



University of Pisa

Master Degree in Physics

---

MS. Thesis

*A minimalist model for the simulation of the  
structure and dynamics of disordered proteins*

Luca Pesce

Supervisor:

**Dr. Valentina Tozzini**



## CONTENTS

---

<b>Introduction</b>	<b>iii</b>
<b>1 The proteins structure</b>	<b>1</b>
1.1 Experimental methods for the proteins structure determination	2
1.2 The protein databases	4
1.3 Amino acids and polypeptides - primary structure	6
1.4 Secondary structure	10
1.5 Tertiary and quaternary structures	20
1.6 Protein folding and Intrinsically disordered proteins	21
1.7 Random coils	25
<b>2 Proteins modeling</b>	<b>27</b>
2.1 Classical molecular dynamics	27
2.1.1 Integration algorithms for equations of motion	29
2.1.2 Constrained dynamics	31
2.1.3 Thermostats	32
2.2 Proteins models	34
2.2.1 Atomistic models	35
2.2.2 Coarse grained models	39
2.3 The $C_{\alpha}$ -based one-bead (minimalist)	44
2.3.1 Description of the model	44
2.3.2 Atomistic to minimalist variables transformation	47
<b>3 Structural dataset preparation</b>	<b>53</b>
3.1 Coil dataset generation	54
3.2 Internal variables statistical distributions	58
3.2.1 Single variable statistical distributions	59
3.2.2 Distribution of geometrical backbone parameters, $\tau$ , $\gamma_1$ and $\gamma_2$	62
3.2.3 Two variables $\theta$ , $\varphi$ distributions	65
3.2.4 Three dimensional distribution ( $\theta_-$ , $\varphi$ , $\theta_+$ )	69

3.3	$\theta, \varphi$ map generation from Ramachandran plot	70
3.4	DisProt: a dataset for disordered proteins	71
<b>4</b>	<b>Minimalist model for unstructured peptides</b>	<b>79</b>
4.1	Model definition	79
4.2	Local potential parametrization	81
4.3	Non local potential optimization	83
4.4	Molecular dynamics simulations	85
	<b>Conclusions and perspectives</b>	<b>93</b>
	<b>Appendices</b>	<b>95</b>
A	IDP's functions	97
B	Secondary structure assignment algorithms	98
C	DL_POLY	100
D	Tricubic interpolation of the 3D maps	101
E	Bias analysis NMR_PDB database	103
F	Angles-torsion potential module in DL_POLY: algorithms and implementation	106
G	Algorithms for the $\theta_-, \varphi, \theta_+$ 3D correlation	108
H	Algorithm for potential parameters optimization	112
	<b>Acknowledgement</b>	<b>119</b>
	<b>Bibliography</b>	<b>121</b>

## INTRODUCTION

---

Proteins are one of the two fundamental classes of biomolecules (the other being nucleic acids), performing most of all the functional roles in living systems. In the last fifty years, a big effort was devoted to the computer simulation of the dynamics of these systems, in order to get a better insight into their structure and behavior, and to complement the experimental studies. However, the size of the system, the time scale reachable in simulations and the accuracy of the representation are limited by the computer performances. Although the Moore law ensured - up to now - the exponential increase in time of the latter, currently, simulations with atomic resolution can address a virus size or very limited portion of a cell on very short time scales while only single proteins can be represented on the macroscopic time scales.

Therefore, in order to study the dynamics of biological systems, coarse grained (CG) models are considered a natural solution to overcome the limits of the atomistic models. CG models address the system at a lower resolution, reducing the number of explicit degrees of freedom and providing a less computationally expensive representation of system. Depending on the level of coarse graining, macroscopic time scales for systems of biologically interesting size can currently be afforded.

The hierarchical organization of the protein structure naturally suggest a possible level of coarse graining, namely that of one interacting center (also called "bead") per amino-acid, being the latter the basic structural unit of a protein. Among "one-bead-models" the subclass of those with the bead placed over  $C_\alpha$  emerges as a good compromise between the simplicity and the possibility of representing accurately the conformation of the backbone and protein secondary structures. The class of  $C_\alpha$ -based one bead models, also called "minimalist", is the focus of this Thesis work.

In the last decade a number of minimalist models were developed, all representing the interactions by means of empirical force fields (FF) consisting of a sum of analytical or numerical terms. Different models differ by the number and composition of the FF terms, and by the parameterization strategy, which can be based over higher level theories (typically atomistic simulations) or on experimental data (i.e. data set of experimental structures, and inclusion of other kind

of macroscopic and thermodynamic informations). As a consequence, the model can be more or less general and transferable. Usually, accuracy and transferability are in conflict: the more bias towards known structures is included, the more structurally accurate, but the less transferable and predictive will be the model. In order to overcome this problem, most of the currently available minimalist models include some *a priori* knowledge of the secondary or tertiary structures within the parameterization, which can be called a "partial bias". Clearly, the general goal is to build a model both accurate and predictive, and therefore, unbiased. In spite of the efforts this problem is still open. Many different recipes including experimental or theoretical information at different levels and in different ways are available, but a rational and standard approach to the problem is still lacking.

The goal of this Thesis is to make some steps along this route. The chosen strategy is to follow a physics based approach, related to the fundamental nature of forces acting within the proteins. Basically, the primary structure of a protein (i.e. sequence and polypeptide chain) is stabilized by covalent chemical bonds, while the secondary structure (e.g. helical or sheet-like structures) is stabilized by specific hydrogen bonds. Higher level structures (tertiary and quaternary) are stabilized by other specific interactions, such as disulphide and salt bridges. Therefore, a possible rational strategy to build a physics based CG model is to start from a model including only the covalent chemistry of the backbone, i.e. to build a model for unstructured proteins, where the network of hydrogen bond interaction is weak, disordered or in some case absent.

Therefore, the specific aim of this work is to build a general minimalist model to be used for unstructured proteins. No hydrogen bonding or other specific interactions are to be included in the FF, and the model is parameterized based on a dataset of unstructured proteins. This strategy is expected to result in a quite general and unbiased model able to reproduce the structure and dynamics of class proteins, namely the "intrinsically disordered proteins" (IDP), which is very interesting *per se*. In addition this model for unstructured proteins is designed to be used as a zero-point approximation over which hydrogen bonding and other interactions can be added in order to build models for structured proteins, in rational and physics-based fashion.

The first chapter of this work provides some basic notions on proteins. After a brief introduction of the experimental techniques related to resolve structured data for proteins, the description of the protein structure is proposed following the hierarchical order. After this the specific case of IDPs is illustrated. Due to the considerable differences between these proteins and the natively folded ones, IDPs require the use of specific experimental and theoretical methods. Random coils, defined as the fragments with the highest disordered contents of IDPs,

are then selected and analyzed in deeper detail. These represent the naturally occurring structural class with the smaller hydrogen bond content.

The second chapter illustrates the most popular simulation approaches for proteins, providing in the first part a description of the classical molecular dynamics, which is applied to a wide range of different models. The atomistic models allows the best accuracy and can provide information for the parametrization of coarser models. Details on the atomistic model are reported in the second section followed by the introduction to the CG models. As said the parametrization of CG models can follow many different strategies. One is the Force Matching (FM) based on the fit of CG forces onto those evaluated from trajectories of atomistic molecular dynamics simulation. Other popular methods fall in the class of the Boltzmann Inversion (BI), in which potentials are evaluated from the distribution of each CG-variables extracted from experimental data or from simulations. The detail on the last approach, the one chosen in this work, are reported in the closing part of the second chapter, which also describes the details related to the one beads model proposed in this work. Despite the large amount of degree of freedom lost in the coarse graining procedure the possibility of representing explicitly the secondary structure of proteins makes this class of models one of the most promising for applications. The final part of this chapter already includes some original results, namely an analysis of the mapping between the internal variables of the atomistic and one bead model for different secondary structures.

The third chapter is devoted to the analysis and selection of the data from the Protein Data Bank (PDB), which represents the main source of structures used as input for the parametrization in this work. A particular care is devoted to the selection of structures with a minimal amount of ordered secondary arrangements, in order to represent with optimal statistics the “unstructured proteins”. This choice is an important original result of this Thesis work, and is based on the selection of the appropriate experimental methods for proteins resolution and secondary structure recognition methods.

The fourth chapter describes the model developed in this work and parametrized on the basis of the dataset previously described. The main limits of the model and strategies to overcome them are also discussed. These involve the statistical relevance of the data, the relative simplicity of the FF used, the absence in it of the amino-acid specificity.

The parametrization of the model is based on the Boltzmann inversion (BI) procedure, basically consisting in deriving a potential related to the inverted logarithm of the internal variables distribution, therefore capable of reproducing those distribution in equilibrated simulations. The procedure here presented is actually an advanced version of the BI, involving multi-variate distribution targeting and the combination with stochastic exploration of the parameters space

for the non local potential. All the details and results are reported. In addition, simulations results are reported, to show the quality of the model and to validate it.

The implementation of these procedure and of the the specific force field for simulation required specific software creation or manipulation, whose technicalities are reported in appendices, together with the algorithmic details, and to some details about structure and function of disordered proteins. The last chapter includes a conclusive summary of possible further developments.



## THE PROTEINS STRUCTURE

---

This chapter provides an introduction to proteins structure and the basic notions used in the development of this work. A brief description of the experimental methods for the determination of the structure of proteins is given in section 1.1. In the subsequent section the database of the protein structures are described. These represent the main source of data supporting this work. Section 1.3 contains the basic notions on amino acids and on their polymeric extension, the polypeptide. Proteins are described according to their hierarchical structure. The amino acid sequence represents their primary structure. The local structural arrangement of each residue is defined as the protein secondary structure and it is described in section 1.4. In this section the main structural patterns and the “Ramachandran map”, a fundamental tool for the secondary structure analysis is illustrated. At the end of this chapter a brief introduction to the algorithms of identification of secondary structure is given, which are subsequently employed in chapter 3. Section 1.5 describes the last levels of the protein structure, namely the tertiary and quaternary ones. A deeper description of the disordered proteins is then given in sec 1.6. This section focus also on the key role of these peculiar proteins within this work. Section 1.7 provides the description of the random coil, which is the state of highest entropy among the disordered ones. Since random coil is characterized by the minimal possible amount of secondary structures, it is here considered as a paradigm to build a minimalist model for destructured proteins, which is the main goal of this work.

## 1.1 EXPERIMENTAL METHODS FOR THE PROTEINS STRUCTURE DETERMINATION

The two mostly used methods for protein structure determination are the X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. In the former technique an X-ray beam scatters across the crystallized sample of the protein and the scattered beams are collected on a screen. In the sample there is a protein in each lattice site. The X-rays are scattered by the electrons of the crystal and are elastically diffused. In some specific direction the scattered rays interfere constructively providing direct information on the reciprocal lattice vectors of the crystal ( $\mathbf{K}$ ). The intensities recorded on the screen provide the form factor  $F_{\mathbf{K}}$  of the protein depending on the diffusion vector, which is related to the electronic density  $f(\mathbf{r}_i)$  by means of Fourier transform. The total form factor can be written as the sum of the atomic form factor ( $f_{\mathbf{K}}^i$ ) with a phase depending on the atomic location within the crystallographic unit cell

$$F_{\mathbf{K}} = \sum_i f_{\mathbf{K}}^i e^{i\mathbf{K} \cdot \mathbf{r}_i}. \quad (1.1)$$

The atomic positions are then determined through an iterative procedure which compare the diffraction pattern with theoretical models.

This method is very efficient but carries several experimental drawbacks. The localization of hydrogen atoms is difficult due to the low electron density. The crystallization of a biological macromolecule implies a non-natural state. The configurational statistic set is reduced because the sample must be cooled down at cryogenic temperatures [1]. In addition it is not possible to evaluate the positions of atoms of the disorder or too mobile regions: the non-regularity of structures inside the crystal cells makes it impossible to have a coherent scattering giving a low signal on electron density map.

The Nuclear Magnetic Resonance (NMR) spectroscopy allows to sample both conformational and functional set of the protein in aqueous solution, which is nearer to the physiological state. NMR is based on the capability of a nuclear spin associated magnetic moment to react to external oscillating magnetic fields. Though quantum classical in nature, the phenomenon can be explained using a semiclassical analogue: when a magnetic moment interacts with the static

magnetic field (with modulus  $B_0$ , oriented toward the  $z$  axis) precesses at the Larmor frequency ( $\omega_L$ ):

$$\omega_L = \gamma B_0 \quad (1.2)$$

$$\gamma = \frac{gZe}{2M} \quad (1.3)$$

where  $g$ , also called  $g$ -factor, relates the angular momentum of the system to the intrinsic magnetic moment and is 1 for classical topic, while it assumes different values for different particles according to the quantum theory.  $M$  is the nuclear mass and  $Ze$  the nuclear charge.  $\gamma$  is named gyromagnetic ratio. The average magnetic moment is parallel to the  $z$  axis. Supplying an orthogonal oscillating magnetic field at the Larmor frequency the whole magnetic moment flips the mean magnetization on the same plane absorbing the electromagnetic energy. Therefore the system will have an adsorption peak at  $\omega = \omega_L$ . Gyromagnetic factors for the commonly used nuclei in NMR spectroscopy of proteins are reported in table 1.

Nucleus	$\gamma$ [ $10^6 \text{rad s}^{-1} \text{T}^{-1}$ ]
$^1\text{H}$	267.513
$^2\text{H}$	41.065
$^{13}\text{C}$	67.262
$^{14}\text{N}$	19.331
$^{15}\text{N}$	-27.116

Table 1: Gyromagnetic ratios of the nuclei used in protein NMR spectroscopy.

However when measured same nuclei show slightly different resonance frequencies depending on several environmental conditions (e.g. solvation, chemical environment), which influences the local magnetic field. The presence or the absence of a bond, any different conformation of the whole structure and other circumstances, alter the resonance frequencies. This deviation is defined as chemical shift  $\sigma$  (generally expressed in ppm);

$$\omega_L = (1 - \sigma)B_0\gamma. \quad (1.4)$$

This feature characterizes each nucleus making it unique within a given molecule. This is precisely the feature which allows to extract information about the molecular structure.

The NMR spectroscopy can be performed using different protocols and techniques. Each technique allows to investigate specific property of the sample and is characterized by a specific timescale. Timescales and main applications of the most used methods in protein are reported in table 2. From these measurements it is possible to extract structural restrains for the analyzed structure. These are subsequently included in molecular dynamics simulations in order to gather a set of low energy structures, which fulfills all the restrains. At variance with X-ray, with this method it is possible to study functional behaviors of a proteins, as, e.g. binding rates, conformational changes frequencies etc.: in fact the chemical environment of nuclei changes during the transition showing separate lineshapes depending whose intensity changes as the transition occurs giving information on kinetic and change rate.

The main drawback of this experimental technique is the limitation on the size of the analyzed molecule. Macromolecules have a larger number of resonances in the same spectral range, generating therefore lower resolution spectra where the resonance assignment is difficult. In addition, imposition of restrains and the use of molecular dynamics simulation to include them in the structure may introduce systematic errors due to the methods used in simulations or some modeler-dependent bias, which must be carefully considered. In spite of these problems, the structures resolved with NMR are to be preferred in cases in which the systems has a large conformational flexibility, which is not appropriately represented in crystallized structures.

## 1.2 THE PROTEIN DATABASES

Most of the experimentally resolved protein structures are deposited in the Protein Data Bank [4] (PDB) a public web database ([www.rcsb.org](http://www.rcsb.org)). The PDB, established in 1971 at Brookhaven National Laboratory, is freely available to the research community. Nowadays it is managed by the Research Collaboratory for Structural Bioinformatics (RCSB), and weekly updated.

The data base collects each structure in a “.pdb” file, with a specific format and labeled with a four-letter alphanumeric code. The submitted structure undergoes to several controls before acceptance for inclusion in the database. Each file contains the coordinates of all resolved atoms of the structure, and, when available, other supporting information such as the "temperature factor",

NMR Parameter	Structural Information	Averaging Time Regime
Chemical shifts	Secondary structure	$\mu\text{s}$ – $\text{ms}$
J-coupling constants	Torsion angles	$\text{ms}$ – $\text{s}$
NOEs	Inter-atom distances	$\text{ps}$ – $\text{ns}$ Longer for exchange
Heteronuclear relaxation rates $R_1$ , $R_2$ , and hetNOE	Global rotational tumbling, local flexibility, global shape	$\text{ps}$ – $\text{ns}$ Longer for exchange ( $R_2$ )
Residual dipolar couplings	Bond vector orientation relative to an external alignment frame	$\text{ms}$ – $\text{s}$
Pseudocontact shifts	Relative orientation and distance to a paramagnetic probe	$\mu\text{s}$ – $\text{ms}$
Paramagnetic relaxation enhancement	Distance to a paramagnetic probe	$\text{ps}$ – $\text{ns}$
Cross-correlated relaxation rates	Projection angles related to torsion angles	$\text{ms}$ – $\text{s}$
Hydrogen/deuterium exchange	Hydrogen bond stability	$\text{s}$ – $\text{days}$
Diffusion constants from diffusion-ordered spectroscopy spectra	Global shape, oligomeric state	Order of diffusion time

Table 2: Table of the main NMR-methods used in protein spectroscopy [2]. The structural information provided by the technique and the timescale of the measurement are shown in the second and third column respectively. The J-coupling is the indirect interaction between two nuclear spins which arises from the chemical bonds connecting the two spins.  $R_1$  and  $R_2$  are the to the spin-lattice and spin-spin relaxation rates. NOE (nuclear Overhauser effect) is the transfer of nuclear spin polarization from one nuclear spin population to another via cross-relaxation [3].

related to average mean squared displacement from the measured position, or the secondary structure arrangement. The repository is accessible via the RCSB website, which also provides useful tools to inquire structures and collect statistical sub-set with user selected properties. The PDB currently contains more than 99818 structures. Among these, 9520 are resolved using NMR spectroscopy. The continuous expansion of the PDB yields bigger and more updated statistics to the scientific community. The obsolete entries are signaled and the superseding structures are reported. In addition NMR-determined molecule generally the set of the best structures (usually about 20), which satisfy the experimental

constraints, is reported in the .pdb file. This yields more statistics and gives a picture of the conformations originally explored by the molecule.

In this work all the experimental data are extracted from the PDB-repository.

### 1.3 AMINO ACIDS AND POLYPEPTIDES - PRIMARY STRUCTURE

Proteins are among the most important biomolecules. They hold many roles ranging from the structural to functional ones. Proteins are finely structured hetero-polymers of the twenty amino acids. They are synthesized during the process called *translation* where the genetic code is expressed by the ribosome.

Amino acids (AA) represent the fundamental unit of proteins. They are composed by an amino group ( $\text{NH}_3^+$ ), a carboxyl group ( $\text{COO}^-$ ) both bonded to an  $\alpha$ -carbon ( $\text{C}_\alpha$ ), and by the side chain (R), which is also bonded to the  $\text{C}_\alpha$ , and the side chains, which characterizes each amino acid differing in length and the chemical content.

The  $\text{C}_\alpha$  is generally bonded with four different ligands forming a chiral center. Therefore, two isomers exist for each AA: L- and D- isomers (see figure 1). These are called “enantiomers” (or “stereoisomers”), related by mirror symmetry. In natural proteins the amino acids are all L-isomer.

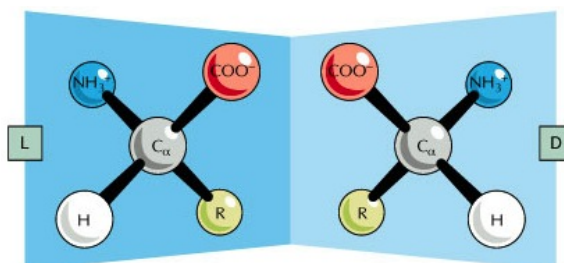


Figure 1: Schematic representation of the L- and D- isomers.

At neutral pH the amino acids are in zwitterionic form with a proton bonded to the amino group and a proton missing in carboxyl group. Therefore, amino acids are generally neutral although with charged extremal groups. Exceptions are the four naturally charged amino acids lysine, arginine (positive), glutamate, aspartate (negative), whose charge, however it is localized on the side chain. The side chain characterizes each amino acid: there are twenty distinct side chains that differing in chemical composition as reported in figure 2.

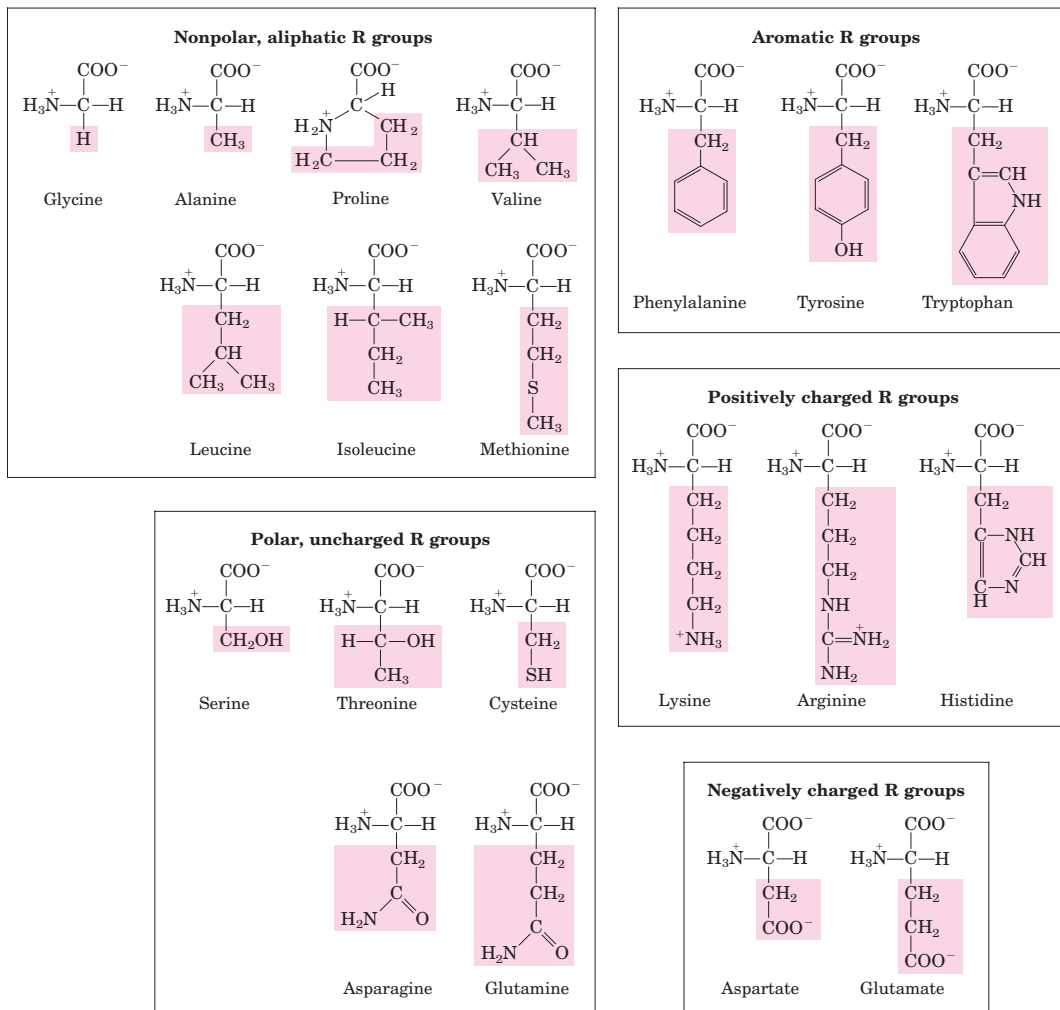


Figure 2: Chemical structure of the twenty natural amino acid and its ionization state at neutral pH. The amino acids are branched into their polarity and ionization state.

An important feature of each amino acid is the polarity. Non-polar amino acids have a low solubility in water. In this class the aliphatic amino acids (alanine, valine, leucine and isoleucine) fall. Although usually classified separately because larger, the aromatic amino acids, i.e. phenylalanine, tryptophan and methionine are also hydrophobic. On the opposite side of the scale, the charged amino acids and in between, the polar uncharged AA, namely, Asparagine, glutamine, serine and threonine, are expected to have strong interactions with water and a good solubility: the first two, thanks to the amine group; the latter two, due to the

large dipole moment and the capabilities to make hydrogen bonds.

The linkage between subsequent amino acids in the proteins and peptides (defined as sequences of 20-100 AA) is named peptide bond. It forms by a condensation reaction between the carboxyl group and amino group of subsequent amino acids operated by the ribosome (see figure 3a). The peptide bond has two chemically resonant forms (reported in figure 3b). One is represented by the neutral state where the oxygen atom is double bonded to the C' atom while there is a single bond between C<sub>α</sub> and N atoms (figure 3b, left). The other form involves the lone pair localized on the nitrogen atom to form a double bond with the carboxyl group while a couple of electrons from the carbon-oxygen bond becomes localized on the oxygen atom leaving a positive charge on the N atom and a partial negative charge on O atom (figure 3b, right). Intermediate forms are possible (fig 3b, center), as indicated by the measured value of bond length of about 1.32Å, which is a mean of the values 1.45Å and 1.25Å of the single and double bond respectively. The mixing of these two resonant forms gives the peptide bond a partial double bond character. As a consequence, the four atoms C<sub>α</sub>-C'-N-C<sub>α</sub> lie in a flat plane and the C'-N bond is not freely rotatable [5]. Therefore, the two isomers corresponding to *cis* and *trans* conformations of the C<sub>α</sub>s with respect to C'-N are not thermally interconvertible. They correspond to 0 and 180deg of the torsion angle ω respectively (C-N torsion angle). The 99.9% of the peptide bonds are in *trans* conformation in natural proteins [6].

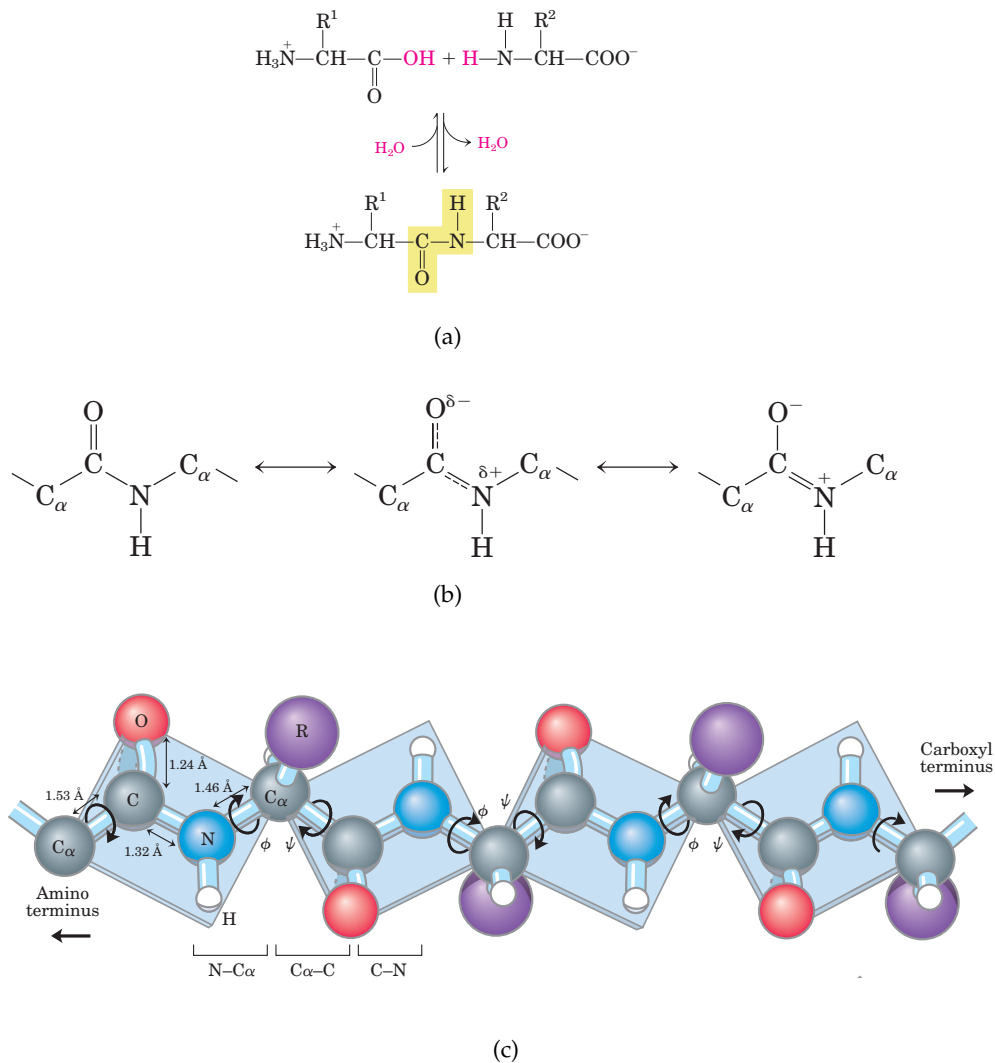
Other than the bond length and torsion angle of the peptide bond, it is important to define all the geometrical quantities related to the protein backbone, defined the chain of C<sub>α</sub>-CO-NH-C<sub>α</sub> atoms. These are illustrated in figure 3c. Their reference values are reported in table 3.

Bond	[Å]	Angle	[deg]
C <sub>α</sub> -C'	1.53	C <sub>αi-1</sub> -C <sub>αi</sub> -N (γ <sub>1</sub> )	15.6
C'-N	1.32	C <sub>αi+1</sub> -C <sub>αi</sub> -C' (γ <sub>2</sub> )	20.2
N-C <sub>α</sub>	1.47	N-C <sub>α</sub> -C' (τ)	111.1

Table 3: Values of the internal geometric parameters of the protein backbone [6][7]. See also figure 3c for the parameters definition.

As a consequence of the planarity of the peptide bond in *trans*-conformation the distance between two consecutive C<sub>α</sub> is 3.8 Å (2.9 Å if in *cis*-conformation). The rigidity of the peptide plane combined with the small variability of the τ, γ<sub>1</sub>





**Figure 3:** (a) Condensation reaction between two amino acids and formation of the peptide bond. (b) The peptide bond resonances. (c) graphical representation of the polypeptide chain in extended conformation. The peptide bond plane is represented in light blue. The torsion angles  $\phi, \psi$  and the backbone geometric parameters  $C_{\alpha}-C, C=O, C-N, N-C_{\alpha}$  distances are indicated, see also table 3.

and  $\gamma_2$  makes the dihedral angles  $\phi$  and  $\psi$  the only two independent geometrical variables determining the backbone structure (fig. 3c).

As previously mentioned, amino acid sequences inside proteins stem from the translation of the genetic code carried out by DNA. The mere sequence represents the first hierarchical structural level of the protein and for this reason it is also

referred as protein primary structure. The polypeptide chain is oriented: positive sign is conventionally positive from N-terminal ending to C-terminal.

In principle, the protein behavior is determined by its sequence. On the basis of the “structure function paradigm”, stating that function is related to the global 3D structure of a protein, this also implies that sequence determines the 3D structure. As it will be clearer forward in this thesis, the paradigm has been recently reconsidered. However, for most of the protein classes, namely the well structured ones, it is still mostly valid. Even in this case, though, the problem of predicting the global structure of a protein from its sequence is still open.

There are however a number of sequence-based estimators of the structure. A large subset of those algorithms belong to the class of Homology-Modeling. These estimators analyze the primary structure of a protein, whose 3D-structure is to be predicted, comparing its sequence with the sequence of proteins with known 3D-structures. The idea under these approach is related to the structure-function paradigm: similar sequences carry out similar functions and due to the structure-function correlation, similar structures. When the sequence homology between target and template is above the 80% homology-modeling shows good results.

#### 1.4 SECONDARY STRUCTURE

The protein secondary structure is the local arrangement of a polypeptide chain. The Ramachandran plot (RP) is one of the most useful tool for the analysis of the protein secondary structure described by Ramachandran in 1968 [5]. It is defined as the scatter plot of the backbone  $\phi_i, \psi_i$  preceding and following the  $i$ -th  $C_\alpha$  evaluated on structural datasets. Since  $\phi$  and  $\psi$  are the internal variables determining the backbone folding, the accumulation region in the RP identify specific ordered local arrangement, defining the ordered secondary structures (see figure 4). For instance, the RP of helical structures accumulate around the region  $\phi = -57\text{deg}$ ,  $\psi = -47\text{deg}$ , because these are the values that the  $\phi, \psi$  dihedral angles assume in the ordered helical structures (figure 5a), while extended structures (figure 5c) and sheets (figure 5b) accumulated around  $\phi = -100\text{deg}$ ,  $\psi = 150\text{deg}$ . The content of the dataset determines the shape of the RP, which on the other way round, can give an immediate idea of the relative content of ordered secondary structures. In addition, it is common to

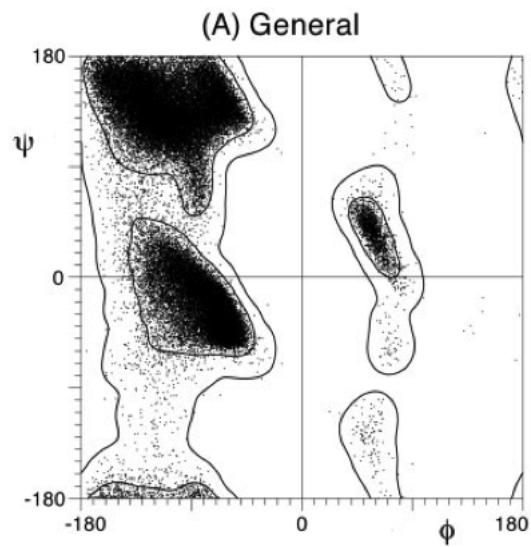


Figure 4: Ramachandran plot of the non proline, non glycine and non pre-proline amino acids taken from a database of 97,368 residues at high resolution X-ray (from [10]).

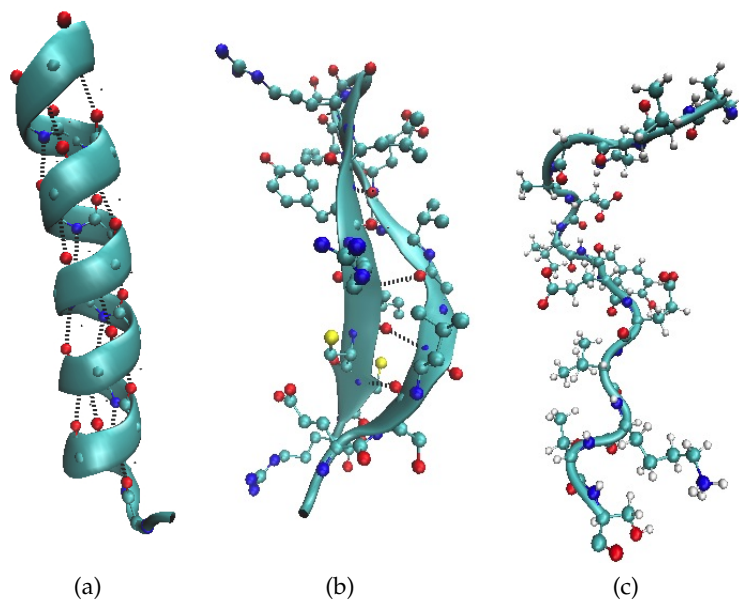


Figure 5: Example of protein secondary structure. For helix (left) and sheet (center) are reported hydrogen bond patterns. Absence of hydrogen bonds in coil structure (right).

build amino-acid specific RPs to identify the secondary structure tendency of a specific (class of) amino-acid (see figures 4 and 8). The case of large dataset it is

common to analyse the density of this quantity therefore in this work also such a plot will be referred as RP.

Figure 6a shows that not all the regions of the RP are accessible. In order to

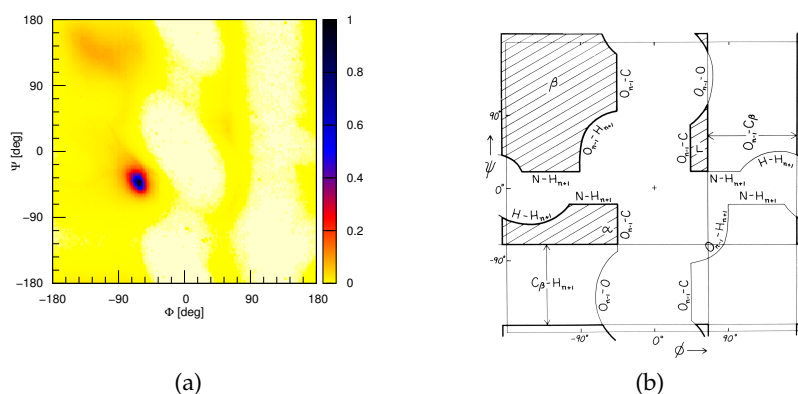


Figure 6: (a)  $(\phi, \psi)$  density distribution from all available PDB-NMR data. (b) “Derivation diagram” for the Ramachandran plot [8]. The boundaries are evaluated considering the hard-sphere model [5] (see text).

better analyze the forbidden areas it is useful to separate the AA in four classes which display different structural namely different RP (as shown in figure 4): glycine, proline, pre-proline<sup>1</sup> and all the others. Considering first the last class, early studies of polypeptide chains only including the steric hindrance by means of hard spheres models were able to recognize the basic shape of the forbidden areas on RP [5]. The radii were set up using a specific contact list, similar to those of the Van der Waals interaction (VdW), for each couple of atoms. The resulting map is reported in figure 6b and is called “Derivation Plot”. The interaction between an amino acid and its nearest neighbor is also reported therein. As indicated in figure 6b, the different steric clashes between the backbone and side chains atoms concur to the boundaries between allowed and forbidden areas, and delimit basically three populated regions, namely the right handed helical region ( $\alpha$ ), the flat structures region ( $\beta$ ) and the left handed helical region (L). Remarkably, the different size of the allowed regions for the left and right handed helices is determined by a complex interplay between the directionality of the polypeptide and chirality of the  $C_\alpha$ .

<sup>1</sup> The pre-proline is the type which collects all the residues preceding a proline along the polypeptide chain.

Figure 7 reports the RP of a simulated tetrapeptide of alanine using an atomistic model. This simulation was performed *ad hoc* for the sake of illustrating the RP (details on the simulation setup reported in section 2.2.1). Due to its small size, this system cannot form any stable secondary structure, therefore it is able to sample all the allowed regions with low structural preference. In addition, thanks to the small size the simulation can give a very large statistics and therefore a high resolution RP. The figure shows the  $\alpha$ ,  $\beta$ , and L basins and their substructures, and the PPII basin in addition, which is mostly favored in polyproline. The same basins appear in the RP derived from experimental dataset [9], reported in figure 4. There are some outliers in the bridging regions between the allowed ones. Their presence can be explained by the fact that in real proteins the structural parameters  $\tau$ ,  $\gamma_s$  and the radii of the side chains do not assume sharp values, rather they have a distribution with a given width [5]. Therefore a realistic derivation plot should have a fuzzy boundaries between allowed and forbidden regions, as, in fact, verified by this simulation and in the experimental RP.

Panel (B) of fig. 8 shows the glycine RP. The observed additional symmetry and smaller forbidden areas in this plot is produced by the achirality of glycine and extremely reduced side chain. Panel (C) illustrates the proline RP. This amino acid has the side-chain linked to the N atom of the backbone forming the pyrrolidine ring (see figure 2). This produces a strong restrain  $\phi$  torsion angle: indeed the RP shows this variable restrained in the interval  $[-90 : -40]$ . The last panel in figure 8 shows the pre-proline RP. As evident from the RP, the geometric restrains of the pyrrolidine ring of proline also acts on the preceding AA, inducing deformations with respect to the generic RP.

As mentioned, the polyproline secondary structures are stabilized by restrains on the backbone geometry induced by the presence of pyrrolidine ring. In addition, the lack of the amide-H atom makes this amino acid unable to create intra-backbone H-bonds therefore the secondary structures stabilized by the H-bonds are less probable. This is however an exception among the ordered secondary structures, which are generally stabilized by the hydrogen bonds (H-bonds) of the backbone. These arise between the interaction of the carboxyl group (C=O) of a given amino acid, which acts as proton acceptor, with the amine group (NH) of another, which acts as proton donor establishing a rather strong and stable linkage between two not subsequent amino-acids in the polypeptide. The energy of this interaction has a large variability ranging in the order 3 – 8kcal/mol [5]. In addition, the C=O and NH groups are polarized, and their alignment within

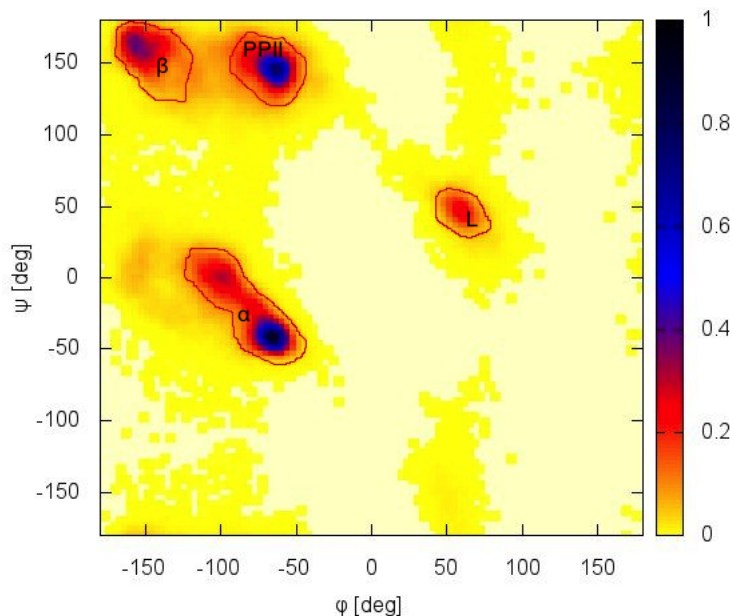


Figure 7: Density of the Ramachandran plot of a simulated tetrapeptide. The cumulative regions are labeled with the name of the specific basin. The contour of each region sketches outline the shape of each secondary structure basin. These data are obtained from an atomistic simulation (section 2.2.1). Details on this simulation are reported in section 2.2.1.

ordered structures may generate strong dipolar interaction which further stabilize the structure. This is for instance the case of helices, in which H-bonds are aligned along the helical axis (see figure 9a and 4).

Hydrogen bonding is responsible for the secondary  $\alpha$ ,  $\pi$ ,  $3_{10}$  helical,  $\beta$ -sheet and turn structures. Among the the H-bonded secondary structures the helices and sheets have more than one H-bond and are differentiated by their regular structure schematically reported in panel (a) and panel (b) figure 9 respectively.

The family of the helices is split into specific structures that differ by number of residues per turn ratio, diameter and circular radius. Helices are usually right handed, although glycine and proline also allow left handed helices. The most stable helix structure is represented by the right handed  $\alpha$ -helix (see the

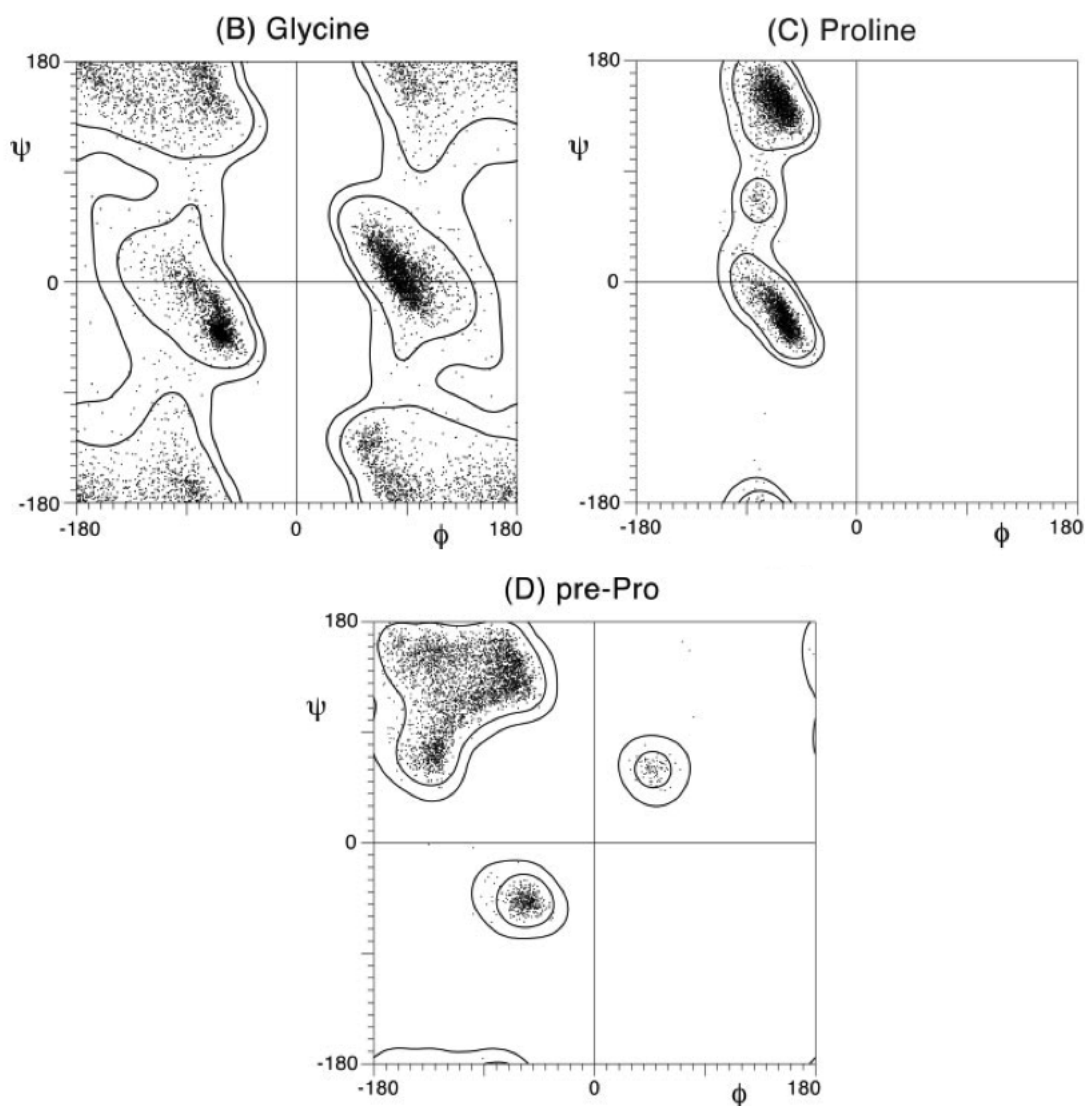
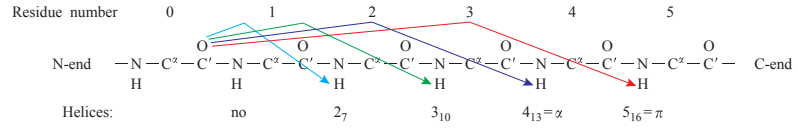
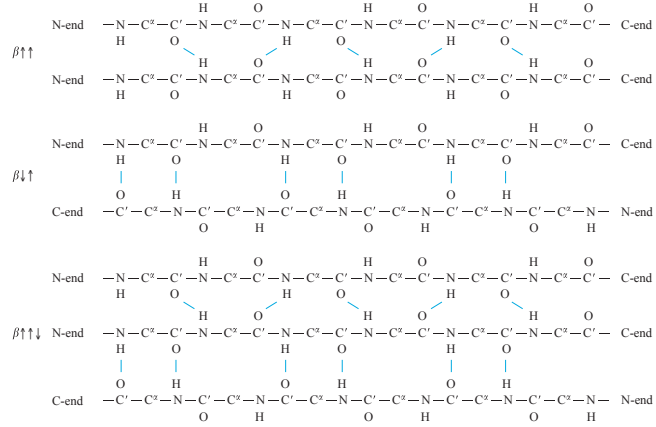


Figure 8: In this figure are reported the Ramachandran plots of proline (C), glycine (B), pre-proline (D) taken from a database of 97,368 residues at high resolution X-ray (from [10]).

prevalence in figure 6a), which is also known as  $3_6$ -helix. Other stable structures are the right handed  $3_{10}$ -helix and  $\pi$ -helix ( $4_4$ -helix) and the left handed  $\alpha$ -helix. Differing in the H-bond pattern  $\alpha$ -helices display H-bond between  $i$ -residue and  $i+4$ -residue where  $3_{10}$  and  $\pi$  helices between  $i/i+3$  and  $i/i+5$  residues respectively. Thanks to their stability  $\alpha$ -helices are the most abundant protein structure. The geometry of helices is summarized in table 4. The  $\alpha$  helix is also represented in



(a)



(b)

Figure 9: (a) Schematic representation of the H-bond pattern of the  $\alpha$ ,  $\pi$ ,  $3_{10}$  helices. (b) The possible H-bonds textures of the conformations of the  $\beta$ -sheet .

figure 5a. Due to a combination of chirality and directionality, left-handed helices are less probable than right-handed (compare  $\alpha$  with L region in the generic RP, fig. 7).

The fundamental element forming the sheets structure is the  $\beta$ -strand. Two or more H-bond can form a  $\beta$ -sheet. The strands can be either parallel or antiparallel oriented (generally considering the orientation of a strand positive from the N-terminal to the C-terminal). The sheets torsional angles are found inside the  $\beta$ -labeled region of the RP (figure 7) (main values reported in table 4).

The structure connecting the strands within the sheets can be either a turn, in the case of consecutive antiparallel strands, or a longer loop structure (see later) that can occur both in antiparallel and parallel strands. The turns are considered independent secondary structures. Turns are generally small fragments of few residues stabilized by H-bonds. The composition of  $\beta$ -strands can be more complex than what is reported in figure 9b. A complete analysis of these structures can be found in [13]. The aggregation of strands influences the stiffness



Struct.	Res/Turn	$\phi$ [deg]	$\psi$ [deg]	$\omega$ [deg]	Pitch (Å)	HBC
$3_{10}$	3.1	-49	-29	180	6.0	i/i+3
$\alpha_R$	3.6	-57	-47	180	5.4	i/i+4
$\alpha_L$	3.6	57	47	180	5.4	i/i+4
$\pi$	4.4	-57	-70	180	5.2	i/i+5
$\beta$ -strand		-120	120	180		
parallel $\beta$ -sheet		-120	113	180		
antiparallel $\beta$ -sheet		-139	135	180		
PPII	3.0	-75	145	180	9.4	
PPI	3.3	-75	160	0	5.6	

Table 4: Geometrical parameters of secondary structure [11][12].

of the structure. Figure 5 panel (b) shows the 3D-representation of an antiparallel  $\beta$ -sheet sketching the H-bond texture with black dashed lines.

The polyproline fragments are found in two regular arrangements which are determined by the conformation of the peptide bond (i.e. cis or trans conformations). As aforementioned the presence of the pyrrolidine ring limit the interval of the  $\phi$  torsion angle. The polyproline I (PPI) is a right-handed helix with peptide bond in cis-configuration whereas the polyproline II (PPII) is a left-handed helix with peptide bond in trans-configuration. The geometrical features of these structures are reported in table 4. The PPII structure can be found also in non-proline AAs. Differently from the other secondary structure PPII does not rely much on dipolar interaction for its stabilization [7]. Shi [14] demonstrated that PPII is the conformation adopted by a protein in a environment with strong denaturants. In this case it is not appropriate call it secondary structure since the protein has lost its natural folding, however this fact indicate that the PPII needs an high accessibility to the solvent, as that found in denatured proteins. For this reason it is expected that long protein fragments in PPII conformation are composed by polar and charged amino acid [14].

The disordered secondary structures are generally found where stabilizing interactions are absent. This provides an high flexibility to the structure. In native (correctly folded) proteins disordered regions are generally found in the loops, connecting two regions with ordered secondary structure. Even in absence of hydrogen bonds, loops have not the same freedom as the unfolded fragments due to the fact that their extremes are bond to the protein. The loop secondary structure is favored either by the presence of locally structured neighbor regions

(such as in beta sheet) or by the whole arrangement of the protein (i.e. tertiary structure of the protein, see next section).

Due to their regularity, ordered secondary structures occupy specific regions of the RP therefore it is possible to distinguish one from the other using this representation. In contrast, disordered structures are expected to be spread approximately in whole allowed region of the RP including those assigned to the regular structures. The unstructured regions in non-ordered native state will be discussed in section 1.6. Figure 5 panel (c) shows an example of disorder structure.

Amino acids generally show propensities for a specific secondary structure, due to the preference for specific torsion angles, peculiar side-chain interactions, steric effects and hydrophobic tertiary contacts [15]. The  $\phi, \psi$ -propensity has been evaluated considering the statistics of a subset of the PDB, called “coil library”, in each structured basin, viz.  $\beta$ ,  $\alpha$ , PPII and the remaining regions. Table 5 shows the propensity of each basin for each type of AA [15]. Since coils are considered the less structured parts of proteins, these numbers indicate the intrinsic propensity of amino acid. Hydrophobic and bulky amino acids like valine and isoleucine prefer the  $\beta$ -conformation in ordered and disordered states. Protonated polar aspartic acid has a high  $\beta$ -basin propensity over PPII, which might reflect the stabilization by the side chain. This table also shows the propensities of each amino acid when it is involved in an ordered secondary structure. The comparison between intrinsic and general propensities can give indication in the model building. For instance, it is interesting to observe that although the alanine and leucine unfolded residues have a high propensity in the PPII basin, they have a high predisposition to fold in the  $\alpha$ -helix structure indicating that the role of the hydrogen bonding in determining the secondary structure is very high, possibly preponderant.

The identification of secondary structures, when the whole protein coordinates set is available, is usually performed by assignment algorithms based on the evaluation of  $\phi \psi$  angles and on the search of hydrogen bonds. Different algorithms differ by the chosen criterion in the assignment of H-bonds, and might lead to slight different results. The most popular are DSSP [16] and STRIDE [17] also important SEGNO [18], PROSS [19] and XTLSSTR [20]. The secondary structure identification has a central role in the determination of the data base used in this work. A detailed description of the algorithms here used DSSP and STRIDE is reported in appendixes B.

Amino Acid	Intrinsic $\phi, \psi$ propensities					Regular secondary structure propensities	
	a/coil	b/coil	p/coil	B/coil	other	$\alpha$ -helix	$\beta$ -strand
Gly	0.33	0.34	0.31	0.32	3.80	0.41	0.64
Ala	1.23	0.78	1.32	1.09	0.39	1.47	0.79
Val	0.89	1.83	0.96	1.33	0.36	0.95	1.73
Leu	1.16	0.82	1.40	1.15	0.35	1.32	1.17
Ile	0.98	1.68	0.99	1.29	0.32	1.13	1.76
Phe	0.93	1.63	0.93	1.23	0.55	1.04	1.39
Tyr	0.85	1.46	1.12	1.26	0.59	0.88	1.52
Trp	1.17	0.90	1.24	1.09	0.48	1.05	1.25
Pro	1.00	0.10	2.29	1.35	0.14	0.46	0.42
Cys	0.87	1.34	1.32	1.33	0.41	0.89	1.18
Met	1.07	1.05	1.23	1.15	0.51	1.37	1.32
Ser	1.29	0.95	1.00	0.98	0.56	0.71	0.93
Thr	1.13	1.39	0.96	1.15	0.43	0.71	1.27
Lys	1.20	1.07	0.94	0.99	0.68	1.10	0.92
Arg	1.09	1.40	0.74	1.03	0.77	1.41	0.71
His	0.93	1.37	0.84	1.07	0.95	0.97	0.86
Asp	1.16	1.18	0.82	0.98	0.80	0.85	0.49
Asn	0.79	1.35	0.60	0.92	1.54	0.78	0.56
Glu	1.45	0.84	0.95	0.90	0.50	1.39	0.78
Gln	1.26	1.07	1.00	1.03	0.48	1.36	0.81

Table 5: Swindells’ propensities of amino acids for the selected structures [15]. The coil-propensities represent the tendency of each amino acid based on the assumption that the structural propensity of a particular amino acid can be derived from the coil library (a more detailed description of the insight of this subset is reported in sec. 1.7). The assumption that these values represent the intrinsic propensity of each amino acids lies in the hypothesis that the context is averaged throughout the dataset.

Experimental determination of the secondary structure contents can be achieved using circular dichroism spectroscopy in the far-UV frequencies (190 – 240 nm) (far-UV CD). This technique measures the difference in the absorbance of left versus right circularly polarized light, and it is therefore sensitive to the chirality of the sample. These measurements are in general expressed as ellipticity<sup>2</sup>. The far-UV light scattering is dominated by the peptide bond absorption [21] whose environmental chirality depends on the secondary structure. Therefore each secondary structure has a characteristic spectrum. This spectroscopy provides

<sup>2</sup> The ellipticity is defined as:

$$\theta = \text{atan}\left(\frac{E_R - E_L}{E_R + E_L}\right) \quad (1.5)$$

where  $E_R$  and  $E_L$  represent the magnitudes of the electric field vectors of the right and left circularly polarized light beams. The ellipticity can be directly measured through absorbance experiments using polarized light.

a quantifiable contents of secondary structure of protein. Similar information is achieved by differential IR Fourier transform (FTIR) spectroscopy analysis of the amide vibrational modes. These vibrations are identified by specific frequency peaks around  $1500 - 1550\text{cm}^{-1}$  and  $2800 - 3000\text{cm}^{-1}$  in the IR spectrum, and are localized on NH and CO groups of the backbone. Therefore their frequencies are extremely sensitive to the type and environment of H-bonds, and each secondary structure has a recognizable spectral signature [22].

### 1.5 TERTIARY AND QUATERNARY STRUCTURES

The three-dimensional space arrangement of the secondary structure and consequently of all atoms in a protein is referred to as the protein's tertiary structure. A protein tertiary structure is often composed by structural domain with defined secondary structure. Whereas the secondary structure define the local conformation of the protein, the tertiary structure includes the long range effects inside the protein. The main interactions involved in the stabilization of tertiary structure are the salt bridges, H-bonds, disulphide bonds, dipolar interactions, hydrophobic and VdW interactions. The salt bridge is the interaction between side-chains with opposite charge, therefore it usually occurs between the charged extremities of aspartic/glutamic acids and lysine/arginine. Salt bridges between polar side chains can occur in specific pH conditions, or between same charge residues if mediated by opposite charge ions (usually metals or halides). The intra-backbone H-bonds are usually all saturated in the stabilization of secondary structures, but additional H-bonds can still be formed by donor-acceptors groups of side chains, producing inter-domain (or even inter-chain) interactions.

The disulphide bonds in proteins are the linkage between two sulphhydryl group contained in the cysteine residues. The formed amino acid-complex is called cystine. The dipolar interactions, between side chains of different domains also play an important role in the stabilization of the tertiary structure.

Finally, the VdW and hydrophobic interactions are the weaker but more ubiquitous interactions, therefore contribute mostly to the stabilization of the protein. Specifically hydrophobicity is a driving force of folding and constitutes the main energetic contribution to protein stability [23] and VdW regulates local contacts, avoiding the protein collapse.

The quaternary structure, if present, is the aggregation of subunit with established tertiary structure. The interactions involved at this level are almost the same of the tertiary structure but in this case they may involve two or more different subunits.

A property which can characterize a tertiary structure is its “globularity”, related to the amount of hydrophobicity and to the formation of the so-called hydrophobic cores, namely delimited regions with high hydrophobicity and low solvent exposure. Experimentally, the hydrophobic fluorescence probes can be used to detect the hydrophobic regions of proteins. These probes are quenched when they come in interaction with the solvent. The ANS hydrophobic probe provides a strong fluorescence with a large blue shift of the peak. The analysis of the fluorescence spectra provides useful information regarding accessibility to the solvent in the protein and the formation of the hydrophobic core.

## 1.6 PROTEIN FOLDING AND INTRINSICALLY DISORDERED PROTEINS

After the translation and the polymerisation processes, the protein is found in a disordered conformation often referred to as unfolded state. It must then perform a structural transition (protein folding) to native low entropy state. The native structure is generally stable, at the macroscopic ( $> 10^2$ s), while the folding process, which is experimentally expected to be completed in microseconds although with a wide range of variability [2]. This relatively short time scales is somehow unexpected: as estimated by Levinthal, a systematic exploration of the conformational space of a protein, could be achieved only within astronomical timescales since the number of conformations to explore is comparably astronomical. This problem is known as Levinthal’s paradox<sup>3</sup>. In order to explain the experimentally observed behaviour, it has theorized that the energy landscape of a protein looks like a deep funnel [2]. The analysis of the pathways followed during the protein folding is one of the hardest and important challenges in theoretical biophysics.

For many years it has been thought that the function of a protein is entirely determined by its native structure and *vice versa*. Namely a specific function

---

<sup>3</sup> In E. Coli cell a protein containing 100 amino acids becomes active in 5 s at 37 °C [6]. Considering that for each residue could take up 10 conformations, the protein has  $10^{100}$  conformations. For an extensive exploration of the available configurations, assuming that it takes the shortest time for atomic scales ( $10^{-13}$ s), it would take  $10^{77}$  years to find the stable conformation.

demands specific spatially defined structure. Thanks to this “structure function paradigm” many prediction were made, from the knowledge of either the structure or the function, including all the theoretical structural models based upon the already mentioned homology modeling method. In the last decade a new theory of the protein functionality arose supported by new experimental observations. The old structure function paradigm has been replaced by the new one, which includes the activity of the intrinsically disordered protein (IDP). IDP is the standard name currently accepted for a class of protein and found with different names [2], characterized by the presence of at least of one natively (intrinsically) disordered region (IDR). The definition of this class besides the standard structural classes represents a revolution in the field of the protein science. The most accepted sub-classification of IDPs recognize their distinction in molten globule (MG), pre-molten globule (pMG) and random coil (RC) natively states differing by the amount of (secondary and tertiary) ordered structures. The MG has disordered tertiary structure but preserves considerable amount of ordered secondary structures organized in a globular form. The pMG has a disordered global arrangement with a residual secondary structure. The RC represents the lowest level of protein order with little residual amount of secondary structure.

The X-ray crystallography is not appropriate in the determination of IDRs, because disordered regions does not provide coherent scattering. Therefore the corresponding atomistic coordinates are usually absent in the X-ray structures, and one can only get information about the location of IDRs along the sequence. The MG state, thanks to the presence of the hydrophobic core, can be observed using the ANS hydrophobic probe (see section 1.5). Figure 10 shows the increase in the fluorescence signal and blue shift passing from the ANS probe alone (continuous line) to structures with increasing presence of hydrophobic cores, ( $\alpha$ -synuclein, RC state, dotted line, caldesmon 636-771 fragment, pMG state, dash-dot line and  $\alpha$ -lactalbumin, MG state, dashed line). This method, however, does not allow to resolve pMG from RC state due to their low hydrophobic content. For their investigation, far UV-CD is more appropriate. Specifically, the correlation plot of ellipticity measured at 222 and 220 nm was shown to distinguish these two states, as shown in figure 11. These two methods, combined with other measuring the presence of secondary structures such as FTIR can identify and classify the IDPs and IDR.

The IDPs are involved in regulation, signaling and control of the metabolic pathways, complementing the functional roles assumed by the other protein

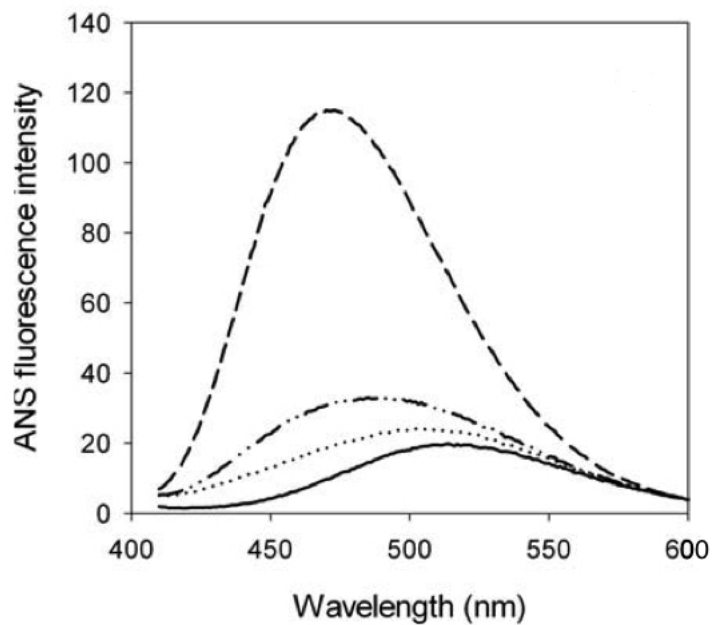


Figure 10: ANS fluorescence spectra measured for free dye (solid line), in the presence of natively disordered coil-like  $\alpha$ -synuclein (dotted line), natively disordered pre-molten globule-like caldesmon 636-771 fragment (dash-dot-dotted line), and molten globule state of  $\alpha$ -lactalbumin (dashed line), extracted from [21].

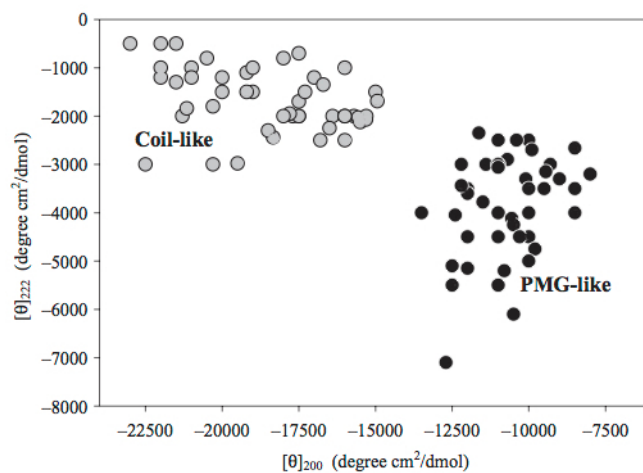


Figure 11: Far-UV CD spectra in terms of double wavelength plot, 222 versus 200 nm, allows the natively unfolded proteins division on coil-like (gray circles) and pre-molten globule-like subclasses (black circles), extracted from [2].

classes. Thanks to their flexibility IDPs have several ways to interact with different molecular targets. This ability is fundamental for many the cellular functions. An interesting observation is that the IDPs have been mainly found in the eukaryotic's rather than in the prokaryotic's and archaea's cells [2]. This suggests that they represent a step forward in the evolution of the cell signaling system. Further details on IDPs' functionality are reported in appendix A.

At variance with the native folding mostly stabilized by the hydrophobicity in IDPs the intrinsic disorder is associated to the high content of charged and polar amino acids, in addition to the small and structure destabilizer glycine and proline with respect of the hydrophobic ones. These prevent or at least hinder the formation of the hydrophobic core. IDPs have low content of Cys amino acid in order that the structural arrangement is not stabilized by the disulphide bridges.

The IDPs cannot be studied considering just one configurational arrangement. The lack of a standard procedure to sample the conformations of IDPs represents an obstacle, which brought contradictory results in the literature [2]. The statistics of IDPs data is low. In fact, X-ray crystallography is useless, and the standard NMR-methods used to sample the conformation of the proteins in native state are not easily transferable to IDPs due to the difficult resonance assignment [24]. The intrinsic motion of this class of proteins does not allow to a simple understanding of the H-NMR spectra. For IDPs the analysis of the  $^{13}\text{C}$ -resonance is more suitable, though this brings additional technical problems [24]. In addition all the available NMR measurements are performed in aqueous solution *in vitro* assuming that the structure does not change appreciably the extremely crowded cellular environment. While this is reasonable for natively folded proteins, IDPs behavior is thought to be in general much affected by the crowding. Therefore the in-cell sampling represents the preferential way of studying the IDPs structure-function relationship. It is however a hard task, due to difficulties in the sample preparation labeling and survival of the cells during the measurements. The identification of a standard method for IDPs structural determination is still under development.

The few available structures of IDPs are collected into the DisProt [25] and IDEAL [26] are databases and classified according to their function.



## 1.7 RANDOM COILS

The reference state for this Thesis work is the random coil, which is, the state with less amount of ordered structures, and therefore the largest amount of disorder. Proteins are not “ideal” random coil (i.e. with normal distribution of internal variables [27]), rather as explained in the previous section, they are associated to given stages of denaturation or unfolding, in the following, it will be referred to an ideal state of random coil in proteins, as a state in which regular structures are (almost) absent. Clearly, this is associated with the absence of hydrogen bonds (especially the intra-backbone ones).

The structural characterization of RC is not straightforward: RC can be experimentally identified only based on spectroscopic data which deliver no structural information. Small peptides (upto 8 residues) are a possible source of structural information for RC, because of their inability of making secondary structure. For these there are spectroscopic data, but few structural one [2], therefore in order to have appropriate statistics and better interpret experiment, these are complemented by atomistic simulations.

In spite of these problems, an RC dataset can be build, based on a “coil library” build from crystallographic data [2][28]. These include a residual amount of helical and other ordered structures, which can be however recognized and subtracted to obtain an RC dataset. The subtraction procedure itself is straightforward, requiring often the elimination not only of the ordered structures but also of the amino-acids flanking them, which include structural correlations. It was shown, however, that the dataset build with this procedure has a good correlation with experimental determination of unstructured peptides, in terms of the NMR-J coupling constant, often considered as an estimator of the regular structure propensity [29].

In [28] it has been evidenced that a bias toward the PPII structures derived from the residues flanking the fragments with ordered secondary structure. These residue following an helical fragment cannot be in the  $\beta$  basin for steric reason, and it is unlike that this residue is in the  $\alpha$  basin, otherwise it should belong to the structured region. Therefore, there is likely to be a bias toward PPII structures, which should be eliminated to have a purely unstructured database. For these reason the flanking residue should be discarded. In addition, the use of non-redundant dataset, obtained inserting a threshold on the maximum sequence homology, prevents from bias toward a specific structure. As already mentioned

(section 1.4 and figure 7) even in the regular structures purged coil dataset a tendency towards recurrent structural arrangement emerges which is considered “intrinsic”. The correct representation of this intrinsic tendency in a simple model is the main goal of this work. It also emerges that even if to a minor extent this tendency is amino-acid type dependent, and non local along the chain. These aspects will be re-considered when the model build in this work is presented in chapter 3.

It is finally to be remarked that the coil dataset includes only short peptides, therefore statistical data therein cannot accurately account for possible long range interactions. However, as it will be clearer in chapter 3, this might be considered an advantage for this work because it allows to separate and better model the short range effects. In spite of all the mentioned problems the coil dataset build as explained currently represent the best dataset for “ideal RC”, defined as the structures with less possible amount of regular secondary structures.

## PROTEINS MODELING

---

In the last decades protein simulations have become an invaluable tool for investigation of protein systems. This chapter provides the bases of classical molecular dynamics simulations in the framework of proteins biophysics which can be applied to any kind of classical model Hamiltonian describing interaction in protein. Two specific cases are subsequently illustrated, namely atomistic and coarse grained models for proteins.

The last section of this chapter focus on the subject of this thesis work, namely the  $C_\alpha$  based one bead models. Although being mainly descriptive, this section also report an elaboration of the Ramachandran plot to build its “minimalist” equivalent, which can be considered the first original result of this Thesis.

### 2.1 CLASSICAL MOLECULAR DYNAMICS

The molecular dynamics of large bio-molecules is which is not easy to access experimentally. Therefore computer simulations can give a great support, by simulating and giving a representation of the dynamical trajectories of molecules. However, obviously, the computational cost of a simulation and therefore the possibility of reaching sufficient time-size scales, is dependent on the level of accuracy at which the system is treated [30]. To this respect, in the following, different possible modeling approaches are reviewed.

The the complete Hamiltonian of a molecular system is generally separated in a nuclear and electronic part using the Born Oppenheimer approximation. The wave functions of electrons is evaluated considering the nuclei in a fixed position in the space. The nuclei interact with each other under the influence of electrons treated within a mean field approach. This is possible because the mass of electrons is much smaller than that of nuclei, and therefore the former usually

adiabatically follow the motion of the latter. The BO approach defines the energy of each electronic level (in particular the ground state) for each coordinate set. These  $U(\mathbf{r}_1, \dots, \mathbf{r}_n)$  are also called potential energy surface (PES) (one for each electronic state) and can be used as effective potential energies for the nuclear motion. In the BO approach they are evaluated by solving the electronic problem at each given nuclear configuration. This approach, defining the class of atomistic *ab initio* methods, is the most accurate and computationally expensive to address a system, because it implies the solution of a Schrödinger like equation for an N-electron system for each atomic configuration encountered.

Alternatively, one can evaluate  $U(\mathbf{r}_1, \dots, \mathbf{r}_n)$  empirically, as will be described in the next section. In any case, however, the classical mechanics for the nuclei is assumed. In fact, while the quantum mechanics is obviously necessary to describe the electrons, quantum effects for the heavier nuclei are expected to be negligible at room temperature<sup>1</sup>. Therefore the Newton's equation of motion of a system of N particles with mass  $m_i$  is to be numerically solved. Each interacting center represents an atom, in this case, but the following formalism stands also for coarse grained classical interacting centers (defined in the next section) representing rigidly moving group of atoms. The Newton's equation of motion, to be numerically integrated, is then:

$$\frac{d\mathbf{r}_i}{dt} = \mathbf{F}_i = m_i \ddot{\mathbf{r}}_i = -\nabla_i U(\mathbf{r}_1, \dots, \mathbf{r}_N), \quad (2.1)$$

The force  $\mathbf{F}_i$  is determined by the gradient of  $U$  with respect to the coordinates of the particle  $i$ . The characterization of the particular system is obtained with the definition of the potential energy function, which, eventually determines the evolution of the system according to the classical dynamics.

The numerical integration of equation 2.1 represents the simplest method to investigate the dynamical properties of a complex system. There are different integration strategies which differ in the quality of the calculation and in the ability to sample a sufficient number of configuration. In the following section the derivation of the integration algorithm used in this work is reported.

---

<sup>1</sup> With the only exception of the lighter nucleus, hydrogen, which in fact is often treated quantum mechanically.

### 2.1.1.1 Integration algorithms for equations of motion

In molecular dynamics the integration of the Newton's equations of motion is implemented by numerical algorithm with discrete time sampling characterized by the time step  $\Delta t$ . Taylor expansion gives a first approximation:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t) \Delta t + \frac{1}{2m_i} \mathbf{F}_i(t) \Delta t^2 + o(\Delta t^3). \quad (2.2)$$

The Verlet approach is obtained summing and subtracting expansions at  $+$  and  $-$   $\Delta t$ :

$$\mathbf{r}_i(t + \Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta t) + \frac{1}{m_i} \mathbf{F}_i(t) \Delta t^2 + o(\Delta t^4). \quad (2.3)$$

Being exact at the fourth order in  $\Delta t$ , this algorithm reduces the integration errors at given  $\Delta t$ . This algorithm is independent from the evaluation of the velocity, but it can still calculated as follow:

$$\mathbf{v}_i(t) = \frac{\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t - \Delta t)}{2\Delta t}. \quad (2.4)$$

In addition, Verlet algorithm (VA) satisfies the important property of the Newton's equations of motion, namely the time reversibility: the algorithms depends only on the even derivatives of the motion, therefore it is invariant for time reversal operations. This algorithm has two variants, which reduce the amount of global error at the expenses of an increase computational costs. These are "leapfrog" (LF) and "Velocity Verlet" (VV):

$$\text{leapfrog} \begin{cases} \mathbf{v}_i(t + \frac{\Delta t}{2}) = \mathbf{v}_i(t - \frac{\Delta t}{2}) + \frac{\mathbf{F}_i(t)}{m_i} \Delta t \\ \mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \frac{\Delta t}{2}) \Delta t \end{cases} \quad (2.5)$$

$$\text{Velocity Verlet} \begin{cases} \mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t) \Delta t + \frac{1}{2m_i} \mathbf{F}_i(t) \Delta t^2 \\ \mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t) + \frac{1}{2m_i} \left( \mathbf{F}_i(t) + \mathbf{F}_i(t + \Delta t) \right) \Delta t \end{cases} \quad (2.6)$$

These two algorithms provides a better evaluation of the velocity allowing a better estimation of the kinetic energy. VV provides a more accurate evaluation of the velocities with respect to LF, therefore it results the most stable scheme among

the numerical integrators [31]. Figure 1 shows graphically the scheme of the VA (a), LF (b) and VV (c) integration algorithms.

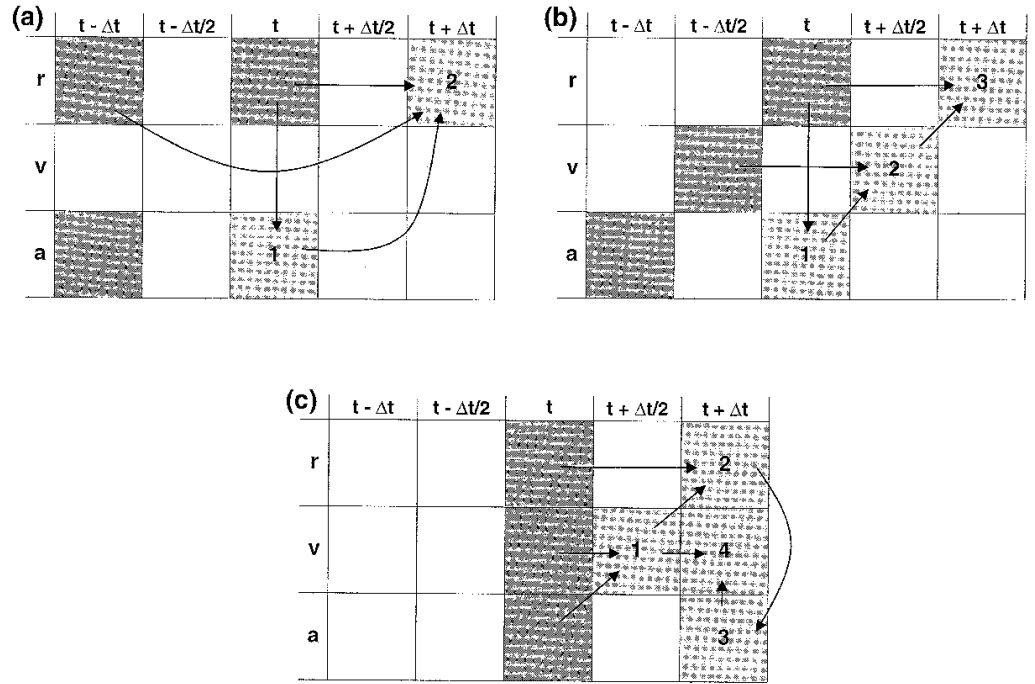


Figure 1: Graphical representation of the Verlet integration algorithm and its variants [31]. Verlet (a), leapfrog (b) and Velocity Verlet (c) integration algorithms. From the first to the third row, the spatial coordinates, the velocities and the accelerations known at each particular time step are evidenced by shaded boxes. For each method the integration path is executed according to the enumeration and arrows, which illustrate the dependencies of each value from the previous one.

In any case the integration error increases with the timestep  $\Delta t$ . The smaller the time steps, the better the integration quality, but the larger number of steps required to obtain the same time length in the trajectory. In general, the best choice is to find the largest time step that yields an accurate integration of the highest frequency modes of the system. Being  $\tau$  their period, a good rule for selecting the time step is [31]

$$\frac{\tau}{\Delta t} \approx 20. \quad (2.7)$$

For instance, in proteins the X-H stretching motion has the highest frequency. Its period is about 10fs ( $\nu = 3000\text{cm}^{-1}$ ). Thus an appropriate time step is represented by  $\Delta t \approx 0.5\text{fs}$  [31]. The highest frequency modes below these are the C=O stretching and lie around  $1500 - 1700\text{cm}^{-1}$ . Therefore, often the distance of H to its binder is constrained at the equilibrium value, so that the H modes frequencies are quenched and  $\sim 1\text{fs}$  timestep can be used, sufficient to integrate the heavy atom frequencies.

### 2.1.2 Constrained dynamics

As said, including holonomic restraints in the system is useful to quench the highest frequencies of the system and allows using larger timesteps. This is the case, e.g. when there are strong chemical bonds described by harmonic potential with very large elastic constant and/or involving light masses. In this case, use an holonomic restrains give a good representation of the behavior, since those modes have very small fluctuations around the equilibrium value. In classical mechanics the holonomic time-independent constraints can be expressed as a set of  $N_c$  linear equation (where  $N_c$  represents the number of constraints):

$$\sigma_k(\mathbf{r}_1, \dots, \mathbf{r}_N) = 0, \quad k = 1, \dots, N_c. \quad (2.8)$$

The equation of motion integrated with the constraints are:

$$m_i \ddot{\mathbf{r}}_i = \mathbf{F}_i + \sum_{k=1}^{N_c} \lambda_k \nabla_i \sigma_k(\{\mathbf{r}_l\}), \quad (2.9)$$

where the  $\lambda_k$  are the Lagrange multipliers, enforcing the constraints. To integrate the constrained equations of motion in the velocity Verlet scheme a new set of Lagrange multipliers is required at each time step. Generally this problem is solved with the SHAKE algorithm scheme [32] in which the Lagrange multipliers are obtained on the fly, correcting iteratively the position obtained with equations 2.6 until the deviation is within a user-defined tolerance, representing the largest accepted deviation of the restrained distance. At the iteration  $j$  the position  $\mathbf{r}_i^{(j)}$  is updated to

$$\mathbf{r}_i^{(j+1)} = \mathbf{r}_i^{(j)} + \frac{1}{m_i} \sum_{k=1}^{N_c} \delta \tilde{\lambda}_k^{(j)} \nabla_i \sigma_k(\{\mathbf{r}_l^{(0)}\}), \quad (2.10)$$

where  $\mathbf{r}_i^{(0)}$  are the coordinates evaluated using the unconstrained integration algorithm, and

$$\delta\tilde{\lambda}_k^{(j)} = -\frac{\sigma_k(\{\mathbf{r}_l^{(j)}\})}{\sum_{i=1}^N (\frac{1}{m_i} \nabla_i \sigma_k(\{\mathbf{r}_l^{(j)}\}) \cdot \nabla_i \sigma_k(\{\mathbf{r}_l^{(0)}\})}. \quad (2.11)$$

Once the convergence is reached, the velocities must be updated according to

$$\mathbf{v}_i(t + \Delta t/2) = \mathbf{v}_i(t) + \frac{\Delta t}{2m_i} \mathbf{F}_i(t) + \frac{\Delta t}{m_i \Delta t} \sum_{k=1}^{N_c} \tilde{\lambda}_k \nabla_i \sigma_k(\{\mathbf{r}_l^{(0)}\}). \quad (2.12)$$

The RATTLE [33] algorithm evaluates the new constrained coordinates following the SHAKE algorithm paradigm and then adjusts the velocities in order to fulfill the  $N_k$  velocity constrains:

$$\dot{\sigma}_k(\{\mathbf{r}_l\}) = 0. \quad (2.13)$$

For this procedure a new set of Lagrangian multipliers is needed. Like in the previous case, the problem is generally resolved using iterative methods.

The SHAKE algorithm is suitable for the Velocity Verlet and leapfrog algorithms whereas the RATTLE can be used only in the Velocity Verlet scheme. The constrained dynamic causes the loss of time reversibility if the full convergence is not reached [34]. The presence of constraint allows to use longer time steps but on the other hand too large time steps causes slow SHAKE or RATTLE convergence. Therefore one must in any case find a compromise.

### 2.1.3 Thermostats

In classical molecular dynamics the statistical definition of the instantaneous temperature of a system is the following [34]:

$$T(t) = \frac{1}{k_b N_{\text{DOF}}} \sum_{i=1}^n m_i v_i^2(t), \quad (2.14)$$

where  $N_{\text{DOF}}$  is the number of degree of freedom of the system and  $k_b$  is the Boltzmann constant. The integration of the Newton's equations of motion allows to sample the microcanonical ensemble of a system's states. Real systems however, usually exchange energy with the environment and are better described by the



canonical ensemble (NVT). This is obtained by coupling a thermostat to the system. One of the simplest consists in through the scaling of velocities. At each timestep

$$v'_i = v_i \gamma, \quad \gamma = \sqrt{\frac{T_t}{T(t)}}, \quad (2.15)$$

$T$  is the instantaneous temperature (eq. 2.14) and  $T_t$  is the target temperature. A refined variant of this method is the Berendsen thermostat in which the scaling factor

$$\gamma_B = \sqrt{1 + \frac{\Delta t}{\tau_T} \left( \frac{T_t}{T(t)} - 1 \right)}. \quad (2.16)$$

The scaling factor depends on the parameter  $\tau_T$  which determines how tightly the system and the thermal bath are coupled. Large values of  $\tau_T$  makes the system uncoupled from the thermal bath, whereas small values produces inappropriate fluctuation. Equation 2.16 returns the 2.15 when  $\tau_T = \Delta t$ . The Berendsen is a refined and more physical velocity scaling, but as the simplest one it does not guarantee the correct sampling of canonical ensemble. The energy fluctuations are asymptotically limited around the target value and not distributed according to the Maxwell-Boltzmann distribution of velocities expected from the kinetic theory [31].

A method to reproduce the canonical ensemble is the Nosé Hoover thermostat. In this method the coupling with a thermal bath is realized adding to the Lagrangian a new variable  $s$  (and its derivative  $\dot{s}$  with respect to time). This coupling is governed by the magnitude of the parameter  $Q > 0$  also called the thermostat "mass". The new variable  $s$  plays the role of the scaling variable, since  $d\bar{t} = s dt$ . Therefore a new set of Lagrangian variables can be represented by:

$$\bar{r} = r, \quad \dot{\bar{r}} = \bar{s} \dot{r}, \quad s = \bar{s}, \quad \dot{s} = \frac{\dot{\bar{s}}}{\bar{s}}. \quad (2.17)$$

The Lagrangian equation of the system can be rewritten as function of the new scaled variable as follow:

$$\mathcal{L} = \sum_i \frac{m_i}{2} \bar{s}^2 \dot{\bar{r}}_i^2 - U(\bar{r}) + \frac{1}{2} Q \dot{\bar{s}}^2 - g k_B T_0 \log(\bar{s}) \quad (2.18)$$

The first term represents the kinetic energy of the real system in which every velocities is scaled by the variable  $\bar{s}$ . The potential function  $U(\bar{r})$  is not modified by the change of variables.  $g$  is the number of degree of freedom, which is

increased by one with respect to the real system. The set of Lagrangian equation of motion is therefore:

$$\ddot{\bar{r}}_i = \frac{\bar{F}_i}{m_i \bar{s}^2} - \frac{2\dot{\bar{s}}\dot{\bar{r}}_i}{\bar{s}}, \quad (2.19)$$

$$\ddot{\bar{s}} = \frac{1}{Q\bar{s}} \left( \sum_i m_i \bar{s}^2 \dot{\bar{r}}_i^2 - g k_B T_0 \right). \quad (2.20)$$

These equations are known as Nosé equations of motion and describe the dynamics of the extended system  $(\bar{s}, \bar{r}, \bar{t})$ . In principle, they realize the Canonical sampling as it can be proven passing through the Hamiltonian representation and the evaluation of partition function [34]. Clearly, however, the quality of the sampling strictly depends on the simulation actual realization. Nosé and Hoover reformulated the equations 2.19 and 2.20 using the variable  $\gamma = \dot{s}/s$  and  $r$

$$\ddot{r}_i = \frac{F_i}{m_i} - \gamma r_i, \quad (2.21)$$

$$\dot{\gamma} = -\frac{k_B N_{df}}{Q} T(t) \left( \frac{g}{N_{df}} \frac{T_0}{T(t)} - 1 \right), \quad (2.22)$$

which assumes a similar form to the Berendsen thermostat with a time dependent  $\gamma$ .

Care must be taken on the choice of the parameter  $Q$ . High values leads to the absence of the heat transfer. Small values may lead to the high frequency transfer causing temperature oscillation. The energy of the real system fluctuates due to the heat transfer.

Although the Berendsen thermostat is not suitable to reproduce the canonical ensemble, it is used to relax a system to the desired temperature thanks to its quick equilibration. The Nosé Hoover approach represents the most used method in classical molecular dynamics [35]. All the the thermostats proposed in this section are suitable to the leapfrog and Velocity Verlet schemes of integration due to the explicit involvement of the velocities.

## 2.2 PROTEINS MODELS

The described equations and algorithms can be applied to any classical Hamiltonian or Lagrangian system. In practice in any case when a potential energy function  $U(r_1, \dots, r_n)$  of the internal variable is defined (either numerically or an-

alytically), and the dynamics of those variables is classic. One of these cases is the already mentioned PES  $U(r_1, \dots, r_n)$  evaluated solving the electronic problem at each nuclear configuration of the system. This approach, using quantum mechanics for electrons and classical dynamics for nuclei interacting with a  $U(r_1, \dots, r_n)$  evaluated at each time step is also called *ab initio* molecular dynamics, because, in principle, the forces acting on nuclei can be derived directly from the basics interactions (electrostatics between electrons and nuclei) only. This approach is very computational expensive. Using massively parallel systems a single protein could be currently addressed, but only at ultrafast timescales, preventing the possibility of studying most of the biological process. A way out is to evaluate the  $U(r_1, \dots, r_n)$  once for all and represent the forces through the sum of empirical terms, generally named the force field (FF). This approach can be applied to atomistic representation as far as to coarser one, and is described in the following.

### 2.2.1 Atomistic models

In the atomistic (or all-atom) empirical models the effect of electronic structure is included into the FF empirical terms. Quantum effects are therefore implicit, and the Hamiltonian is completely classic. This results in the huge simplification of the problem, since a large number of quantum degrees of freedom is eliminated. The computational cost is proportionally reduced, making it possible the simulation of the dynamics of single proteins to the  $\mu\text{s}$  timescales, and to address system as large as viruses. This gain is paid with the introduction of a large number of parameters into the FF (and in the Hamiltonian), whose value must be fitted, i.e. the empirical level of the Hamiltonian is increased, and predictive power and transferability consequently reduced. Parameters can be fixed by fitting each FF term onto its homologue evaluated with *ab initio* calculations on small molecules representing part of the whole system. Alternatively, parameters can be adjusted in order to reproduce experimentally measured observables (i.e. vibrational frequencies, or structural data) or thermo-statistic data (melting temperature/pressure/densities, heat of transition etc.). Different FF may differ in the functional form of the terms and/or in the parameters fitting strategy.

One of the most popular FF for proteins is CHARMM [36], which is here described as an exemplar case. The decomposition of the potential energy in single terms is of course not unique. Therefore the FF terms are generally chosen

the simplest possible describing interactions in a physically sound fashion. For instance, the potential decomposition in CHARMM is the following:

$$\begin{aligned}
 U(\mathbf{R}) = & \sum_{\text{bonds}} K_b (b - b_0)^2 + \sum_{\text{UB}} K_{\text{UB}} (S - S_0)^2 + \sum_{\text{angle}} K_\theta (\theta - \theta_0)^2 \\
 & + \sum_{\text{dihedrals}} K_\varphi (1 + \cos(n\varphi - \delta)) + \sum_{\text{impropers}} K_{\text{imp}} (\chi - \chi_0)^2 \quad (2.23) \\
 & + \sum_{\text{nonbond}} \left\{ \epsilon_{ij} \left[ \left( \frac{R_{\text{min}ij}}{R_{ij}} \right)^{12} - \left( \frac{R_{\text{min}ij}}{R_{ij}} \right)^6 \right] + \frac{q_i q_j}{\epsilon_1 r_{ij}} \right\},
 \end{aligned}$$

where  $K_b$ ,  $K_{\text{UB}}$ ,  $K_\theta$ ,  $K_\varphi$ , and  $K_{\text{imp}}$  are the bond, Urey-Bradley, angle, dihedral angle, and improper dihedral angle force constants, respectively;  $b$ ,  $S$ ,  $\theta$ ,  $\varphi$ , and  $\chi$  are the bond length (fig. 2.23-a), Urey-Bradley 1,3-distance (fig. 2.23-c), bond angle (fig. 2.23-b), dihedral angle (fig. 2.23-d), and improper torsion angle (fig. 2.23-e), respectively, with the subscript zero representing the equilibrium values for the individual terms. The last two terms in eq. 2.23 represent “non bonded” interactions, namely Coulomb and Van der Waals interactions (figure 2.23-f).  $q_i$  in the Coulombic interaction represents the partial charge of the atom  $i$ . Partial charges are assigned to specific atom types, which are atoms involved in specific bonds or functional-groups, in order to better reproduce the electrostatic interactions inside the system.

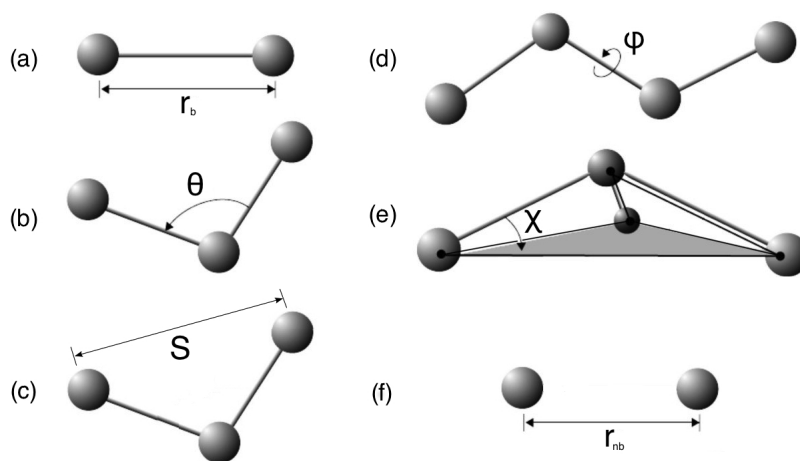


Figure 2: Variable related to the potential function in equation 2.23.

Each atom type have specific force field parameters. Every parameter in the force field must be consistent with the others, therefore, larger the number of atom types, harder the consistency of the parameters. Due to these limitations,

the force fields are generally related to specific issues. For instance, CHARMM22 is the force field oriented toward the simulation of proteins in the CHARMM FF collection [36].

As said, the optimization of parameters may be fitted on the PES calculated on small molecules, or using experimental data from different sources. Both strategies have advantages and disadvantages: using calculated PES allows a more direct and simple fitting procedure, but includes possible systematic inaccuracies due to the non transferability of parameters from small molecules to the extended systems, besides possible inaccuracies of the used calculation method. On the other hand, fitting onto experimental data is technically more difficult, because only values of observables can be compared, which include possible inaccuracies in the calculation of them from simulations. In general, FFs are parametrized through a combination of fitting strategies. For instance CHARMM22 [36] intramolecular interactions (i.e. the first five terms of eq. 2.23) are fitted on structural and vibrational data measured on model compounds. The evaluated terms are subsequently used to refine the intermolecular parameters in order to reproduce the target models. This process is iterated till the convergence is reached [36].

Although sometimes implicit solvents are considered by embedding the atomistic protein model into a dielectric medium with appropriate properties, most often atomistic simulations are performed with explicit solvent, namely including the protein into a water box. In this case water constitutes 70 – 90% of the whole system, therefore the dynamics of the solvent represents the most time consuming operation and the choice of the FF for water molecules has a major in the simulation behavior. There are many models of water differing by the geometry, number of the interaction sites per water molecule (see figure 3), number of constraints and by the potential used for the interactions [37]. Table 1 shows the parameters of the most popular methods.

In this work atomistic simulation were performed *ad hoc* to integrate experimental data for the RP. Specifically, they were used in building figure 7 in section 1.4. Those data were produced by means of a standard protocol atomistic simulation of a tetrapeptide of alanine (see figure 4). The simulation was performed with the GROMACS molecular dynamics package [41][42] using the TIP3P water model and the CHARMM22 FF for the peptide. The simulation protocol here used is reported in table 2.

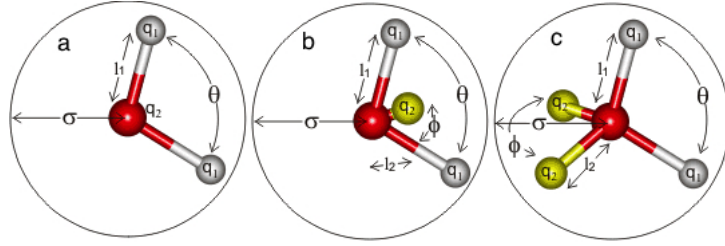


Figure 3: Graphical representation of water models differing in interaction sites per molecule. (a) three sites, (b) four sites, (c) five sites.

Model	Type	$l_1$ [Å]	$\theta$ [deg]	$\phi$ [deg]	$\sigma$ [Å]	$\epsilon/k_B$ [K]	$q_1$ [e]	$l_2$ [Å]
SPC [38]	(a)	1.0	109.47	—	3.1656	78.20	0.41	0
TIP3P [39]	(a)	0.9572	104.52	—	3.1506	76.54	0.417	0
TIP3P/fw [40]	(a)	0.9600	104.5	—	3.1506	76.58	0.417	0
TIP4P [39]	(b)	0.9572	104.52	52.26	3.1540	78.02	0.52	0.15
TIP5P [39]	(c)	0.9572	104.52	109.47	3.1200	80.51	0.241	0.70

Table 1: Main geometrical and energetic parameters of the most popular water models. The second column refers to the number of interaction sites with respect to figure 3. The geometrical variables i.e.  $l_1$ ,  $l_2$ ,  $\theta$ ,  $\phi$  are shown in figure 3. The energetic parameters  $\sigma$  and  $\epsilon$  are Lennard-Jones parameters related to the VdW interaction.

Phase	T[K]	Integrator	$\Delta t$ [fs]	Duration [ns]
Minimization	-	leapfrog	0.1	2
Production run	300	leapfrog	0.1	50

Table 2: Simulation protocol adopted for the atomistic simulation.

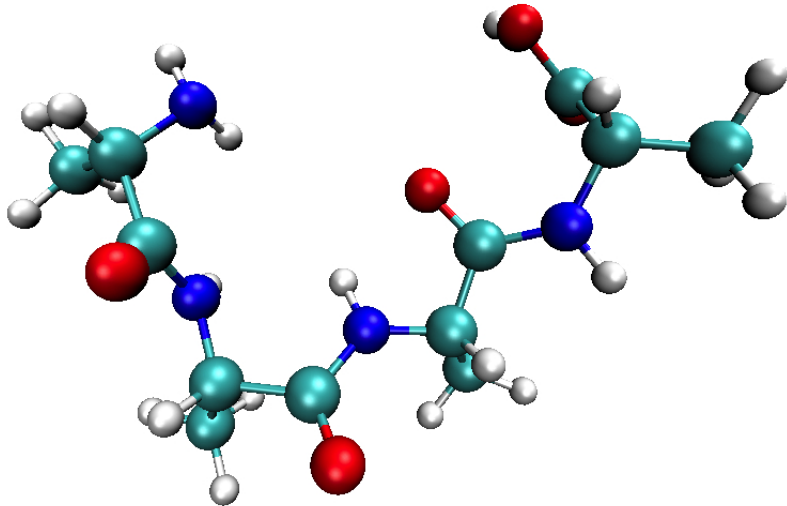


Figure 4: Tetrapeptide of ALA configuration sampled during the simulation (see text).

### 2.2.2 Coarse grained models

The Coarse Grained models (CG) represent a natural way to overcome the time and size limits of atomistic simulations. In these approaches group of atoms are represented by single interacting centers. The interaction between them must be able to reproduce as close as possible the dynamics of the atomistic representation of the system despite the loss of degrees of freedom (DOFs). This approach simplify the representation of the protein thanks to the elimination of a number of DOFs and averaging over the detailed interaction [44]. The CG method reduces the computational costs, allowing to sample the timescales and system sizes not accessible to atomistic representations.

Many strategies of coarse graining are possible. The higher the level of coarse graining, the harder building a FF able to reproduce the structure. In fact, the condensation of DOFs makes it difficult to represent complex interactions between the atoms hidden in the bead.

The CG method begins with the definition of a new set of coordinate  $Q_I$  in terms of the old ones  $q_i$

$$Q_I = Q_I(q_i \in B_I) = \sum_{r_i \in B_i} T_{Ii} q_i. \quad (2.24)$$

The second equality holds in case of a linear relation. In this case,  $T_{Ii}$  is a rectangular matrix, since the dimension of the set  $\{Q_I\}$  is lower than  $\{q_i\}$ . Therefore this transformation is not reversible, in general. The new interacting center representing group of atoms is often called "bead".

Restricting to the protein CG models, the procedure starts with the choice of the number of the beads for amino-acid and their location. The most popular CG models for proteins (reported in table 3) include between 1 and 6 beads per amino acids. Models were associated to different FFs, reported in the last column. The mesoscale level of representation is often separated from the CG, although it merely correspond to a higher level of coarse graining, where the interacting center might represent entire structural domains or proteins. Mesoscale models, though very interesting, go beyond the scope of this work and will no be treated here.

The CG FF terms can be defined either analytically or numerically. Analytical representation give the advantage of being more computationally light, while

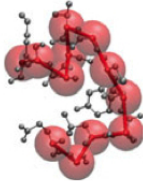

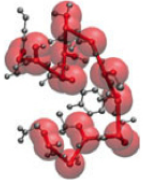
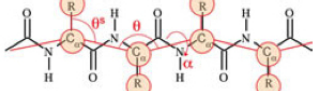
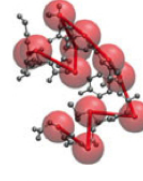
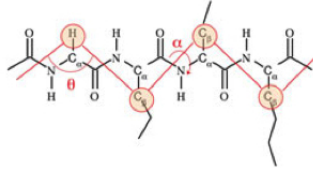
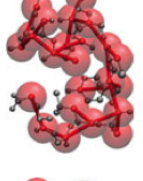
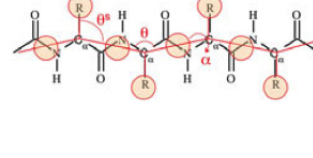
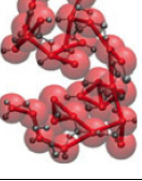
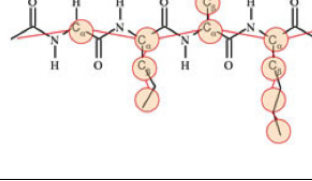
Class	Balls and sticks	Scheme	Name and reference
1 bead $C_{\alpha}$			Sorenson Head-Gordon [43]
2 beads $C_{\alpha}$ (1 bead on side chain)			
1 bead $C_{\beta}$			
2 beads (1 bead on the bb centroid, 1 bead on sc)			UNRES [45], Levitt [44]
1-6 beads $C_{\alpha}$ (0-5 beads on sc)			MARTINI [46]

Table 3: Short list of coarse-grained models with the graphical and schematic representations. The reference of most popular FF is given.



numerical representation have more flexibility in the representation of very complex behavior of the CG interaction.

In any case, the parametrization difficulties arise in the balance between the accuracy and predictive power as aforementioned. This problem, already present in atomistic empirical models, is more evident here due to the higher level of empiricism. Different strategies were adopted to overcome those difficulties. One is to build models completely biased toward a single structure, therefore renouncing to transferability. For instance, the elastic network model (EN, first row in table 4) considers the whole set of interactions as simple harmonic potentials. The spring constants are all equal whereas each equilibrium distance is chosen to reproduce the reference structure. This method was successfully applied to normal mode analysis of proteins. The EN includes a large amount of *a priori* structural information (i.e. the complete 3D structure), therefore its predictive power is low. The spring constant is the only one adjustable parameter, fitted on experimental average fluctuations. More refined models include different spring constant for each interactions and physics-chemistry based cutoff for the interactions (see table 4). The inclusion of physics based parameters into the interactions brings an improvement of the predictive power of CG FFs. The same is obtained eliminating the structural biases, although this brings loss of accuracy and the need of more complex functional form and parametrization strategies (see table 4).

Two popular parametrization strategies are the Boltzmann Inversion (BI) and the Force Matching (FM). The BI method returns statistic based FFs and potentials. Given an internal variable, from its statistical distribution it is possible to evaluate potential of mean force associated with it assuming the Boltzmann statistics. In this way the bias towards a single structure is dropped. Assuming a complete and independent set of internal variables, the total energy can be written as:

$$U(\{Q\}) = \sum_i U^i(\{Q_i\}) \quad (2.25)$$

where each FF term and the relative coordinate is labelled with the index  $i$ . The probability distribution of a variable is related to  $U(\{Q\})$  by:

$$P(Q_i) \propto \int dQ_1 \cdots dQ_{i-1} dQ_{i+1} \cdots dQ_N e^{-\frac{U(\{Q\})}{kT}} = e^{-\frac{U(Q_i)}{kT}} \quad (2.26)$$

where  $N$  is the number of DOF of the system,  $k$  is the Boltzmann constant and  $T$  is the absolute temperature. Equation 2.26 defines in general the potential of mean force. If the variables  $Q_i$  are uncorrelated and no cross-terms are present in 2.25 is exactly the potential term for  $Q_i$ . It follows that:

$$U(Q_i) \propto -k T \log(P(Q_i)). \quad (2.27)$$

The next assumption is that the distribution  $P(Q_i)$  can be evaluated from a statistical set of structures. As mentioned in section 1.1, the available dataset is represented by a collection of PDB entries. There is no physical reason indicating that this is an equilibrium distributed dataset. However, once care is taken in the elimination of redundancies, and possible *a priori* known biases (e.g. too sequentially similar structures, of too specific experimental conditions), and considering that the possibility of resolving a structure is also related to its stability, one might consider the remaining dataset a possible and experimentally accessible representation of a statistical *ensemble*. Moreover using the NMR-derived structures sampled at room temperature, it is possible to obtain a better representation of the dynamic behavior of the system. The origin of the dataset, however, must be always kept in mind when analyzing the result of the Boltzmann inversion based parametrization. Equation 2.27 is more conveniently written as

$$U(Q_i) = -k T \log\left(\frac{P(Q_i)}{P_0(Q_i)}\right) + \text{const.} \quad (2.28)$$

being  $P_0(Q_i)$  the distribution of the variable when no interaction among the “beads” (interactive centers) are present.  $P_0$  can sometimes be evaluated theoretically, independently on the input dataset.  $U$  is sometimes called also “statistical potential” (SP).

When the statistical potential depends on two or more variables, e.g.  $Q_i$  and  $Q_j$ , the distribution  $P_0(Q_i, Q_j)$  can be approximated by  $P_0(Q_i)P_0(Q_j)$  if and only if these two variables are uncorrelated. In general this condition is not satisfied. In addition in general, not even the completeness condition is satisfied, namely, the set of  $Q_i$  does not exhaust the whole description of the system interactions. This implies that the total probability distribution is not separable, and the sum of the terms obtained by applying the BI to single variables give only a first rough approximation of the potential energy. This is then used as the first step of an

iterative process (Iterative Boltzmann inversion, IBI), consisting in producing simulations with the approximate potential, and correcting it with the difference

$$\Delta U = -k T \log \left( \frac{P(Q_i)}{P^j(Q_i)} \right). \quad (2.29)$$

being  $P(Q)$  the target distribution (e.g. evaluated from the experimental dataset) and  $P^j$  the distribution evaluated from the simulation. This algorithm can be repeated until negligible difference between the resulting and the target distributions is reached.

The FM strategy targets the reproduction of the forces observed in the system at all atom resolution. These must be obtained from all-atom trajectories of the system. The quality of the parameterization depends on the quality of the atomistic Force Field used and on the extension of the phase space sampling by the atomistic simulation. The method consists in the minimization of the functional

$$\chi^2(\{\mathbf{F}\}) = \frac{1}{3N} \left\langle \sum_{i=1}^N |\mathbf{F}_I(\{Q(\{q\})\}) - \mathbf{f}_I(\{q\})| \right\rangle \quad (2.30)$$

where  $\mathbf{F}_I$  are the CG forces on CG sites  $I$ , while  $\mathbf{f}_I$  are the forces on the CG sites evaluated from the all atom simulations. The average is performed over the all atom trajectory. The minimization must be obtained tuning the parameters related to  $\mathbf{F}_I$ . This strategy is aimed to obtain a mechanical consistency between the all atom and CG resolution during the simulation.

The FM was shown to obtain very accurate results, especially with numerical potentials [47][48]. However it is showed to atomistic simulations, limited both in accuracy and in the extension of sampling of configurational space [11]. In particular it was shown that the RP obtained by atomistic simulation is very sensitive to the chosen atomistic FF [49]. Therefore in this thesis work the BI related methods were preferred for the parametrization. This choice was also due to availability of a large amount of data which ensures a larger generality and transferability of potentials.

## 2.3 THE $C_\alpha$ -BASED ONE-BEAD (MINIMALIST)

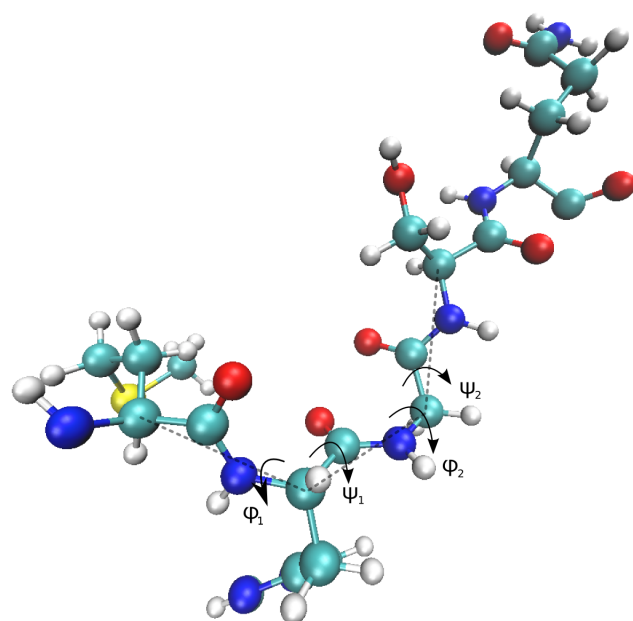
### 2.3.1 *Description of the model*

The one bead per amino acid representation is the most natural, being the basic structural unit of polypeptides. This level of coarse graining transform the polypeptide in a linear chain, simplifying the functional form of the force field. The low number of degree of freedom allows to reach large size and time scales in simulation [11] and, in addition to represent the local arrangement of the backbone, namely the secondary structure [50]. In fact, the protein secondary structure can be expressed as the sequence of the torsion angle couples. In the minimalist representation the dihedral  $\phi$  and  $\psi$  angles are undefined but the secondary structure can be expressed as the sequence of pseudo-dihedral ( $\varphi$ ) and pseudo-bond ( $\theta$ ) angles (see figure 5). It will be shown at the end of this chapter that if the bead is located on the  $\alpha$ -carbon, the representation of secondary structure in terms of the  $\theta$  and  $\varphi$  variables is unique, and the backmapping to atomistic representation is possible. For this reason the  $C_\alpha$  based one bead models (also called minimalist) are privileged among the CG ones, and are the focus of this work.

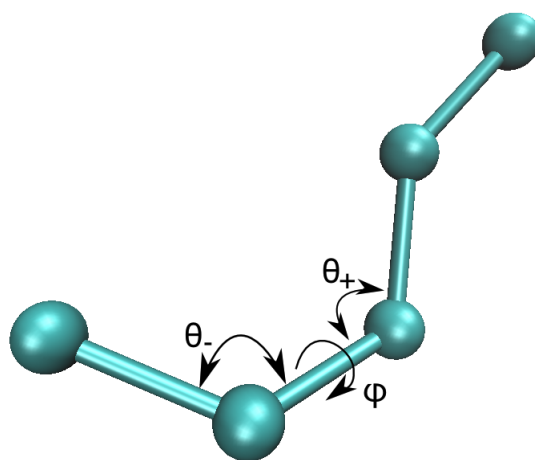
Due to the low number of interacting centers, the FF for minimalist models is described by fewer terms and fewer parameters with respect to the potential function of an atomistic system (compare with equation 2.23 in section 2.2.1):

$$U = U_{\text{bond}}(r_{i,i+1}) + U_{\text{loc}}(\theta_i, \varphi_i) + U_{\text{nb}}(r_{i,j}), \quad (2.31)$$

the first two terms representing the local interactions:  $U_{\text{bond}}$  is the peptide pseudo-bond interaction between two consecutive  $C_\alpha$ , whereas  $U_{\text{loc}}$  is the potential term of the remaining local interactions, responsible for the protein backbone local arrangement. The first term is often described as a stiff harmonic potential with mean value equal to  $3.8\text{\AA}$ , which represents the peptide bond in trans conformation. In order to include the transition in the cis conformation the bond term could be extended [51] adding another deep minimum at  $2.97\text{\AA}$ . However, as previously shown, it is sometimes convenient to substitute stiff interaction with holonomic restrains, which is the strategy adopted in this work. This choice improve the computational performance of the model without appreciable loss of accuracy. The second term of equation 2.31 describes the local backbone arrange-



(a)



(b)

Figure 5: All atom representation of the pentapeptide (a) and its minimalist version (b).

ment while the last term describe the non local interaction, generally between beads separated by at least three beads along the chain. The last term is generally defined as the interaction of two non consecutive beads. This term should include the effects of the side chain interactions, hydrophobic effects, electrostatic

interactions and hydrogen bonds. The latter are particularly difficult to describe within the minimalist models, due to the absence of donor and acceptor explicit representation [52]. Since this work focuses on systems in which hydrogen bonds are virtually absent, the problem of their representation will not be addressed here.

In general the extreme simplification of equation 2.31 might pose some problem in the representation of interactions. However, in spite of the simplification this representation allows to recognize the secondary structure of the simulated system [11][53], which is an important feature in the field of the protein folding and IDP simulations where identification of secondary structure are of the utmost interest.

It is to be observed that equation 2.31 is general enough to describe all the available minimalist models, included the already described EN, those with partial bias such as the Go-like models used to describe the folding kinetics. Table 4 reports a summary of the most popular ones.

It is also to be remarked that also the models generally called “unbiased” generally preserves some form of bias. This manifest in secondary structure dependent parametrization of the  $U_{loc}$  [53][43]. In addition all the aforementioned models do not deliver a very accurate representation of the local arrangement for unstructured proteins. This problem is often bypassed by using functional forms mixing  $\alpha$  and  $\beta$  basins.

The local interactions are often separated in two independent potentials terms [54]

$$U_{loc}(\theta_i, \varphi_i) = U_{ang}(\theta) + U_{dih}(\varphi), \quad (2.32)$$

where  $U_{ang}$  is the bond angle potential term, which is defined as a function of the angle between the three consecutive, and  $C_{\alpha}$ s, and  $U_{dih}$  is the dihedral angle potential term, which is related to the torsion between four consecutive  $C_{\alpha}$ s. The intrinsic amino-acid conformational tendencies are analyzed using the previously defined amino acids classification (figure 8 in section 1.4)

In this work the target is the reproduction of the backbone geometry of the unstructured proteins. The lower statistics of structural data for this class of molecules, makes this task more difficult than for the structured ones. In addition, as said and as it will be clearer in the following, single variable internal distributions are not appropriate targets in this case, because correlations are particularly important. However, the correct reproduction of the intrinsic conformational

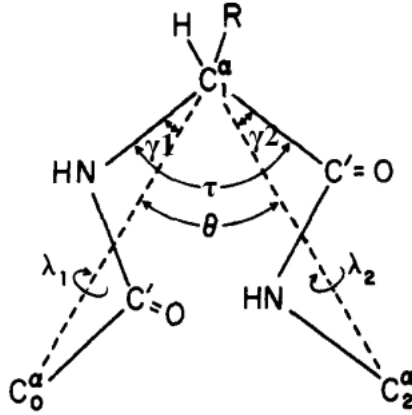


Figure 6: Graphical representation of the  $\gamma_1, \gamma_2$  and  $\tau$  angles and the  $\lambda_1, \lambda_2$  auxiliary variables in a tripeptide; taken from [55].

tendency of the backbone in absence of any specific interactions such as hydrogen bonds, is the main property a minimalist model should have. Therefore this is the main task addressed in this work.

### 2.3.2 Atomistic to minimalist variables transformation

Due to the rigidity of the peptide bond, locating the bead on the  $C_\alpha$  offers a unique advantage of minimalist models with respect to the others, namely the possibility of building a  $(\phi, \psi) \rightarrow (\theta, \varphi)$  mapping between the atomistic and minimalist internal variables, which is unique, i.e. independent on the secondary structure.

As aforementioned the coarse graining process reduces the number of DOF. In fact, as demonstrated by figure 6, a tetrapeptide backbone in atomistic conformation is entirely determined by the quadruplet  $\phi_1, \psi_1, \phi_2, \psi_2$  while the same tetrapeptide is described in terms of three variables  $\theta_-, \varphi$  and  $\theta_+$ . Therefore the  $(\phi, \psi) \rightarrow (\theta, \varphi)$  is a four to three variable mapping. Considering

$$a = \cos(\gamma_2)\cos(\tau) + \sin(\gamma_2)\sin(\tau)\cos(\psi); \quad (2.33)$$

$$b = \cos(\gamma_2)\sin(\tau) + \sin(\gamma_2)\cos(\tau)\cos(\psi); \quad (2.34)$$

$$c = \cos(\gamma_1)\cos(\tau) + \sin(\gamma_1)\sin(\tau)\cos(\phi); \quad (2.35)$$

$$d = \cos(\gamma_1)\sin(\tau) + \sin(\gamma_1)\cos(\tau)\cos(\phi); \quad (2.36)$$

this transformation can be analytically described by [55]

$$\cos(\theta) = a \cos(\gamma_1) + b \sin(\gamma_1)\cos(\phi) - \sin(\gamma_1)\sin(\gamma_2)\sin(\phi)\sin(\psi) \quad (2.37)$$

$$\varphi = (\lambda_2)_{1st} + (\lambda_1)_{2nd} + 180deg, \quad (2.38)$$

where  $\gamma_1, \gamma_2$  and  $\tau$  are the angles defined in the first chapter. The  $\lambda_1$  and  $\lambda_2$  are the angles for rotation of the peptide planes about the virtual bond  $C_0^\alpha \dots C_1^\alpha$  and  $C_1^\alpha \dots C_2^\alpha$  respectively, with respect to the plane containing the three consecutive  $C^\alpha$ s (see figure 6). In order to evaluate  $\varphi$  the  $\lambda_2$  is related to the first residue whereas the  $\lambda_1$  is related to the second peptide (this approximation is valid considering the peptide plane flat,  $\omega = 0deg$ ). These variables are related to the variables  $\gamma_1, \gamma_2, \tau, \phi$  and  $\psi$  with the following relations

$$\tan(\lambda_1) = \frac{-b \sin(\phi) - \sin(\gamma_2)\cos(\phi)\sin(\psi)}{a \sin(\gamma_1) - b \cos(\gamma_1)\cos(\phi) + \cos(\gamma_1)\sin(\gamma_2)\sin(\phi)\sin(\psi)}; \quad (2.39)$$

$$\tan(\lambda_2) = \frac{-d \sin(\psi) - \sin(\gamma_1)\cos(\psi)\sin(\phi)}{c \sin(\gamma_2) - d \cos(\gamma_2)\cos(\psi) + \cos(\gamma_2)\sin(\gamma_1)\sin(\phi)\sin(\psi)}. \quad (2.40)$$

These equations can be simplified in the case of the ordered secondary structures, for which  $\phi_1 = \phi_2$  and  $\psi_1 = \psi_2$ . Under this hypothesis the mapping transforms in the  $\phi - \psi$  to the  $\theta - \varphi$  two-two mapping. Figure 7 shows how the basins of the RP (considering  $\beta$  and PPII a unique one) are mapped onto their corresponding basins in the  $\theta - \varphi$  plane [56]. The butterfly shape, drawn with the gray line on the right-hand side plot, sketches outline the limits of the image of the transformation applied on the whole  $(\phi, \psi)$  space. Therefore, in the range of validity of the approximation of ordered secondary structures, these can still be separated in the minimalist representation. This is a fundamental point, being the practical demonstration that the minimalist representation is a proper one to describe one of the basics structural levels of proteins, namely the secondary structures. It is to be remarked that this possibility descends from having chosen the  $C^\alpha$  as the site for the location of the bead.

Figure 9a reports the full set of correlation plot  $\theta_+, \varphi$  and  $\theta_-, \varphi$  for the main structural basins, obtained from the experimental RP relaxing the assumption of uniform and regular secondary structure, and using the full four to three



mapping expressed in equations 2.37 and 2.38. This is the first original result of this thesis work, and can be considered a refinement of figure 7 [56]. The mapping is the general three to one, spanned by the variables  $\phi$ ,  $\theta_-$ ,  $\theta_+$ , being the two latter bond angles that preceding and following the dihedral along the chain (see figure 5b). Furthermore, two different  $\theta - \phi$  correlation plots must be considered, namely  $(\phi, \theta_-)$  and  $(\phi, \theta_+)$  which are in principle different.

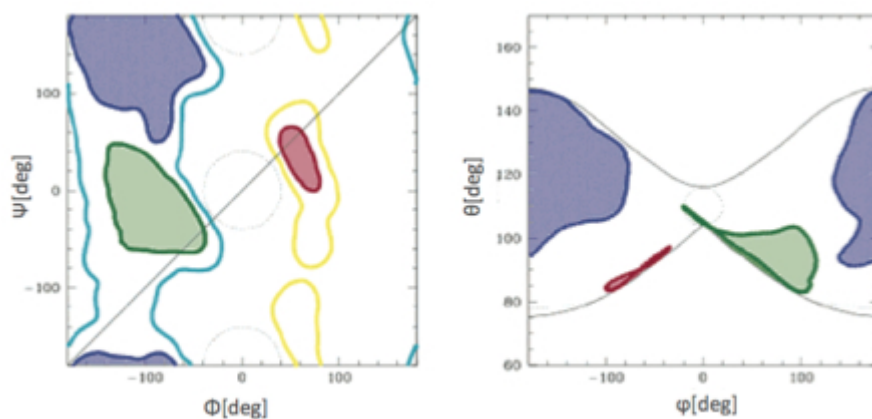


Figure 7: Comparison between the structured regions of the Ramachandran plot (left) and  $\theta, \phi$  correlation plot. In green are reported the right-handed helices, in blue the stranded/extended structures and in red the left-handed helices [56].

These data are sampled from regions, which delimit every basin, by approximating the densities of these region to a constant value. In this way is possible to observe how the transformation modify the density of each basin in the new system. This analysis have been performed following the first algorithm described in appendix G, which allows to reproduce the effects of the  $(\phi, \psi) \rightarrow (\theta, \phi)$  transformations. The regions representing each basin are taken from the RP (figure 8).

Differences between “right” and “left” correlation plots are expected if the directional symmetry along the chain is broken. This is actually the case, due to the slight difference between  $\gamma_1$  and  $\gamma_2$ , which assumes the numerical values reported in table 3 in section 1.3. However, as it can be seen in the regular secondary structures these differences are very small, negligible in the helical basins (plots 9a and 9b), little more evident in the others (plots 9c and 9d). As expected, and as it will be shown in chapter 3, conversely, these difference becomes substantial in the unstructured proteins.

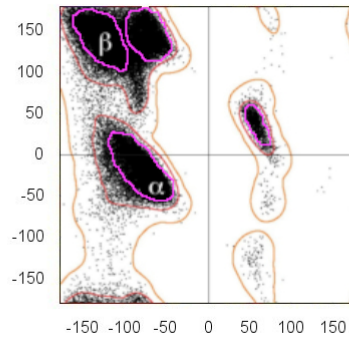
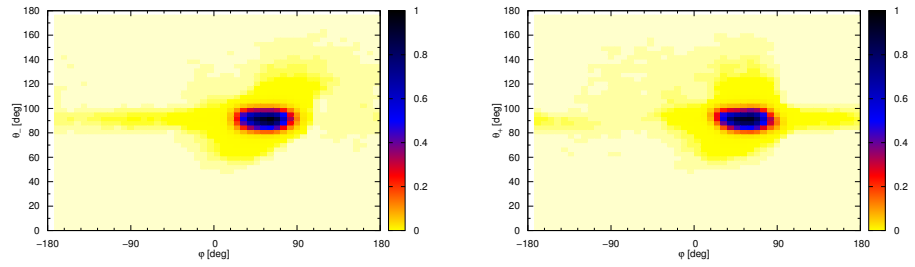


Figure 8: The approximative limits of each basin is superimposed on the RP taken in figure 8 in section 1.4.

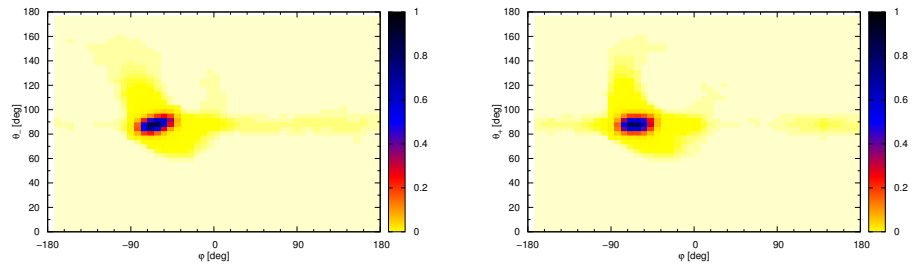
Finally, it is to be observed that right and left  $\theta - \varphi$  plot are projections of  $\theta_- - \varphi, \theta_+ - \varphi$  3D plot, which will be described in detail in chapter 3.

Model	$U_{\text{bond}}$	$U_{\text{loc}}$	$U_{\text{nb,loc}}$	$U_{\text{nb,non-loc}}$	Remarks
Elastic network	Harmonic $\frac{1}{2} k(r_{ij} - r_0)^2$				GNM $r_{\text{cut}} = 6-10$ , $k \sim 1-0.2$ kcal/mol $\text{\AA}^2$ ANM $r_{\text{cut}} = 8-15$ , $k \sim 10-0.9$ kcal/mol $\text{\AA}^2$
Plastic/bimodal networks	Harmonic potential for the single wells Global or local valence-bond like combination				GNM $r_{\text{cut}} = 8$ , $k \sim 0.02$ kcal/mol $\text{\AA}^2$ $r_{\text{cut}} = 13$ , $k = 1$ kcal/mol $\text{\AA}^2$
Heterogeneous EN	Harmonic				In principle, infinite, but $r_{\text{cut}} \sim 15$ for simplicity $k = \text{different for each bond couple}$
Extended/anisotropic network	Harmonic $\frac{1}{2} K(r_{ij} - r_0)^2$	Anharmonic $\frac{1}{2} k_2((r_{ij} - r_0)^2 - a^2) \Theta((r_{ij} - r_0)^2 - a^2)$			$r_{\text{cut}} = 13$ , $K \sim 46$ kcal/mol $\text{\AA}^2$ $k_2 = \text{AA dependent avg } \sim 2$ kcal/mol $\text{\AA}^2$
Chemical EN	Harmonic	Harmonic (1-3 and 1-4 distance-based terms)	Harmonic $\frac{1}{2} k_{\text{val}}(r_{ij} - r_0)^2$		$r_{\text{cut}} = 8$ for $U_{\text{nb,loc}}$
Go models	Harmonic or constraint	Harmonic angle $\frac{1}{2} k_2(r_{ij} - r_0)^2$	LJ 12-6 $\epsilon \left[ \left( \frac{r}{r_0} \right)^{12} - \frac{6}{5} \left( \frac{r}{r_0} \right)^{10} \right]$	Repulsive only	Separated terms for H-bonds, disulfide bridges and salt bridges, with different elastic constants $r_{\text{cut}} = 8$ for $C\alpha$
Partially biased models	Harmonic or constraint	Cosine sum $\sum_{n=1,3} K_n [1 - \cos n(\alpha - \alpha_0)]$			$r_{\text{cut}} = 4$ for side chains $\epsilon = \text{energy unit}$ $k_b = 100\epsilon$ $k_\theta = 20\epsilon$ $K_n \sim \epsilon$
		Harmonic (1-3 and 1-4 distance-based terms) OR Unbiased Harmonic angle $U_{\text{ang}} = \frac{1}{2} k_\theta(\theta - \theta_0)^2$ or $U_{\text{ang}} = \sum_{n=1}^3 k_n \frac{1}{n} (\theta - \theta_0)^n$	Biased Morse $\mu(r_{ij}) = \epsilon[(e^{-\kappa(r-r_0)} - 1)^2 - 1]$	Unbiased Morse $\mu(r_{ij}) = \epsilon[(e^{-\kappa(r-r_0)} - 1)^2 - 1]$	$r_{\text{cut}} \sim 8$ $k_b = 50-100$ kcal/mol $\text{\AA}^2$ $k_\theta \sim 20-50$ kcal/mol $k_\alpha \sim 3$ kcal/mol $\epsilon = \epsilon(\eta) = \text{decreasing from } \sim 5 \text{ to } \sim 0.1$ kcal/mol Parameterization half structure based, half BI based
Unbiased models	Harmonic or constraint	$U_{\text{dih}} = \sum_{n=1,3} K_n [1 - \cos n(\alpha - \alpha_0)]$	Explicit $U_{\text{hb}}$ , anisotropic LJ-like OR dipole-dependent term	LJ-like, single or multiple wells, sometimes anisotropic	Parameterization based on a mix of BI, FM and physical-chemical considerations based

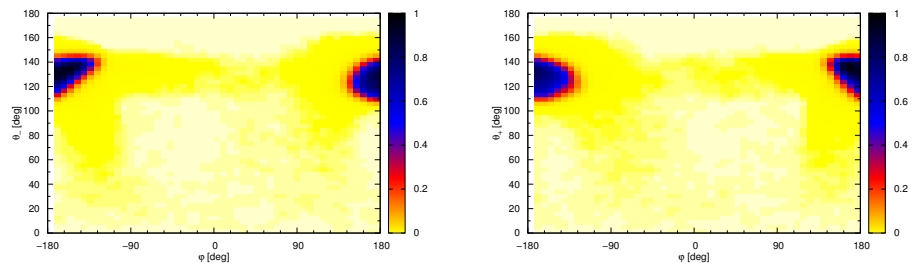
Table 4: A summary of the most popular minimalist force fields. The force fields follow a top-down order from low to the high predictive power [11].



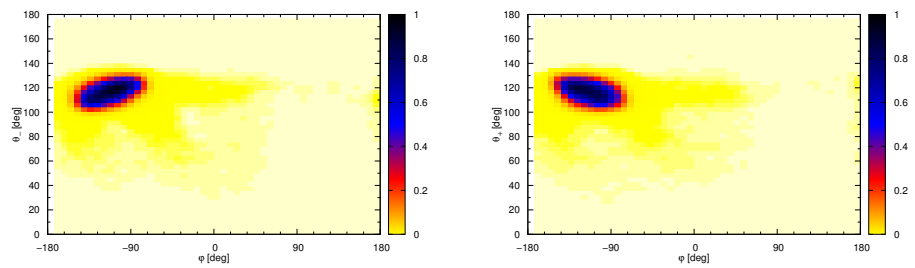
(a)



(b)



(c)



(d)

52  
 Figure 9: Normalized density plot for the main structural basins, mapped from the experimental RP onto the  $\theta$ ,  $\varphi$  planes. They represent the  $\alpha$  (a), the  $\alpha$ -L (b), the  $\beta$  (c) and the pPII (d) regions respectively. See text for the procedure. The color scale bar is on the right of each graph.

## STRUCTURAL DATASET PREPARATION

---

In the previous chapter the generalities of the minimalist models were described. The main focus of this thesis work is to develop a model for unstructured proteins. The functional form chosen for the FF is equation 2.31, namely local interactions, depending on the conformational variables  $\theta$  and  $\varphi$ , are separated from non local interactions, involving all couples separated by more than three beads along the chain. At variance with other models, in this work the potential  $U_{loc}$  is not separated in two terms depending on  $\theta$  and  $\varphi$ . The reason for this choice was outlined in the previous chapter, and will be clearer in this one: it turns out that, specifically for unstructured proteins, the correlations between  $\theta$  and  $\varphi$  variables are not negligible, and a separated potential form is not adequate.

As said, the parametrization strategy will be based on the Boltzmann inversion of structural data. Specifically, the  $\theta, \varphi$  experimental maps will be inverted to obtain a backbone potential including correlations. Therefore, the dataset building assumes a fundamental role in determining the accuracy of the potential. In chapter 1 the difficulties underlying this task were outlined: data for unstructured proteins are more difficult to retrieve, and often biased by the presence of residual ordered secondary structure fragments. This chapter is then entirely devoted to a pure unstructured (coil) dataset building, which will be used in the next chapter for the potential parametrization, and to its analysis. Elements of novelty are introduced in the methodology for these two tasks, specifically in the analysis which put in evidence correlations between internal variables. The coil dataset analysis is also compared with that the largest IDP dataset, having lower level of disorder but larger statistics and biophysical relevance.

### 3.1 COIL DATASET GENERATION

As described in section 1.7, in the literature the “coil libraries” are usually built using the X-ray data assuming that the crystallographic restraints bring negligible structural bias. The NMR data give a larger statistics for unstructured fragments and will be additionally considered in this work. Results from the X-ray and NMR dataset will be continuously compared along this work. It must be kept in mind, however, that also NMR data might include a bias, due to the choice of the specific FF used in simulation to generate structures from NMR restraints.

In any case, care is taken to reduce other sources of bias. For instance, oversampling due to redundancy is reduced filtering out sequence with homology larger than 30% as suggested in the literature [28]. This is particularly important for the X-ray dataset, being those structures near to their equilibrium state. Conversely, the filter is not applied to NMR data, because they provide more out of equilibrium conformational sampling even to the same sequence. These procedure finally lead to two dataset named PDB\_XRAY and PDB\_NMR respectively.

The coil datasets are created by purging all residual secondary structures from the dataset. Therefore, the secondary structure recognition and assignment algorithm represents a crucial choice for the determination of this dataset. STRIDE [17] and DSSP [16] are the most popular secondary structure recognition algorithms using different recognition strategies. In DSSP the recognition is achieved through the identification of the hydrogen bonding pattern whereas STRIDE takes also in consideration the geometry of the system. Both these algorithms do not recognize the PPII secondary structure. In both algorithms, the presence of a H-bond is recognized by empirically evaluating its energy from the geometry of donor-acceptor groups and the strength of the interaction. In the two cases slightly different structural and energetic criteria are used, which, however, does not bring appreciable differences in the datasets. A more relevant difference regards the recognition of a specific secondary structure in DSSP (absent in STRIDE), the “bent”, based on purely structural criteria. This structure is not stabilized by any hydrogen bonds, and therefore there would be no specific reason to eliminate it from the coil dataset. In this work besides the dataset building using basic DSSP, another one is considered DSSP\_S in which the bent are included among the coils.

For each experimental dataset, the disordered structures are evaluated by using both the assignment algorithms STRIDE and DSSP and considering only those

fragments with length greater than six residues. Smaller lengths are discarded in order to reduce possible tail effects. The relative short length of the residues, however, also brings an additional advantage: the non local interactions are statistically minority, therefore the obtained dataset is somewhat ideal to optimize the local interactions.

Figure 1 shows the statistical analysis of the PDB\_NMR (first row) and PDB\_XRAY (second row) made with the aforementioned secondary structure assignment algorithms. In order to cross check how the different selection algorithms perform, cascade selections were performed: the first column shows the secondary structure contents of the coiled fragments revealed by the DSSP algorithm re-analyzed using the STRIDE algorithm. In the second column, the two algorithms are interchanged. The last column shows the secondary structure contents of the coiled fragments revealed by the DSSP\_S algorithm using the STRIDE algorithm. In each graph is reported the total amount of the residues included in fragments labeled as coiled structure.

The largest difference between PDB\_NMR and PDB\_XRAY is revealed by the DSSP\_S assignment. The DSSP algorithm recognizes the lowest number of coiled fragments in these datasets, followed by the STRIDE algorithm as reported by the amount of residues contained in each set. The ratios of the coil dataset sizes, considering as a reference the size of the DSSP dataset, specific for each methods are approximately 1:3:10 for the PDB\_NMR and 1:2:4.7 for the PDB\_XRAY. There is an overall agreement between STRIDE and DSSP, but the STRIDE algorithm provides assignments closer to the secondary structure evaluation published by the authors of the protein structure in the PDB [17], which simply indicates that this is the preferred methods by the experimentalist to assign secondary structure to their experimental determinations.

Panels c and f show the DSSP\_S datasets contain a large bias toward the turn structures when re-analyzed by the STRIDE algorithm results. For the purposes of this work, the most interesting results are achieved by the evaluation of the dataset by using the STRIDE methods (panels b and e). In fact, the DSSP algorithm defines the turn structure as the basic element with a tight hydrogen bond (i.e. with interaction energy higher than 0.5kcal/mol). The STRIDE algorithm adopt the definition given in [57] which identifies a turn structure on the basis of the dihedral angles ( $\phi, \psi$ ) of two consecutive residues. Due to the equation 2.38, this definition is strictly related to a specific regions of the ( $\theta_-, \phi, \theta_+$ ) space. Moreover as stated, the identification of the bent structures is not supported by a specific

local interaction and high flexibility of the overall structure is expected in the coil dataset, therefore the exclusion of these structures based on a geometrical selection criterion is not desired. Therefore considering DSSP-bend assignment negligible by its definition, the low contents of turn structures (from the DSSP analysis) in the STRIDE coiled structures (2% and 4% in NMR and XRAY dataset respectively) reveals that the STRIDE assignment represents the best assignments for the scope of this work. Hereinafter each coil database evaluated in this section is referred as (PDB\_)EXP\_MET, where EXP is either NMR or XRAY and MET can be either DSSP(\_S) or STRIDE.

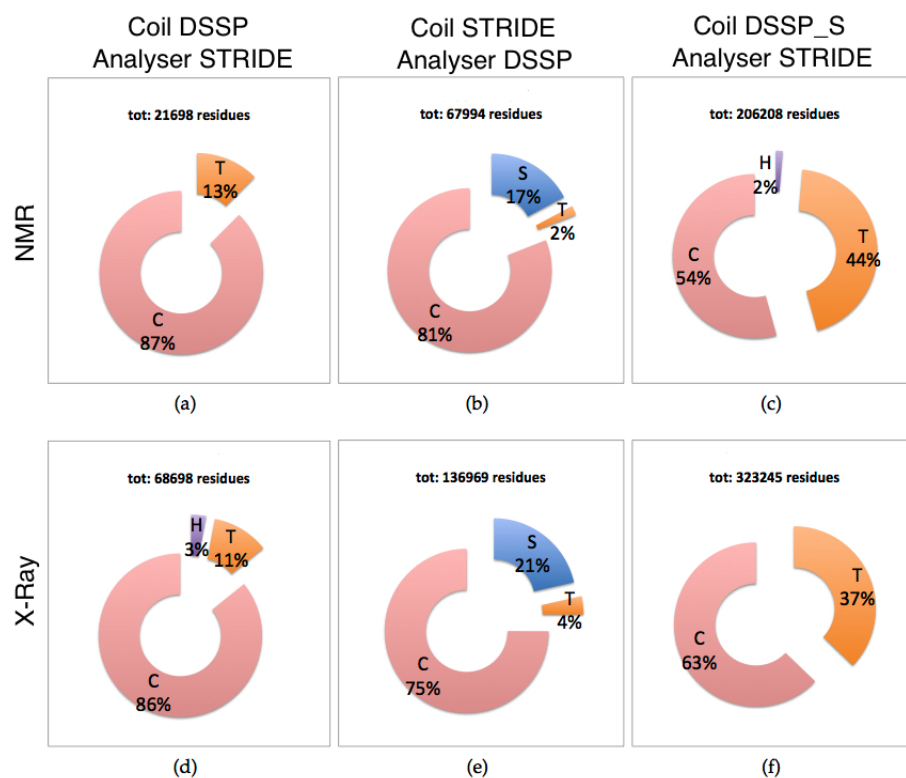


Figure 1: Pie charts on the secondary structure contents of the coiled structures assigned with the DSSP (first column), STRIDE (second column) and DSSP\_S (third column, see text for the definition). The first row contains the data of the PDB\_NMR dataset whereas the second row is related to the PDB\_XRAY dataset. The secondary structure contents reported in each column are the result of the analysis performed with the alternative algorithm on the coiled structure. The labels C, T, H and S are related to the coil, turn,  $\alpha$ -helix and bend secondary structure respectively.



Figure 2 reports the statistics all datasets. As it can be seen, X-ray data include on average shorter sequences, due to the increasing difficulty of the resolution of the diffraction pattern in longer disordered structures. The information contained in this plot is the relative contribution of long fragments with respect to short ones in the dataset, which will be used in the following.

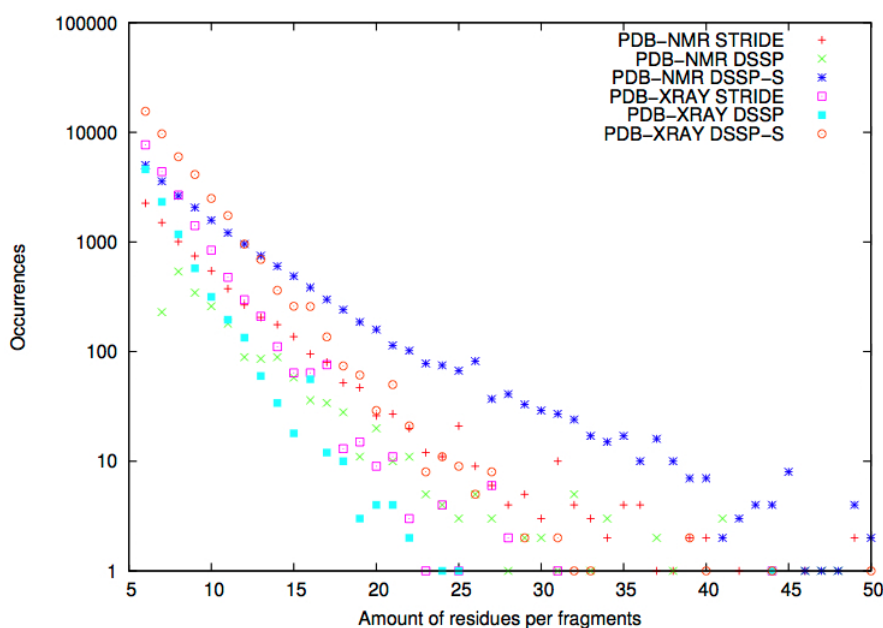


Figure 2: Statistics related to the length of the coiled fragments sampled in the PDB\_XRAY and PDB\_NMR datasets identified by the algorithms DSSP, DSSP\_S and STRIDE.

Table 1 shows the amino acid contents of the coil library evaluated from the experimental datasets by using the STRIDE algorithm and the amino acid content of IDPs [58]. There are some differences between the amino acid contents of IDPs and the STRIDE dataset, but a detailed analysis of these deviations is beyond the scope of this work, also because given the broad classification in residue type, this difference is not expected to affect the result of this work. In any case, the large deviation of the serine content in the NMR-STRIDE dataset with respect to the X-ray-STRIDE was expected due to the disorder promoting action of this AA (see the ranking of the structure destabilizer amino acids in section 1.6) that makes difficult the structure determination using X-ray crystallography. Since the glycine is the amino acid with the highest conformational freedom, it is highly

involved in the formation of the turn secondary structure, which explains its relatively high occurrence in these datasets.

		PDB_XRAY STRIDE [%]	PDB_NRM STRIDE [%]	IDP [58]
<b>ALA</b>	<b>(A)</b>	5.99	5.40	7.15
<b>ARG</b>	<b>(R)</b>	4.98	4.44	4.21
<b>ASN</b>	<b>(N)</b>	4.37	3.43	2.06
<b>ASP</b>	<b>(D)</b>	5.90	4.41	5.05
<b>CYS</b>	<b>(C)</b>	0.99	1.49	0.61
<b>GLU</b>	<b>(E)</b>	6.08	5.82	14.26
<b>GLN</b>	<b>(Q)</b>	3.54	3.20	4.46
<b>GLY</b>	<b>(G)</b>	10.15	13.50	4.31
<b>HIS</b>	<b>(H)</b>	2.67	3.96	1.51
<b>ILE</b>	<b>(I)</b>	4.12	2.90	3.67
<b>LEU</b>	<b>(L)</b>	6.46	4.95	5.44
<b>LYS</b>	<b>(K)</b>	5.70	6.21	10.43
<b>MET</b>	<b>(M)</b>	1.29	2.28	1.30
<b>PHE</b>	<b>(F)</b>	3.21	2.14	1.66
<b>PRO</b>	<b>(P)</b>	11.01	9.05	12.07
<b>SER</b>	<b>(S)</b>	6.99	14.99	6.91
<b>THR</b>	<b>(T)</b>	6.68	5.14	5.14
<b>TRP</b>	<b>(W)</b>	1.16	0.77	0.32
<b>TYR</b>	<b>(Y)</b>	3.10	1.84	1.42
<b>VAL</b>	<b>(V)</b>	5.59	4.09	8.02

Table 1: This table show the amino acid relative amount of the fragments identified as coil by the STRIDE algorithm in the PDB\_XRAY and PDB\_NMR datasets. The last column shows the percentages related to IDPs [58].

### 3.2 INTERNAL VARIABLES STATISTICAL DISTRIBUTIONS

The PDB\_NMR dataset is affected by biases on the  $\phi$  torsion angle of the proline residue (see panels b in figure E.1 in appendix E, for values of  $\phi \approx 75$ deg). This bias should be eliminated. It was evaluated considering the standard deviation (STD) of the  $\phi$  torsion angles of every proline residues present in the PDB file. Since the bias is revealed by an anomalously small STD, it was eliminated by setting a threshold on the STD for acceptance of a structure. Overall, the dataset statistics was decreased of the 17% (see appendix E for details). The filtered

dataset is herein after referred to as “PDB\_NMR” because there is no need to consider the biased structures in the next statistics.

As reported in section 1.4, the RP is amongst the most useful tools to understand the content of a structural dataset. Figure 3 shows the densities of the RP, PDB\_NMR\_STRIDE dataset. The amino acid are further separated in the four classes, already introduced in section 1.4, namely proline (P, top right), glycine (G, bottom left), pre-proline (pP, bottom right) and the remaining (X, top left), showing different structural characters.

The global shape of these plots was described in section 1.4 where the corresponding plots derived from a X-ray coil library [10] were reported. By comparing these plots with those of section 1.4, and with accurate X-ray based RP reported in appendix E, figure E.3, it can be seen that these have generally larger allowed areas. With exclusion of the regions depleted by the  $O_{n-1}$ - $H_{n+1}$  and  $O_{n-1}$ -C steric interactions (see the “derivation diagram” in figure 6b section 1.4), all the regions are explored. This may have several causes. X-ray structures are more restrained, both because of the constraints and for the low temperature of the experimental setup. Therefore the configurational accessible space is more limited than in NMR resolved data. In addition, the NMR-pdb file reports many models, generally between 20 and 40 different models with different conformation of the same structure. This can provide a wider information on the structural properties of the system.

At variance with the X-ray, NMR dataset shows higher occurrences in regions surrounding the depletion regions related to the steric repulsion of the atoms. These regions have been discussed in section 1.4 and graphically sketched in figure 6b. While this broader occupation of the RP in NMR data is expected due to the less restrained environmental conditions with respect to X-ray experimental setup, details of RP analysis are beyond the scope of this work.

### 3.2.1 *Single variable statistical distributions*

In this section the distribution of  $\theta$  and  $\varphi$  variables of the minimalist model are reported. These will be used in chapter 4 for the parametrization. However, more in general, they contain the same information included in the RP, of which can be considered the minimalist counterpart, and give a large amount of structural information. For instance they allow identifying secondary structures, having

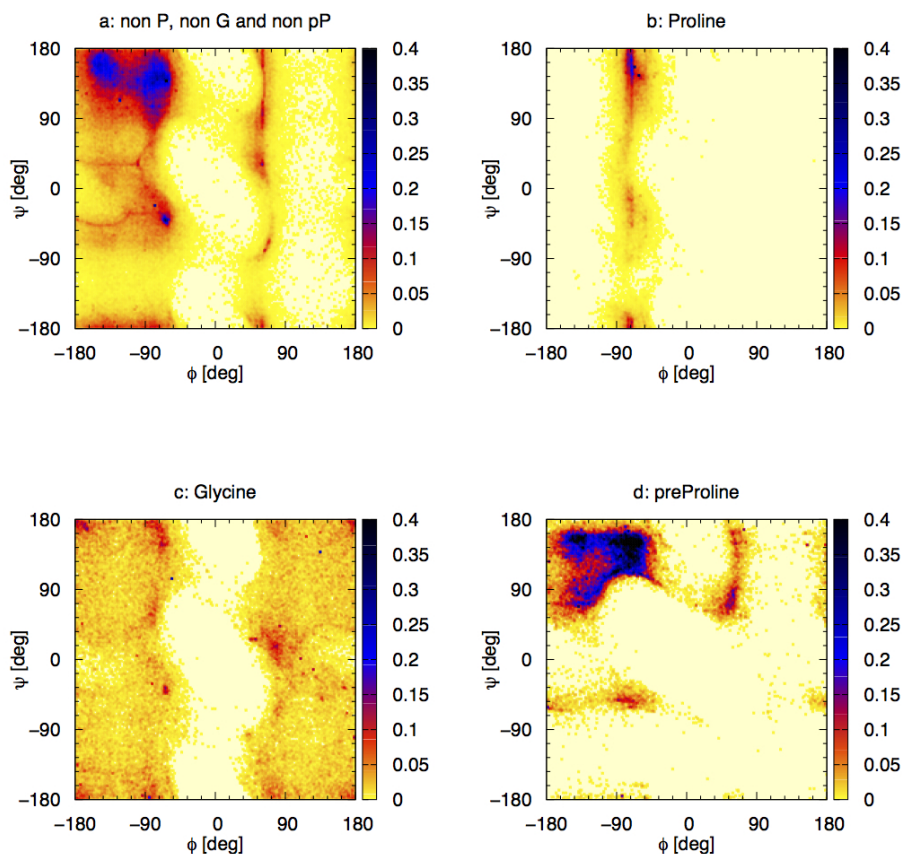


Figure 3: Ramachandran plots of the proline (b), glycine (c), pre-proline (d), and the non proline, non glycine and non pre-proline amino acids (a) of the coiled structures selected using the STRIDE algorithm from the PDB\_NMR dataset avoiding the proline-biased structures (see the text). The color range in each panel have been rearranged up to 40% of the maximum in order to enhance the details.

the bond angle distribution  $P(\theta)$  peaked at  $\sim 90\text{deg}$  (helices) or at  $\sim 120 - 150\text{deg}$  (strands) and the dihedral distribution  $P(\varphi)$  peaked at  $\sim \pm 60\text{deg}$  (helices, right and left-handed) or  $\sim \pm 180\text{deg}$  (strands) [11]. Figure 4 shows the distributions related to the  $\theta$  (left panel) and  $\varphi$  (right panel) variables of the PDB\_XRAY\_STRIDE and PDB\_NMR\_STRIDE datasets. In  $P(\theta)$  a peak at  $\theta \sim 90\text{deg}$  is visible including residual helical-like structures. Analogously residual peaks corresponding to  $\beta$ , PPII ( $\theta$  in  $[120 : 140]\text{deg}$  and  $[110 : 130]\text{deg}$  respectively) are distinguishable although the mixing is significant (see figure 9c and 9d in section 2.3.2).

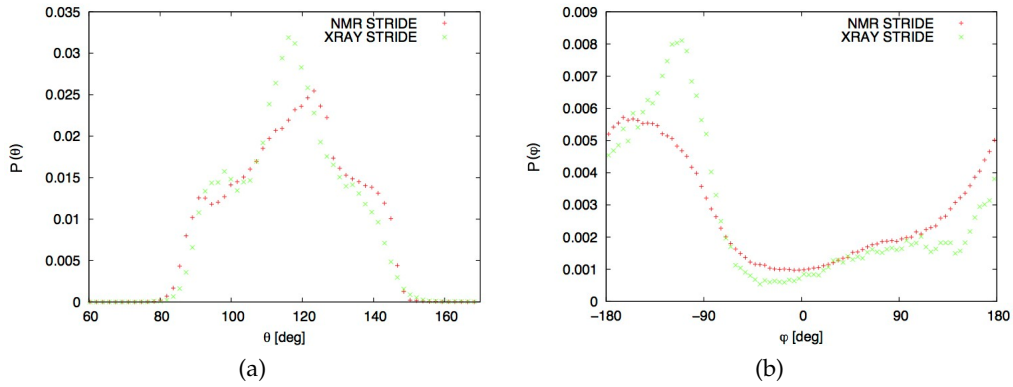


Figure 4: One dimensional distributions of the  $\theta$  angle (a) and  $\phi$  torsion (b) of the coiled structures, contained in the PDB\_NMR (red marks) and PDB\_XRAY (green marks), which are evaluated using the STRIDE algorithm.

The NMR dataset produces smoother distributions due to the above mentioned wider conformational sampling. Conversely, the X-ray dataset reveals a conformational preference for the PPII structures, which are less flexible and therefore more statistically present in the X-ray dataset. Very broad yet still visible structures are present in the  $P(\phi)$  distribution in correspondence of  $\alpha$ -like right handed helices ( $\phi \sim 60\text{deg}$ ), PPII ( $\phi \sim -100\text{deg}$ ) and  $\beta$  ( $\phi \sim 180\text{deg}$ ). It is to be observed that the sign of the  $\phi$  is related to helicity (right-handed is positive whereas left-handed is negative). The fact that the  $\alpha$ -basin propensity is lower than the others is related to the relatively high propensity to make H-bonds in that secondary structure.

Especially for the parametrization sake, it is important to consider also the “non bonded” distribution  $P(r)$ , defined as the distribution function of the spatial distance of beads separated by more than two beads along the chain, i.e. the minimal separation along the chain is defined to minimize the correlation with the other distributions depending on  $\theta$  and  $\phi$ , which involve beads separated up to two beads along the chain.

Due to the short length of the fragments chosen dataset (see figure 2),  $P(r)$  is not statistically well defined. More specifically, the statistical relevance strongly depends on the distance along the chain of the beads: contribution from beads separated by more beads are less statistically represented. This implies that  $P(r)$

at larger  $r$  is more noisy than the short  $r$ . To correct for this problem the following weighted normalization is adopted for  $P(r)$ .

$$\tilde{P}(r) = \frac{1}{N-n+1} \sum_{i=n}^N P_i(r). \quad (3.1)$$

Figure 5a reports the  $\tilde{P}_N(r)$  for different  $N$ . As it can be seen, the different length fragments show peaks at the same locations, representative of the chain induced order, but different statistical weight depending on the length of the fragment. In addition the longer fragment sets show a change in the behavior (a “flex”) around  $50\text{\AA}$ , which could therefore be considered as a sort of coherence length.

For completeness, in this work the model results for non bonded distribution will be compared with the experimental ones of figure 5a.

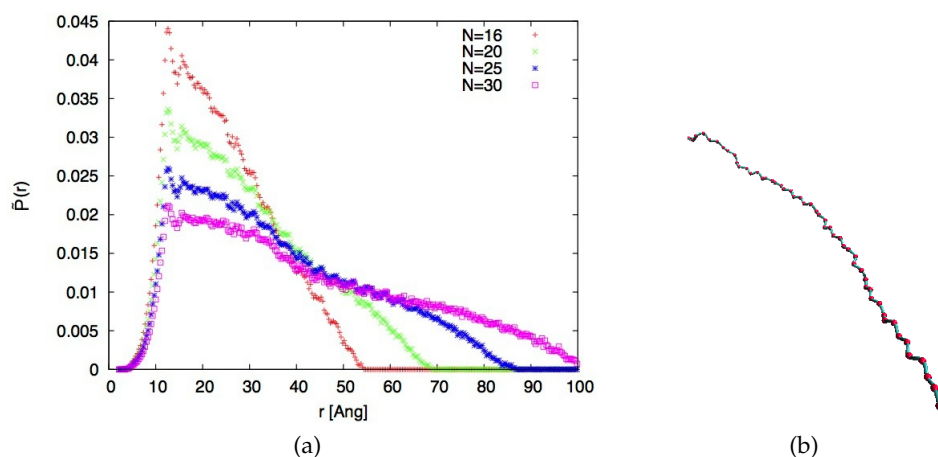


Figure 5: (a) Distance distribution of the PDB\_NMR\_STRIDE dataset using the normalization defined in equation 3.1. For all the distribution  $n = 6$  whereas  $N = \{16, 20, 25, 30\}$ . (b) 3D-rendering of the minimalist representation of the fragment  $81-133$ , 1WGS.

### 3.2.2 Distribution of geometrical backbone parameters, $\tau$ , $\gamma_1$ and $\gamma_2$

The  $\tau$ ,  $\gamma_1$  and  $\gamma_2$  angles, defined in section 2.3.2, are not explicit internal variables of the minimalist model. However, since they enter the relationship between the atomistic backbone variables  $\phi$ ,  $\psi$  with the minimalist counterpart (equations 2.38) an analysis of these variables is of paramount interest for the scope of this work. Figure 6 shows the distribution of these variables extracted

from the PDB\_NMR and PDB\_XRAY datasets. The two datasets show different behaviors: multi-peaked distributions for the NMR dataset whereas single peaked distributions with a Gaussian-like shapes. As for previously discussed cases, the

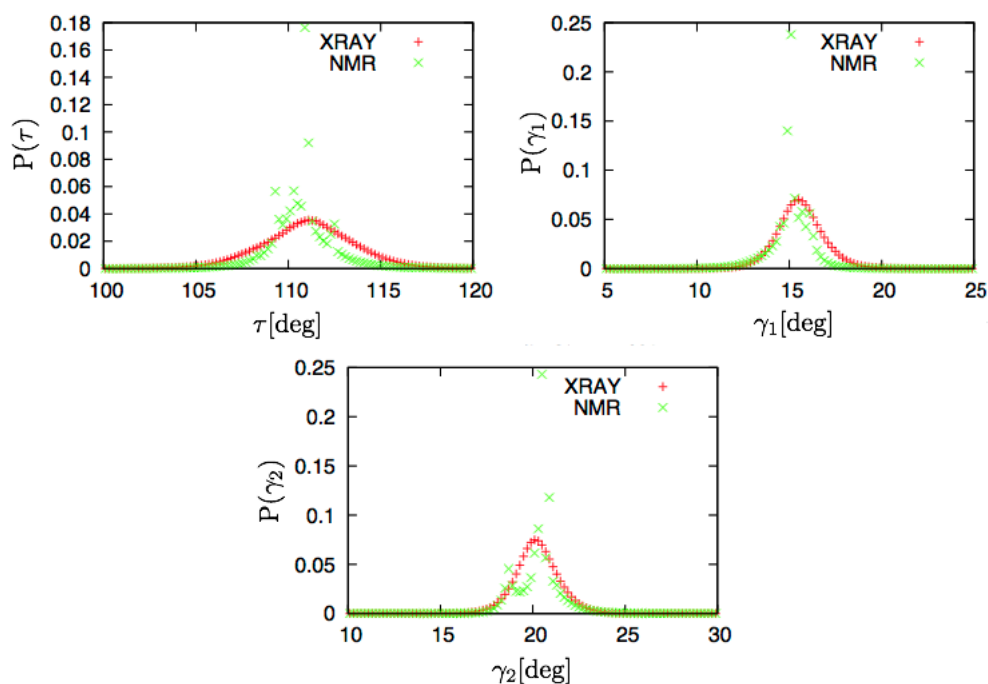


Figure 6:  $\tau$  distribution (top left),  $\gamma_1$  (top right) and  $\gamma_2$  (bottom). For each variable is reported the data extracted from the PDB\_NMR (green) and PDB\_XRAY (red) datasets.

multiple peaks might be attributed to different modules used in the determination of PDB\_NMR data. On the other hand, in this case, since the parameters under consideration involves very local distances and angles, their distribution is not likely to be substantially influenced by crystallographic constrains. Therefore in the following the X-ray derived distributions will be considered the reference ones, for these parameters.

Figure 7 shows the correlation plot for each couple of the  $\tau$ ,  $\gamma_1$  and  $\gamma_2$  variables obtained from the PDB\_NMR dataset. To be noted that, while the correlations between these variables, which are related to the same amino acid, are almost negligible (see panels a, b, c in figure 7), the  $\gamma_2 - \gamma_1'$  plot (figure 7d) ( $\gamma_1$  of the next residue) shows, conversely, a clear correlation. In order to operatively use the previous observation, it is useful to represent them in analytical form. Therefore,

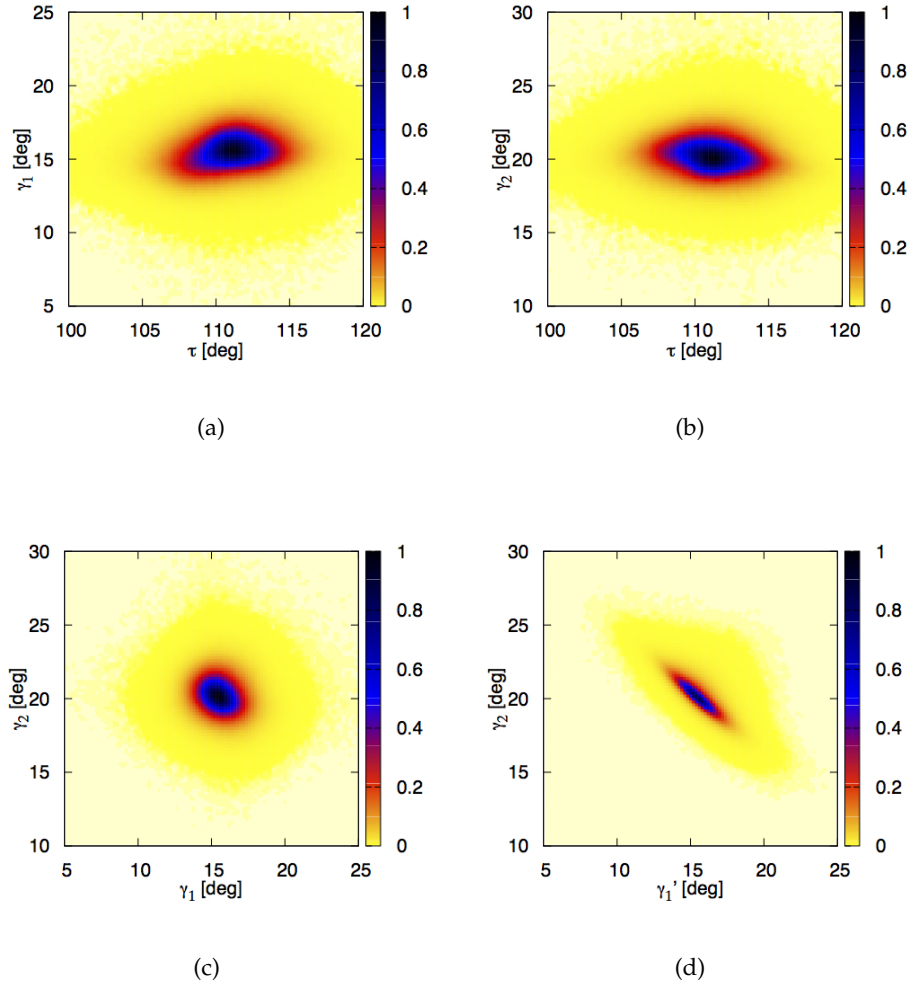


Figure 7: (a) correlation plot between the  $\tau$  and  $\gamma_1$  variables; (b) correlation plot between the  $\tau$  and  $\gamma_2$  variables; (c) correlation plot between the  $\gamma_1$  and  $\gamma_2$ ; (d) correlation plot between the  $\gamma_1'$  and  $\gamma_2$  variables. Data extracted from the PDB\_NMR dataset.  $\gamma_1'$  is the next  $\gamma_1$  angle along the polypeptide.

the  $\tau$  variable distribution is fitted with a Gaussian function (see fitted parameters in table 2). The fitting procedure for  $\gamma_1'$  and  $\gamma_2$  involved the search for a new set of coordinates  $\xi_1$  and  $\xi_2$ , which diagonalize the covariance matrix

$$C = \begin{pmatrix} C_{\gamma_1', \gamma_1'} & C_{\gamma_1', \gamma_2} \\ C_{\gamma_2, \gamma_1'} & C_{\gamma_2, \gamma_2} \end{pmatrix} = \begin{pmatrix} 7.7830 & 4.3357 \\ 4.3357 & 7.0270 \end{pmatrix}. \quad (3.2)$$



The matrix diagonalization lead to:

$$\xi_1 = -0.7371 \gamma_1' - 0.6757 \gamma_2 \quad (3.3)$$

$$\xi_2 = +0.6757 \gamma_1' - 0.7371 \gamma_2 \quad (3.4)$$

$\xi_1$  and  $\xi_2$  are fitted with a Cauchy distribution and a Gaussian distribution respectively. The Cauchy distribution is chosen for  $\xi_1$  in place of the Gaussian one because it provides a model more accurate. Figure 8 shows the results of the fitting procedure. Table 2 reports the values this analysis.

Variable	Function type	mean value [deg]	c [deg]
$\tau$	Gaussian	$111.08 \pm 0.01$	$2.34 \pm 0.09$
$\xi_1$	Cauchy	$-25.03 \pm 0.01$	$0.40 \pm 0.05$
$\xi_2$	Gaussian	$-4.40 \pm 0.01$	$1.48 \pm 0.15$

Table 2: Fitting parameters of the sampled distributions related to the variables  $\tau$ ,  $\xi_1$  and  $\xi_2$ . The Gaussian function is defined as  $f(x) = A \exp(-(x - x_0)^2/2c^2)$ , whereas the Cauchy as  $g(x) = A c/((x - x_0)^2 + c^2)$ , where in both cases the parameter  $x_0$  represent the mean value of the distribution.

### 3.2.3 Two variables $\theta$ , $\varphi$ distributions

In this section the two-variables version of the distribution of section 3.2.1 are evaluated. These are the  $\theta$ ,  $\varphi$  correlation plots, which are considered the minimalist counterparts of the RP, as anticipated in section 2.3.2. Following the same approach followed for the atomistic RP, here those plots are evaluated for separated the separated amino-acid classes defined in section 1.4, namely the proline (P), glycine (G), pre-proline (pP) and all the others (X). As shown in section 2.3.2, in order to completely represent the correlation, the  $\theta_+$ ,  $\varphi$  and  $\theta_-$ ,  $\varphi$  plots must be evaluated separately; this implies that the evaluation of correlations in the minimalist model must involve at least four subsequent "beads" (amino acids). The following algorithm is used for the evaluation of the plots:

- the polypeptide chain is analyzed by subsequent quadruplets;
- the amino acid type of the two central residues define the correlation plot, therefore there would be in principle 16 plots classes, defined by all the possible couples XX, XG etc. However, only 11 are possible, due to the P/pP type definition (see table 3);

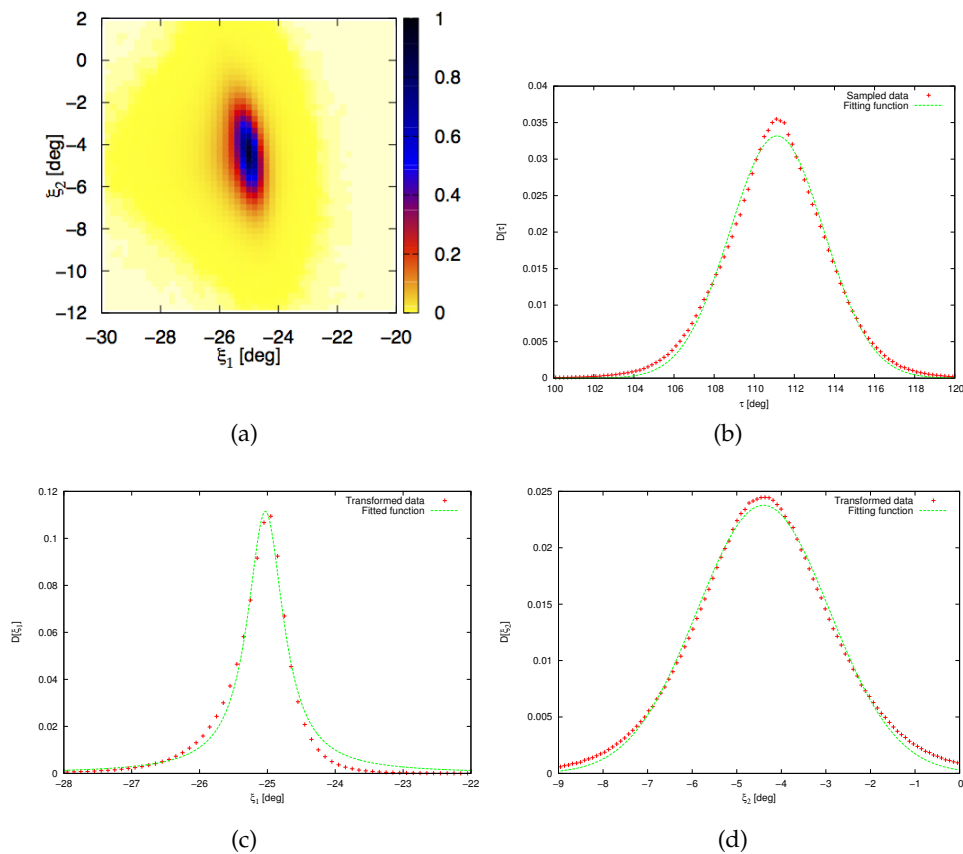


Figure 8: (a) Density of the  $(\xi_1, \xi_2)$  distribution obtained after the variable transformation T. Results fitting procedure on the variables  $\tau$  (b),  $\xi_1$  (c) and  $\xi_2$  (d). The parameters of the function are reported in table 2.

- the third residue characterize the type of the last residue: if the third residue is a pre-proline, the fourth is a proline; the third is an X amino acid or a glycine, then the fourth can be a non proline amino acid (N); if the third amino acid is a proline, then the next can be every amino acid (A).
- the external amino acids (first and fourth) are either determined by their side neighbors (e.g. if the the third is a pP, the the fourth must be a P, if the second is a P, the first is a pP) or considered influent.

Table 3 summarizes the all the allowed types of quadruplets. This classification provides eleven classes of quadruplets.

$\begin{array}{c} \text{3rd} \\ \text{AA} \\ \diagdown \\ \text{2nd} \\ \text{AA} \end{array}$	X	G	P	pP
X	✓	✓		✓
G	✓	✓		✓
P	✓	✓	✓	✓
pP			✓	

Table 3: Central amino acids of the allowed types of quadruplets. The rows reports the type of the second amino acid whereas the columns reports the type of the third amino acid.

Figure 9 and 10 report all the plots, evaluated using the PDB\_NMR\_STRIDE dataset. As expected, plots including G as a central AA (figure 10) shows the largest spread of occupation and smaller forbidden areas (although GG, PG and GP are affected by high noise due to the low statistics), while the plots for generic AA (X) show selective occupation of the secondary structure-like basins (see figure 9 in section 2.3.2 for their location), especially  $\beta$  and PPII (less  $\alpha$  basin), though also the intermediate regions are populated. Conversely, as expected the P-pP plots have a very specific occupation of the PPII basin, due to the conformational restraints due to proline.

All the couples of plots show differences between the each left and right  $\theta, \varphi$  plot. For non homogeneous amino-acids couples, its easy to ascribe these deviation to the different propensity of each amino acids. But when the couple is homogeneous there are still relevant deviations. These can be understood considering equations 2.38 in section 2.3.2. These deviations can be ascribed to the difference of the distribution of  $\phi$  and  $\psi$ , and secondly to the difference between  $\gamma_1$  and  $\gamma_2$ .

The presence of the proline provides a clear preference toward the PPII structures as shown in figure 9 from the second to the fifth row. Moreover the proline limits the exploration of the  $\theta$  space also in the preceding amino acids, as expected from the pP-RP. This effect can be clearly observed in the second row of the XpP-correlation plot. In order to understand the nature of the shapes, these correlation plots are compared with the correlation plots theoretically evaluated from the Ramachandran plots (for details, see section 3.3). In most of the cases the obtained shape reproduces approximately the experimental data, although the low sampling in the most specific cases (when in the site is considered a specific amino acid rather than a set) produce high noise. Therefore every particular

shape can be easily explained with the conformational preference of each amino acid.

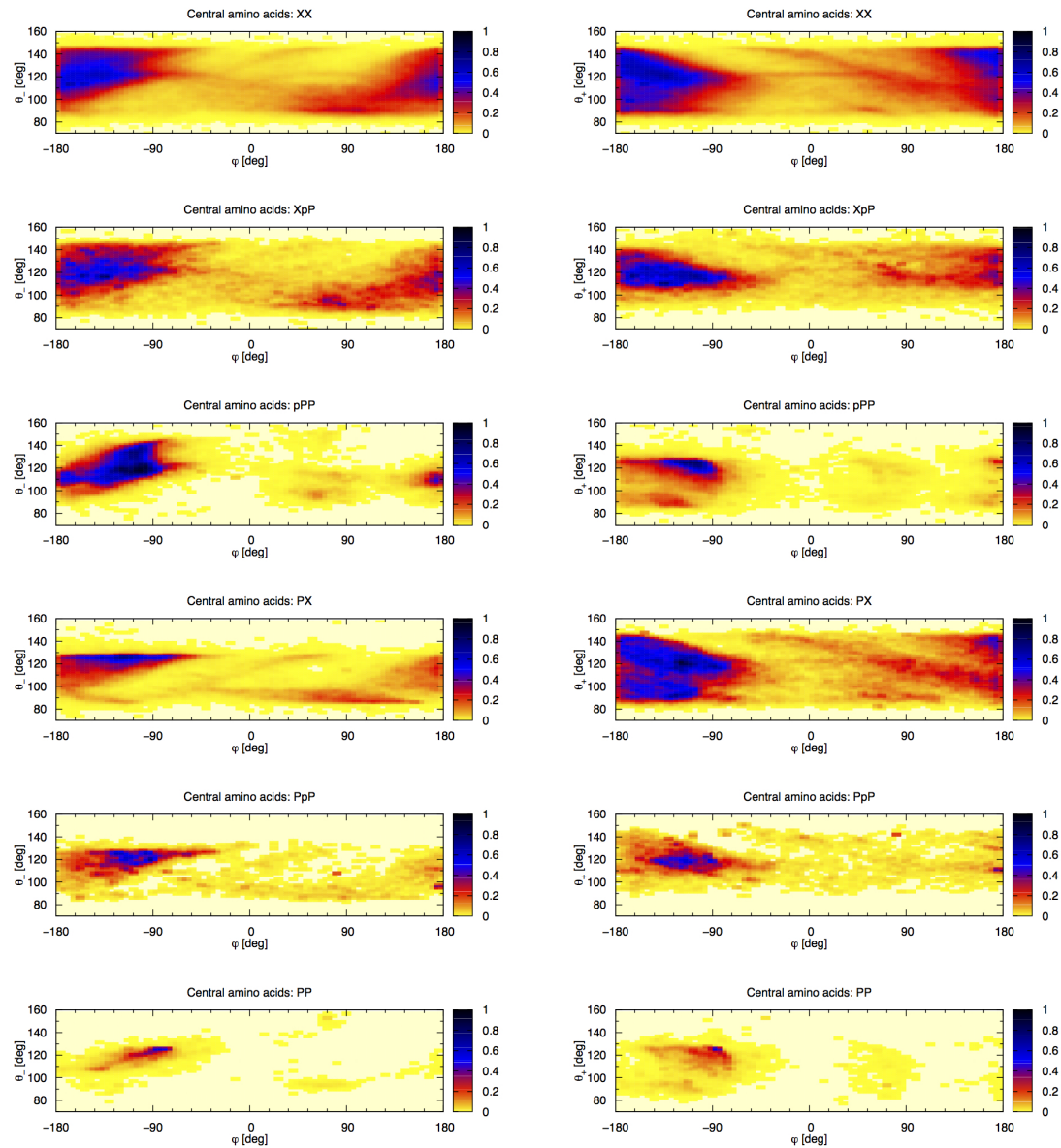


Figure 9: For each row the  $\phi, \theta_-$  (left) and  $\phi, \theta_+$  (right) of the case specified by the upper label. The case is defined by the amino acid composition of the two central C $\alpha$ s of four consecutive C $\alpha$ s. These data are extracted from the PDB\_NMR dataset selecting with the STRIDE algorithm the coiled regions.

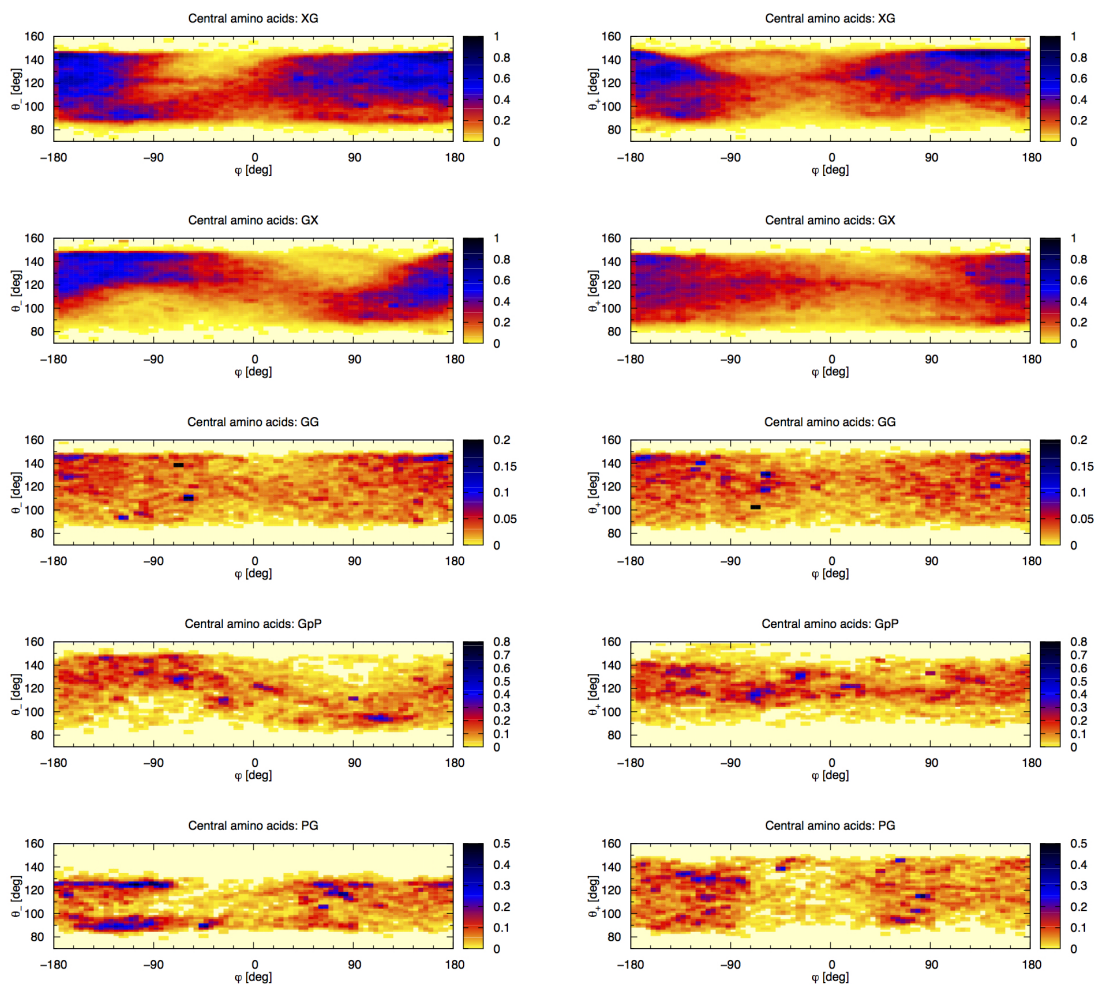


Figure 10: For each row the  $\phi, \theta_-$  (left) and  $\phi, \theta_+$  (right) of the case specified by the upper label. The case is defined by the amino acid composition of the two central  $C\alpha$ s of four consecutive  $C\alpha$ s. These data are extracted from the PDB\_NMR dataset selecting with the STRIDE algorithm the coiled regions. Due to the presence of residual cumulative spots, the color mapping of the GG and GpP have been rescaled in order to make the background motifs visible.

### 3.2.4 Three dimensional distribution ( $\theta_-, \phi, \theta_+$ )

The minimalist model built in this work focus on the reproduction of these correlation plots for unstructured proteins in order to reproduce the highest similarity to real structures. It would represent the highest structure similarity between a minimalist model and the reality. In order to reproduce such complexity

there is the need to extend the knowledge on the dataset analyzing the 3D-distribution (3Dd) of the  $(\theta_-, \varphi, \theta_+)$  variables in the available dataset. Likewise the case of 2D plots, these distributions are evaluated for separate classes of amino acids. Therefore each plot is identified by the specific couple of the central amino acids as illustrated in the correlation plots. The XX quadruplets represent the reference case in the following analysis due to the available high statistics and its extension on the  $(\theta_-, \phi, \theta_+)$  space. Figure 11 shows the slice referred of the 3D-distribution of the XX quadruplet. These slices show the how it is approximate the 2D-distributions indeed each slice differs both for shape and highest counting regions.

All these observations enhance the importance on the utilization of the 3D-distribution showing where the other representations lack of details.

### 3.3 $\theta, \varphi$ MAP GENERATION FROM RAMACHANDRAN PLOT

From the previous sections, it is apparent that the statistics for 2D and 3D maps as directly obtained from the experimental structures dataset, is often rather low, especially for the more specific amino acid classes. This implies that the maps are too noisy to be used e.g. in the Boltzmann inversion-like procedure (see equation 2.28) for the parametrization. In this section, methods are described decrease the noise.

Equations 2.38, 2.40 [55], provide the transformation from the atomistic to the minimalist conformational variables, which is in general a four to three map  $(\phi_1, \psi_1, \phi_2, \psi_2) \rightarrow (\theta_-, \varphi, \theta_+)$ . These formulas can be used to map the RP onto the  $(\theta_-, \varphi, \theta_+)$  plot. The starting RP can be more easily interpolated, bringing a smoother final result. A further smoothening is produced by the use of the analytical representation of the distribution of the transformation parameters  $\tau, \gamma_1, \gamma_2$  (equations 2.38-2.40, section 2.3.2). This process, of course, do not include additional information with respect to the direct reproduction of  $\theta, \varphi$  plots, rather it filters the noise and make the resulting plot easier to handle for the parametrization tasks.

Two different algorithms are used to perform this task. The first one is a Monte-Carlo method, which samples the RP and  $\gamma$ s,  $\tau$  distributions. These values are then used to evaluate the  $\theta_-, \varphi, \theta_+$  map, triplets with which the distribution

is build. The statistics of this map can then be increased at will with longer MC runs, and the noise reduced. Details are given in appendix G.

However, algorithms based only on the  $\phi, \psi$  angles distribution suffer of the not correct recognition of the turns structures, which therefore can bias the final distribution. However, the turns can only definitely be recognized only by the presence of H-bonds, for which detection a full atomistic structure is needed. Therefore an improvement of the above algorithm consists in the atomistic reconstruction from the RP variables. Details are reported in appendix G.

Figure 12 reports the  $\theta, \varphi$  maps evaluated with all the described methods. The first row is the same as figure 9, also reported here for comparison, the second is extracted form the XRAY\_STRIDE, the third and fourth rows are with the two methods described above. As it can be seen, the second method generates a maps that are basically indistinguishable from that directly obtained from the experimental dataset, and with improved statistics. The first method, conversely seems to suffer the bias due to the non proper treatment of the turn structures, which produces an effective oversampling of the  $\alpha$  basin.

The noise reduction is particularly relevant in the case of the glycine class maps, in which the starting statistics is evidently insufficient (e.g. figure 13 shows the plots related to the GG class with the same order of the XX class). In addition, the sampling is sufficient to obtain a well defined 3D map (as reported in figure 11 in section 3.2.4, for the XX class). These maps will be used as reference for the model building described in the next chapter.

### 3.4 DISPROT: A DATASET FOR DISORDERED PROTEINS

The evaluation of a force field for Intrinsically Disordered Proteins (IDPs) (sec. 1.6), in particular for random coil (sec. 1.7), is the main target of this work. The data previously reported are not extracted from a dataset of disordered proteins therefore theoretically they cannot represent the behavior of the the IDPs. Anyway the coil library represent the best starting point for the representation of this set thanks to the large amount of data contained in the PDB.

The principal database of the disordered proteins is the DisProt [25] (Disordered Proteins) introduced in section 1.6. The total amount of disordered regions in this database is 1539 in 694 proteins. As aforementioned the IDRs are generally not resolved on the published structures in the PDB.

The DisProt collect all the information related to the identification of the disorder and the availability of a PDB reference. All the information related to the state of the region, viz. random coil state (also called in the dataset “extended-disorder”), Molten globule state and pre-Molten globule state, and the particular function of the structure, and the derivation of the function (i.e. if it arise from a transition between states or form the state itself), can be found in the DisProt-file. As aforementioned, there are not available PDB-structures sampled using the X-ray crystallography, therefore the only available data are extracted from NMR-spectroscopy. Table 4 reports the correspondence between the DisProt and the PDB. The “Disordered” class, contains all the structure with unclassified

	Disordered	Extended	Molten	Pre-Molten
PDB entries	84	21	6	0

Table 4: Correspondence between the DisProt and the PDB data for each disordered state.

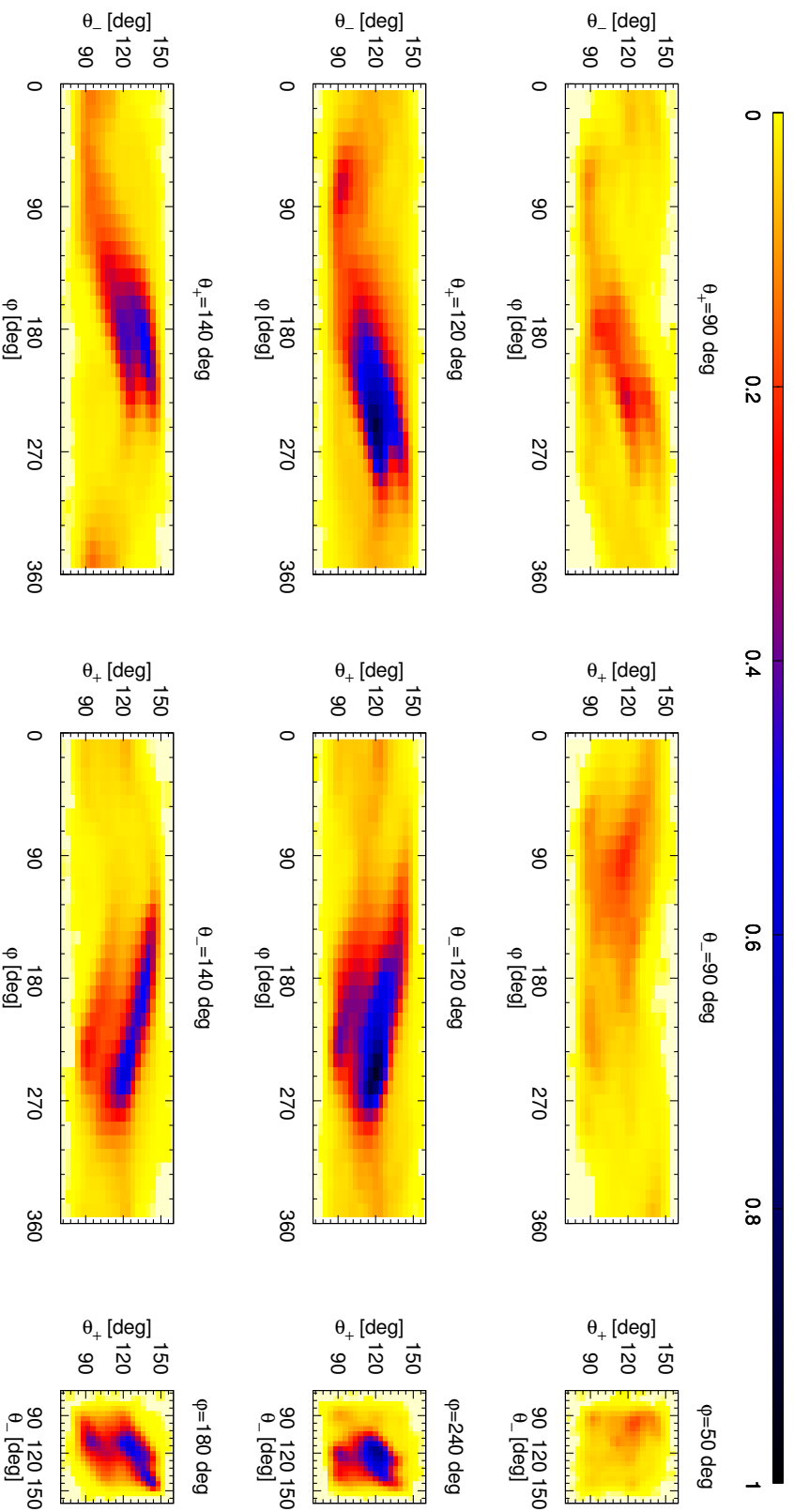
disorder. The number of available PDB data cannot provide a reliable reference database due to the low statistics. Figure 14 reports the correlation plots of the  $\phi$ ,  $\theta$  variables. Despite the high noise caused by the low number of statistical data, some structural preferences can still be observed. Figure 14a shows a significant residual presence of helical structures in the  $\alpha$ -basin. Other structural preferences are observed in the  $\beta$  and PPII basins although less relevant than  $\alpha$ . The Disordered dataset does not allow to recognize the state of the unstructured sample, therefore residual secondary structure may be present as expected by the proteins in the Molten and Pre-Molten globule states (sec. 1.6). Figure 14b shows only the preference toward the PPII basin although no conclusion can be taken due to the statistical irrelevant data.

Despite the low statistics of the structures, the amino acid contents have a more substantial statistics. Table 5 reports the percentage related to the amino acid contents of the two datasets. The data related to IDPs, found in the literature [58], are provided here as reference data. Moreover, the amino acid contents related to the “etropic chain” (particular function of IDPs, see appendix A) is reported. As expected there is a large contents of charged amino acids in each examined case, glycine and proline, which are known as the most structure destabilizers. The reference data have been found consistent with the amino acid contents of the IDR related to the “etropic chain” function. This particular function is found only in the random coil state.



			Disordered	Extended	Entropic Chain	Article [58]
<b>ALA</b>	<b>(A)</b>		8.35	10.84	7.57	7.15
<b>ARG</b>	<b>(R)</b>	+	5.48	5.59	4.50	4.21
<b>ASN</b>	<b>(N)</b>		3.91	3.19	2.20	2.06
<b>ASP</b>	<b>(D)</b>	-	6.35	6.16	4.68	5.05
<b>CYS</b>	<b>(C)</b>		1.08	0.78	0.51	0.61
<b>GLU</b>	<b>(E)</b>	-	7.55	10.02	15.31	14.26
<b>GLN</b>	<b>(Q)</b>		4.61	3.96	5.26	4.46
<b>GLY</b>	<b>(G)</b>		8.35	10.87	6.78	4.31
<b>HIS</b>	<b>(H)</b>		2.23	1.32	1.61	1.51
<b>ILE</b>	<b>(I)</b>		3.66	2.10	3.60	3.67
<b>LEU</b>	<b>(L)</b>		6.96	5.18	5.26	5.44
<b>LYS</b>	<b>(K)</b>	+	6.93	8.95	8.60	10.43
<b>MET</b>	<b>(M)</b>		2.55	2.04	1.25	1.30
<b>PHE</b>	<b>(F)</b>		2.89	2.00	1.69	1.66
<b>PRO</b>	<b>(P)</b>		6.19	8.69	9.77	12.07
<b>SER</b>	<b>(S)</b>		8.51	7.09	6.93	6.91
<b>THR</b>	<b>(T)</b>		5.14	4.14	4.80	5.14
<b>TRP</b>	<b>(W)</b>		1.11	1.17	0.63	0.32
<b>TYR</b>	<b>(Y)</b>		2.32	0.90	1.56	1.42
<b>VAL</b>	<b>(V)</b>		5.83	5.00	7.37	8.02

Table 5: This table show the amino acid percentage of the proteins and regions in the generic disordered state (Disordered) and random coil state (Extended) found in the DisProt database [25]. The third column reports the amino acid contents of the structures identified as entropic chain. The last column shows the percentages related to IDPs [58].



**Figure 11:** Slices of the XX three dimensional distribution of the  $\theta_-$ ,  $\phi$ ,  $\theta_+$  variables from the coiled structure evaluated with the STRIDE algorithm in PDB\_NMR dataset. The value of the orthogonal variable of each panel is reported upon the figure. The color scale of each plot is referred to the value of the voxel with highest value of the three dimensional histogram. Each row shows the content of each slice, which is taken at the reference values of the structured basins i.e.  $\alpha$ , PPII and  $\beta$  (first, second and third row respectively).

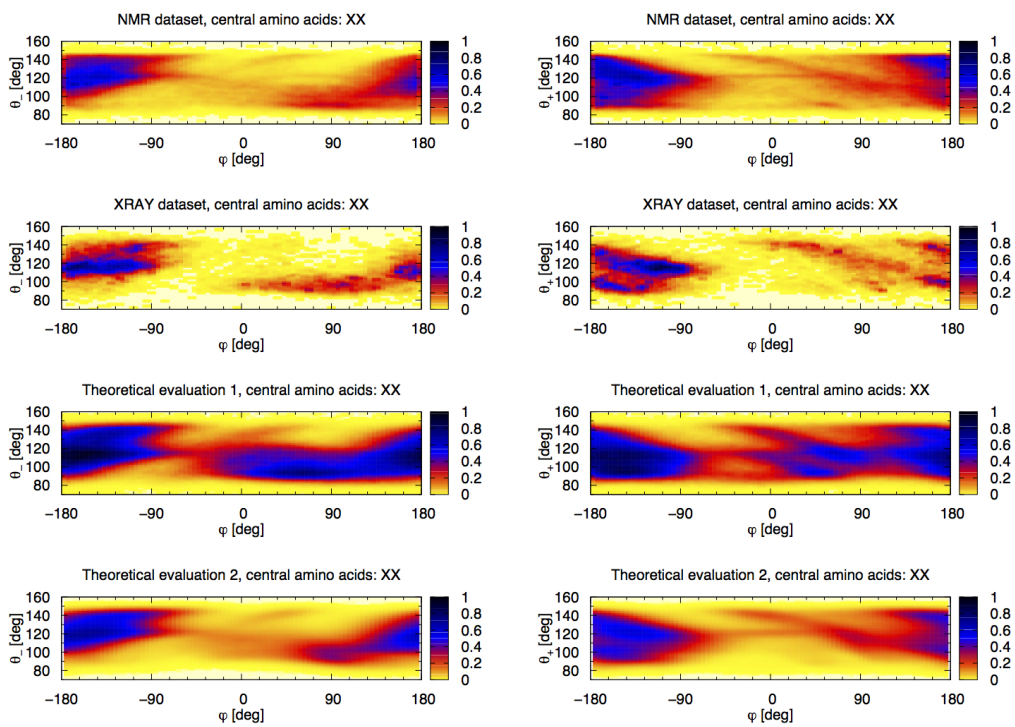


Figure 12: Comparison of the  $\varphi, \theta_-$  (left),  $\varphi, \theta_+$  (right) correlation plots of the XX sequence related to the NMR\_STRIDE dataset (first row), XRAY\_STRIDE dataset (second row), fist reconstruction algorithm (third row) and second reconstruction algorithm (last row).

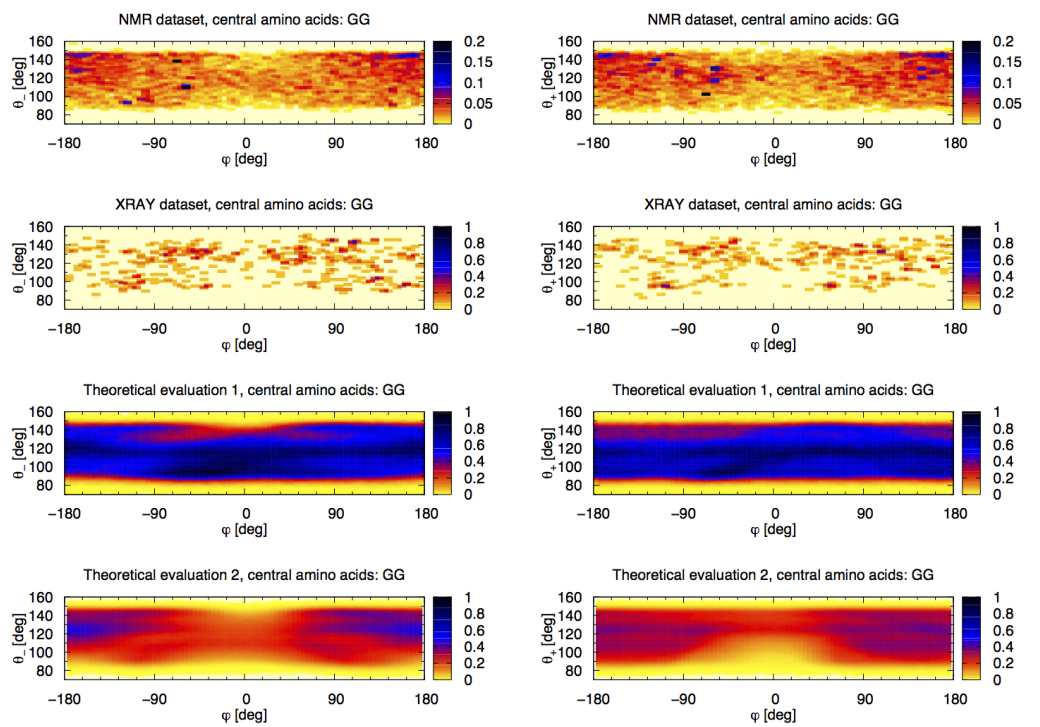
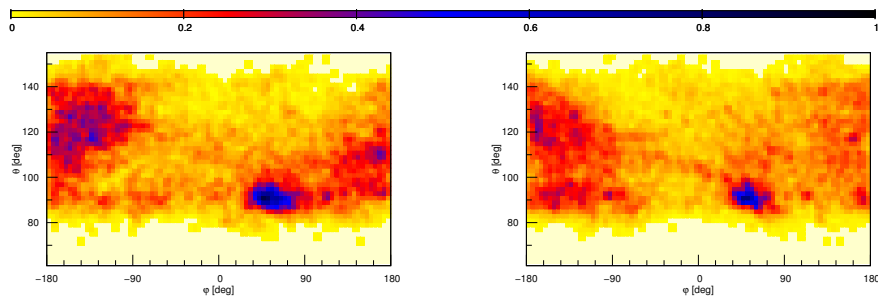
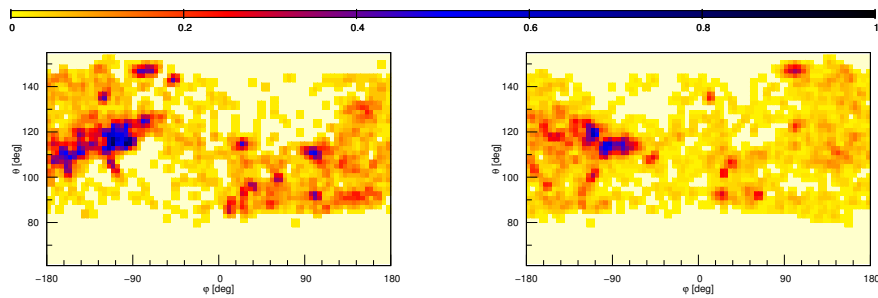


Figure 13: Comparison of the  $\varphi, \theta_-$  (left),  $\varphi, \theta_+$  (right) correlation plots of the GG sequence related to the NMR\_STRIDE dataset (first row), XRAY\_STRIDE dataset (second row), first reconstruction algorithm (third row) and second reconstruction algorithm (last row).



(a)



(b)

Figure 14: The  $\phi, \theta_-$  (left) and  $\phi, \theta_+$  (right) correlation plots of the Disordered (a) and Extended (b) datasets extracted from the DisProt [25].



## MINIMALIST MODEL FOR UNSTRUCTURED PEPTIDES

---

This chapter reports the main original results of this work, namely the parametrization and optimization of a minimalist model for unstructured peptides. The model is based on the reproduction of the 1D, 2D and 3D statistical distribution described in the previous chapter. The potential terms are first described, while the result of the optimization process are subsequently reported. Simulations performed with the model are finally illustrated.

### 4.1 MODEL DEFINITION

The model used in this Thesis is a one-bead  $C_\alpha$ -based model for coiled structures i.e. the minimalist model. The polypeptide is represented as an unbranched beads-chain where each beads is constrained to hold the same distance from its next neighbor along the chain. This distance is  $3.8\text{\AA}$ , which represents the  $C_\alpha$  the separation when the peptide bond is in trans conformation. The mass of the bead is evaluated with the average molecular mass of each amino acid which is about 115 a.u.. As mentioned in chapter 2, the FF includes two terms, describing local and non local interaction respectively (see eq 2.31). As reported in section 2.3.1 the local term is usually separated in terms dependent on  $\theta$  and  $\varphi$  (see equation 2.32). While this description is acceptable for regular secondary structures, it is apparent from the analysis of previous sections 3.2.3 and 3.2.4 that this potential form cannot easily account for  $\theta - \varphi$  correlation, and especially for their direction dependent part, manifesting in the fact that  $\theta_+$ ,  $\varphi$  and  $\theta_-$ ,  $\varphi$  maps are different. In order to account for this, here the following form of the local potential is considered:

$$U_{\text{loc}} = U_{\text{corr}}(\theta_-, \varphi, \theta_+). \quad (4.1)$$

The  $U_{\text{corr}}(\theta_-, \varphi, \theta_+)$  potential term can be evaluated from  $P(\theta_-, \varphi, \theta_+)$  through the BI method (eq. 2.28), illustrated in section 2.2.2, which implies a simple mathematical operation on the P. The numerical accuracy is increased by the elimination of the statistical noise operated by the algorithms described in the previous chapter. However, it requires then knowledge of probability distribution  $P_0(Q_i)$  (eq. 2.28) of the reference state of the non interacting beads. The minimalist model considers each interacting center constrained to its nearest neighbors along the chain at a defined distance. The bond angle and dihedral angle distributions of the non interacting system are:

$$P_0(\theta) = \frac{\sin(\theta)}{2}, \quad P_0(\phi) = 2\pi \quad (4.2)$$

In this case the  $P_0(\theta_-, \varphi, \theta_+)$  is the product of the distributions reference distribution of its elements, therefore:

$$P_0(\theta_-, \varphi, \theta_+) = \frac{\pi}{2} \sin(\theta_-) \sin(\theta_+). \quad (4.3)$$

As the statistical distribution, the  $U_{\text{loc}}$  should depend on the amino acid type. The same classification of AAs in the four classes is used, separating glycine proline and pre-proline from the generic amino acid set. As for the distributions, eleven different classes of potentials can be defined, identified by the following type couples

- XX
- GG
- GpP
- pPP
- GX
- PG
- XpP
- PP
- XG
- PX
- PpP

However, the statistics of data is good only for the XX class. Therefore, though increased by the data improvement, the results for other classes must be considered with care.

In order to take advantage from the DOF reduction, this model considers implicitly the effects of the solvent. Therefore the non bonded term in equation 2.31 must be comprehensive of the solute-solvent and solute-solute non bonded interactions.  $U_{\text{nb}}$  model is built considering the the potential parametrized by a Morse function:

$$U_{\text{Morse}}(r) = A((1 - e^{-b(r-r_0)})^2 - 1), \quad (4.4)$$



where  $A$  is the magnitude of the deep measured from the minimum to the asymptotic value,  $r_0$  is the minimum position and  $b$  is related with the width of the deep indeed the full width at half amplitude ( $\Delta r_{A/2}$ ) follows this relationship:

$$\Delta r_{A/2} = \frac{1}{b} \log\left(\frac{2 + \sqrt{2}}{2 - \sqrt{2}}\right) \approx \frac{1.76}{b}. \quad (4.5)$$

The non bonded potential are in general repulsive at short distances, in order to reproduce the VdW repulsion between the atoms. At larger distance this potential should be able to reproduce the electrostatic interactions of the system which could be either repulsive or attractive depending on the mean amino acid type involved in the interaction.

In order to parametrize such potential, a starting set of parameters reproducing a weak repulsion between each beads is adopted. In section 4.3 the optimization of these parameters is described.

#### 4.2 LOCAL POTENTIAL PARAMETRIZATION

As said, unstructured proteins do not have stabilizing h-bonds. Therefore their conformational tendency, expressed by the minimalist equivalent of the RP, i.e. the  $\theta, \varphi$  maps, is entirely determined by the backbone chemistry and steric interactions. These must be included in the minimalist representation into the local and non local terms. By definition, the non local interaction acts only on beads which are located at least 3 AA apart, and are likely to represent side chains steric and hydrophobic effects, but not the local backbone effects which are included in the local term. Considering that in addition, the statistics of long peptides rapidly decreases as with the peptide length, it is reasonable to deduce that the main effect of the whole peptide structural and dynamical behavior should be imputed to the local term, which is considered the prevalent one, and optimized as the first. The non local term is optimized as secondary and considered a "correction" to the whole system behavior. As said, the local term can be obtained by direct BI from the  $P(\theta_-, \varphi, \theta_+)$ . The densities  $P(\theta_-, \varphi, \theta_+)$  and the result of the BI are in the form of 3D-gridpoint. Therefore, in order to use all the available information, the adopted strategy interpolates all the point of the  $(\theta_-, \varphi, \theta_+)$  space defining the potential function as a stepwise function. Because the potential function  $U(\theta_-, \varphi, \theta_+)$  must be derivable along every direction of the 3D-space, among the interpolation algorithms the tricubic interpolation [64]

results the most suitable. This algorithm, indeed, provides the  $C^1$  continuity. The details on the interpolation algorithm can be found in appendix D.

The interpolation algorithm requires that for each point of the 3D-grid, the values of the potential function, its first derivatives and the mixed derivatives of second and third order are known (the set of needed values is shown in equation D.2). The required derivatives are evaluated through the finite difference method.

In view to reproduce a force field  $U(\theta_-, \varphi, \theta_+)$ , the boundary conditions must be imposed in the algorithm by implementing derivatives matching the topology of the system.  $\varphi$  is periodic of  $2\pi$  so is derivative of  $\varphi$ . For all the values  $\tilde{\varphi}$  and  $\tilde{\theta}_-$  constants, the value of  $U(\tilde{\theta}_-, \tilde{\varphi}, 0\text{deg})$  is constant; this is also true for  $\theta_-$ , and when they are on the other border (180deg). Regarding the derivatives on  $\theta_{+/-} = \{0, 180\}\text{deg}$  must be observed on the supplementary angle of  $\varphi$  with opposite sign. Resuming the periodic boundary conditions for the  $P(\theta_-, \varphi, \theta_+)$  function must be:

$$\begin{aligned} \bullet \forall \theta_+, \theta_-, \quad & U(\theta_-, -180\text{deg}, \theta_+) = U(\theta_-, +180\text{deg}, \theta_+), \\ & \frac{\partial U(\theta_-, -180\text{deg}, \theta_+)}{\partial \varphi} = \frac{\partial U(\theta_-, +180\text{deg}, \theta_+)}{\partial \varphi}; \end{aligned} \quad (4.6)$$

$$\begin{aligned} \bullet \forall \theta_+, \varphi, \tilde{\varphi}, \quad & U(0, \varphi, \theta_+) = U(0, \tilde{\varphi}, \theta_+), \\ & \frac{\partial U(0, \varphi, \theta_+)}{\partial \theta_-} = -\frac{\partial U(0\text{deg}, \varphi \pm 180\text{deg}, \theta_+)}{\partial \theta_-}; \end{aligned} \quad (4.7)$$

$$\begin{aligned} \bullet \forall \theta_+, \varphi, \tilde{\varphi}, \quad & U(180\text{deg}, \varphi, \theta_+) = U(180\text{deg}, \tilde{\varphi}, \theta_+), \\ & \frac{\partial U(180\text{deg}, \varphi, \theta_+)}{\partial \theta_-} = -\frac{\partial U(180\text{deg}, \varphi \pm 180\text{deg}, \theta_+)}{\partial \theta_-}. \end{aligned} \quad (4.8)$$

The last two equations are true also for  $\theta_+$ .

Figure 1 reports the potential function representation of the XX class, represented as an isovalue surface. This figure reports the  $\varphi$  axis in the  $[0 : 360]\text{deg}$  in order to show a central compact core related to the  $\beta$  and PPII basins, which contain most of the energy minima of this potential. Slices of the  $U(\theta_-, \varphi, \theta_+)$  taken on different axes at different values are reported in figure 2. The slices were chosen according to the representative values of each structured basin. It shows

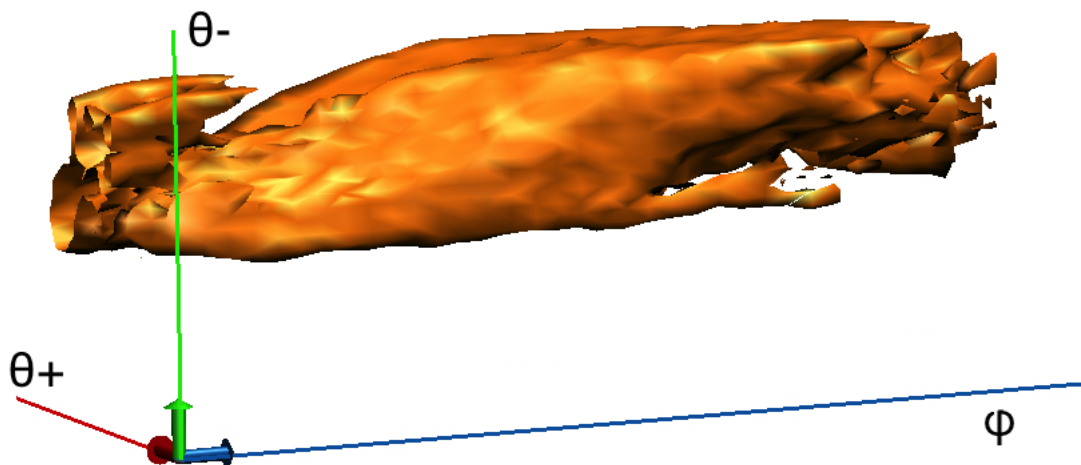


Figure 1: Rendering of isosurface (isovalue  $-1.79\text{kcal/mol}$ ) of the  $U(\theta_-, \varphi, \theta_+)$  for the XX class.  $\varphi$  variable ranges in  $[0 : 360]\text{deg}$ .

that the energy minimum is in proximity to those regions represented in the third row, which is related to the PPII basin.

#### 4.3 NON LOCAL POTENTIAL OPTIMIZATION

Assumed that the non bonded interactions produce correction to the distributions, which are peptide length dependent, the  $U_{nb}$  term was optimized in order to reproduce the differences in the distributions from dataset of different peptide lengths contained in the NMR\_STRIDE dataset. As shown in section 3.2.1, length of the peptides varies in the range between 7 amino-acids (lower limit imposed during the dataset evaluation) and more than 50 amino acids. The total amount of fragments decrease exponentially varying the length of the fragments (figure 2 in section 3.1). For this reason the normalized distance distribution  $\tilde{P}_N(r)$  (equation 3.1) has been used in order to compare the results obtained from the simulation with those related to the experimental dataset.

The use of a single non bonded potential meets certain difficulties related to the different fragment sizes. In longer fragments the contribution of the non bonded interaction will be more important rather than in shorter due to the higher number

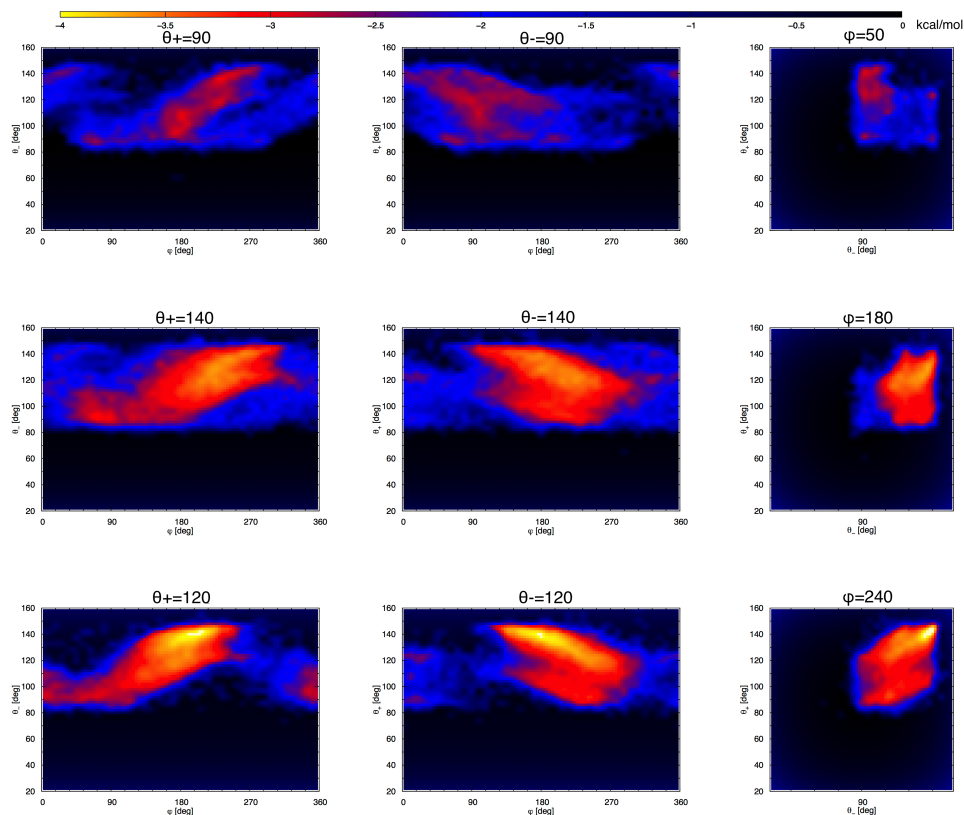


Figure 2: Representation  $U(\theta_-, \varphi, \theta_+)$  after the tricubic interpolation. First column is related for a defined value of the variable  $\theta_+$ . The second column is related for a defined value of the variable  $\theta_-$ . The last column is related to the  $\varphi$  variable. For each row are selected those values related a particular regular secondary structure. The first is related to the  $\alpha$ -helix, the second to the  $\beta$ -sheet and the last to the PPII-helix.

of one-bead one-bead interactions. For this reasons, in the simulation protocol adopted during the modeling of this system, the simulation of polypeptides of different length have been considered. Moreover the contributes obtained from a particular fragment have been conveniently weighted considering the aforementioned length distribution. These weights are reported in table 1.

The optimization algorithm evaluates the best set of parameters in the meaning of the reproduction of the non bonded distribution of the experimental data. Therefore, for each set of parameters a simulation run is performed in order to evaluates the statistics. Each parameter set is drawn form the uniform distribution inside the specific interval which is an input parameter of the algorithm. For each set of sampled parameters, the preliminary check of the potential function,

Length	Weight
17	75
24	18
50	1

Table 1: Weights considered for a specific length during the evaluation of the distributions from the simulations.

Parameter	Starting value	Min	Max
A [kcal/mol]	6.06	0	10
$r_0$ [Å]	43.55	2	150
b [1/Å]	$1.78 \cdot 10^{-03}$	$10^{-5}$	1
$\chi_{\max}^2$	-	-	$10^{-3}$

Table 2: Parameters related to the optimization of the Morse potential.

which is performed evaluating if the mean square root deviations between the derivatives of the function of with the last accepted potential function is below the threshold  $\chi_{\max}^2$ , allows to avoid the simulation of many useless sets of parameters speeding up the whole process. For further details the optimization algorithm for the parameter is discussed in appendix H. In this work, the parameters related to the Morse potential function (equation 4.4) have been optimized considering the values reported in table 2.

The simulations protocol is described in the next session. The set of optimized parameters evaluated after 288 iterations of the algorithm are

$$A = 5.16[\text{kcal/mol}]; \quad r_0 = 139.4[\text{Å}]; \quad b = 5.7 \cdot 10^{-4}[1/\text{Å}]; \quad (4.9)$$

therefore, considering equation 4.4 and 4.5, the optimized function is a soft repulsive potential. This result is consistent with the non globular nature of the unstructured proteins due to the high contents of polar and charged residues.

#### 4.4 MOLECULAR DYNAMICS SIMULATIONS

The simulations performed during the development of the force field follow a common protocol. All the simulations are performed on the DL\_POLY\_Classic software package [35], conveniently modified in order to implement the 3D potential, as reported in appendix F. A brief description of the software package which includes the input and output files is reported in appendix C.

Stage	Phase	T [K]	Integrator	$\Delta t$ [ps]	Duration [ns]
Optimization	Minimization	-	leapfrog	$10^{-3}$	2
	Equilibration	300	leapfrog	$10^{-3}$	0.5
	Production run	300	leapfrog	$10^{-3}$	3
Simulation	Minimization	-	leapfrog	$10^{-3}$	2
	Equilibration	300	leapfrog	$10^{-3}$	3
	Production run	300	leapfrog	$10^{-3}$	15

Table 3: Simulation protocol adopted during the potential optimization and the simulation.

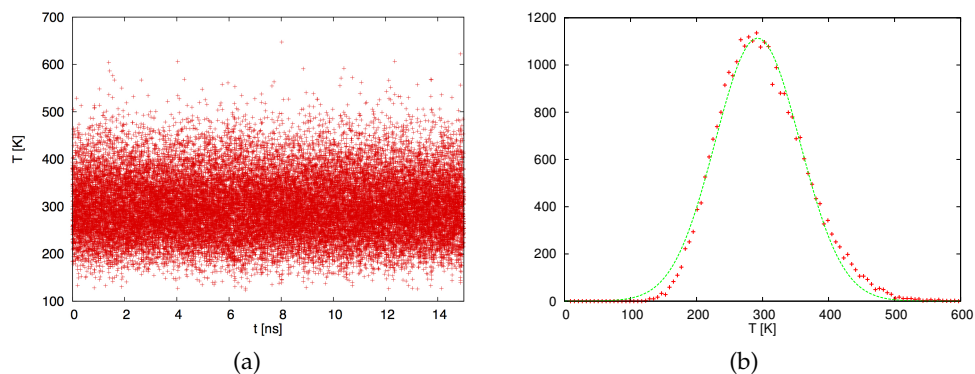


Figure 3: Time evolution and distribution of the temperature obtained during the production of the model with 24 residues. The green line is a Gaussian fit.

For force field model, simulations are performed on systems of different size. The length of the considered fragments are: 17, 24 and 50 residues. Each simulation follows the same protocol, described in table 3.

Figure 3 reports the time evolution (a) and the resulting distribution (b) of the temperature of the production run evaluated using the optimized potential. The Gaussian distribution of temperature indicates a proper thermalization.

Figure 4 shows the distribution of the variables bond angle (a), dihedral angle (b), distance  $r_{1,17}$  (c) and the non bonded  $\tilde{P}_{17}(r)$  (d). The single variable distributions, although they represent a partial picture of the system considering the correlations described in chapter 3, they provide an easily accessible comparison between the experimental data and the simulation results. In general, the simulation result reproduce qualitatively all the distributions. All the peaks corresponding to structural basins are present, located in the right place and approximately of the right height. Quantitatively, it appear to be a compromise

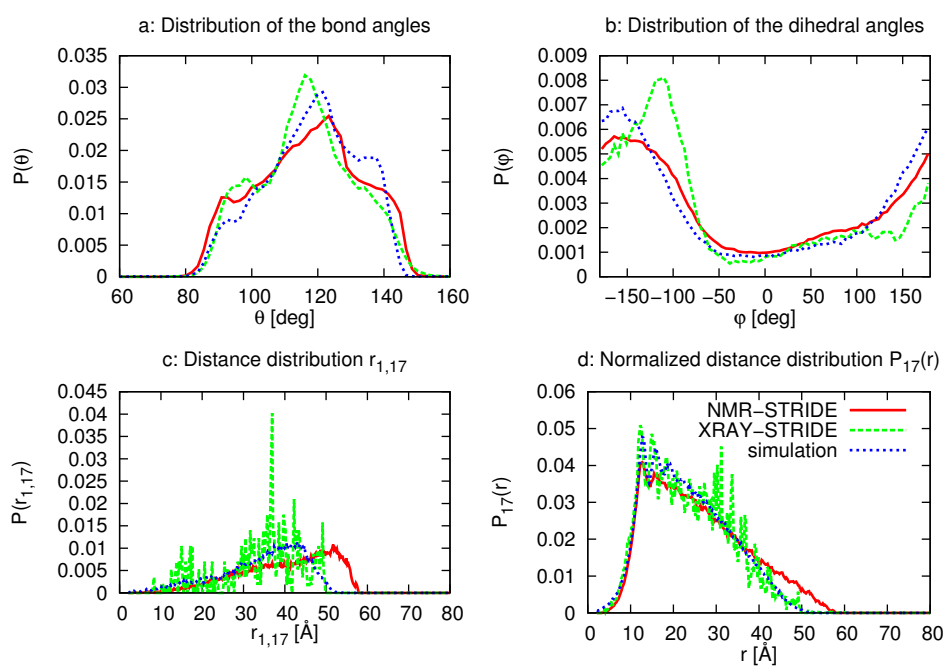


Figure 4: One dimensional distributions of the bond angle (a), dihedral angle (b),  $r_{1,17}$  (c) and non bonded distribution  $\tilde{P}_{20}(r)$  (d). Each plot is related to the NMR\_STRIDE dataset (red line), XRAY\_STRIDE dataset (green line) and the data extracted from the simulation (blue line).

between the XRAY\_STRIDE and NMR\_STRIDE data: the distribution are broader than the X-ray ones, and slightly sharper than NMR data. On other words, the model accounts for all the structural features. The residual discrepancies with experiment can be ascribed to the larger variability of structural and environmental conditions in the case of the NMR, and to the crystallographic constraints in the case of X-ray.

On the technical level, a manipulation of the non bonded interaction parameters could fine tune the height of the shoulders. This can be inferred splitting the average distributions into its components representing the simulation performed with fragments of different length (figure 6, a). It can be observed that the relative height of the peaks depends on the peptide length, and, consequently on the non bonded interaction, which has a different relative weight depending on the length. This can be especially observed in the distribution of the small bond angle interval ( $80\text{deg} < \theta < 110\text{deg}$ ) and in the large bond angle interval ( $130\text{deg} < \theta < 150\text{deg}$ ) consistently with the observed differences in the other distributions. Splitting the PDB\_NMR dataset in partitions with peptides longer and shorter than 18 residues in its higher and lower than 18 residues partitions, namely O18 and U18, it is possible to observe the deviations due to the length of the fragments, which are similar to those aforementioned for the simulated model. Unfortunately there is low statistics for each peptide length and an accurate evaluation of the experimental size effects are not available in this work. However, as said, the relative fine tuning of the height of the peaks might depend also on environmental conditions. Therefore the present determination of the non bonded potential is considered optimal with respect to the available reference data.

As a final validation of the model, figure 7 reports the  $\theta$ ,  $\varphi$  correlation plots, and figure 8 the 3D  $\theta$ ,  $\varphi$ ,  $\theta$  map, and its slices. This is a result that was not produced previously by any model present in the literature. By comparing figure 7 with the first row in figure 9 in section 3.2.3, and in figure 8 with 11 in section 3.2.4, it can be seen that all the feature of the generic amino acid correlation map are reproduced, even the quantitative level. To our knowledge, the present is the only model capable of reproducing all the complex structural details of unstructured proteins.

In summary, in this chapter it was shown that the model built in this thesis work is capable of reproducing complex structural features of the unstructured peptides structure and dynamics. These are (i) the intrinsic conformational tendency,



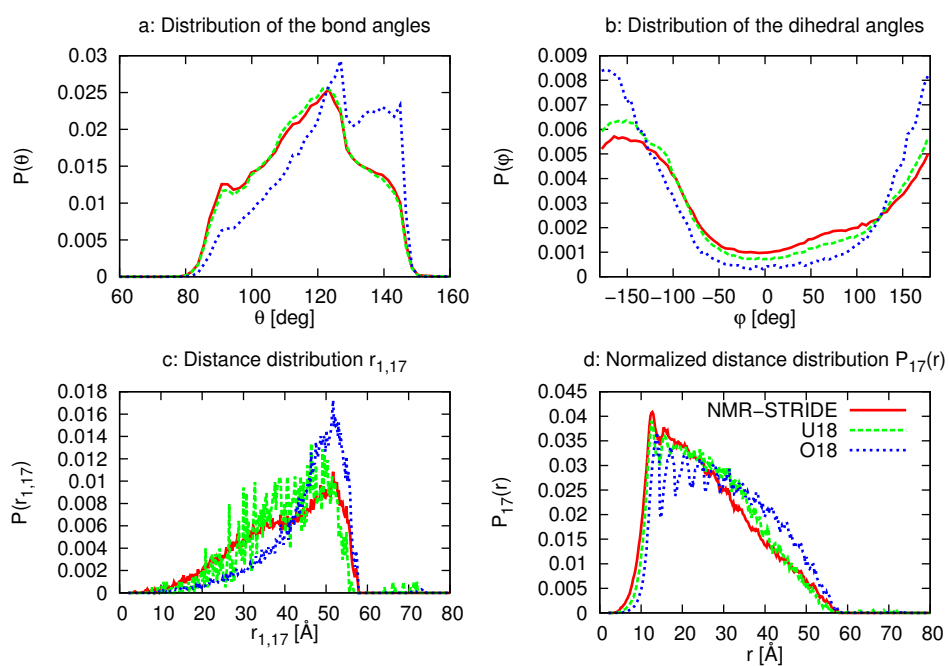


Figure 5: One dimensional distributions of the bond angle (a), dihedral angle (b),  $r_{1,17}$  (c) and non bonded distribution  $\tilde{P}_{17}(r)$  (d). Each plot is related to the NMR\_STRIDE dataset (red line), the U18 dataset (green lines) and the O18 dataset (blue lines).

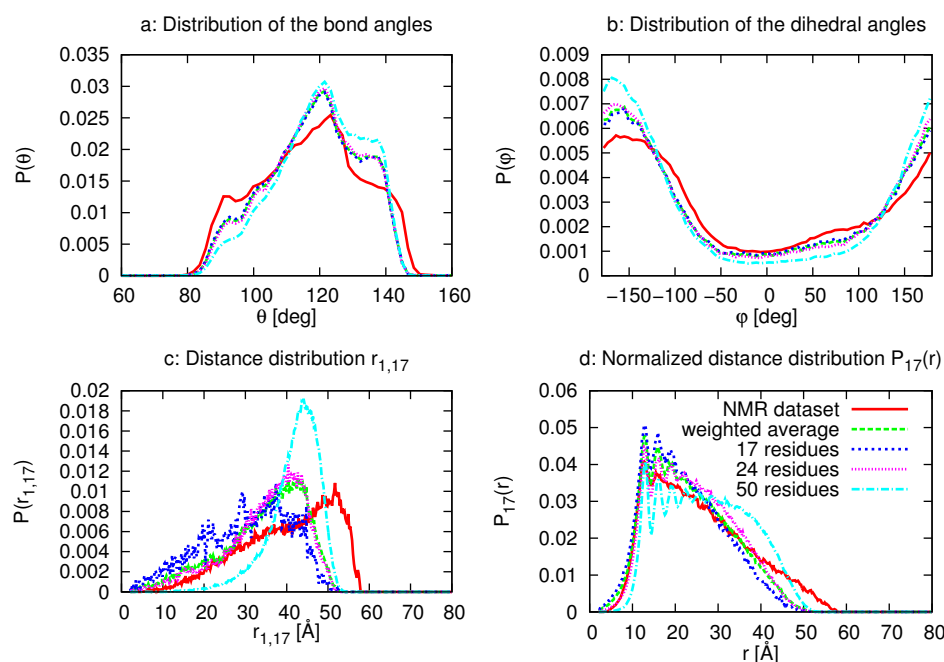


Figure 6: One dimensional distributions of the bond angle (a), dihedral angle (b),  $r_{1,17}$  (c) and non bonded distribution  $\tilde{P}_{17}(r)$  (d). Each plot is related to the NMR\_STRIDE dataset (red line), simulation weighted average (green lines) and the 17 residues (blue lines), 24 residues (violet lines), 50 residues (cyan lines) components.

measured as the relative peaks location and height in the single variable local distribution, (ii) the peptide length dependence of the non bonded interaction relative weigh, (iii) the 3D local variable correlations. All of this is done by a force field composed by two terms, the first representing the local interaction via a three-variable function, the second representing the non local interaction by an extremely simple single variable functional form. The model was implemented in a general purpose MD program and can be used for protein of any length.

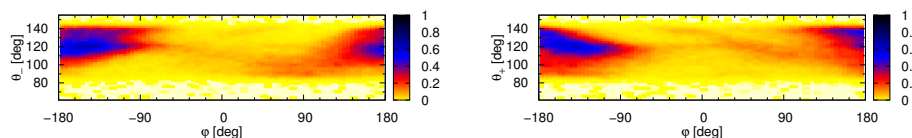


Figure 7: Left and right  $\theta, \phi$  correlation plots evaluated in the simulation.

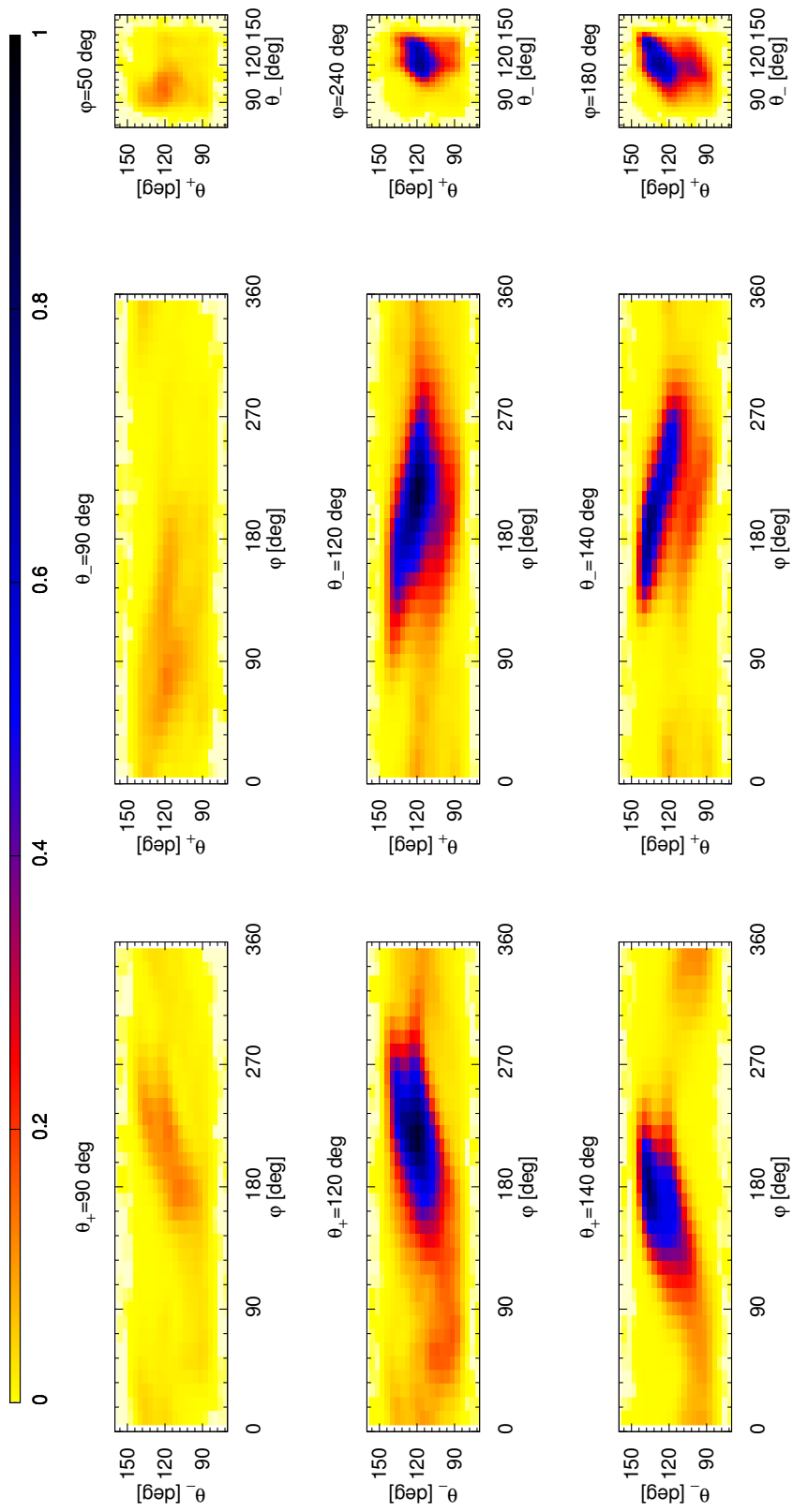


Figure 8: Slices of the XX three dimensional distribution of the  $\theta_-$ ,  $\phi$ ,  $\theta_+$  variables from the simulation result. The value of the orthogonal variable of each panel is reported upon the figure. The colorscale of each plot is referred to the value of the voxel with highest value of the three dimensional histogram. Each row shows the content of each slice, which is taken at the reference values of the structured basins i.e.  $\alpha$ , PPII and  $\beta$  (first, second and third row respectively).



## CONCLUSIONS AND PERSPECTIVES

---

This work returned two classes of results. The main result is reported in chapter 4, and is the complete parameterization (local and non local part) of a minimalist model for unstructured peptides. The model was shown to accurately reproduce the 1D, 2D and 3D distributions of internal variables derived from experimental data both involving local and medium-long range distances.

The conformational tendencies of a polypeptide chain of generic sequence is can address in detail. At variance with previous minimalist models, the peculiar form of the potential is capable of reproducing both the conformational variables correlations and the directionality of the chain. To our knowledge, no CG model reproduces these effects, with the exception of the multi-bead model by Scheraga and co-workers [45], which is however much more complex, including multiple beads for the backbone and for the side chain. Conversely the model here described is a linear chain and all the complex effects are included in the accurate parametrization of the force field, and the peculiar choice of its functional form.

The non local variables distributions are also reproduced, especially their peptide length dependence. This is achieved with a very simple functional form for the non local term, and thanks to the fact that the main physical effects are included in. The final result is an accurate, yet simple and manageable, force field for unstructured peptides and proteins of any length, to be used for generic molecular dynamics simulations.

In the road towards these achievements, several side, yet very important, results were achieved as well. The model parametrization required a very accurate representation of the Ramachandran plot minimalist equivalent. This work has showed that, at variance with structured proteins, for which the 2D  $\theta$ ,  $\varphi$  maps are sufficient to this purpose for unstructured proteins this is represented by the 3D map of the internal variables  $\theta_-$ ,  $\varphi$ ,  $\theta_+$  and/or by the whole set of its 2D projections. These can be experimentally evaluated if a reliable structural dataset is available. Since, however, the structural determination of the unstructured proteins and

peptides is much more difficult than that of structured proteins, the creation of a reliable and sufficient dataset for this proteins class was not straightforward. One solution was to create it by subtraction of ordered secondary structures from the generic dataset. This in turn has led to the reconsideration of the available algorithms for secondary structure detection and assignment, and raised very interesting issues on the definition of unstructured fragments themselves. Side results from this part were improved algorithms for the maps generation and for the selections of secondary structures. In addition, the considered secondary structure assignment algorithms are based on the detection of the hydrogen bond network. Therefore, the subtraction of the ordered structures, ensure the absence of hydrogen bonds from the structures in the resulting dataset.

A second side result, relevant on the technical level, was the fact that the local potential includes correlations and directionality of the chain, via a multiple variable functional form. This is a novelty with respect to previous treatments, and, furthermore, required some technical intervention into the DL\_POLY software for its implementation. These software updates are made available, and could be exploited also for different multi variate potential forms implementations.

This work must be considered within a more general roadmap aimed at building minimalist models for generic proteins. As said, the build model represents a system in which hydrogen bonds are absent. Therefore it can be considered the first step in the physical construction of a model for ordered secondary structures. The obvious step to build physics based and accurate models for secondary structure is to add force field terms representing the hydrogen bond network. This is in fact a natural development of this work.

Other immediate development are the parameterization of specific classes of proteins, namely prolines and glycines, which display rather different conformational tendencies in the RP due to the lack of the  $C_{\beta}$ . The data have already been prepared in this work and the procedure can be immediately repeated to generate the complete model.

Besides the primary results (the model for unstructured proteins, description of the dynamics of some specific cases) this work has returned interesting insight into the whole class of Intrinsically disordered proteins. These elude one of the paradigms of the biomolecular chemistry, namely the relation between structure and function: they do not have a very well defined structure, but they do have a function. Therefore, a reliable model for this class can help redefining this paradigm, including into it dynamical information. Immediately future

development could then regard the direct application of the model to proteins at totally or partially destructured involved in pathological biochemical pathways, such as the  $\alpha$ -synuclein for the Parkinson disease. Finally, for the same reason, this model can shed light on the behavior and function of the unstructured intermediates of the folding process for structured proteins in general. Folding is probably one of the most studied process in the biochemistry, which, however, has still a large number of open issues, which could benefit of a reliable and accurate structural and dynamical model for unstructured proteins.





## APPENDICES

---

### A IDP'S FUNCTIONS

The importance of IDPs cannot be understood without the possible functions that they can perform. Five general functional classes are recognized in IDPs and each class is differentiated for the action mode [58]. For each functional class an example is cited in order to understand the wide spectrum of functions adopted by this type of protein. In *entropic chain* class the protein function stems directly from the protein structural conformations at a given energy. An example of entropic chain is the PEVK domain of the titin. Thanks to the absence of a stable tertiary structure the protein is able to be stretched [59]. In the other four classes the function arises from the *molecular recognition* (MORF). In the *display sites* class proteins bind to their partner(s) transiently in post-translational modification. SNAP-25 belongs to this class; it makes a complex with synaptobrevin and syntaxin in exocytosis of synaptic vesicles with a disorder to order transition in fact it has been observed the growth up of alpha-helical content in the complex structure [60]. The remaining three protein classes viz. *effectors*, *assemblers* and *scavengers*, contain proteins that have a permanent binding partner. The effectors bind and modify the activity of their partner protein. Inhibitors belong to this class. 4E-BP1 has a non folded structure and binds to the initiation factor eIF4E preventing that it binds with eIF4G needed for the translation process [61]. Proteins contained in the second class assemble multi-protein complexes and/or target the activity of attached domains. CITED2 protein is unstructured when free. It folds partially in an extended structure and in an  $\alpha$ -helix structure when it wraps around the TAZ1 domain of CREB-binding protein making a complex [62]. The scavengers store and/or neutralize small ligands. The main example of this class is represented by casein. Casein is an IDP that undergoes to a more folded structure when it binds to  $\text{Ca}^{++}$  [63]. The IDPs represent the actual forefront

of the protein research. The functionalities performed by these proteins are of the main interest for the understanding of the cell biology. Advances on the protein folding pathways are also obtained from the studies on the intermediate states assumed by IDPs. Another aspect of main importance for IDPs is the role that these proteins assume in the development of several neurodegenerative disease [2]. The inhibition of these process, which relays on the nature of the IDPs, is of main interest for the medicine.

## B SECONDARY STRUCTURE ASSIGNMENT ALGORITHMS

One of the target of this thesis is to identify a protein dataset in which the elements have non ordered secondary structure. Algorithms of secondary structure assignment can select the set of structures not involved in ordered secondary structures, which are extracted from the database of structures (PDB). Such algorithm considers a series of criteria which allow to be consistent with the secondary structure definition.

When the atomistic geometry of the protein is available there are algorithms able to assign the local secondary structure either analyzing the geometrical properties or estimating the hydrogen bond topology. The DSSP [16] (Define Secondary Structure of Proteins) is one the most popular algorithm. It first evaluate all the present hydrogen bond on the structure using the following expression:

$$E = q_1 q_2 \left( \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right) * f \quad (B.1)$$

where  $q_1 = 0.42e$  and  $q_2 = 0.20e$  are the partial charges of the dipole C-O and N-H ( $e$  is the electric charge of the electron). The distances are considered in Å and the dimensional factor  $f$  in chemical units is about 332 and  $E$  is in kcal/mol. The algorithm considers a large cutoff for the hydrogen bond (interaction about  $-3\text{kcal/mol}$ ) interaction with  $E < -0.5\text{kcal/mol}$ .

Based on the hydrogen bond definition are defined turns and bridges. A  $n$ -turn correspond to a single hydrogen bond between the C=O group of the residue  $i$  and the group NH of the residue  $i + n$  with  $n = 3, 4, 5$ . A bridge is composed by two non overlapped stretches of three residues each,  $i - 1, i, i + 1$  and  $j - 1, j, j + 1$ . These stretches can form either parallel or antiparallel bridge depending if the following bonds are present either  $[j - 1, i$  and  $i, j - 1]$  or  $[i - 1, j + 1$  and  $j - 1, i +$

1] for the former whereas either [i, j and j, i] or [i - 1, j + 1 and j - 1, i + 1] for the latter.

Once mapped the amino acid sequence it follows the recognition of the cooperative assignment. For a minimal helix there must be at least two consecutive n-turns. It mean that for a  $\alpha$ -helix (HB between C=O of i and NH i + 3) four residues are required (two consecutive 3-turns). The presence of one or more consecutive bridges is considered as a ladder. A sheet is defined as two or more ladders connected by shared residues.

The STRucture IDentifier [17] (STRIDE) is another important secondary structure assignment algorithm. This is a knowledge based secondary structure algorithm in which the propensity of the main secondary structure conformation is considered in support to the determination performed using the H-bond intensities.

The hydrogen bond energy is evaluated considering the empirical formula:

$$E_{hb} = E_r \times E_t \times E_p, \quad (B.2)$$

where the  $E_r$  is the radial contribution of the energy,  $E_t$  and  $E_p$  are the angular contribution. Their expression are reported as follows:

$$E_r = -\frac{2E_m r_m^8}{r^8} - \frac{E_m r_m^6}{r^6} \quad (B.3)$$

$$E_p = \cos^2 p \quad (B.4)$$

$$E_t = \begin{cases} (0.9 + 0.1 \sin(t_i)) \cos(t_o) & 0^\circ < t_i < 90^\circ \\ K_1 (K_2 - \cos^2(t_i))^3 \cos(t_o) & 90^\circ < t_i < 110^\circ \\ 0 & t_i > 110^\circ \end{cases} \quad (B.5)$$

where  $E_m = -2.8 \text{ kcal/mol}$ ,  $r_m = 3.0$ ,  $K_1 = 0.9/\cos^6(110^\circ)$ ,  $K_2 = \cos^2(110^\circ)$ . The  $t_i$  and  $t_o$  are the angular deviation of the H atom from the bisector of the lone pair within the plane of the lone pair orbital and from the plane of the lone pair orbitals respectively.

The configurational propensity is evaluated considering the distribution of the experimentally assigned secondary structures on the Ramachandran Plot.

The helices assignment considers if the intensity of the H-bond between consecutive residues, opportunely weighted with the structural tendency, fulfill the threshold. Regarding the sheet assignment, the formation of two consecutive H-bonded bridge is necessary. Also in this case the evaluation of the presence of the H-bond is weighted with the structural propensity.

The main difference of the two reported algorithms lies in the knowledge based definition of the secondary structure of the STRIDE algorithm. From this point of view it is possible to states that DSSP algorithm represents a first-principle assignment whereas the STRIDE holds a partial knowledge of the previous assignment. The preference between the assignment of the two algorithms depends on the target of the research. It has been demonstrated that STRIDE provides a better assignment of  $\beta$ -sheet than the DSSP. Regarding the helical structures in the  $\alpha$ -basin are better recognized in the DSSP algorithm although they appears to be too fragmented. In section 3.1 it has been shown that the STRIDE coil assignment is preferred over the DSSP assignment simply because it shown the higher correspondence between the algorithms regardless the bend-labeled structures.

## C DL\_POLY

The general purpose toolkit for dynamical molecular simulation DL\_POLY Classic[35] has been used and properly modified during this work.

The simulation setup is done writing the three input files: CONFIG, CONTROL and FIELD. The CONFIG file contains the specifics of the structure of the simulated cell such as vectors of the cell and periodicity directives, and contains the initial state of the simulated system. The types of amino acids contained and their positions are also defined here.

The topologies of the interactions and constrains, and all the related specifications of the force field, are provided in the FIELD input file. At the end of the file the setup of the Van der Waals interactions is reported. These interactions are considered between each couple of two non consecutive atoms at least separated by three bonds. Each interaction is defined by the two specific atom types.

The CONTROL file collects all the directives to run the simulation, such as timesteps, duration, thermostats, cut-offs, and many other specific options.

The output of the program are three files, namely OUTPUT, STATIS and HISTORY. The OUTPUT file contains all the simulation setup and the thermodynamics state of the sampled configurations. The last information is also gathered under another format by the STATIS file. All the trajectory configurations are stored in the HISTORY file.

The last two important files given in output are REVCON and REVIVE, which contains respectively the last sampled configuration and the accumulated statistical data. These files allow to continue the simulation restarting the previous simulation.

#### D TRICUBIC INTERPOLATION OF THE 3D MAPS

This section reports the detailed description of the tricubic interpolation provided by Lekien and Marsden in [64]. Given a function defined on a regular  $\mathfrak{R}^3$  gridpoint, the tricubic interpolation attempt to find the function which interpolates the point with a  $C^1$  isotropic function. Considering the regular cubic cell of

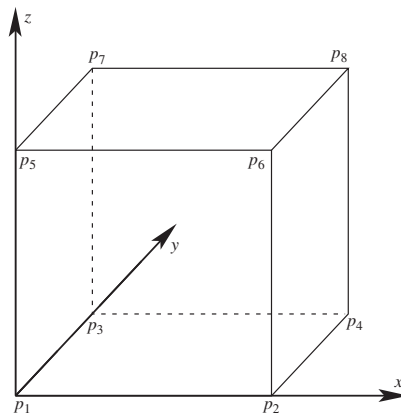


Figure D.1: Element for the interpolation [64].

unitary side, as shown in figure D.1, the function which interpolates the eight points of the cell can be defined as a stepwise function as follow:

$$f(x, y, z) = \sum_{i,j,k=1}^N \alpha_{ijk} x^i y^j z^k. \quad (D.1)$$

The  $C^1$  continuity is provided if and only if the three first-derivatives are continuous through the six faces of the reference cube. Assuming that the values of the derivatives are known at each corner, this provide an amount of 32 constraints (the value and three derivatives for each point). The number of parameters in equation D.1 scales as  $4^N$ . The minimum  $N$  needed to include the 32 constrains is 4, therefore other four constraints for each point must be defined. Higher degree derivatives are chosen in order to favor the smoothness of the function. The complete set of constraints isotropic an linearly independent is the following:

$$\left\{ f(x, y, z), \frac{\partial f(x, y, z)}{\partial x}, \frac{\partial f(x, y, z)}{\partial y}, \frac{\partial f(x, y, z)}{\partial z}, \frac{\partial^2 f(x, y, z)}{\partial x \partial y}, \frac{\partial^2 f(x, y, z)}{\partial x \partial z}, \frac{\partial^2 f(x, y, z)}{\partial y \partial z}, \frac{\partial^3 f(x, y, z)}{\partial x \partial y \partial z} \right\}. \quad (D.2)$$

Stacking the 64 coefficient of equation D.1 in the vector  $\alpha$  as:

$$\alpha_{1+i+4 \cdot j+16 \cdot k} = \alpha_{ijk}, \quad i, j, k = \{0, 1, 2, 3\}, \quad (D.3)$$

and ordering the constraints, in a vector  $\mathbf{b}$ ,

$$\begin{aligned} b_i &= f(p_i), & 1 \leq i \leq 8 & & b_i &= \frac{\partial f(p_{i-8})}{\partial x}, & 9 \leq i \leq 16; \\ b_i &= \frac{\partial f(p_{i-16})}{\partial y}, & 17 \leq i \leq 24; & & b_i &= \frac{\partial f(p_{i-24})}{\partial z}, & 25 \leq i \leq 32; \\ b_i &= \frac{\partial^2 f(p_{i-32})}{\partial x \partial y}, & 33 \leq i \leq 40; & & b_i &= \frac{\partial^2 f(p_{i-40})}{\partial x \partial z}, & 41 \leq i \leq 48; \\ b_i &= \frac{\partial^2 f(p_{i-48})}{\partial y \partial z}, & 49 \leq i \leq 56; & & b_i &= \frac{\partial^3 f(p_{i-56})}{\partial x \partial y \partial z}, & 57 \leq i \leq 64, \end{aligned} \quad (D.4)$$

it is possible to obtain the following linear system:

$$\mathbf{M}\alpha = \mathbf{b}, \quad (D.5)$$

where the  $64 \times 64$  matrix  $\mathbf{M}$  is obtained from the product of the  $i, j, k$  depending on its position. The inversion of the matrix  $\mathbf{M}$  leads to the solution of the  $a_{ijk}$  coefficients.

The input of this program is the vector  $\mathbf{b}$ , therefore the derivatives must be provided as input. In this work the derivatives have been evaluated through the finite difference method (see section 4.2). This interpolation method is extremely simple but each point evaluation requires 64 evaluations, therefore in the framework dynamical simulation it may increase the computational costs.

This program considers the input unit cell as a cube of unitary length, therefore the derivatives given in input must be multiplied for the size of the cell vector on that particular direction. The function evaluation is done through equation D.1 after the the selection of the particular space cell and the coordinates normalization to the unitary cubic cell. The function derivation is easily evaluated from the interpolating function but the result must be divided by the size of the original system cell vector related to the derivation.

## E BIAS ANALYSIS NMR\_PDB DATABASE

As mentioned in section 1.1, the protein space arrangement from the NMR spectroscopy technique is evaluated using the restrained molecular dynamics considering the geometrical constraints taken from the NMR spectra. Sometimes the restraints are applied to the most restrained conformations although no NMR restrain is experimentally observed. Such constraints on the  $\phi$  torsion angle of the proline amino acid are common inside the PDB entries. On 9352 pdb files considered, 2197 files have been considered under these constraints. The bias on the  $\phi$  variable have been evaluated considering all the proline dihedral angles related to a PDB file, which show the standard deviation below two degrees. Figure E.1 and E.2 shows the Resulting Ramachandran Plots resulting from the analysis of the whole "coil library" PDB\_NMR evaluated using the STRIDE algorithm and the same Ramachandran plots avoiding the biased structures respectively.

Each Ramachandran plot has been represented scaling the colors from 0 counts to the 40% of the maximum pixel counts in order to enhance the differences in the low region counts. The direct comparison between the P-RPs (panel (b), figs. E.1,E.1) shows the consistent suppression of the bias toward fixed values

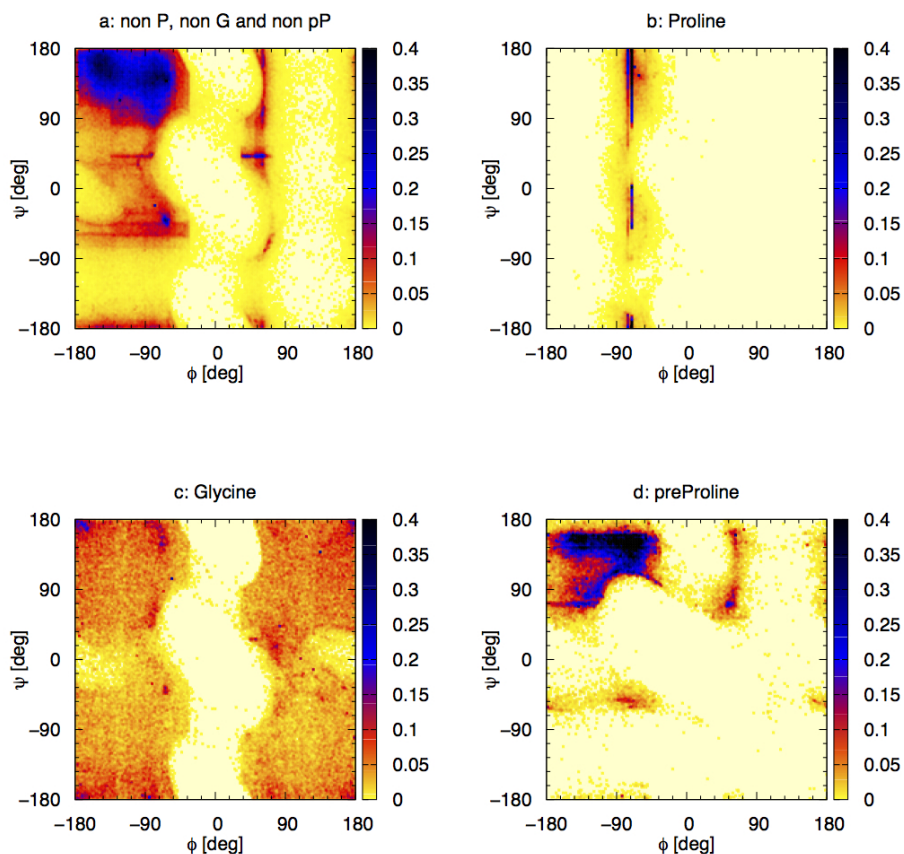


Figure E.1: Ramachandran plots of the proline (b), glycine (c), pre-proline (d), and the non proline, non glycine and non pre-proline amino acids (d) of the coiled structures selected using the STRIDE algorithm from the PDB\_NMR dataset. The color ranges in each panel have been rearranged up to 40% of the maximum in order to enhance the details.

of  $\phi \sim -75\text{deg}$  and  $-69.7\text{deg}$ . Other effects on the X-type of amino acid (a) is represented by the suppression of the horizontal stripes of figure E.1. An unexpected effect is represented by the variation of the color in the  $\beta$  and PPII basins. The presence of a bias toward specific conformation represented by black pixels in both the figures do not allow to scale the colors in the correct manner. These conformation probably are not related to the set of the proline-biased structures revealed. The color scaling can be observed spread in whole RP in the glycine-case (c). Also the pre-proline-RP (d) is affected by the proline bias.



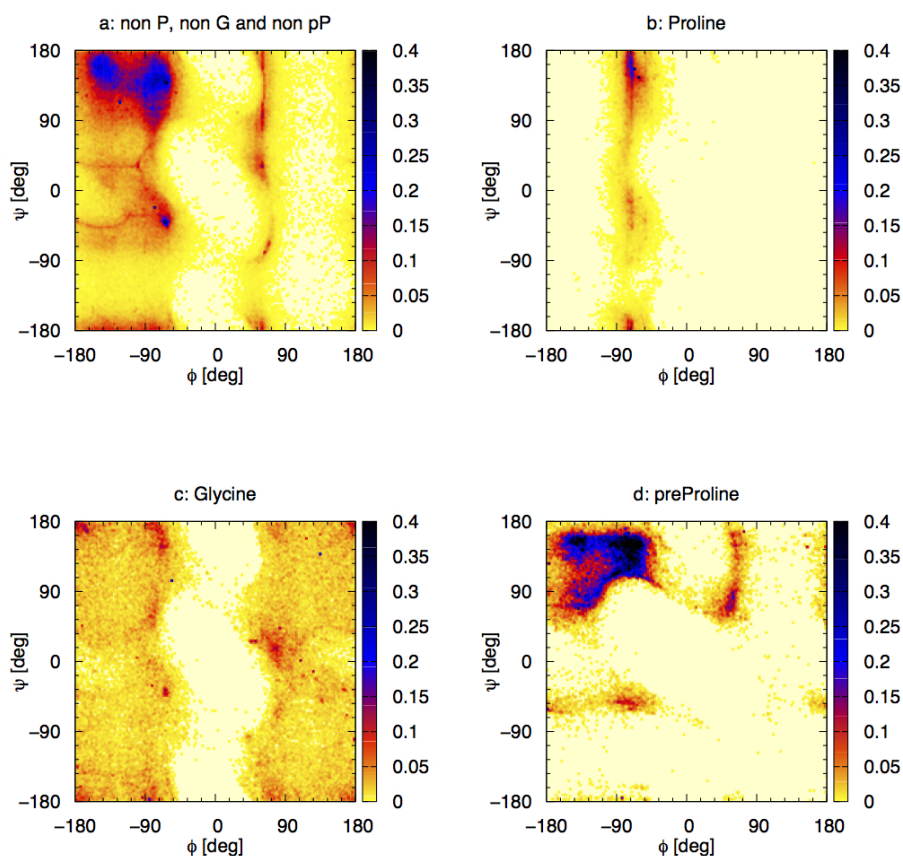


Figure E.2: Ramachandran plots of the proline (b), glycine (c), pre-proline (d), and the non proline, non glycine and non pre-proline amino acids (a) of the coiled structures selected using the STRIDE algorithm from the PDB\_NMR dataset avoiding the proline-biased structures (see the text). The color range in each panel have been rearranged up to 40% of the maximum in order to enhance the details.

Indeed all the relevant black horizontal lines present in figure E.1 are suppressed in figure E.2.

Although this simple analysis is not able to suppress all the conformational preference represented by the the isolated high-counts pixels in each RP, it allows remove to strong bias toward  $\phi \sim -75\text{deg}$  and  $-69.7\text{deg}$  observed in the coiled structures obtained from the complete PDB\_NMR dataset selected using the STRIDE algorithm.

Figure E.3 shows the set of Ramachandran plots for the PDB\_XRAY dataset. Although the RPs derived from the X-Ray dataset is not considered in the building process of the force field (see section 4.1), due to all the observations reported in chapter 1.1, it represents a reference for an unbiased system.

## F ANGLES-TORSION POTENTIAL MODULE IN DL\_POLY: ALGORITHMS AND IMPLEMENTATION

This Thesis is based on the development of a disordered proteins force field for the minimalist protein representation in molecular dynamics simulation. The force field is then implemented in the general purpose molecular dynamics software DL\_POLY Classic [35]. As shown in chapter 3, in the unstructured proteins the correct representation between  $\theta$  on  $\varphi$  internal variables assumes a crucial role. Therefore, here a three variables potential in  $(\theta_-, \varphi, \theta_+)$  is developed. Since DL\_POLY Classic does not provide any module for the implementation of this potential term [35], a proper module was implemented. In the following, the development of the FF is first described, and then the implementation is illustrated. The potential term can be safely developed thanks to the independence of the gradient of these variables. The chosen way to develop this potential term is the following: considering the ordered sequence of beads  $(a, b, c, d)$  (figure F.1) where the first term is in direction of the N-term of the fragment,  $\varphi$  is the torsion angle between the beads,  $\theta_-$  the bond angle between the beads  $a - b - c$ , whereas  $\theta_+$  is the bond angle between  $b - c - d$ . The force on each bead is:

$$\begin{aligned}
 \mathbf{F}_a &= \mathbf{F}_{\theta_-1} + \mathbf{F}_{\varphi1} \\
 \mathbf{F}_b &= \mathbf{F}_{\theta_+1} + \mathbf{F}_{\varphi2} - \mathbf{F}_{\theta_-1} - \mathbf{F}_{\theta_-3} \\
 \mathbf{F}_c &= \mathbf{F}_{\theta_-3} + \mathbf{F}_{\varphi3} - \mathbf{F}_{\theta_+1} - \mathbf{F}_{\theta_+3} \\
 \mathbf{F}_d &= \mathbf{F}_{\theta_+3} + \mathbf{F}_{\varphi4}
 \end{aligned}
 \tag{F.1}$$

where the terms  $\mathbf{F}_{\theta_i}$  and  $\mathbf{F}_{\varphi_i}$  are the forces acting on the  $i$ -th bead involved in the bond angle and diherdal angle interactions respectively. The evaluation of an analytic functional functional for the potential function remains a target for future works. In this first attempt considers a numeric functional form for  $U(\theta_-, \varphi, \theta_+)$  and in this form it was implemented in DL\_POLY Classic. This solution presents the inconvenience of being numerically heavy with respect to analytical potential.

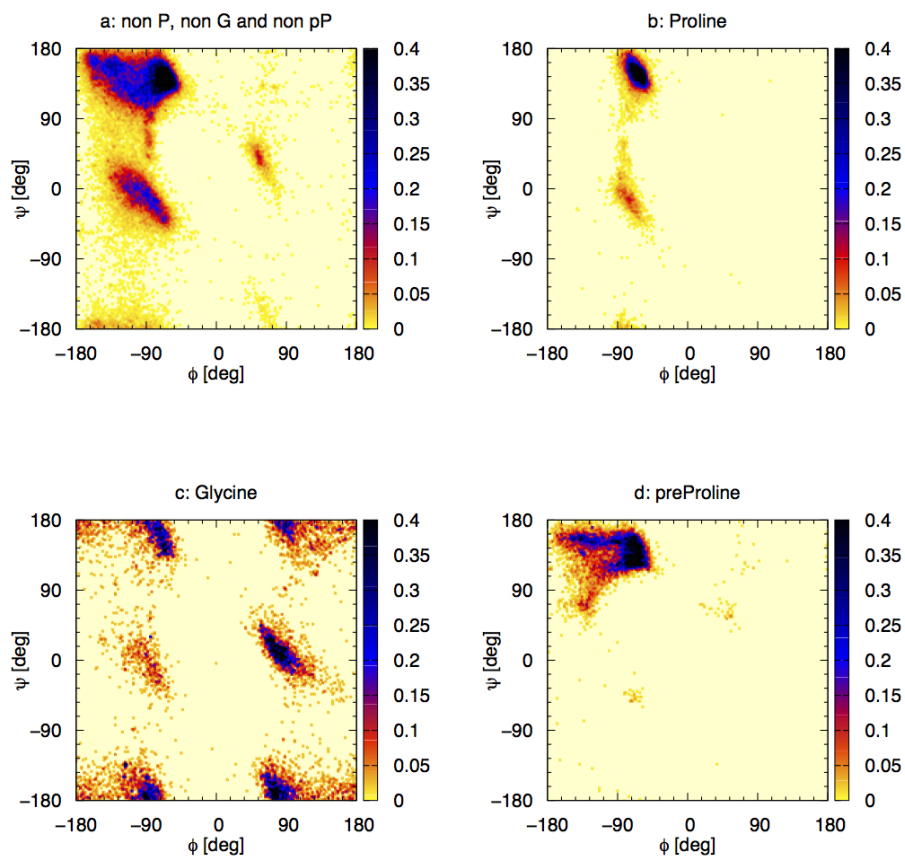


Figure E.3: Ramachandran plots of the proline (b), glycine (c), pre-proline (d), and the non proline, non glycine and non pre-proline amino acids (a) of the coiled structures selected using the STRIDE algorithm from the PDB\_XRAY. The color range in each panel have been rearranged up to 40% of the maximum in order to enhance the details.

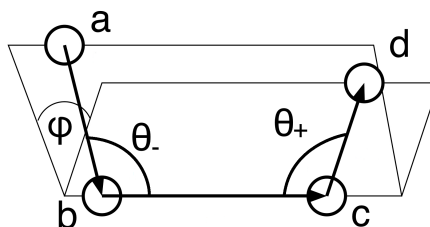


Figure F.1: Four-beads system and its variables.

The development of the interpreter for the tricubic interpolation (see appendix D) have been implemented in DL\_POLY Classic. The interpreter needs as input the output of the tricubic interpolation, which is the set of 64 values for each voxel of the grid-point. These values represent the  $a_{ijk}$  parameters of equation D.1. The whole set of parameters must be known runtime. When the set  $(\theta_-, \varphi, \theta_+)$  is known, the forces must be evaluated from the derivatives evaluated of equation D.1.

The loading in memory and evaluation of the function in a specific point is extremely time expensive with respect to the correspondent obtained with analytic expressions. The time involved represents one limit of this procedure which however for the simulation of small polypeptide represents a valid solution. In future application an equivalent solution obtained through analytic potential will be investigated.

#### G ALGORITHMS FOR THE $\theta_-, \varphi, \theta_+$ 3D CORRELATION

Because in many cases, as show in section 3.2.3, the available statistics do not allow to evaluate a statistical potential through the Boltzmann inversion method, there is the need to elude this limit in order to get a complete set of local interactions required for the completeness of the model. Therefore, two algorithms are show in this section.

The main idea of these two algorithms is to rebuild the  $(\theta_-, \varphi, \theta_+)$  densities, starting from the densities of the RP evaluated in chapter 3. Because the target structures of these algorithms are the unstructured proteins, the basic assumption here is that there is no relationship between consecutive couples of  $\phi, \psi$  required

to sample the  $\theta_-, \varphi, \theta_+$  space (see section 2.3.2). The algorithm takes advantage from the  $4 \rightarrow 3$  mapping which allows to sample conformations which are not already sampled from the database, but which may be physically consistent although no check on the energy is done.

The first algorithm is based on the equations reported in section 2.3.2, assuming moreover that the distribution of the  $\tau, \gamma_1$  and  $\gamma_2$  contained in equations 2.38 and 2.40 are those evaluated in section 3.2.2 in order to obtain a smoother representation of the data as it follows.

The algorithm is proceeds through the following these steps

1. sampling randomly the a couple  $(\phi_1, \psi_1)$  values from the cumulative distribution function of RP densities and a set of  $\gamma_{1-1}, \gamma_{2-1}$ , and  $\tau_1$  from the cumulative of the fitted distribution functions (parameters in table 2, in section 3.2.2, and equations 3.3 and 3.4);
2. sampling randomly the second couple  $(\phi_2, \psi_2)$  as in the first step, and a new set of  $\gamma_{1-2}, \gamma_{2-2}$ , and  $\tau_2$ , with  $\gamma_{1-2}$  correlated to the  $\gamma_{2-1}$ ;
3. evaluating the values of  $\theta_-, \varphi, \theta_+$  from equations 2.37 and 2.38 and store the values;
4. copy the values of the second set of variables on those related to the first step and restart from the second step until the desired number of cycles are completed.

The second algorithm adopts a different scheme. As it will be clear in the next section, the last algorithm need to be improved avoiding the samples related to the turn structures. This can be achieved using the secondary structure assignment algorithm, but the atomistic coordinates of the backbone atoms are needed.

The coordinates might be obtained involving a reconstruction algorithms, such as pulchra [50], although the knowledge on the RP variables, which is the starting point of the algorithm, may be lost. Therefore, a simple algorithm able to reconstruct the backbone atoms position has been developed. This algorithm requires as input the values of all the consecutive couples of dihedral and bond angle, therefore, requires as input the angles  $\nu$  and  $\mu$  representing  $C'NC_\alpha$  and  $C_\alpha NC'$  bond angles respectively (figure G.1). The values of the variables are measured from the PDB\_XRAY dataset are shown in table G.1.

The quite satisfactory results obtained sampling the CG-variables using the first algorithm, let to the use of a different criteria of selection of the coil structures. In

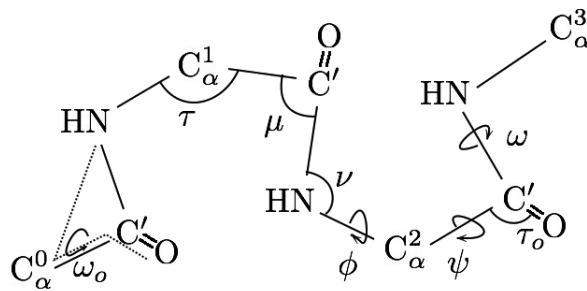


Figure G.1: Reference figure for the set of bond and torsion angles related to the second algorithm.

Variable	Value	Variable	Value
$\tau$	$111.08 \pm 2.34\text{deg}$	$d_{C'N}$	$1.33 \pm 0.01\text{\AA}$
$\nu$	$121.38 \pm 4.07\text{deg}$	$d_{NC^{\alpha}}$	$1.46 \pm 0.01\text{\AA}$
$\mu$	$116.30 \pm 4.26\text{deg}$	$d_{C^{\alpha}C^{\alpha}'}$	$1.52 \pm 0.01\text{\AA}$
$\omega$	$179.12 \pm 11.22\text{deg}$	$d_{C'O}$	$1.23 \pm 0.01\text{\AA}$
$\omega_O$	$180.18 \pm 2.22\text{deg}$	$\tau_O$	$120.54 \pm 0.98\text{deg}$

Table G.1: Geometrical parameters of the protein backbone extracted from the PDB\_XRAY dataset. All the variables are graphically represented in figure G.1.

order to be able to select the unstructured fragments using the secondary structure assignment algorithms all the coordinates of the backbone atoms must be known. The coordinates can may be obtained involving reconstruction algorithms, such as pulchra [50], although the knowledge on the RP variables, which is the starting point of the algorithm, may be lost.

A simple algorithm has been developed in order to reconstruct the backbone atoms for each value of  $\phi, \psi$  sampled. The basic module of this algorithm evaluates the position of the next atom ( $\mathbf{p}_4$ ) in the polymer sequence, starting from the knowledge of the previous three atoms in the sequence ( $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$ ), in order that the next atom assumes specific distance  $d$  from the neighbor atom, angle  $\gamma$  and torsion  $\alpha$  with respect to the preceding two and three atoms respectively. Namely this module can be identified with the linear operator  $N(d, \gamma, \alpha)$ .

$N(d, \gamma, \alpha)$  has been developed through the composition of multiple rotation operation. Starting from the condition in which  $\mathbf{p}_4$  is aligned with  $\mathbf{p}_3$  and  $\mathbf{p}_2$ . Considering a new set of coordinates ( $x', y', z'$ ) referred to the last three atoms as follow:

$$\mathbf{r}_{ij} = \mathbf{p}_j - \mathbf{p}_i, \quad (\text{G.1})$$

$$\mathbf{x}' = \frac{\mathbf{r}_{23}}{|\mathbf{r}_{23}|}, \quad (\text{G.2})$$

$$\mathbf{y}' = \frac{\mathbf{r}_{12} - \frac{\mathbf{r}_{23} \cdot \mathbf{r}_{12}}{|\mathbf{r}_{23}|} \mathbf{r}_{23}}{|\mathbf{r}_{12} - \frac{\mathbf{r}_{23} \cdot \mathbf{r}_{12}}{|\mathbf{r}_{23}|} \mathbf{r}_{23}|}, \quad (\text{G.3})$$

$$\mathbf{z}' = \frac{\mathbf{r}_{23} \times \mathbf{r}_{12}}{|\mathbf{r}_{23} \times \mathbf{r}_{12}|}. \quad (\text{G.4})$$

The unknown position of the fourth atom can be evaluated through  $\mathbf{r}_{34}$ . In the reference system with origin in  $\mathbf{p}_3$  and oriented with the Cartesian axes  $x', y'$  and  $z'$ , the initial condition  $\mathbf{r}_{34}'$  lies on the  $x'$  coordinate. Applying a rotation on the  $x', y'$  plane to  $\mathbf{r}_{34}'$  of the angle  $180\text{deg} - \gamma$ , and then applying the rotation of  $\alpha'$  angle on the  $y', z'$  plane to the result, where

$$\alpha' = \begin{cases} \alpha - 180\text{deg} & \text{if } \alpha \geq 0 \\ \alpha + 180\text{deg} & \text{if } \alpha < 0. \end{cases} \quad (\text{G.5})$$

Evaluating  $\mathbf{r}_{34}$  of the main reference system it is possible to evaluate the position of  $\mathbf{p}_4$  as follow:

$$\mathbf{p}_4 = \mathbf{p}_3 + \mathbf{r}_{34}. \quad (\text{G.6})$$

Using this algorithm it is possible to evaluate the backbone position considering sets of the following values:  $d_{NC_\alpha}$ ,  $d_{C_\alpha C'}$ ,  $d_{C'N}$ ,  $\tau$ ,  $\tau_1$ ,  $\tau_2$ ,  $\phi$ ,  $\psi$ ,  $\omega$ . This allows to evaluate the positions of the  $C'_i$ ,  $N_{i+1}$ ,  $C_{\alpha i+1}$ .

In order to assign the structure using the STRIDE algorithm the position of the oxygen atom is needed. In order to evaluate such position the  $d_{C'O}$  distance,  $C_\alpha C'O$  angle ( $\tau_O$ ) and  $N_{i+1}C_\alpha C'O$  torsion ( $\omega_O$ ) are considered.

Bearing in mind that the evaluation of the  $\theta_-$ ,  $\varphi$ ,  $\theta_+$  maps are the main objective of this analysis, in order to evaluate which variable among  $\tau$ ,  $\tau_1$  and  $\tau_2$  influences the resulting maps, a preliminary evaluation considers the deviations obtained sampling different values of the aforementioned angles. Figures G.2 shows the standard deviations of  $\theta_+$  (left column) and  $\varphi$  (right column) obtained for each couple of  $\phi_2$ ,  $\psi_2$  torsion angles (second couple of angles needed to the evaluation of the  $\theta_-$ ,  $\varphi$ ,  $\theta_+$ , see chapter 2.3) randomly sampling the  $\tau$  (first row),  $\tau_1$  (second row) and  $\tau_2$  (third row). The deviations are evaluated on the basis of 10000 samples for each  $\phi_2$ ,  $\psi_2$  couple. All the plots are represented on the same color scale. The random sampling of the angular variables is accomplished from the knowledge of the distributions of each variable. In this evaluation the known correlation between the  $\tau_{2i}$  and  $\tau_{1i+1}$  has been neglected. The largest deviation is obtained by the  $\tau$  on the  $\theta$  variable reach approximately half dimension of the pixel used in the potential function evaluation, whereas the other are lower, therefore is the only that will be considered in the following. In the evaluation of the CG correlation map these deviations are not determinant in the high counting region although allows to obtain a smoother behavior in the low counting regions. This provide the evaluation of a better defined potential function.

## H ALGORITHM FOR POTENTIAL PARAMETERS OPTIMIZATION

The parametrization of a force field requires the selection of the most representative parameters. In general this choice must be taken in view to reproduce a set of reference results. Regarding the empirical force fields, these results are represented by experimental data. Therefore, the parameter selection and optimization represents the most important step of the process for the application issues.

In this work the parameter optimization is performed with a proprietary software, named CG-autoparam. This useful tool is able to gather the optimal



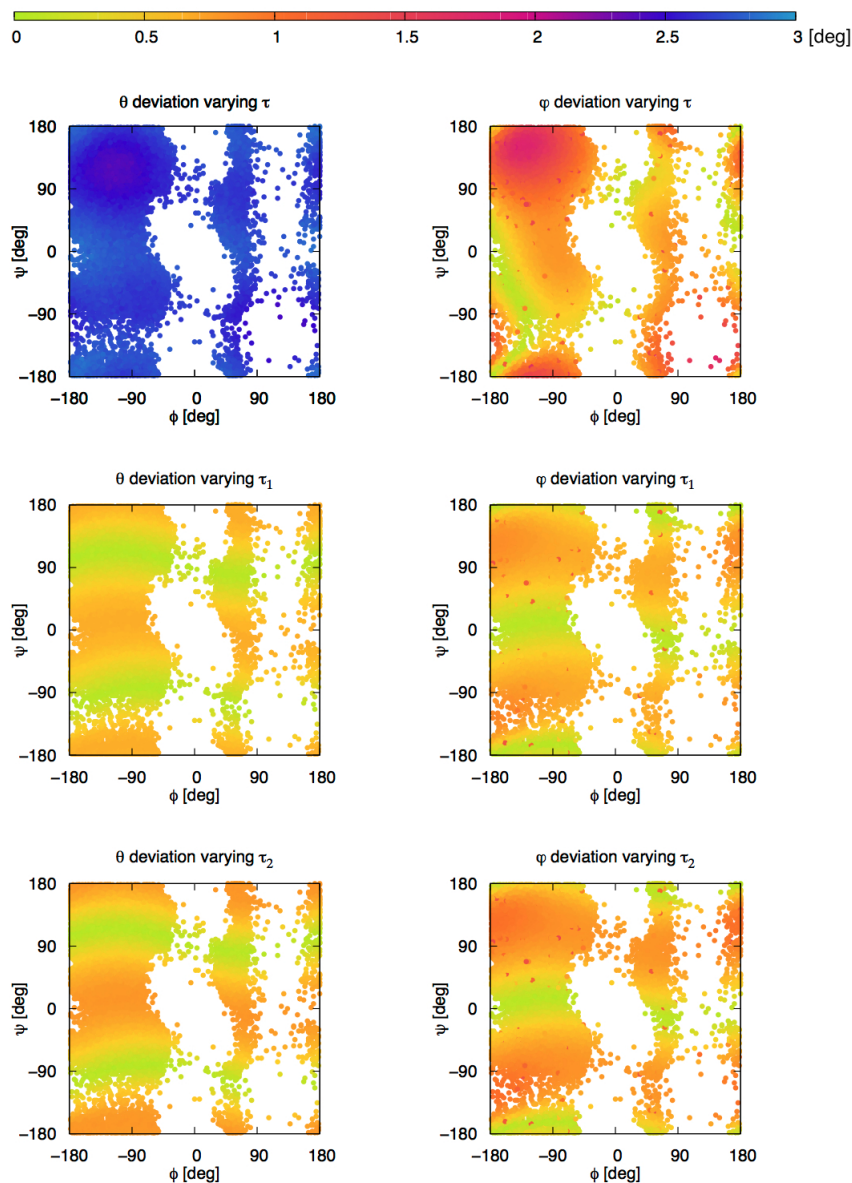


Figure G.2: Standard deviations of  $\theta$  (first column) and  $\phi$  (second column) obtained considering different values of the  $\tau$  (first row),  $\tau_1$  (second row) and  $\tau_2$  (third row) for each value of  $\phi, \psi$ . The RP variable have been sampled randomly from the RP-distribution in the case of the X amino acids (see section 1.4). The angular deviations are referred to the same color scale.

parameters using the IBI (sec. 2.2.2) and a Monte Carlo parameters sampling. Although IBI represents theoretically an efficient method to obtain fast the optimal

parameters it shows many drawbacks due to the many numerical difficulties. Monte Carlo sampling is a sufficient method to reach the optimal parameters and it does not hold problem of IBI. Theoretically it can achieve an extensive sampling of the parameters' space although the procedure is extremely time expensive.

CG-autoparam is optimized for the classical molecular dynamic program "DL\_POLY Classic" [35] (for more information about the simulation toolkit see appendix C). The program accepts the pdb structure of the simulating objects. Given the first set of potential parameters it makes the CONFIG and FIELD input files needed to DL\_POLY-run. The the CONTROL input file should be configured by the user. The FIELD file is made considering that every beads feel the same force field as needed in the  $C\alpha$ -model in this work. The CG-A input is an xml-file containing all the directives of the program.

The program contains every available potential function implemented in DL\_POLY Classic. The applied potentials should be coupled with a specific geometrical distribution.

As aforementioned this program allows to perform cycles of simulations. At the end of each cycle, the obtained distributions are compared with the reference distributions (in this work are set to the experimental results). The comparison between the reference and the simulated distribution is obtained using the Kullback-Liebler divergence [65] (KLdiv) defined for continuous variable as:

$$I(f, g) = \sum_{i=1}^k p_i \log\left(\frac{p_i}{\pi_i}\right), \quad (\text{H.1})$$

where  $f = \{p_1, \dots, p_k\}$  and  $g = \{\pi_1, \dots, \pi_k\}$  and both the distributions are positive and normalized to the unity.  $I(f, g)$  represents the information lost considering  $g$  approximating  $f^1$ . The control parameter of each cycle is the mean KLdiv (mKLdiv) of every function under optimization. The interval of interest on which evaluate the KLdiv is given as input. If the  $i$ -th value is null in one of the distributions, this addend is excluded from the evaluation of the KLdiv. The best iteration is chosen as the set of parameters that gathers the lowest value of

<sup>1</sup> The insight of this formula can be extracted observing the spitted expression:

$$I(f, g) = \sum_{i=1}^k p_i \log(p_i) - \sum_{i=1}^k p_i \log(\pi_i) = E_f(\log(f)) - E_f(\log(g)), \quad (\text{H.2})$$

where the difference of the distribution logarithm's expected values under the reference distribution is obtained.

mKLdiv. For each step the mKLdiv is evaluated and from the comparison with last accepted value, which corresponds to the last accepted set of parameters. If the last accepted KLdiv is lower than simulated KLdiv then this step is accepted, otherwise it is accepted with a probability of:

$$\exp\left(-\gamma \frac{I(f, g_{sim})}{I(f, g_{la})}\right), \quad (H.3)$$

where  $\gamma$  is a scaling factor that regulates the admitted deviations. In this way the parameters are sampled according the Metropolis algorithm. In order to reduce the deviations growing the number of iterations of the algorithm,  $\gamma$  decrease according:

$$\gamma_i = \gamma_0 \delta^i \quad (H.4)$$

where  $i$  is the iteration index,  $\gamma_0 > 0$  is the starting value of the admitted deviation and  $0 < \delta \leq 1$  is the scaling factor. At the beginning of the new iteration, the new set of parameters is sampled uniformly from predefined intervals related to each parameter. The number of iterations of this cycle is given in input.

It is also possible to check the status of distributions with or without assigned potential function term and KLdivs are computed and considered inside the evaluation of mKLdiv.

Modifications have been implemented to the version used in this work in order to minimize the work out time together with the possibility compare two dimensional distribution and to make Monte Carlo parametrizations with two dimensional potential function. These modifications allow the analysis of the  $\phi, \theta_+$  and  $\phi, \theta_-$  that are useful in this work. The comparison between the distributions it has been performed using the root mean square distance (RMSD) between them. The last modification becomes necessary when two distributions are null in different intervals because in the CG-A main release null values, from one or both the distributions, are excluded form the loss function evaluation. Another important change implemented in the code consists in a preliminary function analysis before the simulation run (see fig. H.1). This parameter selection attempts to avoid the run of simulation using sets of worthless parameters. The check of a specific potential function is performed accepting only the set of parameters for which the root mean square difference between the derivatives of the last accepted function and of the sampled function is under a fixed threshold.

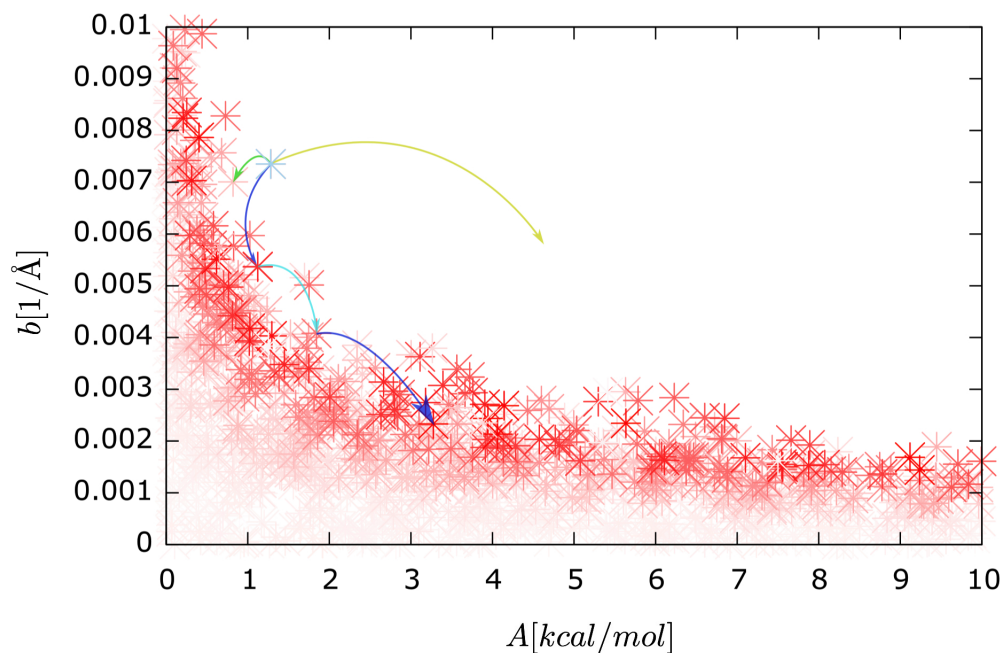


Figure H.1: Example of selective sampling and parameter space exploration (sampling of the Morse potential parameters  $A$  and  $b$ , in eq. 4.4 in section 4.1). The white field represents the regions of the rejected values. The marks are reported by scaling the color from white to red where red represents the minimum of the loss function among the obtained results. It is also represented the sampling process in CG-autoparam. Starting from the light blue mark the space is sampled. The yellow arrow is sampling a region which is avoided by the preliminary selection algorithm. The step represented by the green line is rejected because its loss function value results lower than the starting point's value whereas the blue arrow's step is accepted for the lower value of the loss function for such a configuration. The next step would put in evidence the possibility that during the sampling a loss function higher value may be accepted within the probability reported in eq. H.3.

With this method the role of the last accepted step is of paramount importance because it determines the range of allowed parameter for the next run. Therefore the last-accepted set is the only retained in memory. Modeling the allowed regions in the space of parameters avoids the extensive exploration, which is in general worthless and time expensive.

Another important improvement consists in the possibility to perform many simulations in parallel from the same starting point and then choose the set of parameters which gives the best loss function. In the unmodified version, the only way to speed up the process was to run many unrelated processes. Graphically (see figure H.2) the old process looks like to explore the space with

N-tracer whereas the new version considers many trials for each step. The exploration with n-parallel tracers could be useful in exploring extensively space with multiple minima but it depends on the selection parameter  $\chi_{\max}$  which manages the distance from the previous potential function.

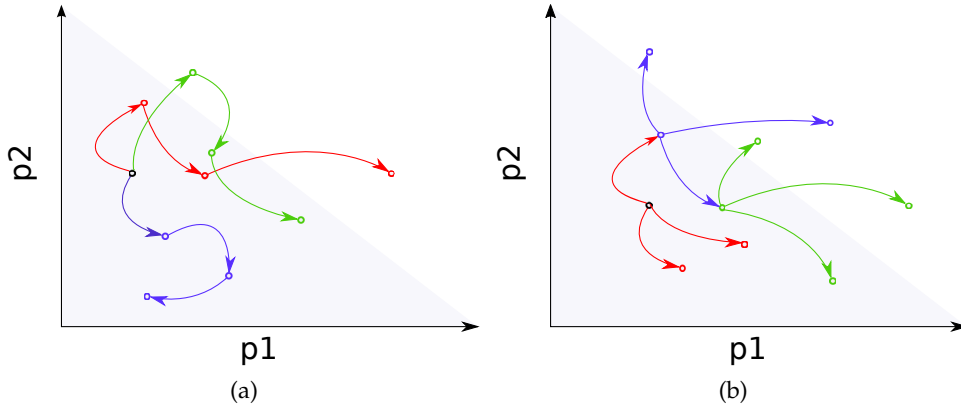


Figure H.2: This figure shows the the parallel strategies available in this work. Panel a shows the sampling process with more tracer whereas panel b shows the sampling performed with a single tracer.

As aforementioned, CG-A is a program oriented to the optimization of the parameters of an analytic function. Regarding the optimization of numerical potential, an easy achievable solution is represented by the iterative Boltzmann inversion (equation 2.29). Theoretically, after having obtained the resultant distribution of the simulation it is possible to apply the IBI, with a simple subtraction of the potential grids and thereafter to perform the parameter evaluation through the interpolation procedure. Although in the best cases it determines only the elongation of the optimization, the use of a scale factor  $\xi \in [0, 1]$ , may represent a wise choice in order to avoid the presence of deep gradient in the new potential function.

Regarding the 3D-potential  $U(\theta_-, \varphi, \theta_+)$ , the full validity of this type of development is diminished by the low number of experimental data relevant to the reference potential needed to achieve a low noise potential reference. Moreover the intended simulation should be long enough to sample in detail the 3D-distribution.



## ACKNOWLEDGEMENT

---

My sincere thanks go to Giulia Spampinato for introducing me to the many tools needed to my thesis work, to Giuseppe Maccari for the support on the code programming, to Riccardo Nifosì for tutoring me in the atomistic simulations, to Andrea Giuntoli and Vito Dario Camiola for the several valuable discussions during the period spent together.

I take this opportunity to express my gratitude to the National Enterprise for nanoScience and nanoTechnology (NEST) for allowing me to accede to all the available resources.





## BIBLIOGRAPHY

---

- [1] M. Kjaergaard, A.B. Nørholm, R. Hendus-Altenburger, S.F. Pedersen, F.M. Poulsen and B.B. Kragelund, *Temperature-dependent structural changes in intrinsically disordered proteins: Formation of  $\alpha$ -helices or loss of polyproline II?*, *Protein Sci.*, **19**, 1555-1564, (2010).
- [2] R. Schweitzer-Stenner, *Protein and peptide folding misfolding and non-folding*, Wiley publication, (2012).
- [3] C.P. Slichter, *Principles of magnetic resonance*, Springer-Verlag Berlin Heidelberg GmbH 1978, section 7.4.
- [4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, *The Protein Data Bank*, *Nucleic Acids Res.*, **28**, 235-242, (2000).
- [5] G.N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *Stereochemistry of polypeptide chain configurations*, *J. Mol. Biol.*, **7**, 95-99, (1963).
- [6] D.L. Nelson and M.M. Cox, *Lehninger-Principles of Biochemistry*, 4th ed., W. H. Freeman, (2004).
- [7] CR. Cantor and PR. Schimmel, *Biophysical chemistry, part I*, 4th ed., W. H. Freeman, San Francisco, (1980).
- [8] N. Mandel, G. Mandel, B.L. Trus, J. Rosenberg, G. Carlson and R.E. Dickerson, *Tuna cytochrome c at 2.0 resolution*, *J. Biol. Chem.*, **252**, 4619-4636, (1977).
- [9] B.K. Ho and R. Brasseur, *The Ramachandran plots of glycine and pre-proline*, *BMC Struct. Biol.*, **5**, 14, (2005).
- [10] S.C. Lovell, I.W. Davis, et al., *Structure validation by  $C\alpha$  Geometry:  $\phi$ ,  $\psi$  and  $C\beta$  Deviation*, *Proteins*, **50**, 437-450, (2003).

- [11] V. Tozzini, *Minimalist models for proteins: a comparative analysis*, Q. Rev. Biophys., **43**, 333-371, (2010).
- [12] C. Mathews, K.E. Van Holde and K.G. Ahern, *Biochemistry - 3rd edn.*, San Francisco: Addison Wesley Longman Inc, (2000).
- [13] J.S. Richardson, *Beta-Sheet Topology and Relatedness of proteins*, Nature, **268**, 495-500, (1977).
- [14] Z. Shi, R.W. Woody and N.R. Kallenbach, *Is polyproline II a major backbone conformation in unfolded proteins?*, Adv. Protein Chem., **63**, 163-240, (2002).
- [15] M.B. Swindells, M.W. MacArthur and J.M. Thornton, *Intrinsic  $\phi, \psi$  propensities of amino acids, derived from the coil regions of known structures*, Nat. Struct. Mol. Biol., **2**, 596-603, (1995).
- [16] W. Kabsch and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*, Biopolymers, **22**, 2577-2637, (1983).
- [17] D. Frishman and P. Argos, *Knowledge-based protein secondary structure assignment*, Proteins, **23**, 566-579, (1995).
- [18] M.V. Cubellis, F. Cailliez, S.C. Lovell, *Secondary structure assignment that accurately reflects physical and evolutionary characteristics*, BMC Bioinformatics, **6**, S8, (2005).
- [19] R. Srinivasan, G.D. Rose, *A physical basis for protein secondary structure*, PNAS, **96**, 14258-14263, (1999).
- [20] S.M. King and W.C. Johnson, *Assigning secondary structure from protein coordinate data*, Proteins, **35**, 313-320, (1999).
- [21] J. Buchner and T. Kiefhaber, *Protein folding handbook, volume 2, chapter 8*, Wiley VCH, (2005).
- [22] CR. Cantor and PR. Schimmel, *Biophysical chemistry, part II, chapter 8-3*, 468-472, 4th ed., W. H. Freeman, San Francisco, (1980).
- [23] P.J. Fleming and G.D. Rose, *Do all backbone polar groups in proteins form hydrogen bonds?*, Protein Sci., **14**, 1911-1917, (2005).

- [24] S. Kosol, S. Contreras-Martos, C. Cedeño and P. Tompa, *Structural characterization of Intrinsically Disordered Proteins by NMR spectroscopy*, *Molecules*, **8**, 10802-10828, (2013).
- [25] M. Sickmeier, J.A. Hamilton, T. LeGall, V. Vacic, M.S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V.N. Uversky, Z. Obradovic, A.K. Dunker, *DisProt: the Database of Disordered Proteins*, *Nucleic Acids Res.*, **35**, D786-D793, (2007).
- [26] S. Fukuchi, S. Sakamoto, Y. Nobe, S.D. Murakami, T. Amemiya, K. Hosoda, R. Koike, H. Hiroaki and M. Ota, *IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature*, *Nucleic Acids Res.*, **40**, D507-D511, (2012).
- [27] R.L. Baldwin and B.H. Zimm, *Are denatured proteins ever random coils?*, *PNAS*, **97**, 12391-12392, (2000).
- [28] A.K. Jha, A. Colubri, M.H. Zaman, S. Koide, T.R. Sosnick, and K.F. Freed, *Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restrained coil library*, *Biochemistry*, **44**, 9691-9702, (2005).
- [29] L.J. Smith, K.A. Bolin, H. Schwalbe, M.W. MacArthur, J.M. Thornton and C.M. Dobson, *Analysis of the main chain torsion angle in proteins: prediction of NMR coupling constants for native and random coil conformations*, *J. Mol. Biol.*, **255**, 494-506, (1996).
- [30] V. Tozzini, *Multiscale modeling of proteins*, *J. Accounts Chem. Res.*, **43**, 220-230, (2010).
- [31] O.M. Becker, A.D. MacKerell, Jr.B. Roux and M. Watanabe, *Computational biochemistry and biophysics*, Marcel Dekker, New York, (2001), chapter 3.
- [32] J.P. Ryckaert, G. Ciccotti and H.J.C. Berendsen, *Numerical integration of the cartesian equation of motion of a system with constraints: molecular dynamics of n-alkanes*, *J. Comp. Phys.*, **23**, 327-341, (1977).
- [33] H.C. Andersen, *Rattle: a "velocity" version of the shake algorithm for molecular dynamics calculations*, *J. Comput. Phys.*, **52**, 24-34, (1983).
- [34] M.E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*, Oxford University Press, (2010).
- [35] W. Smith, T.R. Forester and I.T. Todorov, *The DL\_POLY Classic User Manual*, STFC Daresbury Laboratory, Version 1.9, (2012).

- [36] A.D. MacKerell, D. Bashford, M. Bellott, et al., *All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins*, J. Phys. Chem. B, **102**, 3586-3616, (1998).
- [37] C. Vega, J.L.F. Abascal, M.M. Conde and J.L. Aragones, *What ice can teach us about water interactions: a critical comparison of the performance of different water models*, Faraday Discuss., **141**, 251-276, (2009)
- [38] G.W. Robinson, S.-B. Zhu, S. Singh, and M.W. Evans, *Water in biology, chemistry and physics: experimental overviews and computational methodologies*, World Scientific, Singapore, (1996).
- [39] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey and M.L. Klein, *Comparison of simple potential functions for simulating liquid water*, J. Chem. Phys., **79**, 926-935, (1983).
- [40] E. Neria, S. Fischer and M. Karplus, *Simulation of activation free energies in molecular systems*, J. Chem. Phys., **105**, 1902-1921, (1996).
- [41] H.J.C. Berendsen, D. van der Spoel and R. van Drunen, *GROMACS: A message-passing parallel molecular dynamics implementation*, Comp. Phys. Comm., **91**, 43-56, (1995).
- [42] E. Lindahl, B. Hess and D. van der Spoel, *GROMACS 3.0: A package for molecular simulation and trajectory analysis*, J. Mol. Mod., **7**, 306-317, (2001).
- [43] E.H. Yap, N.L. Fawzi and T. Head-Gordon, *A coarse grained  $\alpha$ -carbon protein model with anisotropic hydrogen-bonding*, Proteins, **70**, 626-638, (2008).
- [44] M. Levitt and A. Warshel, *Computer simulation of protein folding*, Nature, **253**, 694-698, (1975).
- [45] A.V. Rojas, A. Liwo and H.A. Scheraga, *Molecular dynamics with the United-Residue (UNRES) force field. Ab initio folding simulation of multi-chain proteins*, J. Phys. Chem. B, **111**, 293-309, (2007).
- [46] L. Monticelli, S.K. Kandasamy, X. Periole, R.G. Larson, D.P. Tieleman and S.-J. Marrink, *The MARTINI coarse-grained force field: extension to proteins*, J. Chem. Theory and Comput., **4**, 819-834, (2008).
- [47] S. Izvekov and G.A. Voth, *Multiscale coarse graining of liquid-state systems*, J. Chem. Phys., **123**, 134105-134117, (2005).

- [48] W.G. Noid, P. Liu, Y. Wang, J.W. Chu, G.S. Ayton, S. Izvekov, H.C. Andersen and G.A. Voth, *The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models*, *J. Chem. Phys.*, **128**, 244115-244134, (2008).
- [49] H. Hu, M. Elstner and J. Hermans, *Comparison of a QM/MM Force Field and Molecular Mechanics Force Fields in Simulations of Alanine and Glycine "Dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in Water in Relation to the Problem of Modeling the Unfolded Peptide Backbone in Solution*, *Proteins*, **50**, 451-463, (2003).
- [50] P. Rotkiewicz and J. Skolnick, *Fast procedure for reconstruction of full-atom protein from reduced protein representations*, *J. Comput. Chem.*, **29**, 1460-1465, (2008).
- [51] A. Korkut and W.A. Hendrickson, *A force field for virtual atom molecular mechanics of proteins*, *PNAS*, **106**, 15667-15672, (2009).
- [52] G.L.B. Spampinato, G. Maccari and V. Tozzini, *Minimalist model for the dynamics of helical polypeptides: a statistic-based parametrization*, *J. Chem. Theory and Comput.*, **10**, 3885-3895, (2014).
- [53] D. Alemani, F. Collu, M. Cascella and M. Dal Peraro, *A nonradial coarse-grained potential for proteins produces naturally stable secondary structure elements*, *J. Chem. Theory Comput.*, **6**, 315-324, (2010).
- [54] A. Ghavami, E. Van der Geissen and P.E. Onck, *Coarse-Grained potentials for local interactions in unfolded proteins*, *J. Chem. Theory Comput.*, **9**, 432-440, (2013).
- [55] K. Nishikawa, F.A. Momany and H.A. Scheraga *Low-energy structures of two dipeptides and their relationship to bend conformations*, *Macromolecules*, **7**, 797-806, (1974).
- [56] V. Tozzini, W. Rocchia and J.A. McCammon *Mapping all-atom models onto coarse-grained models: general properties and applications in minimal polypeptide model*, *J. Chem. Theory Comput.*, **2**, 667-673, (2006).
- [57] J.S. Richardson, *The anatomy and taxonomy of the protein structure*, *Adv. Protein Chem.*, **34**, 167-339, (1981).

- [58] P. Tompa, *Intrinsically unstructured proteins*, Trends Biochem. Sci., **27**, 527-533, (2002).
- [59] S. Labeit and B. Kolmerer, *Titins: Giant Proteins in Charge of Muscle Ultrastructure and Elasticity*, Science, **270**, 293-296, (1995).
- [60] D. Fasshauer, W.K. Eliason, A. T. Brünger and R. Jahn, *Identification of a Minimal Core of the Synaptic SNARE Complex Sufficient for Reversible Assembly and Disassembly*, Biochemistry-US, **37**, 10354-10362, (1998).
- [61] C.M. Fletcher and G. Wagner, *The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein*, Protein Sci., **7**, 1639-1642, (1998).
- [62] R.N. De Guzman, M.A. Martinez-Yamout, H.J. Dyson and P.E. Wright, *Interaction of the TAZ1 Domain of the CREB-Binding Protein with the Activation Domain of CITED2*, J. Biol. Chem., **279**, 3042-3049, (2004).
- [63] S. Perticaroli, J.D. Nickels, G. Ehlers, E. Mamontov and A.P. Sokolov, *Dynamics and Rigidity in an Intrinsically Disordered Protein,  $\beta$ -Casein*, J. Phys. Chem. B, **118**, 7317-7326, (2014).
- [64] F. Lekien and J. Marsden, *Tricubic interpolation in three dimensions*, Int. J. Numer. Meth. Eng., **64**, 455-471, (2005).
- [65] K.P. Burnham and D. Anderson, *Model selection and multimodel interference*, second edition, Springer, (1998).