the space of possible mappings $\mathcal{M}$.

In the limiting cases of extremely sparse and globular mappings one respectively obtains $\mathcal{E}(M) \approx N$ and $\mathcal{E}(M) \approx N\bar{z}$, where $\bar{z}$ is the atomistic coordination number in Eq. 5.15. As the number of CG sites N increases, however, it is extremely hard for a retained site not to have any retained neighbor within a sphere of radius of order σ, so that the exact scaling of $\mathcal{E}(M)$ on the degree of CG'ing N in the case of sparse mappings will be hardly observed. Thus, it is useful to divide the inner product in Eq. 5.13 by

the average atomistic coordination number (Eq. 5.15), and define

$$\langle \phi_M, \phi_{M'} \rangle_{\bar{z}} = \frac{1}{\bar{z}} \langle \phi_M, \phi_{M'} \rangle. \tag{5.18}$$

Consequently, one has

$$\mathcal{E}_{\bar{z}}(M) = \frac{1}{\bar{z}} \mathcal{E}(M), \tag{5.19}$$

$$\mathcal{D}_{\bar{z}}(M, M') = \frac{1}{\sqrt{\bar{z}}} \mathcal{D}(M, M'), \tag{5.20}$$

while the cosine between two mappings $\cos\theta_{M,M'}$ is not affected by the rescaling. With this choice, globular mappings are now associated to $\mathcal{E}(M)_{\bar{z}} \approx N$, which can always be observed also in the case of low degrees of CG'ing, that is, high N. Note that the definition of $\langle \phi_M, \phi_{M'} \rangle_{\bar{z}}$ in Eq. 5.18 corresponds to a rescaling of the coupling constant J_{ij} in Eq. 5.12 to

$$J_{ij} = \frac{1}{\bar{z}} e^{-r_{ij}^2/4\sigma^2}. \tag{5.21}$$

For notational convenience, in the following I omit the subscript $\bar{z}$ and refer to $\mathcal{E}(M)_{\bar{z}}$, $\langle \phi_M, \phi_{M'} \rangle_{\bar{z}}$ and $\mathcal{D}_{\bar{z}}(M, M')$ as $\mathcal{E}(M)$, $\langle \phi_M, \phi_{M'} \rangle$ and $\mathcal{D}(M, M')$, respectively.

5.1 Exploration of the mapping space

Starting from the definitions introduced in the previous section, let me now perform a quantitative analysis of the high-dimensional space $\mathcal{M}$ of CG representations of a macromolecule that can be constructed through a decimation of its atomistic degrees of freedom. As a testbed system the three-domain enzyme *adenylate kinase* is considered, whose structure is discussed in detail in Chapter 3, Sec. 3.2. The calculations discussed in the previous section require in input only the value of the σ parameter and a static configuration $\mathbf{r}_i,\ i = 1, ..., n$ of the system to determine the set of Gaussian couplings J_{ij}. Here σ is set to 1.9Å (that is, half the separation between two consecutive α carbons), and rely on the *open* crystal conformation of adenylate kinase (PDB code *4ake*), excluding from the analysis all hydrogens composing the biomolecule, resulting in a total of 1656 heavy atoms.

5.1.1 Norm distributions

The numerical results reported in this subsection were obtained by my co-supervisor Roberto Menichetti, whom I here acknowledge for his crucial contribution to this chapter.

Let me first consider the squared norm $\mathcal{E}(M)$ of a CG representation M, as defined in Eq. 5.19. As previously highlighted, this quantity provides information about the spatial homogeneity of a mapping with a given degree of CG N, thus recapitulating how the retained atoms are distributed across the molecular structure, from uniformly scattered ($\mathcal{E}(M) \approx N/\bar{z}$) to mainly concentrated in well-defined, almost atomistic domains emerging out of a severely coarse-grained background ($\mathcal{E}(M) \approx N$).

It is important to stress that mappings belonging to the two aforementioned extreme cases are routinely employed by the CG'ing community in the description of a biomolecular system. In proteins, examples from the homogeneous class include physically-intuitive, residue-based CG representations of the molecule in terms of its α carbons or backbone atoms [23, 207]; homogeneity, on the other hand, is often abruptly broken in chemically-informed, *multiscale* mappings (see Sec. 2.2), in which a higher level of detail, up to the atomistic one, is sharply localised on the biologically/chemically relevant regions of the system—e.g. the active sites of the protein—while the reminder is treated at extremely low resolution [23, 141].

One natural question follows: how representative are these "common" mappings of the diversity of the space $\mathcal{M}$? In other words, how spatially homogeneous are the possible CG descriptions that can be designed for a macromolecule when no prior knowledge about its chemical structure or biological function is exploited to guide the mapping construction?

To answer this question, let me introduce the number of mappings that attain a particular value $\mathcal{E}$ of the squared norm for a given number of CG sites N, which is given by:

$$\Omega_N(\mathcal{E}) = \sum_{M \in \mathcal{M}} \delta(N(M), N)\delta(\mathcal{E}(M), \mathcal{E}) \tag{5.22}$$

Normalizing Eq. 5.22 by the total number of mappings with N sites, Ω_N, the *probability* of having a mapping with given $\mathcal{E}$ and N is defined as:

$$P_N(\mathcal{E}) = \frac{\Omega_N(\mathcal{E})}{\Omega_N}, \tag{5.23}$$

which satisfies the normalisation condition $\sum_{\mathcal{E}} P_N(\mathcal{E}) = 1$ regardless of the number of retained sites. By providing direct insight on the degree of spatial uniformity characterising the ensemble of all possible CG descriptions of a macromolecular system, $P_N(\mathcal{E})$ represents a first important ingredient in the investigation of the structure of the mapping space $\mathcal{M}$. The behaviour of $P_N(\mathcal{E})$ is observed across the decimation mapping space $\mathcal{M}$ of *4ake* for a set of 16 values of N ranging from 53 to 1605, see Table 5.1. However, even

	$\langle\mathcal{E}\rangle_N$		$\sigma_{\mathcal{E},N}$	
N	RS	WL-SP	RS	WL-SP
53	5.41	—	0.31	—
107	14.15	—	0.63	—
214	41.14	40.82	1.32	1.32
321	80.95	—	2.03	—
428	133.58	133.17	2.74	2.74
535	199.04	—	3.45	—
642	277.33	276.93	4.12	4.11
749	368.44	—	4.74	—
856	472.39	471.95	5.29	5.29
963	589.16	—	5.74	—
1070	718.76	718.29	6.06	6.07
1177	861.18	—	6.22	—
1284	1016.43	1016.14	6.16	6.17
1391	1184.51	—	5.79	—
1498	1365.42	1365.05	4.94	4.94
1605	1559.15	—	3.09	—

Table 5.1: Average mapping squared norm $\langle\mathcal{E}\rangle_N$ and associated standard deviation $\sigma_{\mathcal{E},N}$ at different degrees of coarse-graining N, calculated over the mapping space $\mathcal{M}$ of *4ake*. I present random sampling results (RS), as well as those obtained from a saddle-point approximation to the density of states $\Omega_N(\mathcal{E})$ determined through the Wang-Landau method (WL-SP), see text.

restricted to these cases, an exhaustive enumeration of all possible CG representations of the system is unfeasible in practice, as extensively highlighted in the previous chapters (see for example Eq. 3.48).

To overcome this combinatorial challenge, for each degree of CG'ing I generate $\widetilde{\Omega}_{tot} = 2 \cdot 10^6$ uniformly distributed random mappings as strings χ_i, $i = 1, ..., n$ of zeros and ones compatible with the normalisation $\sum_{i=1}^{n} \chi_i = N$. Then the associated squared norm $\mathcal{E}$ is calculated. Results for each N were then binned along the $\mathcal{E}$ axis in intervals of $\delta\mathcal{E} = 0.1$, and the corresponding $P_N(\mathcal{E})$ is estimated as

$$P_N(\mathcal{E}) = \frac{1}{\delta\mathcal{E}} \frac{\widetilde{\Omega}_N(\mathcal{E})}{\widetilde{\Omega}_{tot}}, \tag{5.24}$$

where $\widetilde{\Omega}_N(\mathcal{E})$ is the number of sampled mappings with squared norm falling between $\mathcal{E}$ and $\mathcal{E}+\delta\mathcal{E}$. In this "continuous" limit, the normalisation condition

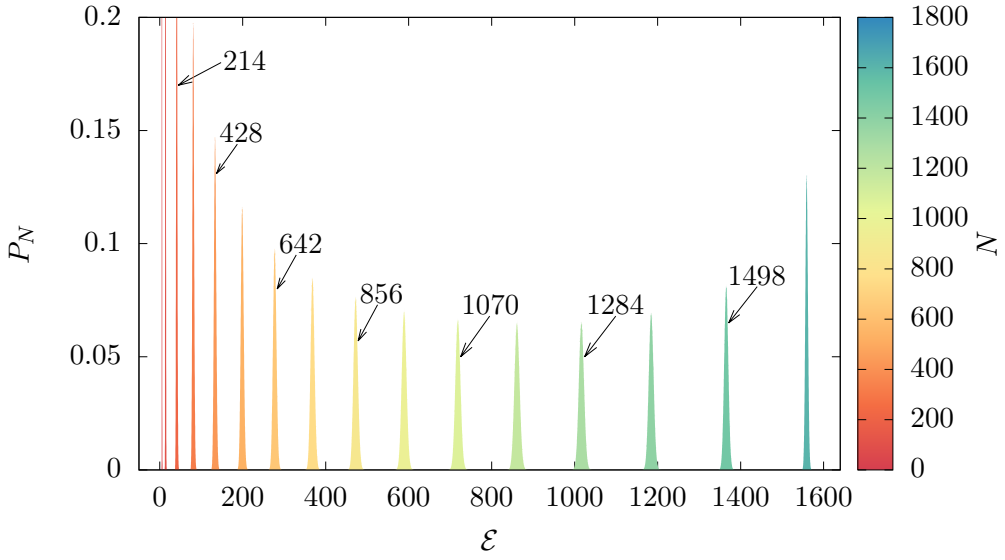


Figure 5.3: Probability $P_N(\mathcal{E})$ of the norm of the mapping $\mathcal{E}$ for *4ake* calculated at various degrees of CG N, as obtained from a random sampling of the mapping space $\mathcal{M}$. Arrows indicate the values of N for which a reconstruction of the density of states $\Omega_N(\mathcal{E})$ through the Wang-Landau algorithm has been performed. Image taken from Ref. [26].

of $P_N(\mathcal{E})$ becomes

$$1 = \sum_{\mathcal{E}} P_N(\mathcal{E})\delta\mathcal{E} \simeq \int d\mathcal{E} P_N(\mathcal{E}). \tag{5.25}$$

The set of distributions $P_N(\mathcal{E})$ obtained from the aforementioned random sampling of the mapping space of *4ake* are displayed in Fig. 5.3. One can observe that, for each value of the CG resolution N, $P_N(\mathcal{E})$ is unimodal and narrowly peaked around its average squared norm,

$$\langle \mathcal{E} \rangle_N = \int d\mathcal{E} P_N(\mathcal{E}) \mathcal{E}, \tag{5.26}$$

$\langle \mathcal{E} \rangle_N$ being an increasing function of N. On the other hand, the standard deviation $\sigma_{\mathcal{E},N}$,

$$\sigma_{\mathcal{E},N} = \left(\int d\mathcal{E} P_N(\mathcal{E}) (\mathcal{E} - \langle \mathcal{E} \rangle_N)^2 \right)^{\frac{1}{2}}, \tag{5.27}$$

is non-monotonic in the degree of CG'ing: starting from extremely small values in the case of few retained atoms (e.g. $N = 53, 107$ and 214), $\sigma_{\mathcal{E},N}$

increases roughly up to $N \approx 3n/4$ and then starts to decrease, reaching zero for $N = n$.[1] These features are further highlighted in Table 5.1 and Fig. 5.4, in which the dependence of $\langle \mathcal{E} \rangle_N$ and $\sigma_{\mathcal{E},N}$ on the degree of CG'ing N is reported as obtained from the distributions $P_N(\mathcal{E})$ in Fig. 5.3.

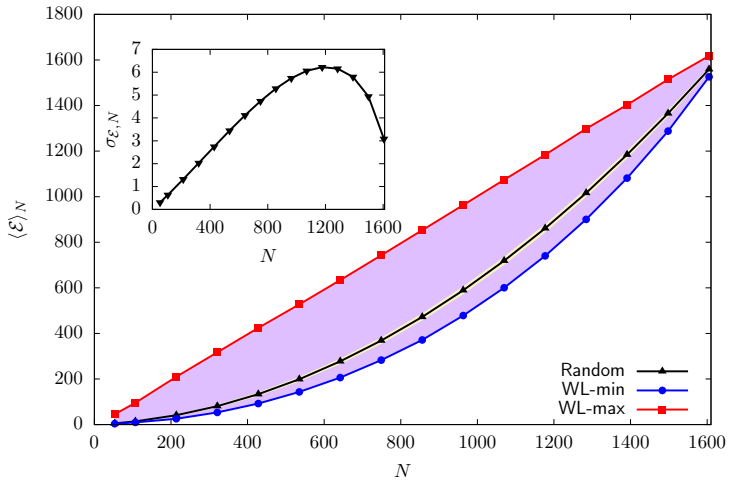


Figure 5.4: Inset: Standard deviation $\sigma_{\mathcal{E},N}$ of the mapping norm $\mathcal{E}$ as a function of the degree of CG'ing N obtained from a random sampling of the mapping space $\mathcal{M}$ of *4ake*. Main plot: N-dependence of the average squared norm $\langle \mathcal{E} \rangle_N$ ("Random", black line) and associated $3\sigma_{\mathcal{E},N}$ confidence interval (khaki area) as obtained from a random sampling of the mapping space of *4ake*, superimposed to the region covered by the set of single-window, preliminary WL runs (purple area). The minimum ("WL-min", blue line) and maximum ("WL-max", red line) squared norms reached by the preliminary runs are highlighted. "WL-max" also corresponds to the scaling $\mathcal{E} \approx N$ obtained in the case of inhomogeneous, globular mappings. Image taken from Ref. [26].

$\langle \mathcal{E} \rangle_N$ quantifies the average spatial homogeneity of the ensemble of CG representations that can be randomly assigned to *4ake* at a specific resolution. As previously discussed, maximally inhomogenous mappings, in which a chiseled chunk of the biomolecule is treated atomistically while the remainder is almost neglected, are characterised by $\mathcal{E} \approx N$. Critically, Fig. 5.4 displays that such linear scaling lies always above the average $\langle \mathcal{E} \rangle_N$ for all degrees of coarse-graining investigated. The deviation between the two curves is non-monotonic, with a maximum obtained for $N = n/2$, and only vanishes for $N \to n$, where mappings become very dense as they collapse towards the

[1] In this case only one possible mapping exists, namely the atomistic one.

atomistic representation. As a consequence, the CG representations one encounters by randomly probing the mapping space $\mathcal{M}$ tend to be "sparse" rather than compact. Furthermore, the difference between the squared norm of the globular case and $\langle \mathcal{E} \rangle_N$ is always (but for $N \approx n$) one or two orders of magnitudes larger than the standard deviation of the corresponding $P_N(\mathcal{E})$, see Fig. 5.4. Inhomogeneous mappings lie extremely far away in the right tails of the distributions displayed in Fig. 5.3, thus constituting an exponentially vanishing subset of the space $\mathcal{M}$.

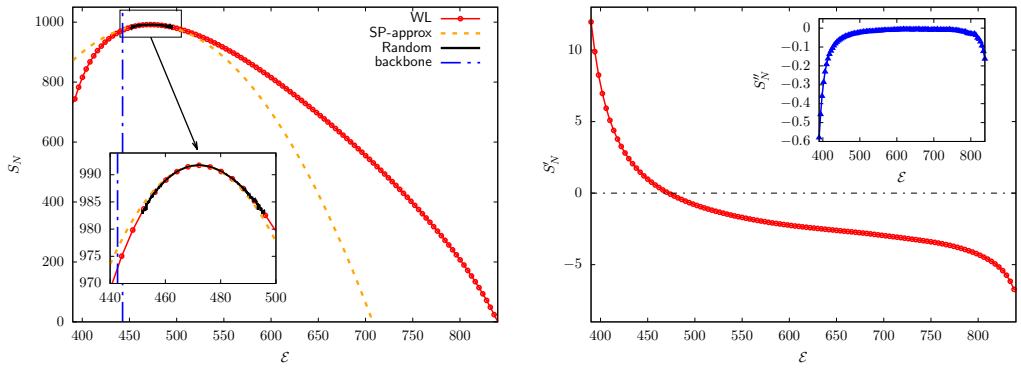


Figure 5.5: *Left*: Logarithm of the density of states $\Omega_N(\mathcal{E})$ of *4ake*, $S_N(\mathcal{E}) = \ln[\Omega_N(\mathcal{E})]$, for $N = 856$. Results are obtained via (*i*) Wang-Landau sampling ("WL", red dotted line), vertically shifting the data so that the minimum of S_N over the range of investigated norms is zero; (*ii*) a saddle-point approximation of the WL predictions ("SP-approx", orange dashed line); and (*iii*) a random drawing of CG representations ("Random", black line), in this latter case shifting the curve so that its maximum coincides with the one of the WL profile. The figure also includes the squared norm associated to the mapping in which all the heavy atoms composing the backbone of *4ake* are retained ("backbone", dashed blue line), a CG representation that is commonly employed when CG'ing a protein [23, 207]. *Right*: First (main plot) and second (inset) derivatives $S'_N(\mathcal{E})$ and $S''_N(\mathcal{E})$ of the entropy $S_N(\mathcal{E})$ determined via WL sampling for $N = 856$. Image taken from Ref. [26].

The suppression of the statistical weight associated to high-norm, globular CG representations of *4ake* in the space of all possible ones is not surprising, and is solely dictated by entropic effects. Indeed, at least for small and intermediate N, it is extremely unlikely that a completely random selection of retained atoms across the biomolecule results in their dense confinement within sharply-defined spatial domains of the system, just as it is unlikely for a gas to occupy only a small fraction of the volume in which it is enclosed.

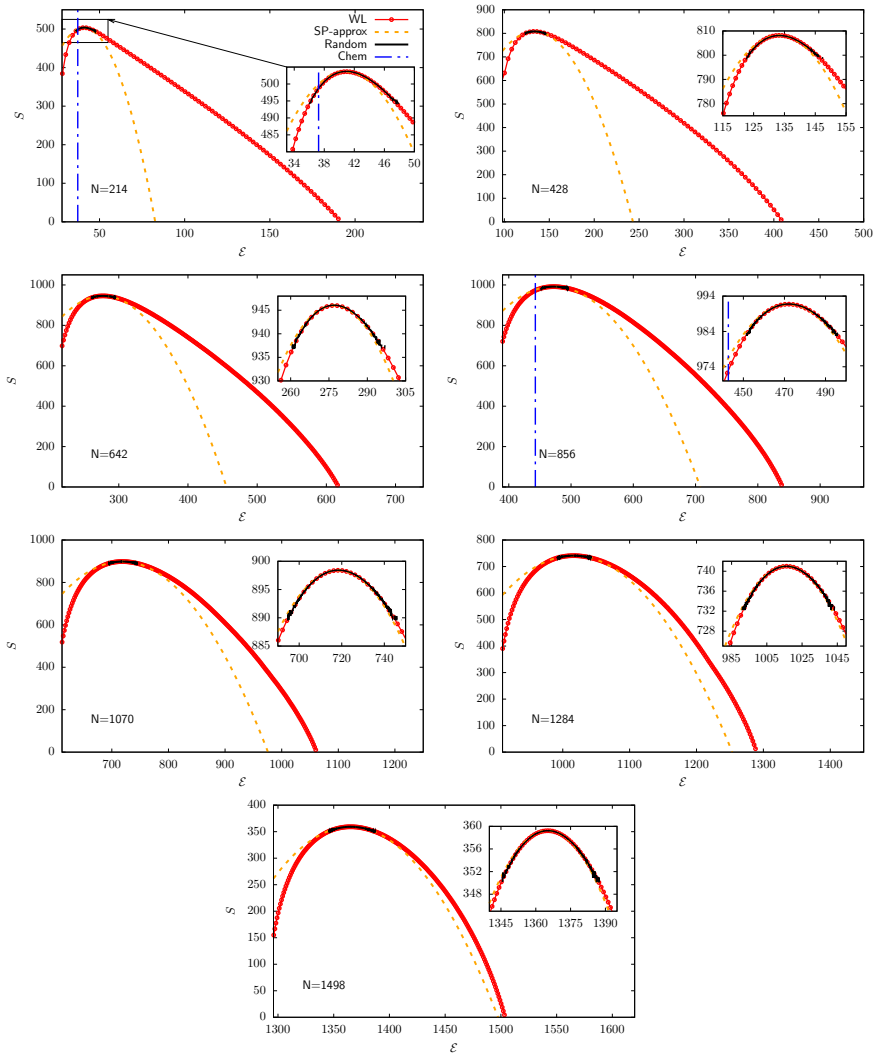


Figure 5.6: For each N, the figure shows the values of the entropy $S_N(\mathcal{E})$ obtained via (*i*) Wang-Landau sampling ("WL", red dotted lines), shifting the data so that the minimum of S_N over the range of investigated norms is zero; (*ii*) a saddle-point approximation of the WL predictions ("SP-approx", orange dashed lines); and (*iii*) a random drawing of CG mappings ("Random", black lines), in this latter case shifting the curve so that its maximum coincides with the one of the corresponding WL profile. In the case of $N = 214$ (resp. $N = 856$), the subfigure contains the squared norm associated to the C_α (resp. backbone) mapping ("Chem", blue dashed line), a CG representation routinely employed in protein CG [23, 207]. Image taken from Ref. [26].

Interestingly, this latter analogy can be pushed further by noting that the squared norm $\mathcal{E}(M)$, see Eq. 5.19 and 5.14, is akin to the negative configurational energy of a lattice gas living on the irregular grid defined by the protein conformation, whose particle interact via a hard-core, short-range potential followed by an attractive Gaussian tail. In this context, the selection operators $\chi_{M,i} = 0, 1,\ i = 1, ..., n$ of a mapping M with N retained atoms can be interpreted as the set of occupation numbers describing a distribution of the N particles of the gas on the n available lattice sites. It follows that compact CG representations of *4ake*, located in the large-$\mathcal{E}$ limit of $P_N(\mathcal{E})$, are just as challenging to randomly sample within the space $\mathcal{M}$ as are the low-energy configurations of the gas in which the N particles spontaneously occupy only a fraction of the available volume.

The strongly entropy-driven distribution of mappings calls for the introduction of enhanced sampling techniques to boost the exploration of $\mathcal{M}$, such as the algorithm proposed by Wang and Landau (WL) [205, 206, 226, 227]. For each CG resolution N, the aim is to obtain a *uniform* sampling of the possible mapping norms $\mathcal{E}$ across the space $\mathcal{M}$, in contrast to the set of narrowly-peaked probability distributions displayed in Fig. 5.3. In this context, the acceptance probability of Eq. 4.4 is translated into the following expression:

$$\begin{aligned} W(M \to M') &= \min\left[1, \frac{\Omega_N(\mathcal{E}(M))}{\Omega_N(\mathcal{E}(M'))}\right] \\ &= \min\left[1, \exp\left(-[S_N(\mathcal{E}(M')) - S_N(\mathcal{E}(M))]\right)\right], \end{aligned} \quad (5.28)$$

where $\Omega_N(\mathcal{E})$ is the density of states defined in Eq. 5.23 while $S_N(\mathcal{E}) = \ln[\Omega_N(\mathcal{E})]$ is the corresponding microcanonical entropy. $\Omega_N(\mathcal{E})$ is a priori unknown and can be determined by means of the self-consistent, iterative procedure illustrated in Chapter 4, Sec. 4.3.

As discussed in Sec. 4.3, knowledge of the sampling boundaries proves extremely beneficial to the accuracy and rate of convergence of the Wang-Landau self-consistent scheme [239]. Hence, for each value of N investigated, an exploratory, non-iterative WL run is initially performed in order to approximately locate the minimum and maximum mapping norms $\mathcal{E}_{min}(N)$ and $\mathcal{E}_{max}(N)$ achievable at that specific CG resolution, and consequently bound the support of the corresponding $\Omega_N(\mathcal{E})$.

Mapping norms visited by the set of preliminary WL runs extend, for all values of N, over a significantly wider range compared to the one obtained by random sampling (see Fig. 5.4). Remarkably, the maximum norm $\mathcal{E}_{max}(N)$ exhibits a linear dependence on N that is fully compatible with the one associated to globular CG representations, $\mathcal{E}_{max}(N) \approx N$, highlighting that

the preliminary WL runs succeed in exploring this entropically suppressed region of the mapping space. Furthermore, Fig. 5.4 displays that the minimum norm $\mathcal{E}_{min}(N)$ identified by the preliminary runs lies always below the average $\langle \mathcal{E} \rangle_N$ for all values of N. In contrast to globular mappings, CG representations living in this low $\mathcal{E}$ limit are *maximally homogeneous*, that is, retained atoms are scattered throughout the molecular structure as uniformly as possible. This class constitutes another exponentially vanishing subset of the mapping space: in the gas picture, it would correspond to the ensemble of configurations in which the particles are *regularly* distributed within the available volume.

Having approximately identified the range of norms achievable for *4ake* at each CG resolution, the associated densities of states $\Omega_N(\mathcal{E})$ can be determined via the iterative WL scheme described in Sec. 4.3. To speed-up convergence of the algorithm, for each N the range of norms $[\mathcal{E}_{min}, \mathcal{E}_{max}]$ is slightly reduced with respect to the one predicted by the explorative WL runs. This interval is then divided into a set of overlapping windows in which independent WL simulations are performed [206]. The resulting partial densities of states are combined *a posteriori* to determine the cumulative $\Omega_N(\mathcal{E})$ up to a global multiplicative factor, or, in this case, the entropy $S_N(\mathcal{E}) = \ln[\Omega_N(\mathcal{E})]$ up to an additive constant.

WL estimates of the entropy $S_N(\mathcal{E})$ are presented in Fig. 5.5 for $N = 856$, while results for all the other degrees of CG are reported in Fig. 5.6. The behaviour of S_N is non-monotonic in $\mathcal{E}$ in all cases, exhibiting a unique maximum as the mapping norm moves from the left to the right boundary of the range of investigated ones—that is, in transitioning from extremely homogeneous to maximally globular CG representations. As $\Omega_N(\mathcal{E}) = \exp[S_N(\mathcal{E})]$, this result confirms how these two limiting classes of mappings constitute regions of exponentially vanishing size within the broad space $\mathcal{M}$. At the same time, the overall shape of S_N strongly depends on the degree of CG: while for high N entropy profiles are nearly symmetric around their maximum, they become increasingly skewed as fewer and fewer atoms are employed to represent the macromolecule. This asymmetry becomes apparent by performing, for each CG resolution, a quadratic expansion of S_N around its maximum,

$$S_N(\mathcal{E}) \simeq S_N(\tilde{\mathcal{E}}(N)) + \frac{1}{2} S''_N(\tilde{\mathcal{E}}(N))(\mathcal{E} - \tilde{\mathcal{E}}(N))^2, \tag{5.29}$$

where $\tilde{\mathcal{E}}(N)$ is the norm at which the first derivative S'_N of the entropy vanishes, and $S''_N(\tilde{\mathcal{E}}(N))$ is the corresponding second derivative—the dependence of S'_N and S''_N on $\mathcal{E}$ being displayed in Fig. 5.5 for $N = 856$. The accuracy of this parabolic, symmetric approximation in reproducing the exact S_N over

the whole $\mathcal{E}$-range increases with the number of retained atoms, see Fig. 5.5 and 5.6, especially as far as the limit of high mapping norms is concerned.

Finally, it is interesting to test the predictions of WL sampling against the results obtained via a completely random exploration of the mapping space. To this end, Fig. 5.5 and Fig. 5.6 include a comparison between the WL entropies S_N and their random counterparts S_N^{ran}, the latter defined as $S_N^{ran}(\mathcal{E}) = \ln[P_N(\mathcal{E})] + C_N$, where $P_N(\mathcal{E})$ are the probability densities presented in Fig. 5.3 and the constants C_N are set so that the maxima of S_N^{ran} and S_N coincide. For each value of N the two profiles are in perfect agreement, thus confirming the accuracy of the self-consistent WL scheme in determining the density of states of a system. Critically, results for S_N^{ran} only extend over a very narrow range of mapping norms, centered around the value $\tilde{\mathcal{E}}(N)$ for which the maximum of the entropy is attained. It is therefore largely unfeasible, by randomly drawing CG representations, to exhaustively explore the mapping space $\mathcal{M}$ of a macromolecule. In this respect it is worth to inspect the position, on the $\mathcal{E}$ axis, of the C_α and backbone mappings (which in *4ake* retain $N = 214$ and $N = 856$ sites, respectively), two reduced representations that are routinely employed for CG'ing proteins [23, 207]. These turn out to be located in the vicinity of the class of "prototypical" random ones, for which the entropy S_N reaches its maximum; however, their intrinsic regularity, dictated by the position of the retained sites on the peptide chain, makes these mappings slightly more homogeneous than the random ones, see Fig. 5.5 and Fig. 5.6.

To provide a more quantitative measure of the consistency between random and WL sampling results, for each degree of CG'ing the average and variance of the mapping norm, see Eqs. 5.26 and 5.27, are recalculated starting from the WL entropies S_N. These are used to compute $P_N(\mathcal{E})$ making use of a saddle-point approximation of Eq. 5.23, namely

$$P_N(\mathcal{E}) = \frac{\Omega_N(\mathcal{E})}{\Omega_N} = \frac{\exp[S_N(\mathcal{E})]}{\int d\mathcal{E} \exp[S_N(\mathcal{E})]} \simeq$$
$$\left(\frac{|S_N''(\tilde{\mathcal{E}}(N))|}{2\pi} \right)^{\frac{1}{2}} \exp\left[\frac{1}{2} S_N''(\tilde{\mathcal{E}}(N))(\mathcal{E} - \tilde{\mathcal{E}}(N))^2 \right], \tag{5.30}$$

where in the last step of Eq. 5.30 the quadratic expansion of S_N defined in Eq. 5.29 is used. Within the saddle point approximation, one has $\langle \mathcal{E} \rangle_N = \tilde{\mathcal{E}}(N)$, $\tilde{\mathcal{E}}(N)$ being the position of the maximum of S_N, and $\sigma_{\mathcal{E},N} = |S_N''(\tilde{\mathcal{E}}(N))|^{-\frac{1}{2}}$: these predictions are found to be in perfect agreement with their random sampling counterparts, results being presented in Table 5.1.

5.1.2 Inner product distributions

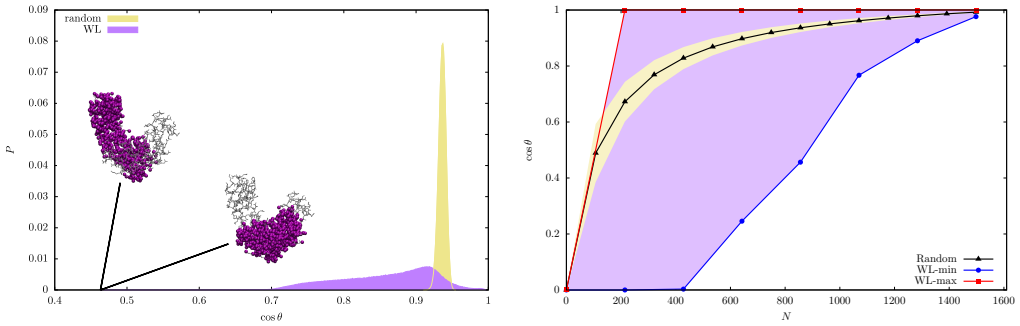
In this section, the mapping space $\mathcal{M}$ is described from the perspective of the inner product between its elements. Following the same scheme of Sec. 5.1.1, the analysis is restricted to the cosine between mappings that are constrained to share the same resolution N. To fulfil this purpose one can compute the probability $P_{NN}(\cos\theta)$ of observing a value of $\cos\theta$ provided that this constraint is satisfied:

$$P_{NN}(\cos\theta) = \frac{\Omega_{NN}(\cos\theta)}{\Omega_N^2}, \tag{5.31}$$

that is, the ratio between the number of mapping pairs whose cosine is equal to $\cos\theta$, $\Omega_{NN}^2(\cos\theta)$, and the total number of possible pairs Ω_N^2. Now it is possible to investigate how the average *degree of parallelism* between two mappings changes when considering randomly selected representations or more peculiar elements of $\mathcal{M}$.

In this section two data sets are compared, each one containing 10^6 elements: the first is obtained by computing the cosine between two mappings in which the retained sites have been picked randomly; the second data set is constructed in a more sophisticated manner, making use of the WL sampling scheme to collect mappings that uniformly span the range of accessible values of $\mathcal{E}$, which is known from the previous section (see Fig. 5.4). More specifically, a non-iterative, single-window WL exploration as in Sec. 5.1.1 over this range is started and, when all the reference bins have been visited at least once, a mapping is saved every 1656 Monte Carlo moves. Mappings are saved in different macro-bins, each one covering an interval of amplitude 20 (in terms of units of $\mathcal{E}$). Sampling ends when 5000 mappings are saved in each box, without considering the convergence of the WL algorithm. The data set is then generated by computing the cosine (Eq. 5.16) between randomly selected pairs of mappings extracted through this procedure. Importantly, the WL sampling scheme produces a pool of potentially correlated mappings and the chance of collecting similar elements of $\mathcal{M}$ cannot be excluded.

Fig. 5.7(a) shows the histograms between the two data sets for $N = 856$. While the random cosine distribution displays a narrow peak around its average value, the WL histogram is more distributed, reflecting the increased diversity of the data set. Indeed, the latter histogram spans values that range from ~ 1, obtained when two mappings are perfectly parallel, to 0.457, when two mappings are as orthogonal as possible given the properties of the lattice, defined by the protein conformation, and this number of retained sites. Fig. 5.7(a) also includes a graphical rendering of the two maximally orthogonal mappings, which possess a high value of $\mathcal{E}$ ($\mathcal{E} = 847.32$ and $\mathcal{E} = 843.82$, respectively) and cover different regions of the enzyme structure.



(a) (b)

Figure 5.7: (a): histogram of cosine values extracted from random CG mappings (yellow) and WL CG mappings (purple, see main text) for *4ake* with $N = 856$ sites. Elements of $\mathcal{M}$ with the lowest value of the cosine ($\cos\theta = 0.457$) are shown; such value corresponds to an angle of 63.25 degrees. (b): range of cosine values covered by the two data sets when N is changed. The dotted black line shows the average value of $\cos\theta$ over the different random data sets and the yellow region represents the points within 3σ from the mean. The red (blue) dotted lines report the maximum (minimum) values of $\cos\theta$ inside WL data sets, respectively. Image taken from Ref. [26].

In Fig. 5.7(b) these considerations are extended to different values of N, namely those employed in Sec. 5.1.1. The random distribution is always confined in a narrow interval of values of $\cos\theta$, while WL data sets are capable of spanning a much wider range. In particular, for sufficiently small values of N, it is possible to retrieve maximally parallel ($\cos\theta = 1$) and maximally orthogonal ($\cos\theta = 0$) mappings inside the WL dataset. Reaching orthogonality is made possible by the fact that, at such low values of N, it is possible to confine retained sites in two separate regions of the protein structure.

5.2 Lattice gas analogy and phase transitions

The numerical results reported in this section were obtained by my supervisor Raffaello Potestio, whom I here acknowledge for his crucial contribution to this chapter.

As anticipated in Sec. 5.1, the reduced representation discussed in the present work, in which a mapping is defined in terms of a *decimation* of the atoms available on the molecular structure, suggests the analogy with a lattice gas. Such analogy is a classic of statistical mechanics, which enables one to map an Ising model to a gas of interacting particles, thus making it

manifest that the spontaneous magnetisation in the former and the liquid-gas phase transition in the latter belong to the same universality class [243]. Here, the consequences of the lattice gas interpretation of mappings are investigated in order to tackle the issue of characterising the mapping space from a different perspective.

The role of the energy can be played by the norm of the mapping:

$$E(M) = -\mathcal{E}(M). \tag{5.32}$$

In analogy with a lattice gas, if two retained sites are close to each other, they feel an attractive interaction, thereby reducing the energy. Making use of Eq. 5.32 one can thus write

$$\Omega_N(E) = \Omega_N(-\mathcal{E}). \tag{5.33}$$

Let me now consider a system governed by the lattice Hamiltonian in Eq. 5.32 at equilibrium with a reservoir at temperature $T = \beta^{-1}$. The partition function of such system can be expressed in terms of $\Omega_N(E)$ *via*

$$\mathcal{Z}_N(\beta) = \int dE\ e^{-\beta E}\Omega_N(E) \equiv \int dE\ e^{-(\beta E - S_N(E))}, \tag{5.34}$$

where the relation $S_N(E) = \ln \Omega_N(E)$ is used to define the entropy. Eq. 5.34 allows to compute the dimensionless Helmholtz free energy as

$$\beta F_N(\beta) = -\ln \mathcal{Z}_N(\beta) = -\ln \int dE\ e^{-(\beta E - S_N(E))}. \tag{5.35}$$

While the logarithm of the integral can be theoretically and numerically cumbersome to compute, it is possible to obtain a reasonable estimate of βF_N through a saddle point approximation. Specifically, one can expect that the integral is approximately equal to the largest integrand, so that

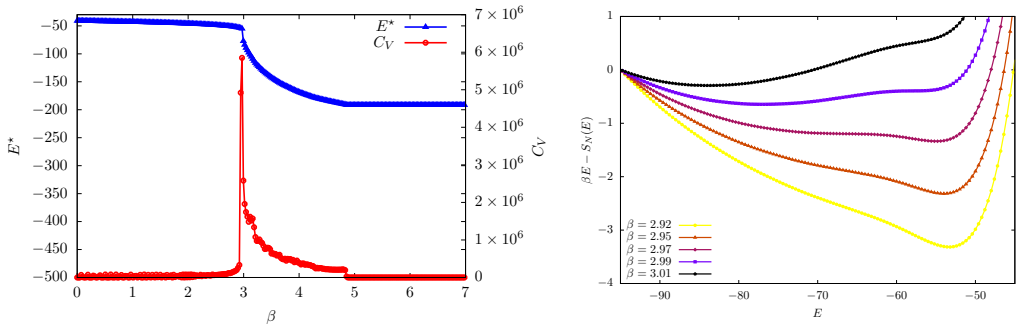
$$\int dE\ e^{-(\beta E - S_N(E))} \simeq C\ \max_E \left(e^{-(\beta E - S_N(E))}\right), \tag{5.36}$$

where C is an immaterial constant. This approximation provides a definition of the free energy that is equivalent to the Legendre-Fenchel transform:

$$\beta F_N(\beta) \simeq \min_E \left(\beta E - S_N(E)\right). \tag{5.37}$$

The thermodynamics of the lattice gas at thermal equilibrium can thus be retrieved computing Eq. 5.37 for a given value of N at all values of β.

It is particularly instructive to investigate the temperature dependence of $E^\star$, defined as the value of the energy for which $\beta E - S(E)$ reaches its



(a) (b)

Figure 5.8: (a): heat capacity C_V (red circles, right ordinate) and value of the energy $E^\star$ corresponding to the minimum of the free energy (blue triangles, left ordinate) as functions of β for the system with $N = 214$. $E^\star$ decreases monotonically with β, indicating that higher temperatures correspond to higher values of the average internal energy of the lattice gas, as expected; however, a jump discontinuity in $E^\star$ appears in correspondence of the same value β_{gl} for which the heat capacity features a sharp peak, suggesting the occurrence of a first order phase transition separating two distinct phases: a gas (low β) from a liquid (high β) for the lattice gas model, and, correspondingly, a sparse phase from a dense, localised phase in the case of CG mappings. (b): Helmholtz free energy βF of the lattice gas as a function of the energy for different values of β. The curves have, in general, a unique absolute minimum; however, as β increases, a metastable minimum appears that, for a particular value of β, becomes degenerate. The presence of a small but appreciable barrier between the two minima makes the position of the absolute minimum, $E^\star$, shift abruptly from one to the other, as can be seen in Fig. 5.9, thus making $E^\star(\beta)$ discontinuous. Image taken from Ref. [26].

minimum. In Fig. 5.8a (blue curve, left ordinate) this function is reported for $N = 214$: note that $E^\star = E^\star(\beta)$ decreases monotonically, i.e., the lower the temperature, the lower the value of the energy–which corresponds to higher values of the mapping norm. At a particular value β_{gl} of the inverse temperature, however, $E^\star$ drops abruptly, thus suggesting the occurrence of a first-order, discontinuous phase transition.

To gain further insight, the shapes of $\beta E - S(E)$ for values before and after β_{gl} are computed. These functions, reported in Fig. 5.8b, indeed show two minima separated by a relatively low barrier; increasing β, the absolute minimum shifts from the right to the left, crossing a point for which the two are essentially degenerate. This is the point of coexistence of two distinct "phases" of this lattice gas: a low density one corresponding to distributed

mappings (high energy), and one ascribable to more dense, compact conglomerates of sites (low energy). The critical nature of the transition from one regime to the other is confirmed by the inspection of the heat capacity, computed as

$$C_V = -\beta^2 \frac{\partial^2(\beta F)}{\partial \beta^2} \tag{5.38}$$

and reported in Fig. 5.8a (red curve, right ordinate). The sharp, asymmetric peak in C_V, located at the value β_{gl} of the inverse temperature, shows that the lattice gas crosses a phase transition between a gas and a liquid phase.

A crucial role in this behaviour is played by the number of coarse-grained sites. In fact, as N increases, the system acquires the possibility of crossing a second phase transition: for example, in the case of $N = 1070$, besides the gas-liquid one, it is possible to observe a second, even sharper discontinuity in $E^\star$ for a value of the inverse temperature $\beta_{ls} > \beta_{gl}$. This temperature separates the liquid from the solid phase: when the lattice gas particles are sufficiently many, and the temperature sufficiently low, the system can "freeze" in particularly dense mappings with very low entropy. Also in this case, the inspection of the heat capacity (Fig. 5.9) supports the interpretation of this as a phase transition. Finally, if the number of sites is too large (e.g. $N = 1498$) no transition is observed, see Fig. 5.9.

5.3 Topology

Here, the distance $\mathcal{D}$ (Eqs. 5.8 and 5.19) between members of $\mathcal{M}$ is discussed with the aim of showing, once again, that a *peculiar* choice of retained CG sites, i.e., one impossible to obtain with random sampling, displays non-trivial statistical properties that reflect in the topological organisation of the mapping space.

5.3.1 Topology of the mapping norm space

Without loss of generality, this analysis can be restricted to the case $N = 214$, which is the number of amino acids of *4ake*. Here a data set of mappings is generated via the WL sampling protocol explained in Sec. 5.1.2; in this case, the range of values of $\mathcal{E}$ is narrower and only 10 macro-bins of amplitude 20 are explored. The data set is constructed by randomly selecting 100 elements for each of the macro-bins, resulting in 1000 CG mappings that homogeneously span the accessible values of $\mathcal{E}$. The sketch map algorithm [244, 245] is employed to embed these points from the high-dimensional

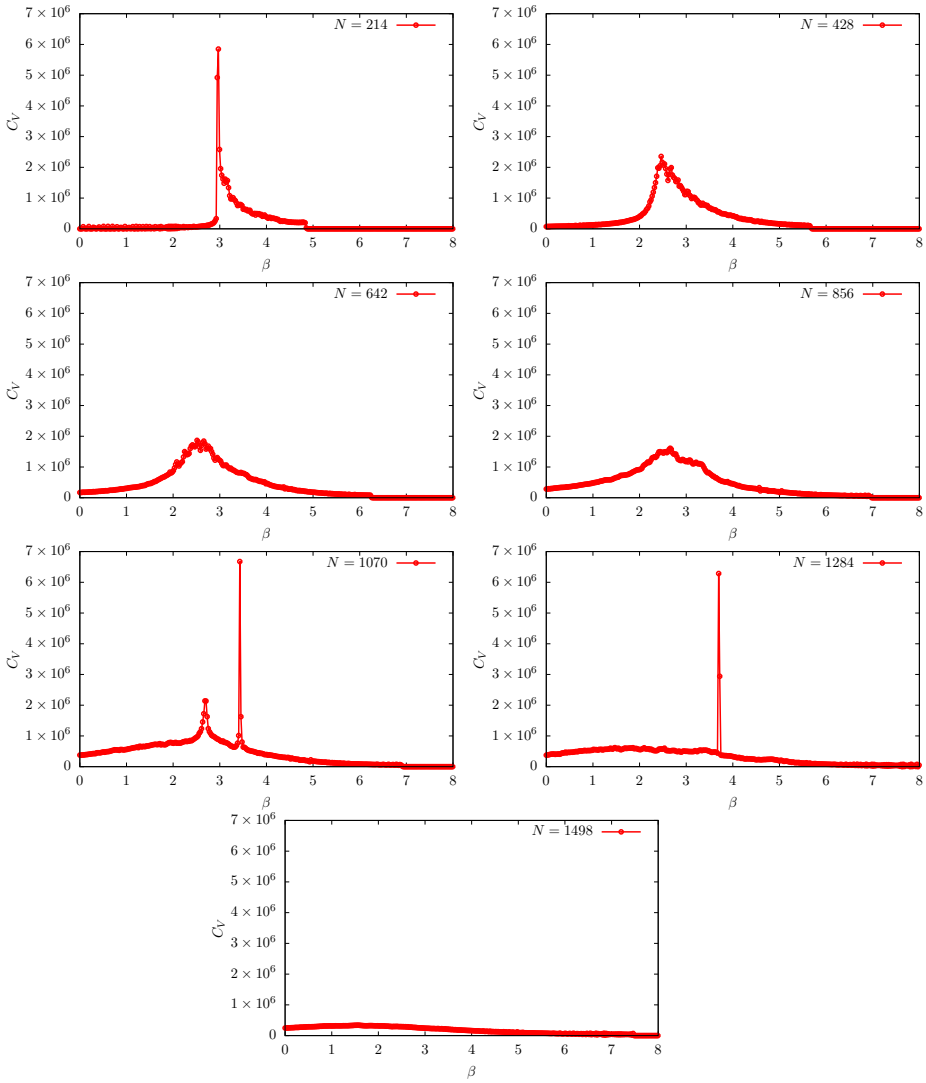


Figure 5.9: Dependence of the heat capacity C_V on the inverse temperature β for the lattice gas analogue of the mapping norm of *4ake* calculated at several degrees of CG. Sharp peaks in C_V at high (resp. low) values of N suggest the presence of a solid-liquid (resp. liquid-gas) transition in the system. It should be noted that the scales of β and C_V are the same in all plots. Image taken from Ref. [26].

space of mappings $\mathcal{M}$ into a two-dimensional plane, however preserving as faithfully as possible the relative distances among them—that is to say that nearby points in the mapping space are mapped onto nearby points on the 2D space, see Fig. 5.10. The two critical parameters of the algorithm are σ_d and σ_D, which modulate how *far* and *close* points are in the low (high) resolution space [244]. To provide the reader with a feeling of the impact that these parameters have on the structure of the low-dimensional representation, Fig. 5.10(a-b) reports the embeddings obtained for a low and high value of σ_d and σ_D.

In the first case, reported in Fig. 5.10(a) and referring to low values of the σ parameters, data points are in general very sparse and uniformly distributed on the plane, with the exception of a group of points that accumulate in a denser cluster: these are particularly compact mappings localised in a specific region of the molecule. Such mappings remain close to each other even when the σ parameters are increased, thus "squeezing" all points in the low-D embedding, see Fig. 5.10(b). In this latter scenario, points corresponding to low-$\mathcal{E}$, uniform mappings collapse in a small region of the embedding space. Furthermore, a third group of points corresponding to compact mappings appears, distinct from the ones previously discussed, and absent in the low-σ embedding.

The high-σ embedding thus highlights two relevant features: first, the presence of specific regions with qualitatively distinct mapping properties; these are either sparse, but necessarily similar one to the other (Fig. 5.10(d)), or dense, with atoms localised in different domains of the molecule (Figs. 5.10(c,e)). The distance among the latter is necessarily large, since the retained sites cover non-overlapping regions.

The second relevant feature is that different groups of points, associated to qualitatively distinct types of mappings, can be connected one to the other only "passing through" a third one, as in the case of mapping *c* going to *e* through *d*. This is suggestive of the presence of *routes* in mapping space that join points having the same value of the norm, which however cannot be connected through "iso-$\mathcal{E}$" paths: in order to transform mappings such as that in *c* into that in *e* through a sequence of single-site changes (i.e. one retained atom is discarded, a formerly discarded one is now retained) one cannot but increase or decrease the value of the norm.

5.3.2 Topology of mapping entropy space

The stochastic minimisation of the mapping entropy S_{map} (Eq. 3.47) carried out in Sec. 3.3 outputs a pool of optimised solutions, each one being the result of an independent simulated annealing run. Sec. 3.3.2 shows how CG

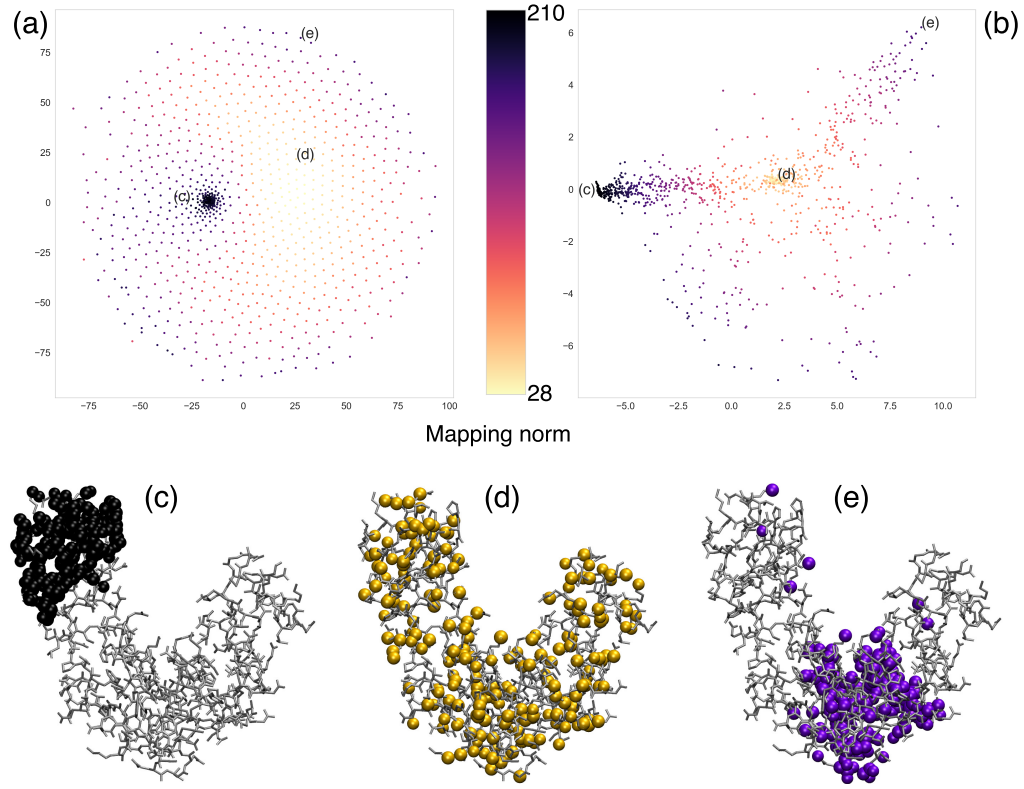


Figure 5.10: *Top*: topology of the mapping space $\mathcal{M}$ in 2D obtained with the sketch map algorithm [244, 245]. The algorithm requires six parameters, namely σ_d, a_d, b_d in the low resolution space and σ_D, a_D, b_D in the original, high resolution one. Here $\sigma_D = \sigma_d = 2$ in (a) and $\sigma_D = \sigma_d = 20$ in (b), while $a_d = b_d = 2$ and $a_D = b_D = 5$ in both cases. Mappings are depicted with different colors depending on their norm $\mathcal{E}$. A different choice for σ_D and σ_d results in a completely different 2D embedding (see Ref. [244] for a detailed explanation). *Bottom*: three different mappings located in three separated regions of the plane in (a) and (b). Mappings in subfigures (c) and (e) possess very high values of $\mathcal{E}$ and are localised in different domains of the protein. Note that sparse mappings, such as the one in subfigure (d), are clustered in the same region in (b) but not in (a). Image taken from Ref. [26].

representations with comparable values of S_{map} are neatly *separated* from each other, namely it is impossible to transform one mapping into another keeping the value of S_{map} low. This analysis, however, does not exclude that the selected CG representations share common structural features.

In this subsection the distance $\mathcal{D}$ is employed to investigate the structural similarity between the aforementioned mappings, that is, to understand if low-S_{map} representations are closer to each other than randomly selected ones according to the newly introduced metric. More specifically, the data set of 1968 CG mappings of *4ake* with $N = 214$ illustrated in Sec. 4.1 is employed, which covers a wide range of values of S_{map}; the relations among these mappings are then quantified in terms of their distance $\mathcal{D}$, taking the enzyme crystal structure as a reference. With respect to this, it is worth keeping in mind that $\mathcal{D}$ intimately depends on this reference, and mappings that lie close to each other when a given structure is considered might turn out to be closer or further away from each other when a different conformation is used.

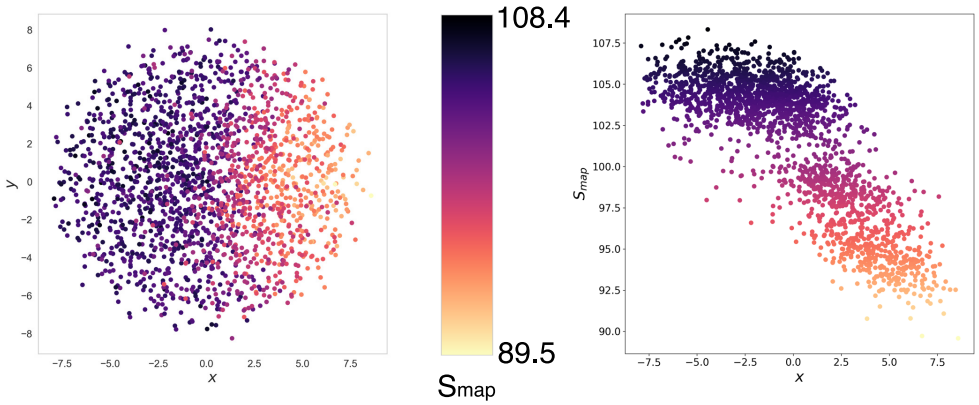


Figure 5.11: Application of the sketch map algorithm to a distance matrix obtained calculating $\mathcal{D}$ (Eqs. 5.8 and 5.19) over the data set of 1968 mappings described in Chapter 4. The x component separates very well the data points according to their value of S_{map}, thus proving that informative mappings can be distinguished among the elements of $\mathcal{M}$ according to a measure of geometrical similarity such as $\mathcal{D}$. The parameters fed to the algorithm are the following: $\sigma_D = \sigma_d = a_D = b_D = 5$, $a_d = b_d = 2$. Image taken from Ref. [26].

Fig. 5.11 shows that the two-dimensional embedding obtained through the application of the sketch map algorithm separates the CG mappings according to a gradient of S_{map}. In particular, the x component of the sketch

map and the mapping entropy S_{map} display a clear anticorrelation. The results suggest that highly informative mappings, characterised by low values of S_{map}, share geometrical features that are not present in less informative (high-S_{map}) representations. In other words, the peculiar resolution distribution found in low-S_{map} mappings separates them from other elements of $\mathcal{M}$. The relevant features highlighted by the mapping entropy thus reverberate in the merely structural characterisation given by the mapping distance.

5.4 Extension of the theory to equilibrium sampling

Insofar, the analysis of the mapping space has relied on a definition of a scalar product between CG representations based on a single, static structure of the reference protein. However, proteins are not static objects, but rather flexible entities which, in a typically aqueous environment, undergo fluctuations and deformations. It is therefore natural to extend the proposed metric to incorporate such structural variability.

The high-resolution (i.e. atomistic) system, constituted by the protein (whose atomic coordinates are indicated with $\mathbf{r}$) and its environment (indicated with $\mathbf{s}$), is assumed to be subject to an interaction potential $u(\mathbf{r}, \mathbf{s})$, such as the one in Eq. 1.1. In the canonical ensemble the probability density to sample a given configuration is proportional to the Boltzmann weight:

$$p_r(\mathbf{r}, \mathbf{s}) = \frac{e^{-\beta u(\mathbf{r},\mathbf{s})}}{Z} \tag{5.39}$$

where $Z = \int d\mathbf{r} d\mathbf{s}\ e^{-\beta u(\mathbf{r},\mathbf{s})}$ is the configurational partition function of the system.

The norm $\mathcal{E}$ of a mapping in Eq. 5.5 only depends on a single conformation of the molecule under examination; however, one can straightforwardly extend the definition of $\mathcal{E}$ to account for the whole conformational space sampled by the system:

$$\begin{aligned}
\langle \mathcal{E} \rangle &= \int d\mathbf{r} d\mathbf{s}\ p_r(\mathbf{r}, \mathbf{s})\ \mathcal{E}(\mathbf{r}) \\
&= \int d\mathbf{r} d\mathbf{s}\ p_r(\mathbf{r}, \mathbf{s})\ \frac{1}{\bar{z}(\mathbf{r})} \left(\sum_{i,j=1}^{n} e^{-r_{ij}^2/4\sigma^2} \chi_{M,i} \chi_{M,j} \right) \\
&= \sum_{i,j=1}^{n} \langle J_{ij} \rangle \chi_{M,i} \chi_{M,j}.
\end{aligned} \tag{5.40}$$

Note that the average is carried out both over the protein and environment degrees of freedom; at the same time, for mappings that only retain protein degrees of freedom, the couplings J_{ij}—and thus the norm $\mathcal{E}$—only depend on the latter. The linearity of the norm with respect to the couplings allows one to first compute their thermal average, that is,

$$\langle J_{ij} \rangle = \int d\mathbf{r} d\mathbf{s}\; p_r(\mathbf{r}, \mathbf{s}) J_{ij}(\mathbf{r}) = \int d\mathbf{r} d\mathbf{s}\; \frac{e^{-\beta u(\mathbf{r},\mathbf{s})}}{Z} \frac{1}{\bar{z}(\mathbf{r})} e^{-r_{ij}^2/4\sigma^2}, \tag{5.41}$$

and subsequently employ them for the calculation of norms, scalar products, and distances, in the same manner as it was done insofar. In this case, however, the resulting values entail information about the conformational space sampled by the whole system, including the environment, described in terms of a high-resolution model.

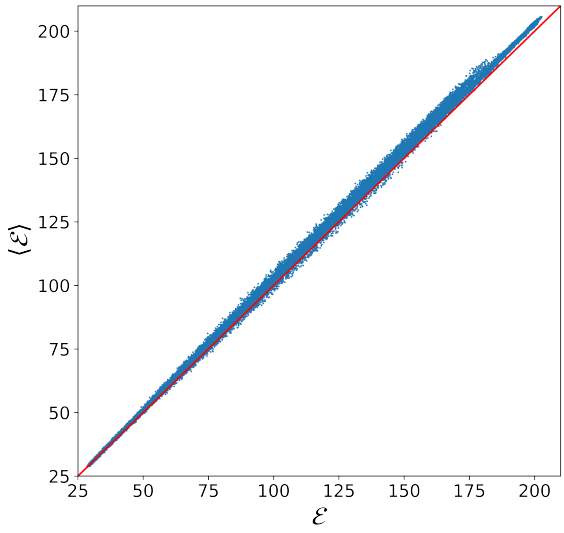


Figure 5.12: Scatter plot of the single-conformation mapping norm $\mathcal{E}$, calculated on the crystal structure of *4ake*, against its canonical average $\langle \mathcal{E} \rangle$ (Eq. 5.40) for 5×10^4 CG mappings with $N = 214$. The red straight line with slope one serves as a guide to the eye. The Pearson correlation coefficient is 0.9997. Image taken from Ref. [26].

Fig. 5.12 displays a comparison between the value of $\mathcal{E}$ computed on the crystal structure of *4ake* and its canonical average $\langle \mathcal{E} \rangle$ obtained through MD sampling. Each point in the plot represents a $\mathcal{E}$–$\langle \mathcal{E} \rangle$ pair out of 5×10^4 mappings with $N = 214$ extracted so as to homogeneously span all the

possible values of $\mathcal{E}$, see Sec. 5.1.2. The ensemble average is performed over 10^4 configurations of the 200 ns long NVT simulation described in Sec. 3.2.

Points are very narrowly dispersed along the diagonal, with a Pearson correlation coefficient very close to unity. This suggests that, at least in this case, the canonical average of $\mathcal{E}$ is robust to structural changes: this behaviour can be ascribed to the fact that at the outset of the simulation the protein is in its native state and, due to the strong constraints present in the molecule, the local environment of each atom generally performs small-amplitude fluctuations about a well-defined average. In this particular case the couplings computed explicitly accounting for the energetics of the system do not induce significant deviations in the value of the norm with respect to the static-structure values; it is hence reasonable to expect that the same will hold for the metric and topological properties of the mapping space.

However, this consistency will not be observed when secondary and tertiary structures heavily change, as in the case of protein folding: the value of $\mathcal{E}$ calculated over the unfolded polypeptide chain will not match its canonical average performed over a sample containing folded configurations.

5.5 Conclusions

This chapter aims at defining a measure of distance between two low-resolution representations of a macromolecule, and to "explore" the metric space induced by it. The metrics illustrated here has been employed to quantify the number, dissimilarity, and structural features of different mappings of a macromolecule in a static conformation. In this context, the proposed tool can be leveraged to perform a molecular-wise comparison between the several existing mapping prescriptions, that is, a quantitative version of the qualitative analysis carried out in Chapter 2 and in Ref. [23], in which CG representations were examined by visual inspection.

A further interesting aspect of this chapter is represented by its connections to a recent work by Foley et al. [156], in which the authors identify a phase transition between reduced representations on the basis of their *spectral quality*, a quantity related to the sum of the eigenvalues of the covariance matrix obtained integrating exactly a Gaussian network model (GNM). In the context of this chapter, the phase transition appears when looking at the mapping norm $\mathcal{E}$, which is not immediately interpretable as a measure of *quality* but rather of *sparsity* of a CG representation. What kind of relationship exist between the information entailed by these two observables? Is it possible to observe a phase transition when considering a more complex measure of quality, such as the mapping entropy?

Chapter 6

Resolution, Relevance and Mapping Entropy

Chapters 3 and 4 are devoted to the mapping entropy (Eq. 3.6), an observable capable of measuring the intrinsic information content of a coarse-grained representation. In order for this quantity to be computed from a fully atomistic MD trajectory, few approximations are employed, the most important one being the truncated cumulant expansion in Eq. 3.31.

Such approximations are necessary due to the incapacity of calculating the ingredients involved in the definition of the mapping entropy, which is the canonical average of the logarithm of $p_r/\bar{p}_r$ (see Eq. 3.10). Both these probability distributions are complicated to extract because of their high dimensionality and the numerical instabilities associated to the explicit calculations of the exponentials.

This chapter shows a way to retrieve an approximate, "frequentist" version of the atomistic probability without explicitly taking into account the Boltzmann weight of each configuration. This is obtained relying on the resolution-relevance framework proposed by Marsili and coworkers [246, 247, 27].

Given a trajectory of a protein containing L sampled configurations, they are lumped into an appropriate number of microstates C ($C < L$), each one populated by k atomistic configurations. Each choice of this "binning" defines a probability distribution over the sample, whose Shannon entropy can be immediately evaluated as:

$$H_s = -\sum_{i=1}^{C} \frac{k_i}{L} \ln\left(\frac{k_i}{L}\right). \tag{6.1}$$

This quantity, also called *resolution* [247], is a measure of the level of detail employed to describe the sample. A very gross description (few microstates,

wide binning) corresponds to a sharply peaked probability distribution, i.e., almost all instances fall inside the same bin, thus leading to very low values of resolution ($k_i \sim L,\ k_j = 0\ \forall j \neq i$). On the contrary, extremely detailed descriptions such that each data point is associated to a different microstate (or bin) assign a uniform probability to the bins ($\frac{1}{L}$), thus leading to the highest possible value of resolution, namely $H_s \sim \ln L$. Intuitively, one can state that the featurisations of the sample corresponding to these two extremal values of resolution are not particularly informative. Nevertheless, the resolution is, on average, monotonic with C, and it is impossible to make sense of this intuition based only on the value of H_s.

A number of works by Marsili and coworkers [246, 247, 27] have related the resolution to another information-theoretical quantity, that is, the entropy of the distribution of the frequencies, or *relevance*:

$$H_k = -\sum_{\mathcal{K}} \frac{km_k}{L} \ln\left(\frac{km_k}{L}\right) \tag{6.2}$$

where $\mathcal{K}$ is the set of unique frequencies observed in the sample of C microstates and m_k is the number of times the frequency k occurs:

$$m_k = \sum_{i=1}^{C} \delta_{k,k_i}. \tag{6.3}$$

The relevance is a Shannon entropy that measures the information content of the frequency distribution. In Ref. [27] Cubero et al. state that "a featurisation of the sample in terms of the frequency k provides a *minimally sufficient representation* of it, in the sense that the total information content of a sample can be divided as:

$$H_s = H_k + H_{s|k} \tag{6.4}$$

where $H_{s|k}$ is a measure of noise:

$$H_{s|k} = \sum_{\mathcal{K}} \frac{km_k}{L} \ln m_k \tag{6.5}$$

In absence of prior information, H_k is the *maximal* number of bits per unit point that has to be used to estimate the underlying generative process, and $H_{s|k}$ is a measure of noise."

More specifically, $H_{s|k}$ measures the ambiguity of the representation employed, i.e., the number of different classifications that produce the same distribution of frequencies m_k.

The intuitive lack of information associated to the two extremal values of resolution showcased above can be rationalised now, as they correspond to zero relevance, since $km_k = L$ in both cases. In particular, when the resolution is zero all configurations are lumped inside the same microstate i, thus implying that $k_i = L$ and $m_L = 1$; analogously, the maximum value of the resolution corresponds to a single configuration per microstate, namely $k_i = 1 \; \forall \; i$ and $m_1 = L$. The non-negativity of the entropy combined with Rolle's theorem allows one to conclude that the relevance displays a maximum.

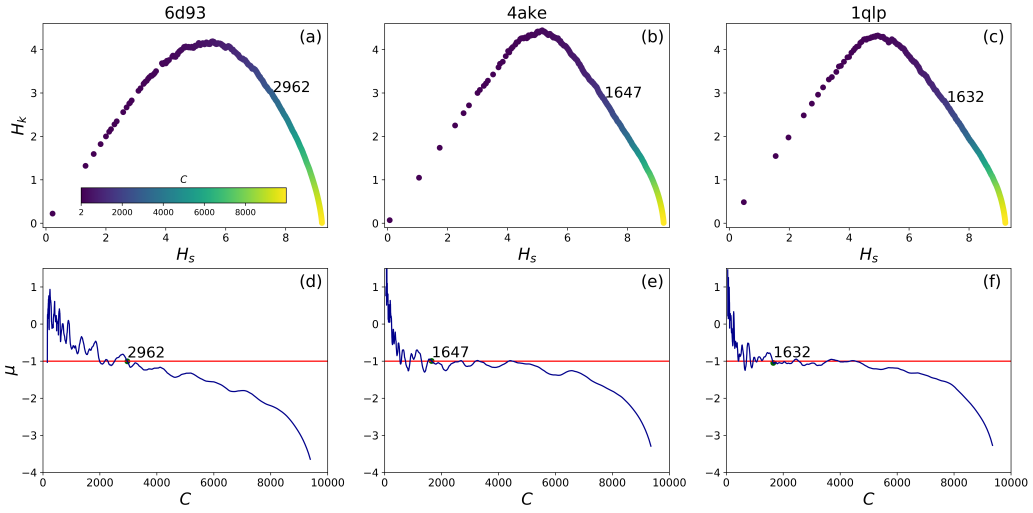


Figure 6.1: (a-c): Resolution-relevance plots for the trajectories of *6d93*, *4ake*, and *1qlp*. Each *all-atom* trajectory of $L = 10001$ frames has been clustered in 2000 different values of C, starting from $C = 2$ and ending with $C = 9997$, with an intermediate step equal to 5. The usual UPGMA algorithm (see Sec. 3.2.1) is applied over the all-atom RMSD matrix in order to perform the clustering. Each data point is colored according to the value of C employed to discretise the original trajectory. (d-f) : local slope μ of the H_s-H_k curve over all the spectrum of possible values of C. μ is computed by iteratively performing a linear regression over all values of the curve such that the resolution falls into an interval of amplitude $\frac{\ln L}{50}$. Such resolution window is iteratively moved from right to left by a factor $\frac{\ln L}{1000}$, until points with $H_s = H_k \sim 0$ are found.

When the resolution is low (well-sampled regime), m_k is almost never different from one, $H_{s|k}$ is negligible, and the relevance is almost equal to the resolution ($H_k \sim H_s$). At high values of H_s the situation changes, as the relevance *decreases* with a non-constant slope μ ($\mu = \mu(H_s)$) with increasing

resolution. In particular, Marsili and coworkers [248, 27] identify the point with $\mu = -1$ as especially interesting, since it provides the optimal trade-off between resolution and relevance. Starting from the rightmost point of a typical H_s-H_k plot (see Fig. 6.1(a-c)), where $H_s = \ln L$ and $H_k = 0$, it is possible to slowly reduce the resolution with which the system is treated, thus increasing the relevance. Going from right to left in the plot one can observe how, at the beginning, a reduction of one bit in resolution is compensated by a gain higher than one bit in relevance, until the critical point $\mu = -1$ is reached. After $\mu = -1$, increasing the relevance of one bit implies a major reduction in resolution.

With the aim of exploiting this separation between the regimes of lossless and lossy compression, I deem it appropriate to consider this critical point as the threshold that must be employed to extract microstates from the sample. This allows one to choose an optimal number of bins, $\bar{C}$, which is used to create an atomistic histogram of the collected configurations, where the probability of each microstate (bin) i is now given by the number of times it is observed in the sample:

$$p_r(\mathbf{r}_i) = \frac{k_i}{L} \tag{6.6}$$

The trajectory can now be reduced to $\bar{C}$ atomistic configurations, whose probability is dictated by Eq. 6.6. A CG representation $\mathbf{M}$ will induce an additional clustering on such reduced trajectory, lumping $\bar{C}$ microstates in K macrostates. At this point, the probability of a CG macrostate $\mathbf{R}$ ($p_R(\mathbf{R})$) can be immediately calculated by inserting Eq. 6.6 into Eq. 1.7:

$$\begin{aligned} p_R(\mathbf{R}) &= \int d\mathbf{r}\ p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \\ &\sim \sum_i \frac{k_i}{L}\delta(\mathbf{M}(\mathbf{r}_i) - \mathbf{R}). \end{aligned} \tag{6.7}$$

Knowing p_r, p_R and the multiplicity of the microstates mapping onto each macrostate ($\Omega_1(\mathbf{R}) = \sum_i \delta(\mathbf{M}(\mathbf{r}_i) - \mathbf{R})$), it is possible to calculate the mapping entropy in its original formulation:

$$\begin{aligned} S^{KL}_{map} &= \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})}\right] \\ &\sim \sum_i \frac{k_i}{L} \ln\left[\frac{k_i}{L\ \bar{p}_r(\mathbf{r})}\right] \end{aligned} \tag{6.8}$$

where the smeared probability of the microstate, $\bar{p}_r(\mathbf{r})$, is defined (Eq. 3.7) as the CG probability of the macrostate divided by $\Omega_1(\mathbf{R})$. For the sake of clarity the Boltzmann constant k_B is here omitted.

Fig. 6.1(a-c) reports the H_s-H_k dependence for the 200-ns long simulations discussed in Chapter 3 and in Ref. [24]. The three different proteins are associated to markedly different plots: *6d93* trajectory displays a flat maximum of the relevance, which remains constant over a wide range of values of H_s and C. The other two protein simulations are mapped onto a more bell-shaped curve, with the one concerning *4ake* displaying a steep decrease in H_k right after the maximum. The nature of this difference is certainly related to the hidden structure of the sample and to the properties of the clustering algorithm used to divide it in different bins. Further work is needed to rationalise this observation and to value its possible "relevance" (sic).

The calculation of the optimal $\bar{C}$ that separates the region with $\mu < -1$ from that with $\mu > -1$ is showed in Fig. 6.1(d-f). While the choice of the trade-off point is quite unambigous for *6d93* and *1qlp*, the curve of *4ake* displays several fluctuations in μ [1]. In this context, let me choose the first value of C after which $\mu < -1$ for a consistent set of values of C. Each trajectory of L snapshots is then converted into its reduced counterpart of $\bar{C}$ frames by choosing the first configuration of the sample belonging to each bin.

Let me now proceed to the calculation of the mapping entropies for a set of decimation mappings proper to the three proteins of interest, employing the original formula with the Kullback-Leibler divergence (Eq. 6.8). At first, the parameters reported in Chapter 3, Tab. 3.1 are employed to cluster the *reduced* trajectory of each protein into CG macrostates; then, the values of $\bar{p}_r(\mathbf{r})$ are calculated for each of the $\bar{C}$ atomistic microstates. Finally, it is possible to carry out the sum in Eq. 6.8, obtaining the overall value of S^{KL}_{map}.

Fig. 6.2 reports the comparison of the values of S^{KL}_{map} calculated according to this method with the ones of Σ, obtained using Eq. 3.47 for mappings with $N = N_\alpha$. The two data sets of 4968 and 1968 mappings employed in Chapter 4 for *6d93* and *4ake*, respectively ($N_\alpha = 31$, $N_\alpha = 214$), are taken as reference, while for *1qlp* ($N_\alpha = 372$) I consider the set of 548 random and optimised representations described in Chapter 3. Fig. 6.2 shows that a good correspondence exists between the two data sets of values S_{map} for *6d93*, while the bigger proteins *4ake* and *1qlp* do not display any significant correlation. A possible motivation for this behaviour could be the fact that the tamapin protein is already equilibrated after 200 ns of simulation, while the other two proteins are not sampling the equilibrium distribution. Moreover, configurations of the tamapin protein extracted every 20 ps are more decorre-

[1] The reason why *4ake*'s trajectory shows such a peculiar behaviour could be connected to the fact that the enzyme is continuously transitioning between two stable states throughout the simulation, thus inducing a peculiar frequency distribution at different values of C.

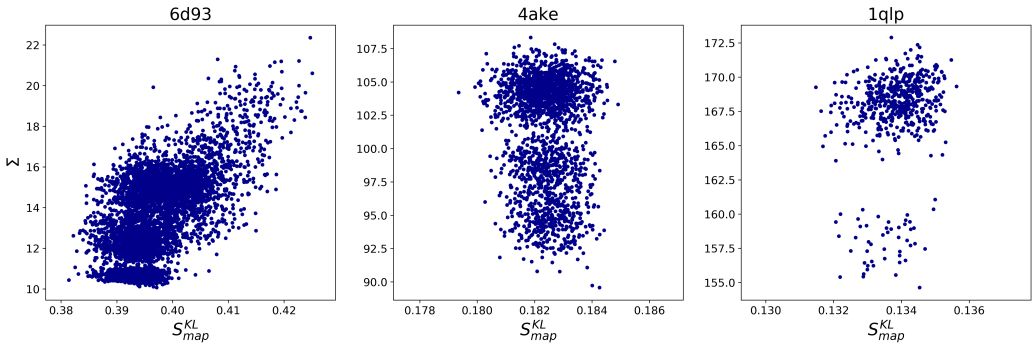


Figure 6.2: Comparison of the values of mapping entropy calculated using the original Kullback-Leibler formula (S^{KL}_{map}, Eq. 6.8) and the approximated expression of Eq. 3.47 (Σ). The small protein tamapin displays a clear correlation between the two expressions, which is totally absent for adenylate kinase and weaker in the case of $\alpha - 1$ antitrypsin. The Pearson correlation coefficients for the three cases are 0.62, -0.10, and 0.21, respectively.

lated than those of *4ake* and *1qlp*: sampling highly correlated structures can bias the probability distributions considered in this section. Furthermore, it is important to underline how the nature of the energy considered in the calculation of Σ can possibly play a role in this difference: indeed, Σ is computed employing only the protein-protein interaction energy, thus neglecting protein-solvent and solvent-solvent effects. Such approximation can give rise to a bias towards exposed regions, where the interactions are not properly screened. One of the strengths of S^{KL}_{map} is represented by the fact that the solvent contribution is taken into account more accurately by the probability. Overall, further work is needed to assess the nature of this discrepancy.

To conclude this section, 48 mapping optimisations are run for the *6d93* protein, exactly as explained in Sec. 3.3 and Ref. [24], but employing S^{KL}_{map} (Eq. 6.8) as the observable to be minimised. As in Sec. 3.3.3, one can perform a basic statistics over the pool of low-S^{KL}_{map} mappings by using the conservation probability of each atom, defined as the fraction of times it is included inside an optimised solution. Fig. 6.3 reports a comparison between the values of such probability distributed all over *6d93* obtained after this optimisation (P^{KL}_{cons}, Fig. 6.3(a)) or in the previous case (P_{cons}, Fig. 6.3(b), see Fig. 3.5). The first thing to notice in the figure is that P^{KL}_{cons} is more distributed along the protein sequence than its energy-based counterpart, due to the fact that, being solely based on the protein structure, the probability is smoother than the energy, as the latter can display stronger local fluctuations. Second, it is possible to see how the terminal atoms of the arginine residues (ARG6,

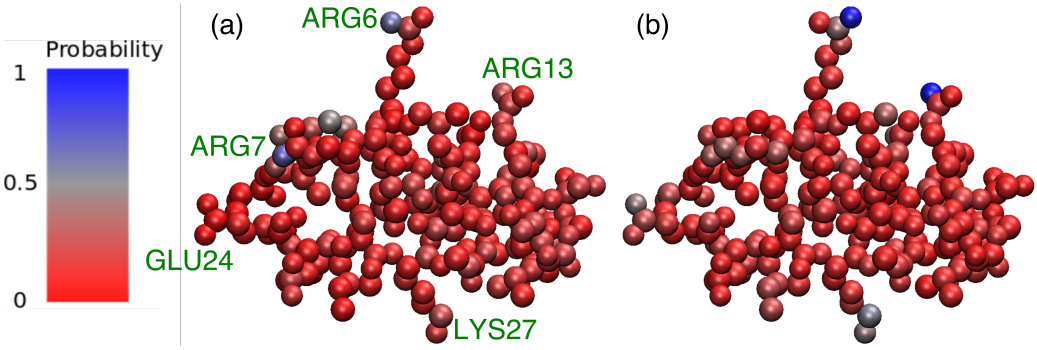


Figure 6.3: Comparison of the values of P^{KL}_{cons} (a) and P_{cons} (b) for mappings obtained minimising S^{KL}_{map} (a) or Σ (b). The optimised solutions employed in (a) display an average Z score (Eq. 3.51) of -3.81 ± 0.32, more negative than that reported in Tab. 3.3, corresponding to the low-Σ mappings used in (b).

ARG7, ARG13) are particularly conserved: even if the energetic fluctuations are not explicitly considered, the structural fluctuations proper of these atoms make them crucial for a highly informative mapping. The side chain of ARG6 is a paradigmatic example of this concept: in Sec. 3.3.3 (Fig. 6.3(b)) it is showed that an optimal CG mapping of *6d93* must contain the NH1 atom ($P_{\rm cons}(\mathrm{NH1, ARG6}) = 0.92$). Here, the atom with highest *importance* is NH2 ($P^{KL}_{\rm cons}(\mathrm{NH2, ARG6}) = 0.60$), but all the other atoms in the terminal region of the arginine display a non-negligible value of $P^{KL}_{\rm cons}$, namely 0.10, 0.23, 0.08 for NE, CZ, and NH1, respectively. The sum of these probabilities with the one associated to NH2 gives 1.02: except for two (resp. one) cases in which there are two (resp. zero) atoms of this region in the optimal mapping [2], all the remaining 46 optimal solutions contain exactly one atom in the terminal region of ARG6. As for ARG7 and ARG13, they display a similar behaviour, with the majority of the optimisations retaining one atom of their side chain terminus. In particular, the NH2 atom of ARG7 shows the highest value of $P^{KL}_{\rm cons}$ ($P^{KL}_{\rm cons}(\mathrm{NH2, ARG7}) = 0.67$).

Another interesting difference emerging from Fig. 6.2 concerns the re-

[2]By inspecting the optimisation run leading to the absence of any atom $\in$ {ARG6-NE, ARG6-CZ, ARG6-NH1, ARG6-NH2}, I notice that the minimum of the mapping entropy is obtained in the late stage of the Simulated Annealing protocol (step number 14347 out of 20000): after that, only 113 swapping moves are proposed that contain one of ARG6 terminal atoms. Of those, only 76 are unique, which amount at the 61% of the possible ones ($N_\alpha \times 4$). Therefore, it is possible that one of the remaining 48 combinations would have led to an accepted move and, consequently, to the inclusion of one of the atoms into the optimal mapping.

duced values of conservation probabilities assigned to terminal atoms of residues GLU24 and LYS27; while these atoms were usually part of low-Σ mappings, they are almost never present in the CG representations built minimising S_{map}^{KL}. GLU24 and LYS27 are charged residues, and the energetic fluctuations proper to the terminal atoms can be huge, especially when the considered energies are not screened by the solvent, as in Chapter 3. This is a further proof that S_{map}^{KL} is less biased towards solvent-exposed, charged residues than Σ.

	GLU24-CD	GLU24-OE1	GLU24-OE2	LYS27-CE	LYS27-NZ
P_{cons}	0.27	0.21	0.44	0.52	0.44
P_{cons}^{KL}	0.00	0.00	0.02	0.17	0.27

Table 6.1: Differences between the values of conservation probabilities for the terminal atoms of residues GLU24 and LYS27. The difference is striking especially for GLU24, as its terminal atoms are never conserved in the Kullback-Leibler-based optimisation.

Overall, it is possible to conclude that Fig. 6.2 (a) and Fig. 6.2 (b) are quite similar, with P_{cons}^{KL} that is, on average, more evenly distributed over the full structure, displaying a tendency to reduce the probability weight assigned to terminal atoms of charged residues with respect to P_{cons}.

In this subsection the properties of relevance and resolution are exploited in order to extract a set of atomistic microstates out of a Molecular Dynamics trajectory, each one weighted with its own approximated probability. The strategy presented here is general and unsupervised, being applicable to tasks other than the calculation of the mapping entropy; in this respect, it is worth mentioning that the proposed method can be employed to construct microstates of Markov State Models [174, 249] in a completely hyperparameter-free manner.

6.1 Discrete models

In the previous section I discussed how it is possible to recover an approximate atomistic probability distribution from a MD trajectory using the optimal trade-off between resolution and relevance. Here, the case of two discrete systems is analysed, in which a non-trivial, non-uniform atomistic probability p_r is immediately available, since the number of states accessible to the system, C, is smaller than the number of sampled states L.

6.1.1 A system of non-interacting spins

The numerical calculations associated to this model were carried out by my colleague Roi Holtzman, whom I here acknowledge for his crucial contribution to this chapter.

The first model system contains $n = 20$ non-interacting spins, each characterised by its probability to be in the "up" state. These spins are partitioned into two subsets of biased and non-biased spins. The first 10 spins are biased in a linear descending order according to $p_i(\sigma_i = 1) = 1 - (i-1)/20$ for $1 \leq i \leq 10$, while the last 10 spins are unbiased, namely $p_i(\sigma_i = 1) = 0.5$ for $11 \leq i \leq 20$, see Fig. 6.4.

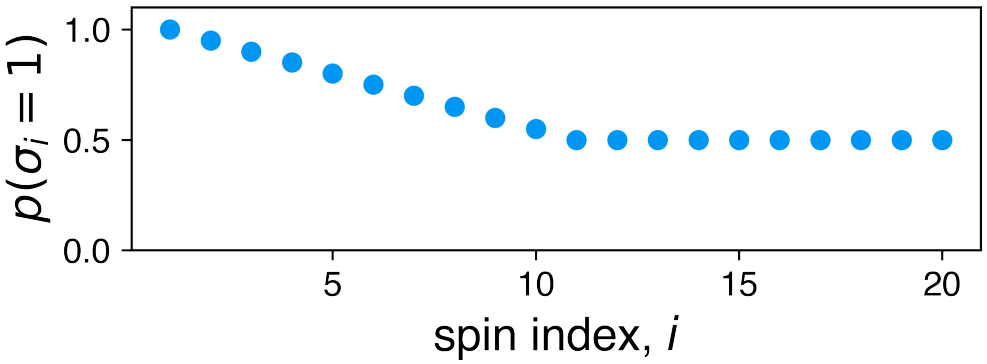


Figure 6.4: Probability of extracting the *up* configuration of each spin. The first spin is always sampled in the up state.

A simple calculation reveals how the number of possible states of the system is $2^{20} \sim 6 \times 10^6$, although many of them are almost impossible to observe due to their vanishing probability weight. As an example, half of these states will be never accessed, since it is impossible to observe the first spin in the *down* configuration. $L = 10^5$ fully atomistic configurations of the system are extracted by sampling the state of each spin according to its probability p_i, thus obtaining an *empirical* atomistic probability of sampling each state $\mathbf{r}$:

$$\begin{aligned} p_r(\mathbf{r}) &= \frac{1}{10^5} \sum_{j=1}^{10^5} \delta(\mathbf{r}_j - \mathbf{r}) \\ \lim_{L\to\infty} p_r(\mathbf{r}) &\to \prod_i^n p_i \end{aligned} \tag{6.9}$$

where the product in the second term represents the *analytical* probability, that would coincide with the empirical one in the case of infinite sampling.

I am now interested in the properties of the coarse-grained representations of this simple spin system, that is, those selections of N spins out of n. For each decimation CG representation, the corresponding resolution, relevance and mapping entropy are calculated and reported in Fig. 6.5(a-d).

One interesting aspect revealed by Fig. 6.5(a) is the range of resolution and relevance values for different numbers N of retained spins. CG mappings such that N is close to n display little variations in resolution and relevance, while an intermediate coarse-graining is associated with a wide range of values. Fig. 6.5(b) reports the results for the CG representations retaining only $N = 10$ sites. Configurations are ordered in a clustered structure that can be captured by introducing a rank for each representation, which quantifies the balance between biased and non-biased spins. The rank of a single spin σ_i is given by

$$\tilde{r}(\sigma_i) = \begin{cases} +1, & \text{if } 1 \leq i \leq 10 \\ -1, & \text{if } 11 \leq i \leq 20, \end{cases} \tag{6.10}$$

and the rank for a CG representation $\mathbf{M} = (\sigma_1, \ldots, \sigma_N)$ is given by the average of the rank over all retained spins, that is

$$r(\mathbf{M}) = \frac{1}{N} \sum_{j=1}^{N} \tilde{r}(\mathbf{M}_j). \tag{6.11}$$

The rank satisfies $-1 \leq r(\mathbf{M}) \leq 1$, and it is $r(\mathbf{M}) = 0$ when the number of retained biased spins equals the number of retained non-biased spins; in the limit case of $r(\mathbf{M}) = 1$ (resp. $r(\mathbf{M}) = -1$) all retained spins are biased (resp. unbiased).

Fig. 6.5(b) shows that CG configurations with positive rank provide higher relevance values, while low-rank CG configurations decrease in relevance and possess a higher resolution. This is a consequence of the uniform distribution of states induced by retaining all the unbiased spins, which is a maximum entropy distribution (high resolution, see Eq. 6.1) associated to a very low information content in the space of frequencies k (all the states are sampled with more or less the same frequency). In order to have higher relevance it is sufficient that the coarse-graining procedure induces macrostates with different populations, so it is sufficient that *the majority* of the spins are biased.

In Figs. 6.5(c,d) the dependence of the mapping entropy on the resolution is investigated. In contrast to the relevance, which tends to zero in the two limit cases of low and high resolution, the mapping entropy is monotonically decreasing with the resolution; when all of the spins are retained the CG probability is exactly equal to the atomistic distribution and there is no

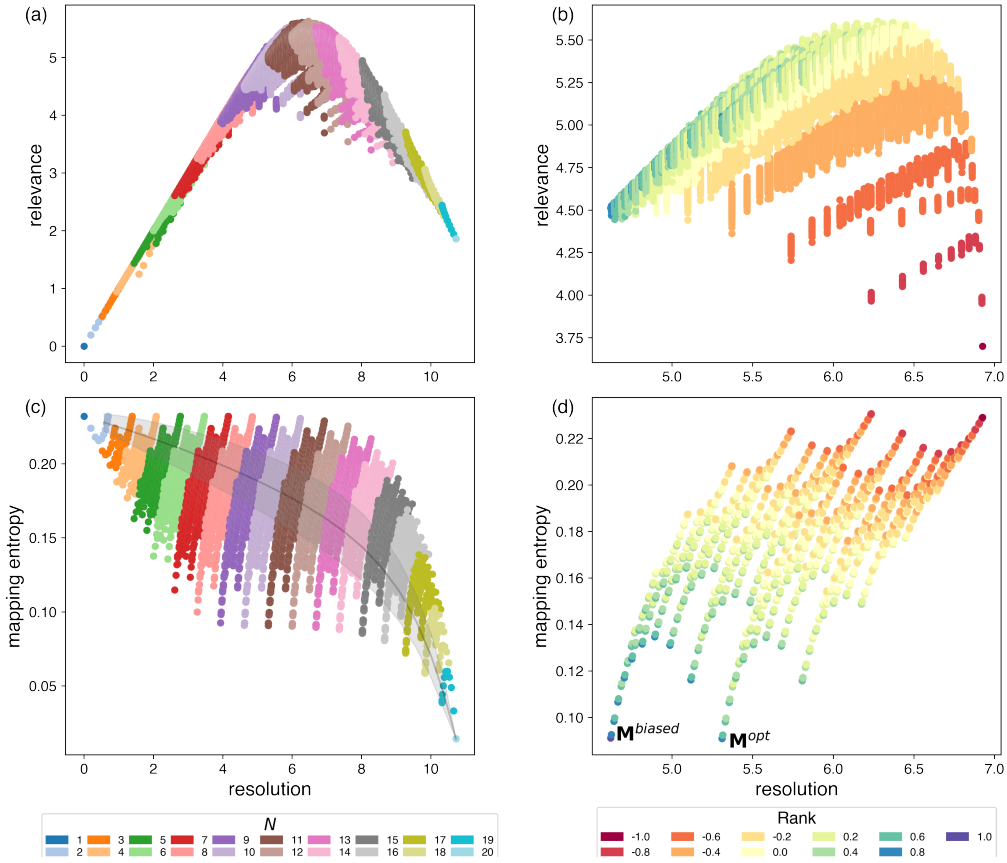


Figure 6.5: (a-d): resolution, relevance and mapping entropy for different coarse-grained representations of the system of non-interacting spins. (a) and (c) show how relevance and mapping entropy vary with increasing resolution. Each data point is depicted according to the number of conserved sites N. (b) and (d) report the values of relevance and mapping entropy in the case of $N = 10$, respectively. Points are colored according to their rank, as defined in Eq. 6.11.

coarse-graining in place; on the other hand, if only one spin is retained, the resulting CG probability is as far as it can be from the full-system probability. For some intermediate values of N it is possible to observe a large range of mapping entropy values, which depend on the specific choice of the CG configuration.

Figure 6.5(d) shows that minimal entropy values are obtained for high rank CG configurations, that is, those displaying non-uniform probabilities. This is in contrast to the insensitivity of the relevance to the rank, as long as it is non-negative. A closer look into the minimum mapping entropy values of Fig. 6.5(d) allows one to observe that the CG configuration (bottom left) with maximum rank ($\mathbf{M}^{\text{biased}}$) is not the absolute minimum of the mapping entropy, being overcome by a representation with $H_s \sim 5.3$, corresponding to a CG mapping in which the first spin is replaced by the 14-th. This is a consequence of the fact that the first spin ($p_1 = 1$) is not informative at all about the state of the overall system, while the 14-th spin can provide a minimal advantage due to the finiteness of sampling. More specifically, knowing the state of the first spin is not beneficial for decreasing the mapping entropy, as in all the atomistic microstates it is in the up state. Knowing the state of the 14-th spin provides a little advantage, given by the fact that it allows to *resolve* 2 more states s_1 and s_2 (it can be either up or down): this distinction provides an approximated $\bar{p}$ for these two states that is slightly more accurate than the (uniform) one obtained by retaining the first spin. Put differently, the fact that finite sampling gives similar but not equivalent probabilities for CG macrostates is tantamount to modifying the non-biased spins to slightly biased spins. This is a manifestation of the statistical nature of the empirical observation used to compute these quantities: in the case of fully analytical calculations (infinite sampling, see Eq. 6.9), the values of mapping entropy obtained by retaining all the spins from 2 to 9 and one of the eleven others would be exactly equal ($\bar{p}(s_1) = \bar{p}(s_2)$).

These considerations allow one to rationalise a feature of Fig. 6.5(c), namely the fact that the minimum value of the mapping entropy remains constant for a wide range of numbers of sites, that is, from $N = 9$ to $N = 16$, approximately. When $N = 9$, the minimum of this quantity is obtained for the CG mapping that retains the spins with index $i \in [2, 10]$, and adding other spins to this representation does not guarantee a substantial decrease in the mapping entropy, which is only obtained when the mapping gets closer to the fully atomistic representation ($N \geq 17$). At the same time, some mappings with $N = 18$ exist such that the associated mapping entropy is higher than the minimum value obtained when $N = 9$: these are coarse-grained representations that do not retain two of the biased spins.

In conclusion, a discrete system whose constituents are completely in-

dependent is analysed with the help of resolution, relevance and mapping entropy. These three variables shed light on some intrinsic features of the model, thus making them a promising candidate analysis tool for more complex systems.

6.1.2 A discrete model of financial markets

In the last section of this chapter an approximate model of financial markets is considered, in which the constituent elements are certainly interacting with a functional form that is completely unknown a priori. Common stock market indexes, such as NASDAQ-100, FTSE MIB, DAX 30, are usually defined as a function of the value of the X most traded stocks, or the ones with the highest market capitalisation. As an example the NASDAQ-100 index considers the largest non-financial companies listed on the Nasdaq stock market [250]. It is well-known [251, 252] that changes in the composition of such indexes have an impact on the stock prices, temporarily favoring the stocks that are added to the index.

These indexes can be considered as coarse-grained mappings of the high-resolution system, i.e., the full stock market, to a lower number of degrees of freedom. The natural question that arises is the following: are these indexes always appropriate to coarse-grain the full market? Can one find a different subset of stocks that bring more information about the high-resolution system?

The analysis is restricted to the stocks with the highest market capitalisation (at the date 1/10/2021) in the NASDAQ-100 index, which are described in Tab. 6.2. The first model, *m1*, considers the 10 non-underlined entries of the table as the high-resolution system, while in the second model (*m2*), the two remaining stocks are included, namely CSCO and NTES. The values of these stocks are investigated over a ten year time window, for a total of 2225 days of sampling considered. For each day, a stock can assume three discrete values (see Fig. 6.6), namely $+1$ if the stock value increases during the day, 0 if it's stationary and -1 if it decreases. In this way the full market is mapped to a system of *interacting* spins with 3^{10} (3^{12}) available configurations. As in the non-interacting case discussed in the previous section, many of these are impossible to observe in a pool of real configurations: imagine for example how unlikely it is that 12 stocks of this importance are stationary in the same day. Indeed, it is possible to observe only 630 (1148) configurations of the system in the available sampling. As in the previous subsection, it is possible to define the atomistic probability as the number of times a full-system configuration is observed divided by the number of days (Eq. 6.6).

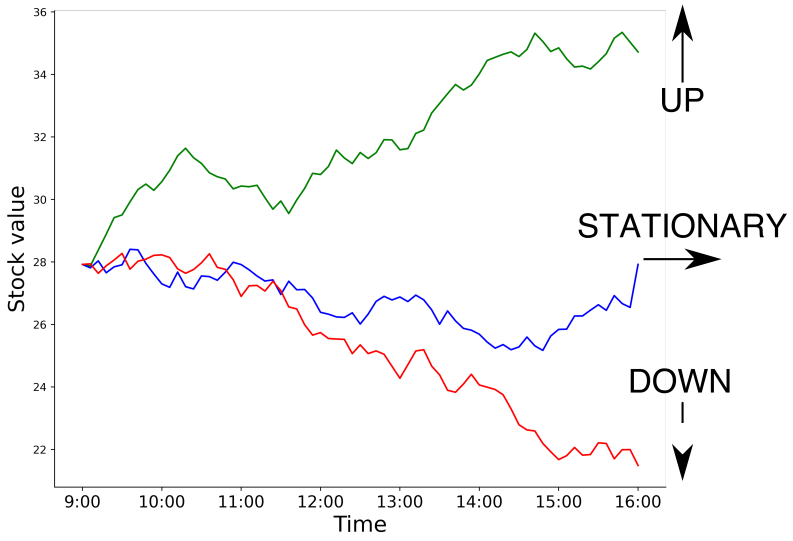


Figure 6.6: A pictorial representation of the prescription used to build the data set of the discrete model of financial markets illustrated in this section. If the stock value V grows (decreases) during the day with respect to its starting value, that is, if $V_{final} > V_{start}$ ($V_{final} < V_{start}$), a spin *up* (*down*) is assigned to the company for the specific date. If the two values coincide ($V_{final} = V_{start}$), the date is labelled as *stationary* for the considered stock.

At this point one aims at exhaustively investigating the behaviour of resolution, relevance and mapping entropy for all the 2^9 (2^{11}) CG decimation mappings that can be defined on the two models. Fig. 6.7 reports the values of these three quantities for all the coarse-grained representation of the two high-resolution systems.

First, it is possible to notice how the baseline value of the mapping entropy for highly coarse-grained mappings ($N = 1$) is higher in the case of *m1*. This is a consequence of the sampling, as in the case of *m1* there are fewer spins and a lower number of atomistic configurations (630).

Second, for each value of $N \neq 1, n$ there exist two clouds of points separated by a huge gap in resolution. A manual inspection of the data shows that, at fixed N, the low-resolution cloud of mappings is characterised by a common trait: all these representations retain both GOOG and GOOGL. As expected, these two stocks are highly interacting and correlated, displaying the same value in the 94.3% of the selected time window. Therefore, it is reasonable that, for each N, a representation containing both Google stocks

Symbol	Name	# ↓	# →	# ↑
AAPL	Apple Inc. Common Stock	1060	5	1160
ADBE	Adobe Inc. Common Stock	1017	5	1203
ADI	Analog Devices, Inc. Common Stock	1090	13	1122
CSCO	Cisco Systems, Inc. Common Stock	1022	29	1174
GOOG	Alphabet Inc. Class C Capital Stock	1069	1	1155
GOOGL	Alphabet Inc. Class A Common Stock	1075	2	1148
IDXX	IDEXX Laboratories, Inc. Common Stock	977	8	1240
MSFT	Microsoft Corporation Common Stock	1048	24	1153
NFLX	Netflix, Inc. Common Stock	1110	1	1114
NTES	NetEase, Inc. American Depositary Shares	1095	4	1126
NVDA	NVIDIA Corporation Common Stock	1078	15	1132
TSLA	Tesla, Inc. Common Stock	1111	3	1111

Table 6.2: Nasdaq stocks considered in this subsection. # ↓, # →,and # ↑ represent the number of down, stationary and up "spins" for each stock during the available sampling time, respectively. CSCO and NTES are absent in *m1* and are included in *m2*. Data were downloaded using *yfinance* [253], a python package to download Yahoo! finance data. Companies for which there are no data for all the considered dates were immediately excluded from the analysis.

provides a low-resolution coarse-graining of the system, comparable to the resolution of a coarse-grained representation with $N-1$ stocks. In principle, this observation does not have an impact on the mapping entropy. In Fig. 6.7(c-d) it is possible to appreciate how the choice of the model influences the average value of mapping entropy of the two clouds. When considering *m1* (Fig. 6.7(c)), mappings containing both Google stocks (on the left cloud at constant N) display an average mapping entropy equal or lower than the one of the other mappings. The situation changes when observing Fig. 6.7(d), where the data come from *m2*: since two additional stocks are included, the atomistic probability is less biased by the presence of Google instances and the mapping entropy of representations containing both GOOG and GOOGL is consistently higher than that of the other mappings. Intuitively, one of the two Google stocks possesses a high level of information about the system, but the inclusion of both of them in a coarse-grained description of the full market is redundant.

The third interesting aspect revealed by an inspection of Fig. 6.7(c-d) is that all the mappings retaining TSLA and NVDA display a value of mapping entropy higher than the average (see Fig. 6.8). In particular it is possible to observe that, in both models, (i) when $N = n-2$ the mapping

with lowest mapping entropy is the one that *does not contain* TSLA and NVDA; (*ii*) when $2 \leq N \leq n-2$ the mapping with highest mapping entropy retains TSLA and NVDA. Hence, retaining these two stocks is detrimental for a correct coarse-grained description of the market. A possible explanation to this behaviour can be found in the fact that both of them have been of marginal importance to the market for a vast majority of the sampling time (10 years), having experienced an exponential growth only in the latest years. In the case of TSLA, the corresponding company operates in a field that is neatly separated from the other stocks reported in Tab. 6.2.

As for maximally informative mappings, that is, those with minimal mapping entropy, it is possible to observe that GOOG, MSFT, and NFLX appear to be always conserved in particularly informative representations. In particular, when $3 \leq N \leq n-1$, the mappings displaying the lowest value of mapping entropy at fixed N always include the combination of these three stocks in both models. The reason behind the high informativeness of these companies can be attributed to their long-time, dominant presence in the stock market.

Let me conclude this analysis by noting that the interacting case does not display the flatness in the mapping entropy minima that was observed in Fig. 6.5(c). In this context, adding a new site to an optimal coarse-grained mapping always results in a gain of information about the high-resolution system.

This section shows that, even considering two simple models, the mapping entropy proves to be an invaluable instrument for the understanding of complex systems, being able to differentiate between informative and non-informative features in a precise and completely unsupervised manner. It is my opinion that this mathematical quantity can be employed in a huge variety of scenarios, either as a feature selection and ranking algorithm or as an analysis tool, able to provide the user with objective, unbiased data on the information content possessed by its constituent elements.

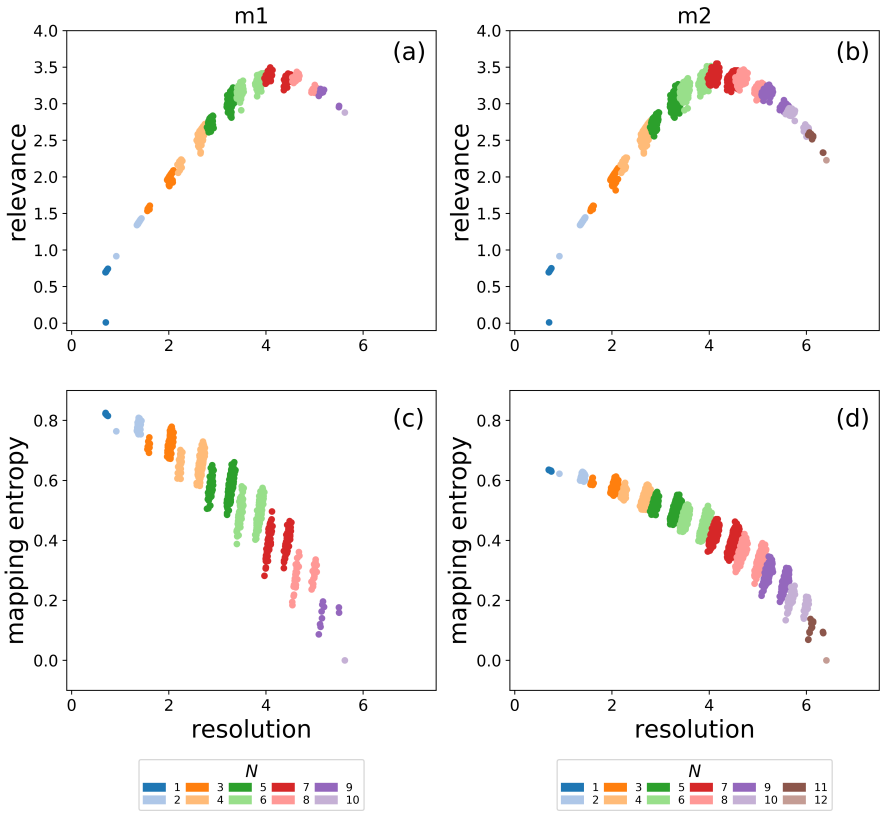


Figure 6.7: Resolution, relevance (a-b) and mapping entropy (c-d) for the two models. Mappings in *m2* can reach high values of resolution because adding information (two stocks) allows to define a higher number of atomistic configurations out of the available sampling. In (a-b) there exists a CG mapping with $N = 1$ possessing a very low value of relevance ($H_k \sim 0$); this is the mapping that retains TSLA stock: by chance, the number of spins in the up and down configurations coincide (see Tab. 6.2).

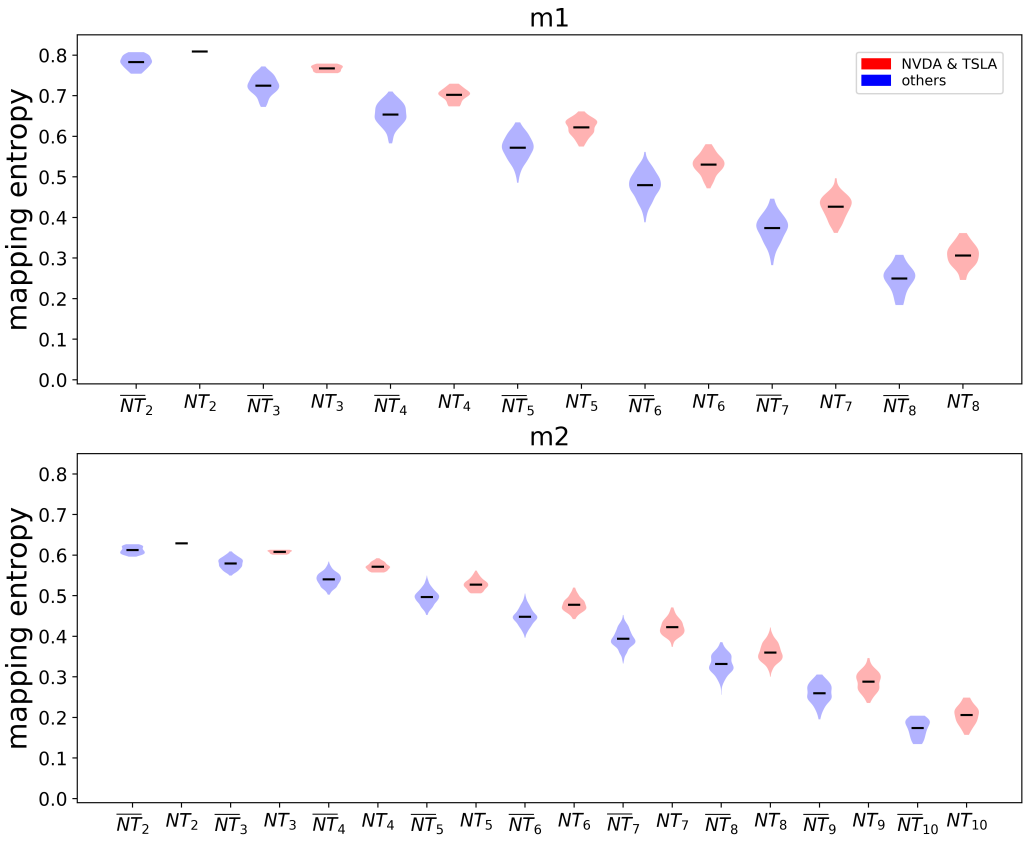


Figure 6.8: Violin plot of the values of mapping entropy for coarse-grained mappings that either retain (red distributions, labelled with NT) or not (blue distributions, labelled with $\overline{NT}$) both TSLA and NVDA stocks. The number of sites N, ranging from 2 to $n-2$, is indicated as a subscript of each entry. Black bars indicate the mean values of each distribution, showing how the combination of TSLA and NVDA is detrimental for the mapping. The plot is created using Matplotlib [254], and distributions are smoothed applying the "scott" criterion to Gaussian Kernel Density Estimation.

Chapter 7

EXCOGITO: an EXtensible COarse-GraIning TOol

In this seventh and last chapter of this work I provide a brief description of EXCOGITO, a software suite built to perform the calculations illustrated in this thesis, especially those discussed in chapters 3, 5, and 6. EXCOGITO is available at the following address: https://github.com/potestiolab/excogito.

Given its modular architecture, EXCOGITO is specifically designed to be easily extensible, so as to possibly incorporate different coarse-graining algorithms, paying special attention to those that employ a non-trivial reduced representation of the system. The program is entirely written in C, with few python scripts to help the user setting up and preprocessing the initial data. In particular, the user is guided in a step-by-step procedure throughout the generation of a *.ini* parameter file containing mandatory and optional parameters that must be provided to the core routines of EXCOGITO (Tab. 7.1).

Currently, EXCOGITO contains the following nine subprograms fulfilling nine different tasks:

1. *optimize*: a mapping optimisation run resulting in `Ncores` (Tab. 7.1) local minima. The *approximated* mapping entropy (Eq. 3.41) is minimised in the space of coarse-grained mappings with the Monte Carlo Simulated Annealing algorithm (see Sec. 3.3). The user can choose the number of Monte Carlo steps, start temperature and decay parameter of Eq. 3.50 (Tab. 7.1). The number of minima `Ncores` has to be lower or equal to the number of CPU cores of the employed architecture, since each of them carries out a single optimisation;

2. *random*: generation of `n_mappings` (Tab. 7.1) and measurement of the corresponding values of S_{map} (Eq. 3.41). This task is useful when one

Parameter	Description	Type	Mandatory
`frames`	number of frames in the trajectory	int	all
`atomnum`	number of atoms in the system	int	all
`cgnum`	number of CG sites	int	all
`nclust`	number of CG macrostates	int	C0 - C3
`n_mappings`	number of mappings	int	R-D
`MC_steps`	number of SA steps in task optimize	int	O
`rotmats_period`	SA steps between two alignments	int	O
`t_zero`	start temperature for task optimize	float	no
`criterion`	criterion for clustering (see Sec. 7.1)	int	O-R-M
`distance`	cophenetic distance threshold	float	C1
`max_nclust`	higher number of CG macrostates	int	C2
`min_nclust`	lower number of CG macrostates	int	C2
`Ncores`	number of cores to employ	int	no
`decay_time`	temperature decay in SA	float	no
`rsd`	use rsd instead of rmsd	int	no
`stride`	distance between pivot points	int	C3

Table 7.1: List of parameters of EXCOGITO. In the "mandatory" column, *all* (*no*) indicates parameters that are always (never) mandatory, while O, R, and M refer to parameters that are mandatory only for *optimize*, *random*, and *measure* (including the *kl* counterparts) tasks, respectively. C0, C1, C2, C3, C4 correspond to the different clustering criteria (Sec. 7.1): for example, if the selected `criterion` is 2, parameters `min_nclust` and `max_nclust` must be present.

wants to compare the values of mapping entropy of optimal mappings to those of coarse-grained representations randomly drawn from the mapping space, as it is done in chapters 3 and 4;

3. *measure*: the user provides a mapping to EXCOGITO in the form of a text file (a prototype is available in the examples) and the associated mapping entropy (Eq. 3.41) is computed;

4. *norm*: given a mapping and a full-atom MD trajectory, the time-evolution of the squared mapping norm (Eqs. 5.14 and 5.19) is calculated;

5. *cosine*: given two mappings and a full-atom MD trajectory, the time-evolution of the cosine (Eq. 5.16) between them is calculated;

6. *distance*: given a set of `n_mappings` (Tab. 7.1) coarse-grained map-

pings and a single configuration of a biomolecule, the distance matrix between them is computed using Eq. 5.19. Such matrix can be employed for several purposes, such as the calculation of the sketch maps as in Sec. 5.3;

7. *optimize_kl*: analogous to *optimize*, but using the Kullback Leibler version of the mapping entropy (Eq. 6.8). More specifically, the user provides EXCOGITO with a set of atomistic configurations, together with the associated, non-uniform probabilities. A further clustering on this set of microstates divides the conformational space in CG macrostates, each one possessing a probability given by Eq. 6.7. For each microstate, $p_r(\mathbf{r}) \ln\left(\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})}\right)$ measures the discrepancy between its probability and the smeared one;

8. *random_kl*: the Kullback-Leibler version of task *random*;

9. *measure_kl*: the Kullback-Leibler version of task *measure*;

7.1 Clustering the conformational space

As the computation of the mapping entropy implies the identification of coarse-grained macrostates out of a pool of atomistic microstates, all subprograms except *norm*, *cosine* and *distance* require a prescription to perform this clustering procedure. The latter is specified by the parameter `criterion` (Tab. 7.1), which can assume four values, each one associated to a slightly different choice for the clustering. Note that the software extensibility is evident in this context, as one can independently choose and deploy another clustering algorithm, that is incorporated in EXCOGITO specifying another value for the `criterion` parameter.

All the available criteria employ average linkage (Eq. 3.45), agglomerative hierarchical clustering (UPGMA [198], see Sec. 3.2.1) to divide the space of configurations sampled by MD in configurational clusters. The algorithm is written following the implementation [199, 200] provided by *scipy* [201, 202]. As described in chapter 3, the RMSD^{CG} (Eq. 3.44) is used as similarity measure between different conformations.

Essentially, these four criteria employ different prescriptions to cut the dendrogram.

- `criterion` = 0 analogously to the *maxclust* criterion in *scipy*, a fixed number of coarse-grained macrostates is retrieved. The dendrogram is

cut when the number of clusters matches the input parameter `nclust` (Tab. 7.1);

- `criterion` = 1, corresponding to the *distance* criterion in *scipy*: the number of coarse-grained macrostates is not fixed, but rather determined by the cophenetic distance. More specifically, the algorithm cuts the dendrogram when MD configurations in each cluster possess a cophenetic distance lower than the input parameter `distance` (Tab. 7.1). This choice is effectively employed in Sec. 3.4 in order to observe the scaling of S_{map} with the number of CG sites. In the latter context the `rsd` parameter must be set to 1 to exploit the unweighted RMSD as a similarity measure between CG structures;

- `criterion` = 2, that is the iteration of `criterion` = 0 for five integers between input parameters `min_nclust` and `max_nclust` (Tab. 7.1). This prescription is used to compute Σ (Eq. 3.47) in Refs. [24, 25] (see chapters 3 and 4), with the purpose of increasing the robustness of the Simulated Annealing procedure devoted to the mapping optimisation.

  A pictorial representation of criteria 0, 1, and 2 is sketched in Fig. 7.1.

- `criterion` = 3: a fast version of `criterion` = 0 that can be used only when a continuous trajectory is provided in input. In this case, the algorithm computes the pairwise $\mathrm{RMSD}^{\mathrm{CG}}$ matrix between a subset of the overall configurations of the trajectory, that is, one every `stride` (Tab. 7.1) configurations. For example, if `frames` = 101 and `stride` = 50, only "pivot" configurations number 1, 51 and 101 are considered in the pairwise alignments. Subsequently, standard hierarchical clustering applied to this reduced matrix assigns the coarse-grained macrostate to each *pivot* configuration. Then, the remaining data points are labelled using a simple prescription: if the previous and following pivot configurations possess the same label, the latter is assigned to all the intermediate structures. Instead, if the two pivot points have been labelled differently by the clustering algorithm, each intermediate structure is assigned to the same cluster of the closer pivot, that is, the one corresponding to the lower $\mathrm{RMSD}^{\mathrm{CG}}$. This approximation guarantees a substantial speed-up to the overall calculation, as the computation of the $\mathrm{RMSD}^{\mathrm{CG}}$ matrix and the following clustering are the most cumbersome tasks, scaling quadratically with the number of frames of the trajectory. More specifically, given a certain value of `frames`, f, and `stride`, s, the overall number of pairwise alignments,

N_a, in the worst case scenario is given by:

$$N_a = \frac{N_p(N_p - 1)}{2} + 2(f - N_p) \tag{7.1}$$

where $N_p = \frac{f}{s} + 1$ is the total number of pivot points. As for the clustering procedure, its high computational cost ($\mathcal{O}(f^2 log f)$) (see Sec. 3.2.1) makes this `criterion` extremely appealing. As an example, $s = 10$ corresponds to a speed-up factor approximately equal to 300. This procedure is schematically illustrated in Fig. 7.2, where the computational gain arising by employing this `criterion` is made evident by the shrinkage of both RMSD^{CG} matrix and dendrogram.

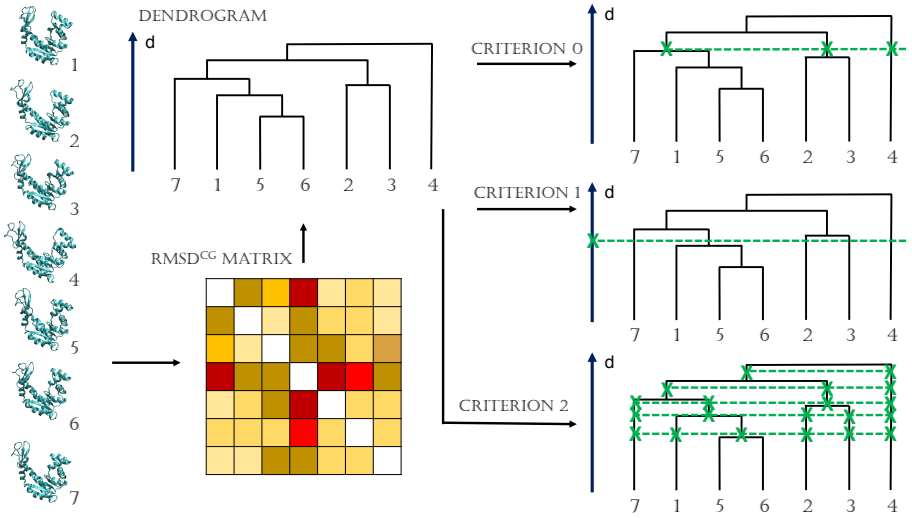


Figure 7.1: Schematic representation of criteria 0, 1, and 2 for conformational clustering. These are equivalent in the first stage of the procedure, where a RMSD^{CG} matrix is calculated between all the configurations (`frames`, see Tab. 7.1) of a full-atom MD trajectory, observed through the glasses of a CG mapping. From this typically large matrix, the full dendrogram is constructed using the average linkage prescription. Then, conformational clusters can be selected in three manners, namely 0) cutting the dendrogram when `nclust` (equal to 3 in this case) leaves are present; 1) cutting the dendrogram when a certain value of cophenetic `distance` (on the ordinate) is reached, irrespectively of the number of leaves; 2) applying the procedure 0 for a set of 5 evenly spaced values of the number of clusters ($\{2, 3, 4, 5, 6\}$ in this case), determined by parameters `min_nclust` and `max_nclust` (2 and 6 in this figure).

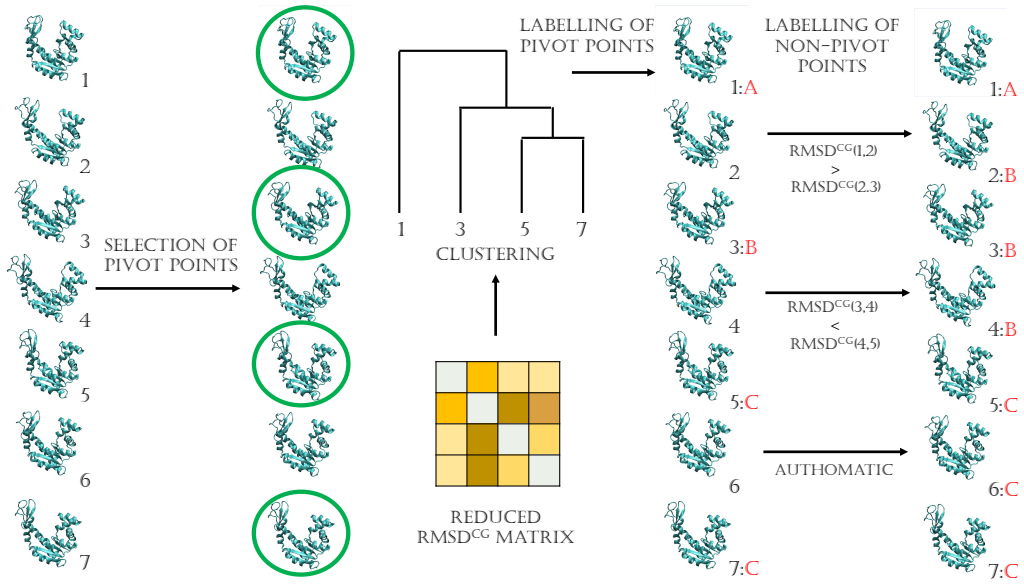


Figure 7.2: Graphical description of `criterion` 3 for an accelerated clustering of the conformational space. The `stride` parameter (Tab. 7.1) is equal to 2 in this case, meaning that 4 pivot points are considered. The reduced RMSD^{CG} matrix and dendrogram are computed taking into account only the coordinates of the selected conformations. Upon clustering, labels of the non-pivot points are assigned based on their proximity with respect to the two closest pivots. If the latter share the same label, as it is for configurations 5 and 7 in this example, the intermediate structures are automatically labelled.

7.2 Usage, supported platforms, and requirements

The README file of EXCOGITO provides all the necessary details to compile and run the calculations. In addition, the PDF documentation created with *doxygen* is available at
https://github.com/potestiolab/excogito/blob/master/docs/refman.pdf.

EXCOGITO must be compiled with *CMake* (minimum version 3.15) [255]. The software has been tested for several compilers, such as *gcc* [256] (versions 4.8.5, 5.4.0, 7.5.0, and 9.1.0), Intel C compiler *icc* [257] (version 19.0.0.20181018) and *Clang* [258] (version 10.0.1.10010046). The only mandatory requirement is to have the *openmp* library [259] installed on the machine. *Openmp* is already included in the majority of compilers, such as *gcc* and *icc*.

Conclusions

In this work a series of approaches have been proposed, whose aim is to shed some new light on the concept of mapping in the field of coarse-grained modelling of proteins.

Nowadays, low-resolution, CG models represent powerful tools to examine biological systems whose relevant time and length scales exceed those easily reachable by more detailed, atomistic simulations. Indeed, the reduction of the number of degrees of freedom operated by the CG procedure guarantees a substantial speed-up of the calculations. In this context, particular attention has been dedicated to the construction of accurate CG force fields, that is, effective interactions among the "survived" degrees of freedom that allow the CG model to reproduce the properties of the high-resolution system, where the latter is observed at low resolution. On the contrary, the choice of which degrees of freedom "deserve" to be included in the CG model, namely the mapping, has not been investigated with the same intensity: in most cases the mapping is imposed by the user based on system knowledge, chemical intuition, and trial and error procedures.

In this thesis I have illustrated the properties of the mapping entropy, an information-theoretical measure that determines how distant the probability distribution of the low-resolution system, dictated by the choice of the CG mapping, is from the atomistic reference. Such *Kullback-Leibler* distance naturally implies a ranking among CG mappings, with the ones possessing low mapping entropy, i.e., retaining the maximum amount of information about the system, that should be preferred by the modeller. Following this consideration, I have proposed a strategy to unsupervisedly optimise the CG mapping over a small data set of candidate proteins. Importantly, the solutions of this optimisation problem show significant similarities, most notably the fact that amino acids retained with high probability are those known to be crucial for the biological function of the protein and for the interactions with the substrate. This amounts at saying that "I learned something about the atomistic system thanks to (good) CG mappings", which is something that can be considered a change of paradigm in the field of CG modelling:

the mapping should not be considered anymore as a mere, external prescription fed to the model, as an astute choice of the CG representation alone can already provide valuable information about the properties of the fully atomistic system.

The computational burden associated to the computation of the mapping entropy for biomolecules makes the exhaustive exploration of this observable's landscape hardly doable. I showed how modern deep learning techniques allow one to accurately approximate this quantity, guaranteeing a dramatic speed-up with respect to the reference calculations and making a thorough exploration of the space of mappings affordable. This quasi-comprehensive inspection of the mapping space called for the introduction of tools to quantify the statistical properties of and the relationships among its elements: in this thesis, a purely structural measure of distance between CG mappings was introduced, which enables to measure and rationalise the differences between intrinsically separated CG representations.

The mapping entropy is a powerful tool to rank reduced representations of systems other than biological molecules. In this respect, preliminary studies carried out on toy models of spin systems and financial markets show that, once again, an analysis of low-mapping entropy CG representations sheds light on the behaviour of the high-resolution systems. In this context, the intrinsically multi-body nature of the mapping entropy leads to the emergence of multi-body correlations between the existing variables. These results lay the foundations for the application of the mapping entropy in data science, either as a feature selection algorithm or as a novel instrument of analysis of complex data sets. The first use is analogous to the mapping definition in CG, that is, a smart prescription to be implemented *prior to the modelling*. The second application is more intriguing, as it suggests that the process of dimensionality reduction *per se* can provide information on high-dimensional data sets.

It is my opinion that the application of the mapping entropy outside of the realm in which it was conceived could pave the way to the use of coarse-grained methodologies as analysis tools. In a world full of data, where the production and storage of tremendous quantities of information are increasingly cheaper, the elegance and generality of CG models promise to be extremely useful for making sense of them.

Appendix

On figures and plots

The figures reported in this manuscript have been generated using Matplotlib [254] or gnuplot [260] for standard plots, and Visual Molecular Dynamics (VMD) [261] for figures containing biological structures.

On the analysis of data

All the calculations reported in this thesis have been carried out using C for data generation (see Chapter 7) and python for data analysis. In the latter context, the packages MDAnalysis [262, 263] and MDTraj [264] have been employed for dealing with biomolecular data, while I strongly relied on Numpy [265], Pandas [266] and Scipy [199, 200] for pre- and post-processing tasks.

Bibliography

[1] K. Wüthrich, "Protein structure determination in solution by nmr spectroscopy.," *Journal of Biological Chemistry*, vol. 265, no. 36, pp. 22059–22062, 1990.

[2] A. Ilari and C. Savino, "Protein structure determination by x-ray crystallography," *Bioinformatics*, pp. 63–87, 2008.

[3] K. Murata and M. Wolf, "Cryo-electron microscopy for structural analysis of dynamic biological macromolecules," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1862, no. 2, pp. 324–334, 2018.

[4] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics. i. general method," *The Journal of Chemical Physics*, vol. 31, no. 2, pp. 459–466, 1959.

[5] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nature structural biology*, vol. 9, no. 9, pp. 646–652, 2002.

[6] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí, "Molecular dynamics simulations: advances and applications," *Advances and applications in bioinformatics and chemistry: AABC*, vol. 8, p. 37, 2015.

[7] S. Bottaro and K. Lindorff-Larsen, "Biophysical experiments and biomolecular simulations: A perfect match?," *Science*, vol. 361, no. 6400, pp. 355–360, 2018.

[8] A. N. Naganathan and V. Muñoz, "Scaling of folding times with protein size," *Journal of the American Chemical Society*, vol. 127, no. 2, pp. 480–481, 2005.

[9] A. Gershenson, S. Gosavi, P. Faccioli, and P. L. Wintrode, "Successes and challenges in simulating the folding of large proteins," *Journal of Biological Chemistry*, vol. 295, no. 1, pp. 15–33, 2020.

[10] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12562–12566, 2002.

[11] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling," *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, 1977.

[12] L. Maragliano and E. Vanden-Eijnden, "A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations," *Chemical physics letters*, vol. 426, no. 1-3, pp. 168–175, 2006.

[13] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical physics letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[14] J. G. Kirkwood, "Statistical mechanics of fluid mixtures," *The Journal of chemical physics*, vol. 3, no. 5, pp. 300–313, 1935.

[15] A. Liwo, M. Khalili, and H. A. Scheraga, "Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains," *Proceedings of the National Academy of Sciences*, vol. 102, no. 7, pp. 2362–2367, 2005.

[16] C. Clementi, "Coarse-grained models of protein folding: toy models or predictive tools?," *Current opinion in structural biology*, vol. 18, no. 1, pp. 10–15, 2008.

[17] L. F. Signorini, C. Perego, and R. Potestio, "Protein self-entanglement modulates successful folding to the native state: A multi-scale modeling study," *The Journal of Chemical Physics*, vol. 155, no. 11, p. 115101, 2021.

[18] Y. Wang, A. Rader, I. Bahar, and R. L. Jernigan, "Global ribosome motions revealed with elastic network model," *Journal of structural biology*, vol. 147, no. 3, pp. 302–314, 2004.

[19] H. Gohlke and M. F. Thorpe, "A natural coarse graining for simulating large biomolecular motion," *Biophysical Journal*, vol. 91, no. 6, pp. 2115–2120, 2006.

[20] S. Takada, "Coarse-grained molecular simulations of large biomolecules," *Current opinion in structural biology*, vol. 22, no. 2, pp. 130–137, 2012.

[21] W. G. Noid, "Perspective: Coarse-grained models for biomolecular systems," *The Journal of chemical physics*, vol. 139, no. 9, p. 09B201_1, 2013.

[22] M. S. Shell, "The relative entropy is fundamental to multiscale and inverse thermodynamic problems," *The Journal of chemical physics*, vol. 129, no. 14, p. 144108, 2008.

[23] M. Giulini, M. Rigoli, G. Mattiotti, R. Menichetti, T. Tarenzi, R. Fiorentini, and R. Potestio, "From system modeling to system analysis: The impact of resolution level and resolution distribution in the computer-aided investigation of biomolecules," *Frontiers in Molecular Biosciences*, vol. 8, 2021.

[24] M. Giulini, R. Menichetti, M. S. Shell, and R. Potestio, "An information-theory-based approach for optimal model reduction of biomolecules," *Journal of chemical theory and computation*, vol. 16, no. 11, pp. 6795–6813, 2020.

[25] F. Errica, M. Giulini, D. Bacciu, R. Menichetti, A. Micheli, and R. Potestio, "A deep graph network–enhanced sampling approach to efficiently explore the space of reduced representations of proteins," *Frontiers in Molecular Biosciences*, vol. 8, 2021.

[26] R. Menichetti, M. Giulini, and R. Potestio, "A journey through mapping space: characterising the statistical and metric properties of reduced representations of macromolecules," *The European Physical Journal B*, vol. 94, no. 10, p. 204, 2021.

[27] R. J. Cubero, J. Jo, M. Marsili, Y. Roudi, and J. Song, "Statistical criticality arises in most informative representations," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 6, p. 063402, 2019.

[28] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers: Original Research on Biomolecules*, vol. 22, no. 12, pp. 2577–2637, 1983.

[29] E. Engel and R. M. Dreizler, *Density functional theory*. Springer, 2013.

[30] M. P. Nightingale and C. J. Umrigar, *Quantum Monte Carlo methods in physics and chemistry*. No. 525, Springer Science & Business Media, 1998.

[31] A. V. Onufriev and S. Izadi, "Water models for biomolecular simulations," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 8, no. 2, p. e1347, 2018.

[32] J. Huang and A. D. MacKerell Jr, "Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data," *Journal of computational chemistry*, vol. 34, no. 25, pp. 2135–2145, 2013.

[33] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb," *Journal of chemical theory and computation*, vol. 11, no. 8, pp. 3696–3713, 2015.

[34] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. De Groot, H. Grubmüller, and A. D. MacKerell, "Charmm36m: an improved force field

for folded and intrinsically disordered proteins," *Nature methods*, vol. 14, no. 1, pp. 71–73, 2017.

[35] P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states," *Proceedings of the National Academy of Sciences*, vol. 115, no. 21, pp. E4758–E4766, 2018.

[36] P. H. Nguyen, A. Ramamoorthy, B. R. Sahoo, J. Zheng, P. Faller, J. E. Straub, L. Dominguez, J.-E. Shea, N. V. Dokholyan, A. De Simone, *et al.*, "Amyloid oligomers: A joint experimental/computational perspective on alzheimer's disease, parkinson's disease, type ii diabetes, and amyotrophic lateral sclerosis," *Chemical Reviews*, vol. 121, no. 4, pp. 2545–2647, 2021.

[37] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1, pp. 19–25, 2015.

[38] J. C. Phillips, D. J. Hardy, J. D. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, *et al.*, "Scalable molecular dynamics on cpu and gpu architectures with namd," *The Journal of chemical physics*, vol. 153, no. 4, p. 044130, 2020.

[39] D. E. Shaw, J. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, *et al.*, "Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer," in *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 41–53, IEEE, 2014.

[40] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," *Science*, vol. 334, no. 6055, pp. 517–520, 2011.

[41] D. E. Shaw, P. J. Adams, A. Azaria, J. A. Bank, B. Batson, A. Bell, M. Bergdorf, J. Bhatt, J. A. Butts, T. Correia, *et al.*, "Anton 3: twenty microseconds of molecular dynamics simulation before lunch," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–11, 2021.

[42] https://foldingathome.org.

[43] M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, *et al.*, "Sars-cov-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome," *Nature Chemistry*, pp. 1–9, 2021.

[44] A. Singharoy, C. Maffeo, K. H. Delgado-Magnero, D. J. Swainsbury, M. Sener, U. Kleinekathöfer, J. W. Vant, J. Nguyen, A. Hitchcock, B. Isralewitz, *et al.*, "Atoms to phenotypes: molecular design principles of cellular energy metabolism," *Cell*, vol. 179, no. 5, pp. 1098–1111, 2019.

[45] T. E. Ouldridge, A. A. Louis, and J. P. Doye, "Structural, mechanical, and thermodynamic properties of a coarse-grained dna model," *The Journal of chemical physics*, vol. 134, no. 8, p. 02B627, 2011.

[46] M. Zgarbová, J. Sponer, M. Otyepka, T. E. Cheatham III, R. Galindo-Murillo, and P. Jurecka, "Refinement of the sugar–phosphate backbone torsion beta for amber force fields improves the description of z-and b-dna," *Journal of chemical theory and computation*, vol. 11, no. 12, pp. 5723–5736, 2015.

[47] R. Galindo-Murillo, J. C. Robertson, M. Zgarbová, J. Sponer, M. Otyepka, P. Jurecka, and T. E. Cheatham III, "Assessing the current state of amber force field modifications for dna," *Journal of chemical theory and computation*, vol. 12, no. 8, pp. 4114–4127, 2016.

[48] T. Sun, V. Minhas, N. Korolev, A. Mirzoev, A. P. Lyubartsev, and L. Nordenskiöld, "Bottom-up coarse-grained modeling of dna," *Frontiers in Molecular Biosciences*, vol. 8, 2021.

[49] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries, "The martini force field: coarse grained model for biomolecular simulations," *The journal of physical chemistry B*, vol. 111, no. 27, pp. 7812–7824, 2007.

[50] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, "The martini coarse-grained force field: extension to proteins," *Journal of chemical theory and computation*, vol. 4, no. 5, pp. 819–834, 2008.

[51] M. Tuckerman, *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2010.

[52] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, "The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models," *The Journal of chemical physics*, vol. 128, no. 24, p. 244114, 2008.

[53] K. M. Lebold and W. Noid, "Dual approach for effective potentials that accurately model structure and energetics," *The Journal of chemical physics*, vol. 150, no. 23, p. 234107, 2019.

[54] T. T. Foley, M. S. Shell, and W. G. Noid, "The impact of resolution upon entropy and information in coarse-grained models," *The Journal of chemical physics*, vol. 143, no. 24, p. 12B601_1, 2015.

[55] N. J. Dunn, T. T. Foley, and W. G. Noid, "Van der waals perspective on coarse-graining: Progress toward solving representability and transferability problems," *Accounts of chemical research*, vol. 49, no. 12, pp. 2832–2840, 2016.

[56] K. M. Kidder, R. J. Szukalo, and W. Noid, "Energetic and entropic considerations for coarse-graining," *The European Physical Journal B*, vol. 94, no. 7, pp. 1–29, 2021.

[57] J. F. Rudzinski and W. Noid, "Coarse-graining entropy, forces, and structures," *The Journal of chemical physics*, vol. 135, no. 21, p. 214101, 2011.

[58] C. N. Likos, "Soft matter with soft particles," *Soft matter*, vol. 2, no. 6, pp. 478–498, 2006.

[59] G. D'Adamo, R. Menichetti, A. Pelissetto, and C. Pierleoni, "Coarse-graining polymer solutions: A critical appraisal of single-and multi-site models," *The European Physical Journal Special Topics*, vol. 224, no. 12, pp. 2239–2267, 2015.

[60] M. Dijkstra, R. van Roij, and R. Evans, "Phase diagram of highly asymmetric binary hard-sphere mixtures," *Physical Review E*, vol. 59, no. 5, p. 5744, 1999.

[61] A. Soper, "Empirical potential monte carlo simulation of fluid structure," *Chemical Physics*, vol. 202, no. 2-3, pp. 295–306, 1996.

[62] A. P. Lyubartsev and A. Laaksonen, "Calculation of effective interaction potentials from radial distribution functions: A reverse monte carlo approach," *Physical Review E*, vol. 52, no. 4, p. 3730, 1995.

[63] S. Izvekov and G. A. Voth, "A multiscale coarse-graining method for biomolecular systems," *The Journal of Physical Chemistry B*, vol. 109, no. 7, pp. 2469–2473, 2005.

[64] S. Izvekov and G. A. Voth, "Multiscale coarse graining of liquid-state systems," *The Journal of chemical physics*, vol. 123, no. 13, p. 134105, 2005.

[65] A. Chaimovich and M. S. Shell, "Coarse-graining errors and numerical optimization using a relative entropy framework," *The Journal of chemical physics*, vol. 134, no. 9, p. 094112, 2011.

[66] M. S. Shell, "Coarse-graining with the relative entropy," *Advances in chemical physics*, vol. 161, pp. 395–441, 2016.

[67] W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, and O. Hahn, "Simulation of polymer melts. i. coarse-graining procedure for polycarbonates," *Acta Polymerica*, vol. 49, no. 2-3, pp. 61–74, 1998.

[68] R. Potestio, C. Peter, and K. Kremer, "Computer simulations of soft matter: linking the scales," *Entropy*, vol. 16, no. 8, pp. 4199–4245, 2014.

[69] J. Mullinax and W. Noid, "Generalized yvon-born-green theory for molecular systems," *Physical review letters*, vol. 103, no. 19, p. 198104, 2009.

[70] J. Mullinax and W. G. Noid, "A generalized-yvon- born- green theory for determining coarse-grained interaction potentials," *The Journal of Physical Chemistry C*, vol. 114, no. 12, pp. 5661–5674, 2010.

[71] F. Müller-Plathe, "Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back," *ChemPhysChem*, vol. 3, no. 9, pp. 754–769, 2002.

[72] A. P. Lyubartsev and A. Laaksonen, "Osmotic and activity coefficients from effective potentials for hydrated ions," *Physical Review E*, vol. 55, no. 5, p. 5689, 1997.

[73] A. Lyubartsev, A. Mirzoev, L. Chen, and A. Laaksonen, "Systematic coarse-graining of molecular models by the newton inversion method," *Faraday discussions*, vol. 144, pp. 43–56, 2010.

[74] G. A. Papoian, *Coarse-grained modeling of biomolecules*. CRC Press, 2017.

[75] V. Ruhle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, "Versatile object-oriented toolkit for coarse-graining applications," *Journal of chemical theory and computation*, vol. 5, no. 12, pp. 3211–3223, 2009.

[76] B. Hess, C. Holm, and N. van der Vegt, "Modeling multibody effects in ionic solutions with a concentration dependent dielectric permittivity," *Physical review letters*, vol. 96, no. 14, p. 147801, 2006.

[77] A. Mirzoev and A. P. Lyubartsev, "Systematic implicit solvent coarse graining of dimyristoylphosphatidylcholine lipids," *Journal of computational chemistry*, vol. 35, no. 16, pp. 1208–1218, 2014.

[78] N. Korolev, D. Luo, A. P. Lyubartsev, and L. Nordenskiöld, "A coarse-grained dna model parameterized from atomistic simulations by inverse monte carlo," *Polymers*, vol. 6, no. 6, pp. 1655–1675, 2014.

[79] Y.-L. Wang, A. Lyubartsev, Z.-Y. Lu, and A. Laaksonen, "Multiscale coarse-grained simulations of ionic liquids: comparison of three approaches to derive effective potentials," *Physical Chemistry Chemical Physics*, vol. 15, no. 20, pp. 7701–7712, 2013.

[80] A. Mirzoev and A. P. Lyubartsev, "Magic: Software package for multiscale modeling," *Journal of chemical theory and computation*, vol. 9, no. 3, pp. 1512–1520, 2013.

[81] W. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, "Multiscale coarse-graining and structural correlations: Connections to liquid-state theory," *The Journal of Physical Chemistry B*, vol. 111, no. 16, pp. 4116–4127, 2007.

[82] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé, and C. Clementi, "Machine learning of coarse-grained molecular dynamics force fields," *ACS central science*, vol. 5, no. 5, pp. 755–767, 2019.

[83] B. E. Husic, N. E. Charron, D. Lemm, J. Wang, A. Pérez, M. Majewski, A. Krämer, Y. Chen, S. Olsson, G. de Fabritiis, *et al.*, "Coarse graining molecular dynamics with graph neural networks," *The Journal of Chemical Physics*, vol. 153, no. 19, p. 194101, 2020.

[84] Y. Chen, A. Krämer, N. E. Charron, B. E. Husic, C. Clementi, and F. Noé, "Machine learning implicit solvation for molecular dynamics," *The Journal of Chemical Physics*, vol. 155, no. 8, p. 084101, 2021.

[85] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[86] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[87] P. Espanol and I. Zuniga, "Obtaining fully dynamic coarse-grained models from md," *Physical Chemistry Chemical Physics*, vol. 13, no. 22, pp. 10538–10545, 2011.

[88] I. Bilionis and N. Zabaras, "A stochastic optimization approach to coarse-graining using a relative-entropy framework," *The Journal of chemical physics*, vol. 138, no. 4, p. 044313, 2013.

[89] S. P. Carmichael and M. S. Shell, "A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly," *The Journal of Physical Chemistry B*, vol. 116, no. 29, pp. 8383–8393, 2012.

[90] T. Sanyal and M. S. Shell, "Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation," *The Journal of chemical physics*, vol. 145, no. 3, p. 034109, 2016.

[91] A. Pritchard-Bell and M. S. Shell, "Smoothing protein energy landscapes by integrating folding models with structure prediction," *Biophysical journal*, vol. 101, no. 9, pp. 2251–2259, 2011.

[92] G. R. Bowman, D. L. Ensign, and V. S. Pande, "Enhanced modeling via network theory: Adaptive sampling of markov state models," *Journal of chemical theory and computation*, vol. 6, no. 3, pp. 787–794, 2010.

[93] K. M. Lebold and W. Noid, "Dual-potential approach for coarse-grained implicit solvent models with accurate, internally consistent energetics and predictive transferability," *The Journal of chemical physics*, vol. 151, no. 16, p. 164113, 2019.

[94] T. Dannenhoffer-Lafage, J. W. Wagner, A. E. Durumeric, and G. A. Voth, "Compatible observable decompositions for coarse-grained representations of real molecular systems," *The Journal of chemical physics*, vol. 151, no. 13, p. 134115, 2019.

[95] R. C. Oliver, D. J. Read, O. G. Harlen, and S. A. Harris, "A stochastic finite element model for the dynamics of globular macromolecules," *Journal of Computational Physics*, vol. 239, pp. 147–165, 2013.

[96] R. Welch, S. A. Harris, O. G. Harlen, and D. J. Read, "Kobra: a fluctuating elastic rod model for slender biological macromolecules," *Soft Matter*, vol. 16, no. 32, pp. 7544–7555, 2020.

[97] R. E. Amaro and A. J. Mulholland, "Multiscale methods in drug design bridge chemical and biological complexity in the search for cures," *Nature Reviews Chemistry*, vol. 2, no. 4, pp. 1–12, 2018.

[98] L. Darrè, M. R. Machado, A. F. Brandner, H. C. Gonzàlez, S. Ferreira, and S. Pantano, "Sirah: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics," *Journal of chemical theory and computation*, vol. 11, no. 2, pp. 723–739, 2015.

[99] M. R. Machado, E. E. Barrera, F. Klein, M. Sóñora, S. Silva, and S. Pantano, "The sirah 2.0 force field: altius, fortius, citius," *Journal of chemical theory and computation*, vol. 15, no. 4, pp. 2719–2733, 2019.

[100] A. B. Poma, M. Cieplak, and P. E. Theodorakis, "Combining the martini and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins," *Journal of chemical theory and computation*, vol. 13, no. 3, pp. 1366–1374, 2017.

[101] P. C. Souza, S. Thallmair, P. Conflitti, C. Ramírez-Palacios, R. Alessandri, S. Raniolo, V. Limongelli, and S. J. Marrink, "Protein–ligand binding with the coarse-grained martini model," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[102] D. Alemani, F. Collu, M. Cascella, and M. Dal Peraro, "A nonradial coarse-grained potential for proteins produces naturally stable secondary structure elements," *Journal of chemical theory and computation*, vol. 6, no. 1, pp. 315–324, 2010.

[103] E. Spiga, D. Alemani, M. T. Degiacomi, M. Cascella, and M. Dal Peraro, "Electrostatic-consistent coarse-grained potentials for molecular simulations of proteins," *Journal of chemical theory and computation*, vol. 9, no. 8, pp. 3515–3526, 2013.

[104] P. Derreumaux, "From polypeptide sequences to structures using monte carlo simulations and an optimized potential," *The Journal of chemical physics*, vol. 111, no. 5, pp. 2301–2310, 1999.

[105] J. Maupetit, P. Tuffery, and P. Derreumaux, "A coarse-grained protein force field for folding and structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 69, no. 2, pp. 394–408, 2007.

[106] F. Sterpone, P. H. Nguyen, M. Kalimeri, and P. Derreumaux, "Importance of the ion-pair interactions in the opep coarse-grained force field: parametrization and validation," *Journal of chemical theory and computation*, vol. 9, no. 10, pp. 4574–4584, 2013.

[107] T. Bereau and M. Deserno, "Generic coarse-grained model for protein folding and aggregation," *The Journal of chemical physics*, vol. 130, no. 23, p. 06B621, 2009.

[108] A. Voegler Smith and C. K. Hall, "α-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model," *Proteins: Structure, Function, and Bioinformatics*, vol. 44, no. 3, pp. 344–360, 2001.

[109] M. Cheon, I. Chang, and C. K. Hall, "Extending the prime model for protein aggregation to all 20 amino acids," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 14, pp. 2950–2960, 2010.

[110] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, "Awsem-md: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing," *The Journal of Physical Chemistry B*, vol. 116, no. 29, pp. 8494–8503, 2012.

[111] H. Wu, P. G. Wolynes, and G. A. Papoian, "Awsem-idp: a coarse-grained force field for intrinsically disordered proteins," *The Journal of Physical Chemistry B*, vol. 122, no. 49, pp. 11115–11125, 2018.

[112] A. Liwo, M. Baranowski, C. Czaplewski, E. Gołaś, Y. He, D. Jagieła, P. Krupa, M. Maciejczyk, M. Makowski, M. A. Mozolewska, *et al.*, "A unified coarse-grained model of biological macromolecules based on mean-field

multipole–multipole interactions," *Journal of molecular modeling*, vol. 20, no. 8, p. 2306, 2014.

[113] A. K. Sieradzan, M. Makowski, A. Augustynowicz, and A. Liwo, "A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. i. backbone potentials of coarse-grained polypeptide chains," *The Journal of chemical physics*, vol. 146, no. 12, p. 124106, 2017.

[114] A. Liwo, C. Czaplewski, A. K. Sieradzan, E. A. Lubecka, A. G. Lipska, Ł. Golon, A. Karczyńska, P. Krupa, M. A. Mozolewska, M. Makowski, *et al.*, "Scale-consistent approach to the derivation of coarse-grained force fields for simulating structure, dynamics, and thermodynamics of biopolymers," *Progress in molecular biology and translational science*, vol. 170, pp. 73–122, 2020.

[115] F. Sterpone, P. Derreumaux, and S. Melchionna, "Protein simulations in fluids: Coupling the opep coarse-grained force field with hydrodynamics," *Journal of chemical theory and computation*, vol. 11, no. 4, pp. 1843–1853, 2015.

[116] D. Rapaport, "Molecular dynamics simulation of polymer chains with excluded volume," *Journal of Physics A: Mathematical and General*, vol. 11, no. 8, p. L213, 1978.

[117] A. Bellemans, J. Orban, and D. Van Belle, "Molecular dynamics of rigid and non-rigid necklaces of hard discs," *Molecular Physics*, vol. 39, no. 3, pp. 781–782, 1980.

[118] M. Chen, X. Lin, W. Zheng, J. N. Onuchic, and P. G. Wolynes, "Protein folding and structure prediction from the ground up: The atomistic associative memory, water mediated, structure and energy model," *The journal of physical chemistry B*, vol. 120, no. 33, pp. 8557–8565, 2016.

[119] R. D. Hills and C. L. Brooks, "Insights from coarse-grained gō models for protein folding and dynamics," *International journal of molecular sciences*, vol. 10, no. 3, pp. 889–905, 2009.

[120] S. Takada, "Gō model revisited," *Biophysics and Physicobiology*, vol. 16, pp. 248–255, 2019.

[121] M. H. Kim and M. K. Kim, "Review: Elastic network model for protein structural dynamics," *JSM Enzymol Protein Sci*, vol. 1, no. 1, p. 1001, 2014.

[122] Y. Togashi and H. Flechsig, "Coarse-grained protein dynamics studies using elastic network models," *International journal of molecular sciences*, vol. 19, no. 12, p. 3899, 2018.

[123] J.-W. Chu and G. A. Voth, "Coarse-grained modeling of the actin filament derived from atomistic-scale simulations," *Biophysical journal*, vol. 90, no. 5, pp. 1572–1582, 2006.

[124] D. Sept and F. C. MacKintosh, "Microtubule elasticity: connecting all-atom simulations with continuum mechanics," *Physical review letters*, vol. 104, no. 1, p. 018101, 2010.

[125] Y. Zhang, Z. Cao, and F. Xia, "Construction of ultra-coarse-grained model of protein with a gō-like potential," *Chemical Physics Letters*, vol. 681, pp. 1–6, 2017.

[126] Y. Zhang, Z. Cao, J. Z. Zhang, and F. Xia, "Double-well ultra-coarse-grained model to describe protein conformational transitions," *Journal of Chemical Theory and Computation*, vol. 16, no. 10, pp. 6678–6689, 2020.

[127] M. G. Saunders and G. A. Voth, "Coarse-graining of multiprotein assemblies," *Current opinion in structural biology*, vol. 22, no. 2, pp. 144–150, 2012.

[128] M. F. Hagan and R. Zandi, "Recent advances in coarse-grained modeling of virus assembly," *Current opinion in virology*, vol. 18, p. 36, 2016.

[129] A. Arkhipov, P. L. Freddolino, K. Imada, K. Namba, and K. Schulten, "Coarse-grained molecular dynamics simulations of a rotating bacterial flagellum," *Biophysical journal*, vol. 91, no. 12, pp. 4589–4597, 2006.

[130] A. Arkhipov, P. L. Freddolino, and K. Schulten, "Stability and dynamics of virus capsids described by coarse-grained modeling," *Structure*, vol. 14, no. 12, pp. 1767–1777, 2006.

[131] H. D. Nguyen, V. S. Reddy, and C. L. Brooks Iii, "Invariant polymorphism in virus capsid assembly," *Journal of the American Chemical Society*, vol. 131, no. 7, pp. 2606–2614, 2009.

[132] J. M. Grime, J. F. Dama, B. K. Ganser-Pornillos, C. L. Woodward, G. J. Jensen, M. Yeager, and G. A. Voth, "Coarse-grained simulation reveals key features of hiv-1 capsid self-assembly," *Nature communications*, vol. 7, no. 1, pp. 1–11, 2016.

[133] J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth, "The theory of ultra-coarse-graining. 1. general principles," *Journal of chemical theory and computation*, vol. 9, no. 5, pp. 2466–2480, 2013.

[134] A. Davtyan, J. F. Dama, A. V. Sinitskiy, and G. A. Voth, "The theory of ultra-coarse-graining. 2. numerical implementation," *Journal of chemical theory and computation*, vol. 10, no. 12, pp. 5265–5275, 2014.

[135] J. F. Dama, J. Jin, and G. A. Voth, "The theory of ultra-coarse-graining. 3. coarse-grained sites with rapid local equilibrium of internal states," *Journal of chemical theory and computation*, vol. 13, no. 3, pp. 1010–1022, 2017.

[136] R. J. Gowers and P. Carbone, "A multiscale approach to model hydrogen bonding: The case of polyamide," *The Journal of Chemical Physics*, vol. 142, no. 22, p. 224907, 2015.

[137] E. Villa, A. Balaeff, L. Mahadevan, and K. Schulten, "Multiscale method for simulating protein-dna complexes," *Multiscale Modeling & Simulation*, vol. 2, no. 4, pp. 527–553, 2004.

[138] E. Villa, A. Balaeff, and K. Schulten, "Structural dynamics of the lac repressor–dna complex revealed by a multiscale simulation," *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6783–6788, 2005.

[139] P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano, "A coarse grained model for atomic-detailed dna simulations with explicit electrostatics," *Journal of chemical theory and computation*, vol. 6, no. 5, pp. 1711–1725, 2010.

[140] A. C. Fogarty, R. Potestio, and K. Kremer, "A multi-resolution model to capture both global fluctuations of an enzyme and molecular recognition in the ligand-binding site," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. 12, pp. 1902–1913, 2016.

[141] R. Fiorentini, K. Kremer, and R. Potestio, "Ligand-protein interactions in lysozyme investigated through a dual-resolution model," *Proteins: Structure, Function, and Bioinformatics*, 2020.

[142] M. Neri, C. Anselmi, M. Cascella, A. Maritan, and P. Carloni, "Coarse-grained model of proteins incorporating atomistic detail of the active site," *Physical review letters*, vol. 95, no. 21, p. 218102, 2005.

[143] M. Neri, C. Anselmi, V. Carnevale, A. V. Vargiu, and P. Carloni, "Molecular dynamics simulations of outer-membrane protease t from e. coli based on a hybrid coarse-grained/atomistic potential," *Journal of Physics: Condensed Matter*, vol. 18, no. 14, p. S347, 2006.

[144] M. Leguèbe, C. Nguyen, L. Capece, Z. Hoang, A. Giorgetti, and P. Carloni, "Hybrid molecular mechanics/coarse-grained simulations for structural prediction of g-protein coupled receptor/ligand complexes," *PloS one*, vol. 7, no. 10, p. e47332, 2012.

[145] T. Tarenzi, V. Calandrini, R. Potestio, and P. Carloni, "Open-boundary molecular mechanics/coarse-grained framework for simulations of low-resolution g-protein-coupled receptor–ligand complexes," *Journal of chemical theory and computation*, vol. 15, no. 3, pp. 2101–2109, 2019.

[146] R. Potestio, S. Fritsch, P. Espanol, R. Delgado-Buscalioni, K. Kremer, R. Everaers, and D. Donadio, "Hamiltonian adaptive resolution simulation for molecular liquids," *Physical review letters*, vol. 110, no. 10, p. 108301, 2013.

[147] K. Korshunova and P. Carloni, "Ligand affinities within the open-boundary molecular mechanics/coarse-grained framework (i): Alchemical transformations within the hamiltonian adaptive resolution scheme," *The Journal of Physical Chemistry B*, vol. 0, no. 0, p. null, 2021. PMID: 33443434.

[148] P. Doruker, R. L. Jernigan, and I. Bahar, "Dynamics of large proteins through hierarchical levels of coarse-grained structures," *Journal of computational chemistry*, vol. 23, no. 1, pp. 119–127, 2002.

[149] O. Kurkcuoglu, R. L. Jernigan, and P. Doruker, "Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions," *Polymer*, vol. 45, no. 2, pp. 649–657, 2004.

[150] K. Eom, S.-C. Baek, J.-H. Ahn, and S. Na, "Coarse-graining of protein structures for the normal mode studies," *Journal of computational chemistry*, vol. 28, no. 8, pp. 1400–1410, 2007.

[151] H. Jang, S. Na, and K. Eom, "Multiscale network model for large protein dynamics," *The Journal of chemical physics*, vol. 131, no. 24, p. 12B623, 2009.

[152] Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfaendtner, and G. A. Voth, "A systematic methodology for defining coarse-grained sites in large biomolecules," *Biophysical journal*, vol. 95, no. 11, pp. 5073–5083, 2008.

[153] Z. Zhang, J. Pfaendtner, A. Grafmüller, and G. A. Voth, "Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models," *Biophysical journal*, vol. 97, no. 8, pp. 2327–2337, 2009.

[154] J. Mullinax and W. G. Noid, "Extended ensemble approach for deriving transferable coarse-grained potentials," *The Journal of Chemical Physics*, vol. 131, no. 10, p. 104110, 2009.

[155] J. F. Rudzinski and W. G. Noid, "Investigation of coarse-grained mappings via an iterative generalized yvon–born–green method," *The Journal of Physical Chemistry B*, vol. 118, no. 28, pp. 8295–8312, 2014.

[156] T. T. Foley, K. M. Kidder, M. S. Shell, and W. Noid, "Exploring the landscape of model representations," *Proceedings of the National Academy of Sciences*, vol. 117, no. 39, pp. 24061–24068, 2020.

[157] P. Koehl, F. Poitevin, R. Navaza, and M. Delarue, "The renormalization group and its applications to generating coarse-grained models of large biological molecular systems," *Journal of Chemical Theory and Computation*, vol. 13, no. 3, pp. 1424–1438, 2017.

[158] P. Diggins IV, C. Liu, M. Deserno, and R. Potestio, "Optimal coarse-grained site selection in elastic network models of biomolecules," *Journal of chemical theory and computation*, vol. 15, no. 1, pp. 648–664, 2018.

[159] Z. Zhang and G. A. Voth, "Coarse-grained representations of large biomolecular complexes from low-resolution structural data," *Journal of chemical theory and computation*, vol. 6, no. 9, pp. 2990–3002, 2010.

[160] A. V. Sinitskiy, M. G. Saunders, and G. A. Voth, "Optimal number of coarse-grained sites in different components of large biomolecular complexes," *The Journal of Physical Chemistry B*, vol. 116, no. 29, pp. 8363–8374, 2012.

[161] M. Li, J. Z. Zhang, and F. Xia, "A new algorithm for construction of coarse-grained sites of large biomolecules," *Journal of Computational Chemistry*, vol. 37, no. 9, pp. 795–804, 2016.

[162] M. Li, J. Z. Zhang, and F. Xia, "Constructing optimal coarse-grained sites of huge biomolecules by fluctuation maximization," *Journal of Chemical Theory and Computation*, vol. 12, no. 4, pp. 2091–2100, 2016.

[163] Z. Wu, Y. Zhang, J. Z. Zhang, K. Xia, and F. Xia, "Determining optimal coarse-grained representation for biomolecules using internal cluster validation indexes," *Journal of Computational Chemistry*, vol. 41, no. 1, pp. 14–20, 2020.

[164] K. Hinsen, "Analysis of domain motions by approximate normal mode calculations," *Proteins: Structure, Function, and Bioinformatics*, vol. 33, no. 3, pp. 417–429, 1998.

[165] T. Aleksiev, R. Potestio, F. Pontiggia, S. Cozzini, and C. Micheletti, "Pisqrd: a web server for decomposing proteins into quasi-rigid dynamical domains," *Bioinformatics*, vol. 25, no. 20, pp. 2743–2744, 2009.

[166] R. Potestio, F. Pontiggia, and C. Micheletti, "Coarse-grained description of protein internal dynamics: an optimal strategy for decomposing proteins in rigid subunits," *Biophysical journal*, vol. 96, no. 12, pp. 4993–5002, 2009.

[167] G. Polles, G. Indelicato, R. Potestio, P. Cermelli, R. Twarock, and C. Micheletti, "Mechanical and assembly units of viral capsids identified via quasi-rigid domain decomposition," *PLoS Comput Biol*, vol. 9, no. 11, p. e1003331, 2013.

[168] G. Morra, R. Potestio, C. Micheletti, and G. Colombo, "Corresponding functional dynamics across the hsp90 chaperone family: insights from a multiscale analysis of md simulations," *PLoS Comput Biol*, vol. 8, no. 3, p. e1002433, 2012.

[169] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, "Stability of graph communities across time scales," *Proceedings of the national academy of sciences*, vol. 107, no. 29, pp. 12755–12760, 2010.

[170] Y.-L. Chen and M. Habeck, "Data-driven coarse graining of large biomolecular structures," *PloS one*, vol. 12, no. 8, p. e0183057, 2017.

[171] L. Boninsegna, R. Banisch, and C. Clementi, "A data-driven perspective on the hierarchical assembly of molecular structures," *Journal of chemical theory and computation*, vol. 14, no. 1, pp. 453–460, 2018.

[172] W. Wang and R. Gómez-Bombarelli, "Coarse-graining auto-encoders for molecular dynamics," *npj Computational Materials*, vol. 5, no. 1, pp. 1–9, 2019.

[173] R. Banisch and P. Koltai, "Understanding the geometry of transport: Diffusion maps for lagrangian trajectory data unravel coherent sets," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 3, p. 035804, 2017.

[174] G. R. Bowman, V. S. Pande, and F. Noé, *An introduction to Markov state models and their application to long timescale molecular simulation*, vol. 797. Springer Science & Business Media, 2013.

[175] L. P. Kadanoff, "Notes on migdal's recursion formulas," *Annals of Physics*, vol. 100, no. 1-2, pp. 359–394, 1976.

[176] J. V. José, L. P. Kadanoff, S. Kirkpatrick, and D. R. Nelson, "Renormalization, vortices, and symmetry-breaking perturbations in the two-dimensional planar model," *Physical Review B*, vol. 16, no. 3, p. 1217, 1977.

[177] E. T. Jaynes, "Gibbs vs boltzmann entropies," *American Journal of Physics*, vol. 33, no. 5, pp. 391–398, 1965.

[178] S. Park and K. Schulten, "Calculating potentials of mean force from steered molecular dynamics simulations," *The Journal of chemical physics*, vol. 120, no. 13, pp. 5946–5961, 2004.

[179] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *The Journal of chemical physics*, vol. 126, no. 1, p. 014101, 2007.

[180] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *Journal of Applied physics*, vol. 52, no. 12, pp. 7182–7190, 1981.

[181] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "Gromacs: fast, flexible, and free," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[182] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99sb protein force field," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 8, pp. 1950–1958, 2010.

[183] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh ewald method," *The Journal of chemical physics*, vol. 103, no. 19, pp. 8577–8593, 1995.

[184] M. Mayorga-Flores, A. Chantôme, C. M. Melchor-Meneses, I. Domingo, G. A. Titaux-Delgado, R. Galindo-Murillo, C. Vandier, and F. del Río-Portilla, "Novel blocker of onco sk3 channels derived from scorpion toxin tamapin and active against migration of cancer cells," *ACS Medicinal Chemistry Letters*, vol. 11, no. 8, pp. 1627–1633, 2020.

[185] P. Pedarzani, D. D'hoedt, K. B. Doorty, J. D. Wadsworth, J. S. Joseph, K. Jeyaseelan, R. M. Kini, S. Gadre, S. Sapatnekar, M. Stocker, *et al.*, "Tamapin, a venom peptide from the indian red scorpion (mesobuthus tamulus) that targets small conductance ca2+-activated k+ channels and after-hyperpolarization currents in central neurons," *Journal of Biological Chemistry*, vol. 277, no. 48, pp. 46101–46109, 2002.

[186] C. D. Gati, M. R. Mortari, and E. F. Schwartz, "Towards therapeutic applications of arthropod venom k+-channel blockers in cns neurologic diseases involving memory acquisition and storage," *Journal of toxicology*, vol. 2012, 2012.

[187] C. Müller, G. Schlauderer, J. Reinstein, and G. E. Schulz, "Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding," *Structure*, vol. 4, no. 2, pp. 147–156, 1996.

[188] Y. E. Shapiro, E. Kahana, and E. Meirovitch, "Domain mobility in proteins from nmr/srls," *The Journal of Physical Chemistry B*, vol. 113, no. 35, pp. 12050–12060, 2009.

[189] P. C. Whitford, O. Miyashita, Y. Levy, and J. N. Onuchic, "Conformational transitions of adenylate kinase: switching by cracking," *Journal of molecular biology*, vol. 366, no. 5, pp. 1661–1671, 2007.

[190] S. L. Seyler and O. Beckstein, "Sampling large conformational transitions: adenylate kinase as a testing ground," *Molecular Simulation*, vol. 40, no. 10-11, pp. 855–877, 2014.

[191] E. Formoso, V. Limongelli, and M. Parrinello, "Energetics and structural characterization of the large-scale functional motion of adenylate kinase," *Scientific reports*, vol. 5, no. 1, pp. 1–8, 2015.

[192] J. Wang, C. Peng, Y. Yu, Z. Chen, Z. Xu, T. Cai, Q. Shao, J. Shi, and W. Zhu, "Exploring conformational change of adenylate kinase by replica exchange molecular dynamic simulation," *Biophysical journal*, vol. 118, no. 5, pp. 1009–1018, 2020.

[193] C. F. Scott, R. W. Carrell, C. B. Glaser, F. Kueppers, J. H. Lewis, R. W. Colman, *et al.*, "Alpha-1-antitrypsin-pittsburgh. a potent inhibitor of human plasma factor xia, kallikrein, and factor xiif.," *The Journal of clinical investigation*, vol. 77, no. 2, pp. 631–634, 1986.

[194] T. Nukiwa, M. L. Brantly, F. Ogushi, G. A. Fells, and R. G. Crystal, "Characterization of the gene and protein of the common alpha 1-antitrypsin normal m2 allele.," *American journal of human genetics*, vol. 43, no. 3, p. 322, 1988.

[195] M. Luisetti and N. Seersholm, "α1-antitrypsin deficiency· 1: Epidemiology of α1-antitrypsin deficiency," *Thorax*, vol. 59, no. 2, p. 164, 2004.

[196] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[197] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 34, no. 5, pp. 827–828, 1978.

[198] R. R. Sokal, "A statistical method for evaluating systematic relationships.," *Univ. Kansas, Sci. Bull.*, vol. 38, pp. 1409–1438, 1958.

[199] E. Jones, T. Oliphant, P. Peterson, *et al.*, "Scipy: Open source scientific tools for python," 2001.

[200] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van

Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[201] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics*, vol. 17, no. suppl_1, pp. S22–S29, 2001.

[202] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[203] S. Kirkpatrick, "Optimization by simulated annealing: Quantitative studies," *Journal of statistical physics*, vol. 34, no. 5, pp. 975–986, 1984.

[204] V. Černỳ, "Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm," *Journal of optimization theory and applications*, vol. 45, no. 1, pp. 41–51, 1985.

[205] F. Wang and D. Landau, "Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram," *Physical Review E*, vol. 64, no. 5, p. 056101, 2001.

[206] F. Wang and D. P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states," *Physical review letters*, vol. 86, no. 10, p. 2050, 2001.

[207] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, "Coarse-grained protein models and their applications," *Chemical reviews*, vol. 116, no. 14, pp. 7898–7936, 2016.

[208] P. Kunzmann and K. Hamacher, "Biotite: a unifying open source computational biology framework in python," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–8, 2018.

[209] N. Andreotti, E. Di Luccio, F. Sampieri, M. De Waard, and J.-M. Sabatier, "Molecular modeling and docking simulations of scorpion toxins and related analogs on human skca2 and skca3 channels," *Peptides*, vol. 26, no. 7, pp. 1095–1108, 2005.

[210] V. Quintero-Hernández, J. Jiménez-Vargas, G. Gurrola, H. Valdivia, and L. Possani, "Scorpion venom components that affect ion-channels function," *Toxicon*, vol. 76, pp. 328–342, 2013.

[211] B. Ramírez-Cordero, Y. Toledano, P. Cano-Sánchez, R. Hernández-López, D. Flores-Solis, A. L. Saucedo-Yáñez, I. Chávez-Uribe, L. G. Brieba, and F. del Río-Portilla, "Cytotoxicity of recombinant tamapin and related toxin-like peptides on model cell lines," *Chemical research in toxicology*, vol. 27, no. 6, pp. 960–967, 2014.

[212] T. T. Thach, T. T. Luong, S. Lee, and D.-K. Rhee, "Adenylate kinase from streptococcus pneumoniae is essential for growth through its catalytic activity," *FEBS open bio*, vol. 4, pp. 672–682, 2014.

[213] M. Bellinzoni, A. Haouz, M. Graña, H. Munier-Lehmann, W. Shepard, and P. M. Alzari, "The crystal structure of mycobacterium tuberculosis adenylate kinase in complex with two molecules of adp and mg2+ supports an associative mechanism for phosphoryl transfer," *Protein science*, vol. 15, no. 6, pp. 1489–1493, 2006.

[214] J. Reinstein, A.-M. Gilles, T. Rose, A. Wittinghofer, I. Saint Girons, O. Bârzu, W. K. Surewicz, and H. H. Mantsch, "Structural and catalytic role of arginine 88 in escherichia coli adenylate kinase as evidenced by chemical modification and site-directed mutagenesis," *Journal of Biological Chemistry*, vol. 264, no. 14, pp. 8107–8112, 1989.

[215] A. Akbari, "Phenylglyoxal," *Synlett*, vol. 23, no. 06, pp. 951–952, 2012.

[216] Y. Matsunaga, H. Fujisaki, T. Terada, T. Furuta, K. Moritsugu, and A. Kidera, "Minimum free energy path of ligand-induced transition in adenylate kinase," *PLoS computational biology*, vol. 8, no. 6, p. e1002555, 2012.

[217] M. Gur, J. D. Madura, and I. Bahar, "Global transitions of proteins explored by a multiscale hybrid methodology: application to adenylate kinase," *Biophysical journal*, vol. 105, no. 7, pp. 1643–1652, 2013.

[218] R. Halder, R. N. Manna, S. Chakraborty, and B. Jana, "Modulation of the conformational dynamics of apo-adenylate kinase through a π–cation interaction," *The Journal of Physical Chemistry B*, vol. 121, no. 23, pp. 5699–5708, 2017.

[219] M. Schapira, M.-A. Ramus, S. Jallat, D. Carvallo, M. Courtney, *et al.*, "Recombinant alpha 1-antitrypsin pittsburgh (met 358—-arg) is a potent inhibitor of plasma kallikrein and activated factor xii fragment.," *The Journal of clinical investigation*, vol. 77, no. 2, pp. 635–637, 1986.

[220] C. Taggart, D. Cervantes-Laurean, G. Kim, N. G. McElvaney, N. Wehr, J. Moss, and R. L. Levine, "Oxidation of either methionine 351 or methionine 358 in α1-antitrypsin causes loss of anti-neutrophil elastase activity," *Journal of Biological Chemistry*, vol. 275, no. 35, pp. 27258–27265, 2000.

[221] M. C. Owen, S. O. Brennan, J. H. Lewis, and R. W. Carrell, "Mutation of antitrypsin to antithrombin: α1-antitrypsin pittsburgh (358 met—arg), a fatal bleeding disorder," *New England Journal of Medicine*, vol. 309, no. 12, pp. 694–698, 1983.

[222] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, *et al.*, "The rosetta all-atom energy function for macromolecular modeling and design," *Journal of chemical theory and computation*, vol. 13, no. 6, pp. 3031–3048, 2017.

[223] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.

[224] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[225] D. Bacciu, F. Errica, A. Micheli, and M. Podda, "A gentle introduction to deep learning for graphs," *Neural Networks*, vol. 129, pp. 203–221, 2020.

[226] M. S. Shell, P. G. Debenedetti, and A. Z. Panagiotopoulos, "Generalization of the wang-landau method for off-lattice simulations," *Physical review E*, vol. 66, no. 5, p. 056703, 2002.

[227] L. Y. Barash, M. Fadeeva, and L. Shchur, "Control of accuracy in the wang-landau algorithm," *Physical Review E*, vol. 96, no. 4, p. 043307, 2017.

[228] D. Gfeller and P. De Los Rios, "Spectral coarse graining of complex networks," *Physical review letters*, vol. 99, no. 3, p. 038701, 2007.

[229] M. A. Webb, J.-Y. Delannoy, and J. J. De Pablo, "Graph-based approach to systematic molecular coarse-graining," *Journal of chemical theory and computation*, vol. 15, no. 2, pp. 1199–1208, 2018.

[230] Z. Li, G. P. Wellawatte, M. Chakraborty, H. A. Gandhi, C. Xu, and A. D. White, "Graph neural network based coarse-grained mapping prediction," *Chemical science*, vol. 11, no. 35, pp. 9524–9531, 2020.

[231] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. suppl_1, pp. i47–i56, 2005.

[232] A. Fout, J. Byrd, B. Shariat, and A. Ben-Hur, "Protein interface prediction using graph convolutional networks," in *Advances in neural information processing systems*, pp. 6530–6539, 2017.

[233] W. Torng and R. B. Altman, "Graph convolutional neural networks for predicting drug-target interactions," *Journal of Chemical Information and Modeling*, vol. 59, no. 10, pp. 4131–4149, 2019.

[234] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[235] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, JMLR Workshop and Conference Proceedings, 2011.

[236] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[237] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the trade*, pp. 55–69, Springer, 1998.

[238] D. Landau, S.-H. Tsai, and M. Exler, "A new approach to monte carlo simulations in statistical physics: Wang-landau sampling," *American Journal of Physics*, vol. 72, no. 10, pp. 1294–1302, 2004.

[239] T. Wüst and D. Landau, "The hp model of protein folding: A challenging testing ground for wang–landau sampling," *Computer Physics Communications*, vol. 179, no. 1-3, pp. 124–127, 2008.

[240] D. Seaton, T. Wüst, and D. Landau, "A wang-landau study of the phase transitions in a flexible homopolymer," *Computer Physics Communications*, vol. 180, no. 4, pp. 587–589, 2009.

[241] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Physical Review B*, vol. 87, no. 18, p. 184115, 2013.

[242] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," *Physical Chemistry Chemical Physics*, vol. 18, no. 20, pp. 13754–13769, 2016.

[243] P. Beale, *Statistical Mechanics*. Elsevier Science, 2011.

[244] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," *Proceedings of the National Academy of Sciences*, vol. 108, no. 32, pp. 13023–13028, 2011.

[245] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Demonstrating the transferability and the descriptive power of sketch-map," *Journal of chemical theory and computation*, vol. 9, no. 3, pp. 1521–1532, 2013.

[246] M. Marsili, I. Mastromatteo, and Y. Roudi, "On sampling and modeling complex systems," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013, no. 09, p. P09003, 2013.

[247] A. Haimovici and M. Marsili, "Criticality of mostly informative samples: a bayesian model selection approach," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2015, no. 10, p. P10013, 2015.

[248] J. Song, M. Marsili, and J. Jo, "Resolution and relevance trade-offs in deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2018, no. 12, p. 123406, 2018.

[249] E. Suárez, R. P. Wiewiora, C. Wehmeyer, F. Noé, J. D. Chodera, and D. M. Zuckerman, "What markov state models can and cannot do: Correlation versus path-based observables in protein-folding models," *Journal of Chemical Theory and Computation*, vol. 17, no. 5, pp. 3119–3133, 2021.

[250] https://www.nasdaq.com/market-activity/quotes/nasdaq-ndx-index.

[251] P. N. Afego, "Effects of changes in stock index compositions: A literature survey," *International Review of Financial Analysis*, vol. 52, pp. 228–239, 2017.

[252] E. N. Biktimirov and Y. Xu, "Asymmetric stock price and investor awareness reactions to changes in the nasdaq 100 index," *Journal of Asset Management*, vol. 20, no. 2, pp. 134–145, 2019.

[253] https://github.com/ranaroussi/yfinance.

[254] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 03, pp. 90–95, 2007.

[255] https://cmake.org/.

[256] R. M. Stallman and G. DeveloperCommunity, *Using The Gnu Compiler Collection: A Gnu Manual For Gcc Version 4.3.3*. Scotts Valley, CA: CreateSpace, 2009.

[257] https://www.intel.com/.

[258] https://clang.llvm.org/.

[259] https://www.openmp.org/.

[260] http://www.gnuplot.info/.

[261] W. Humphrey, A. Dalke, and K. Schulten, "Vmd: visual molecular dynamics," *Journal of molecular graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[262] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "Mdanalysis: a toolkit for the analysis of molecular dynamics simulations," *Journal of computational chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011.

[263] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domanski, D. L. Dotson, S. Buchoux, I. M. Kenney, *et al.*, "Mdanalysis: a python package for the rapid analysis of molecular dynamics simulations," tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2019.

[264] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "Mdtraj: a modern open library for the analysis of molecular dynamics trajectories," *Biophysical journal*, vol. 109, no. 8, pp. 1528–1532, 2015.

[265] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[266] W. McKinney *et al.*, "pandas: a foundational python library for data analysis and statistics," *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.