UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Fisica, Astrofisica e Fisica Applicata
Dipartimento di Fisica

Corso di Dottorato in Fisica, Astrofisica e Fisica Applicata
Ciclo XXIX

Computational and theoretical studies of the conformational properties of peptides in their denatured state

Settore Scientifico Disciplinare FIS/03

Supervisore: Professor Guido TIANA

Coordinatore: Professor Francesco RAGUSA

Tesi di Dottorato di:
Roberto MELONI

Anno Accademico 2016/2017

Commission of the final examination:

External Member:
Alessandro Laio

External Member:
Antonio Trovato

Internal Member:
Giovanni Onida

Final examination:

Date *January 19, 2017*

Università degli Studi di Milano, Dipartimento di Fisica, Milano, Italy

Cover illustration:

Roberto Meloni

MIUR subjects:

FIS/03 - Fisica della materia

PACS:

87.15.-v - Biomolecules: structure and physical properties

Contents

List of Figures

List of Tables

Motivation

Since the development of first computers, scientists have tried to exploit their ability of performing very fast calculations, in order to simulate those systems whose dynamics, nor even the equilibrium states, could not be resolved analytically [1]. The concurrent exponential growth of computational power through the decades and the evolution of more and more efficient simulation methods allowed people, over the years, to address problems of always increasing complexity.

Biological molecules, like proteins, are a typical example of a complex system, as they are usually constituted of dozens to hundreds of thousands atoms, which interact in a non–trivial way by means of both long–range and short–range, attractive and repulsive forces [2, 3]. As a consequence of such a variety of interactions, from an energetic point of view proteins are a highly frustrated collection of atoms, whose equilibrium and dynamical properties need to be studied using those simulation techniques mentioned before [4, 5, 6, 7].

One of these techniques is classical Molecular Dynamics where, for each atom of a system, the trajectory in space is computed as a function of the time, via the resolution of Newton's equations of motion. Unless otherwise specified, Molecular Dynamics will be the exploited technique throughout all the following chapters.

In the field of the physics of biomolecules, an open question concerns the characterization of the conformational properties of the denatured state of proteins, namely the collection of all the disordered phases populated by the polipeptides during their thermal motion inside the cell [8]. Indeed, the study of the denatured state is an extremely important task, as it is critical for determining the folding kinetics of proteins and their thermodynamic stability [9], their ability to cross lipid bilayers or their turnover in the cell, [10] or even their topology when proteins are finally folded into the native state, after being synthetized by the ribosomes inside the cell [11, 12, 13, 14, 15].

In laboratory, the characterization of the denatured state is far from being straightforward. Actually, under biological conditions – namely at $T = 300\,\mathrm{K}$, in water – the denatured state is metastable, due to its exceedingly short half–life (which typically ranges from milliseconds to some seconds for most proteins). This feature severely limits the feasibility of most used experimental techniques, such as Nuclear Magnetic Resonance (NMR), X–Ray Crystallography, Fluorescence Resonance Energy Transfer (FRET) or Small–Angle X–Ray Scattering (SAXS), which do not possess a sufficiently high time–resolution to "take a photograph" of the proteins when they still are in their denatured phase.

A typical approach is then to stabilize the denatured state, raising the temperature, lowering the pH of the solution or adding to the solute some chemical agents – such as urea or guanidine chloride – whose net effect is the conversion of the denatured, metastable state into the equilibrium one via a partial or total destruction of native contacts [13, 16].

Interestingly, despite their daily use in the laboratories all over the world, their effect at a molecular level is not completely clear. For example, it is yet unclear whether they interact directly with the atoms of the protein [17, 18, 19], or if the denaturants affect the solvent, perturbing the hydrophobic effective force, which plays an important role in the stabilization of the proteins [20, 21, 22, 23, 24, 25].

Moved by the motivation mentioned before, in the Chapter 1 of the present thesis we address this open question by means of Molecular Dynamics simulations, since the ability of the technique to characterize systems at atomic level is exactly what is required to (or at least try to) give an overall picture of the molecular mechanisms behind the chemically–induced denaturation of proteins.

In doing so, we had to foresee a twofold problem: the necessity of a thorough exploration of the phase space and the need for a fast overcoming of energetic barriers. Indeed, if equilibrium properties are investigated, the simulation needs to be at convergence, namely it should be "long enough" to allow the system to sample a statistically–relevant portion of its phase space, in order to extract reliable equilibrium information from the simulated trajectory. This is a dramatic requirement when it comes to the study of the denatured state, which has an extremely high entropy. Moreover, when performing Molecular Dynamics of biomolecules, a typical issue one has to face is the presence of an highly–frustrated energy landscape, which shows local minima separated by high energy barriers at any scale. This feature has the effect of slowing–down the sampling of the phase space, since the simulated system has a finite probability of remaining stuck in a local energy minimum for a non–negligible amount of simulation time. A plethora of enhanced–sampling methods has been developed through the decades to overcome one of or both these two problems: Umbrella Sampling [26], Simulated Annealing [27], Replica– or Bias–Exchange [28, 29], Metadynamics [30] are a not–exhaustive list of techniques devoted to a faster exploration of the phase space. Among these, Metadynamics seemed to be the most suitable for our purposes and we chose to apply it on our simulations, in combination with a Bias–Exchange approach.

A critical choice for the implementation of a Metadynamics simulation is the selection of the Collective Variable against which the dynamics is to be biased, namely that low–dimensional function – which we will generically refer to as Y – of the high–dimensional microscopic coordinates $\mathbf{r}$ of the system: $Y = Y(\mathbf{r})$. Compared to the *in vitro* or the *in vivo* framework, it is not always straightforward how to make such a choice in *in silico* experiments. Indeed, in the design of experiments on proteins, the selection of Y is essentially determined by the technique itself: the fluorescence intensity of tryptophanes in FRET, the nuclear chemical shifts in NMR spectra, the intensity of scattered X–rays in SAXS or the ellipticity in Circular Dichroism (CD) are examples of low–dimensional reductions "naturally" performed in laboratories. On the other side, when one performs a simulation, he or she can choose to compute and analyse – or even bias, as in the case of Metadynamics or of other enhanced–sampling techniques – practically any function of the microscopic coordinates $\mathbf{r}$. In the study of the equilibrium properties of a system, often the only feature required from a CV is the ability to identify the relevant states of the system. For instance, in the case of protein folding, a "good" Collective Variable Y should assume different values when the protein is in its native state, in the denatured state and, when relevant, in intermediate states. Collective Variables like the fraction of native contacts q or the root mean square deviation ($RMSD$) of the atomic positions with respect to those of the native conformation are usually able to perform such a discrimination [31]. When one needs to describe the time–dependent, dynamical properties of a system, on the other side, the choice is more troublesome [32, 33]. A common assumption is that the Collective Variable Y follows an overdamped Langevin of kind

$$\frac{dY}{dt} = D^{(1)}(Y) + \sqrt{2D^{(2)}(Y)} \cdot \eta_t \tag{1}$$

where η_t is a stochastic, gaussian–distributed noise with moments $\overline{\eta_t} = 0$ and $\overline{\eta_t \eta_{t'}} = \delta(t-t')$ and where $D^{(1)}(Y)$ and $D^{(2)}(Y)$ are called "drift" and "diffusion" coefficients, respectively. The former, $D^{(1)}(Y)$, can be related to the gradient of the free energy $F(Y)$ plus a correction, depending on $D^{(2)}(Y)$, which is needed to allow the probability $p(Y)$ to evolve in time towards the Boltzmann distribution [34]. On the other side, $D^{(2)}(Y)$ is a position–dependent coefficient which controls the diffusion of Y within its low–dimensional space [35]. The knowledge of the effective $D^{(1)}(Y)$ and $D^{(2)}(Y)$ of a Collective Variable Y would be then very rich of information on the system and would provide a valuable tool for analyses and predictions, for example to apply Arrhenius equation to estimate reaction rates or to model the dynamics as transitions between discrete states.

Formally, an equation similar to Eq. (1) can always be written for any Collective Variable, although a bad choice of Y results in functions $D^{(1)}(Y)$ and $D^{(2)}(Y)$ depending not only on Y, but on the whole history of the system [35]. In general, the evaluation of $D^{(1)}(Y)$ and $D^{(2)}(Y)$ to discern whether Eq. (1) holds or not for a chosen Collective Variable Y is not an easy task. Indeed, several works tried to estimate $D^{(1)}(Y)$ and $D^{(2)}(Y)$ for generic time series characterized by an uncontrollable sampling rate, by correction terms [36, 37], by iterative procedures [38] or by evaluating the adjoint Fokker–Planck operator [39, 40]. However, these methods cannot be applied in the case of Molecular Dynamics simulations, where the minimum time period can be small as an integration timestep, which is smaller than any other process involved in the microscopic dynamics.

Assuming to know the Collective Variable Y, an efficient algorithm for the back–calculation of drift and diffusion coefficients from Molecular Dynamics simulations was developed using a Bayesian approach [41, 42] and then applied to protein folding [43, 44]. Using a maximum–likelihood principle [45], the drift and diffusion coefficients could be obtained as average of Molecular Dynamics trajectories, and a criterion for the choice of the sampling rate of the trajectories was introduced to minimize time correlations of noise.

In the Chapter 2 of the present thesis, we investigate the validity of the framework defined by Eq. (1) and, in particular, whether it is possible to define the drift and diffusion coefficients $D^{(1)}$ and $D^{(2)}$ as functions only of Y. In other words, we studied the legitimacy of the hypothesis at the basis of refs. [41, 42, 43, 44, 45]. Since we are interested in facing the problem from a computational perspective, we did not study directly the validity of true Langevin equation (1), but its finite–differences counterpart, defined within the scheme of a standard integrator at finite time step Δt. In fact, it is the finite–difference dynamic equation what one usually calculates in Molecular Dynamics simulations.

CHAPTER 1

Molecular dynamics simulations of peptides in water and in a denaturant solution

1.1 Introduction

The study of the disordered phases of proteins and peptides is an important, although complicated, task. The denatured state of structured proteins is critical for determining their folding kinetics and thermodynamic stability, their ability to cross lipid bilayers, and their turnover in the cell [10]. In the case of intrinsically disordered proteins, disordered states are directly involved in biological function [46]. Fluorescence and Circular–Dichroism spectroscopy provide coarse information about non-native states, whereas NMR techniques can refine it to the amino-acid length-scale. However, for structured proteins the conformational characterization of the denatured state requires its stabilizations, typically with denaturants like urea or guanidine chloride [9] (GndCl). The natural question one is pushed to ask is then what is the effect of these denaturants on the thermodynamic and structural properties of the polypeptidic chain. In particular, one is usually interested in the properties of the (metastable) denatured state in water, that is under chemical conditions that are more similar to the biological ones. Thus, studying the effect of chemical denaturants can be relevant for interpreting the results of experiments conducted in urea and GndCl, in order to extrapolate information on the biological denatured state.

The mechanism that allows urea and GndCl to stabilize the denatured state of proteins has been discussed for forty years. The particularly low viscosity of urea solutions raised the suggestion that it affects the hydrogen bonding of the water, decreasing the effective hydrophobic interaction which stabilizes proteins [47]. Although this could be the case, calorimetric experiments suggest that the main factor which destabilizes the native state of proteins ia a direct interaction with the denaturant molecules [48]. Also molecular dynamics simulation point towards a direct interaction of chemical denaturants with the protein backbone [17, 22, 19]. Hydrogen-exchange experiments indicate that urea can interact with the protein building hydrogen bonds mainly with its backbone, but no hydrogen bonds are detected in the case of GndCl [18]. Thus, urea and GndCl seem to act according to different mechanisms. This difference has consequences on the kinetics of protein chains. The different viscosity and association propensity to a poly-dipeptide was shown to be the cause of the different rate constants of the end-to-end diffusion in urea and GndCl [21]. Unfolding simulations of protein L in these two denaturants highlighted a different order in the disruption of its secondary structure elements [49].

Different denaturants are then expected to have different effect in determining the non-native states of proteins. This is apparent in the case of GB1, one of the most widely characterized proteins with biochemical techniques. GB1 follows a two state behavior in urea [50, 51] but it displays an intermediate in GndCl [52, 53]. From the structural point of view, GB1 shows essentially no residual secondary structure in 7.4M urea [50], while in GndCl its second hairpin has some residual structure [52]. The scenario is still different if GB1 is denatured by mutating an amino acid and lowering the pH, thus under conditions expectedly closer to the biological (metastable) denatured state.

In order to investigate the molecular effect of the chemical denaturant on proteins, we carried out molecular dynamics simulations of the helical segment and of the second hairpin of GB1 in urea and GndCl at equilibrium, and we compared them with simulations conducted in water.

The same fragments of GB1 were characterized experimentally by CD and NMR. Fragment 41-56, corresponding to the second hairpin was shown to be structured in water [54]. Upon addition of 6M urea, it still retains 40% native population [55]. The fragment 21-40, corresponding to the central helix of the protein, is mainly unstructured in water, but its CD spectrum further shifts towards random-coil values if 6M urea is added. Nuclear Overhauser Effect signals indicate that in water its N-terminal region populates the beta region of dihedral space, while the C-terminal is in the alpha region [55]. The residual helical population was estimated from its ellipticity is 9% [56].

A large number of simulations were described in the literature to investigate the equilibrium properties of the fragment corresponding to the second hairpin of GB1 in water [57, 58, 59, 60], to the extent that it has become the sand box to test routinely new algorithms. According to all these calculation, this fragment displays a clean two-state behavior in water. The equilibrium sampling of the fragment corresponding to the alpha-helix of GB1 in water highlights a more complicated free-energy landscape [49], in which the metastable alpha-helix competes not only with a random-coil state, but also with other types of helices and with a hairpin state.

The comparison of the free-energy landscapes of the hairpin fragment of GB1 in urea, GndCl and water was reported on the basis of Hamiltonian-exchange simulations [61]. According to these calculations, urea disrupts completely the native region and stabilizes a state that resembles a random coil, while guanidine chloride has a milder effect, maintaining the structure it has in simulations in water. A random-coil behavior in urea was also found in parallel-tempering metadynamics simulations [19]. In the present work, we simulated the fragment corresponding to the helix of GB1 in water, urea and GndCl, comparing the associated free-energy landscapes, and we studied the interaction between the solvent and the peptide. Moreover, we extended our previous calculations concerning the fragment corresponding to the second hairpin fragment of GB1 [61] to study also in this case the interactions between the peptide and the denaturant.

1.2 Model & algorithms

The segment $22 - 38$ of protein–G B1 domain (pdb code 1PGB) was modeled with the Amber99 potential, as modified in ref. [62]. The parameters for urea were those of Amber99, while those for GndCl were those developed in ref. [49]. The model for segment 41-56 is identical to that reported previously [61]. The simulations were carried out with the bias-exchange metadynamics algorithm [29], implemented in Plumed 2 [63] for Gromacs 5.0 [64]. A total of five different environmental conditions was studied for the α–helix:

- pure water;
- water + urea (2M);
- water + urea (5M);
- water + GndCl (2M);
- water + GndCl (4M)

whereas the β–hairpin was instead simulated in three cases:

- pure water;
- water + urea (5M);
- water + GndCl (4M)

Initially, both the α–helix and the β–hairpin were unfolded at $T = 800\,\mathrm{K}$, in aqueous environment. The unfolded structures were then inserted in a dodecahedric box and solvated at first with the appropriate number of molecules of denaturant, then with Tip3p water molecules to fill the remaining empty volume. Ions

System	Volume	# mol. den.	# mol. H_2O	# ions	den. conc.
(α)–water	$75\,\mathrm{nm}^3$	-	2380	1 Na^+	-
(α)–urea (2M)	$85\,\mathrm{nm}^3$	98 urea	2375	1 Na^+	1.92 M
(α)–urea (5M)	$85\,\mathrm{nm}^3$	245 urea	1948	1 Na^+	4.78 M
(α)–gnd (2M)	$85\,\mathrm{nm}^3$	98 Gnd^+	2228	1 Na^+, 98 Cl^-	1.92 M
(α)–gnd (4M)	$85\,\mathrm{nm}^3$	196 Gnd^+	1875	1 Na^+, 196 Cl^-	3.83 M
(β)–water	$88\,\mathrm{nm}^3$	-	2774	3 Na^+	-
(β)–urea (5M)	$88\,\mathrm{nm}^3$	265 urea	1946	3 Na^+	5 M
(β)–gnd (4M)	$88\,\mathrm{nm}^3$	200 Gnd^+	1862	3 Na^+, 200 Cl^-	3.77 M

Table 1.1: Number of molecules for the simulations of the α–helix and the β–hairpin.

$Na^+$and Cl^-were finally added to ensure charge neutrality; all the parameters concerning the number of molecules are listed in Tab. 1.1. After a $5\,\mathrm{ns}$ equilibration run in NVT regime, simulations were carried out at $T = 300\,\mathrm{K}$, coupled with a v-rescale thermostat ($\tau = 0.1\,\mathrm{ps}$); electrostatic interaction was evaluated with PME algorithm; an integration timestep $\Delta t = 2\,\mathrm{fs}$ was used, keeping the bond lengths constant via LINCS algorithm. Each of the eight systems was structured in a five-replicas, bias-exchange well-tempered metadynamics scheme; the Collective Variables biased were the following

r.0) degree of helicity R_α , described in [65];

r.1) degree of β content R_β , *ibid.*;

r.2) radius of gyration R_g ;

r.3) end–to–end distance d_{ee} ;

r.4) unbiased

Exchanges were attempted every $50\,\mathrm{ps}$; the hills used had height=$0.3\,\mathrm{kJ\,mol^{-1}}$ and width=0.2 c.v. units, deposited every 500 timesteps with a biasfactor=10. Each replica was run for $2\,\mu\mathrm{s}$. Subsequent analyses were performed using Gromacs and Plumed 2 tools, METAGUI [66], APBS [67] and in-house software; images and graphs were created with VMD [68], tachyon [69], gnuplot [70] and xmgrace [71].

The CD spectrum is predicted as a linear combination of the standard spectra [72], weighted by the probabilities of α, β and coil structures calculated with STRIDE [73] on the unbiased replica. The chemical shift are calculated using SPARTA [74] on the nine conformations displayed in Fig. 1.3 and weighted by their Boltzmann factor evaluated in terms of R_α , R_β and R_g .

1.3 Results

1.3.1 The effect of urea and GndCl on the states of fragment 22-38

We carried out the five simulations of the fragment $22-38$, corresponding to the helix of GB1, in water, 2M urea, 5M urea, 2M GndCl and 4M GndCl until the convergence of free energy profiles was reached. To check it, we adopted the following criterion: if a free energy profile did not significantly change between the windows $t' = 1.5\,\mu\text{s}$ and $t = 2\,\mu\text{s}$, at least in the low–energy regions of the Collective Variable (which are supposedly those subject to bigger variations in long runs, when further minima are discovered), then the profile was considered at convergence. In Fig. 1.1 we reported the difference in free energy between the profiles drawn at t and t', as a function of the free energy itself. Indeed, the low–energy regions are subject to a variation in free energy $\Delta F \leq 1\,\mathrm{k_B T}$ for $F < 5\,\mathrm{kJ\,mol^{-1}}$ and, anyhow, $\Delta F \leq 2\,\mathrm{k_B T}$ for $F < 10\,\mathrm{kJ\,mol^{-1}}$, which we find acceptable for the study of a state characterized by metastability. Two qualitative examples of the variation of profiles are displayed in Fig. 1.2, where $F(R_\alpha)$ and $F(R_g)$ for the simulation in water are shown at times t' and t.

The free energy profile of this fragment in water, as a function of the helicity R_α and of the gyration radius R_g is displayed in the left panel of Fig. 1.3. Its features are in agreement with those calculated previously in [49], namely a metastable helical state which is $\approx 5\,\mathrm{k_B T}$ above the random coil, and a variety of local minima corresponding to different degree of formation of the helix. Interestingly, it includes a partially-formed β–hairpin state, whose existence is highlighted by the presence of a local minimum at $R_\beta \approx 0.5$ in the free energy profile as function of R_β as showed in the Fig. 1.4. In the free-energy profile we identified nine local minima separated by at least $kT = 2.5\,\mathrm{kJ\,mol^{-1}}$ from the surrounding. The minima are marked with crosses, and some of the associated conformations are plotted in the figure. The peptide does not exhibit a clear two-state behavior, as expected from the Zimm–Bragg theory [75]. Besides the fully formed helix (labelled as H1), there are other two states in the region $R_\alpha > 0.5$ in which only the C-terminus is helical. In the other half of the plot one can identify at least five disordered states with varying gyration radius, and the β–hairpin.

In the right panels of Fig. 1.3 we show how the free energy landscape are changed upon addition of 5M urea and 4M GndCl, respectively (landscapes in milder denaturing conditions are displayed in Fig. 1.5 . Urea has the strongest effect on the helical state; the fully formed helix (H1) is lost, while the free energy of states H2 and H3 is raised of $12\,\mathrm{k_B T}$ and $5\,\mathrm{k_B T}$, respectively. Also the free energy of coil states at intermediate values of R_α is raised of several $k_B T$. The global minimum is squeezed towards the value $R_\alpha \to 0$, and it is rather flat at values of R_g between 0.6 and $1.4\,\mathrm{nm}$. Also the hairpin state B1 essentially disappears, as shown in Fig. 1.5 A. On the other side, guanidine chloride has a smaller effect on the peptide. The free energy of the helical state H1 is raised by $5\,\mathrm{k_B T}$, and the rest of the landscape at $R_\alpha > 0$ is raised by few $k_B T$, maintaining a pattern of local minima similar to that in water. Also the β–hairpin state B1 is raised by some kT but it is still detectable, as can be seen in Fig. 1.5 B.

1.3.2 Experimental observables associated with fragment 22-38

From the ensemble of conformations generated by the simulation one can calculate some macroscopic observables and compare them with those measured in experiments. Moreover, one can use these data to evaluate if their interpretation according to the standard tools is compatible with the underlying conformational properties of the peptide, which are known. In Fig. 1.6 A we reported the CD spectrum of the peptide in the five simulated solutions. Similarly to the experimental findings reported in [56], the curves recorded in water and at high urea concentration (6M in the experiment, 5M in the simulations) are similar, the latter displaying an upward shift in the region around $220\,\mathrm{nm}$. Overall, the curve in water is very similar to that in 2M GndCl; that in 4M GndCl is similar to that in 2M urea, displaying a more pronounced minimum at $195\,\mathrm{nm}$, and this minimum decreases even more at 5M urea.

The secondary-structure profiles calculated from the same ensemble of conformations used to predict the CD spectra are displayed in Fig. 1.8. Under all conditions the helicity is concentrated in the C-terminal half of the peptide, in agreement with the corresponding sequence-based propensities [56]. The overall helicity of the peptide is comparable in water and in 2M GndCl, and decreases moving to 2M urea, 4M GndCl and reached its minimum at 5M urea. On the other hand, a residual β–hairpin is apparent in water, but diminishes in all the other denaturants.

Analysis of CD spectra associated with residual secondary structure is always cumbersome. The complexity of the conformational space of the peptide makes its interpretation even worse. All the curves are dominated by the coil component. The similarity between the curves in water and 2M urea and between those in 2M urea and 4M GndCl actually derives from different combination of α and β components. Indeed the de-convolution of the predicted CD spectra with different standard tools gives different secondary–structure propensities, as summarized in Tab. 1.2.

Also the secondary chemical shifts predicted from the simulation and displayed

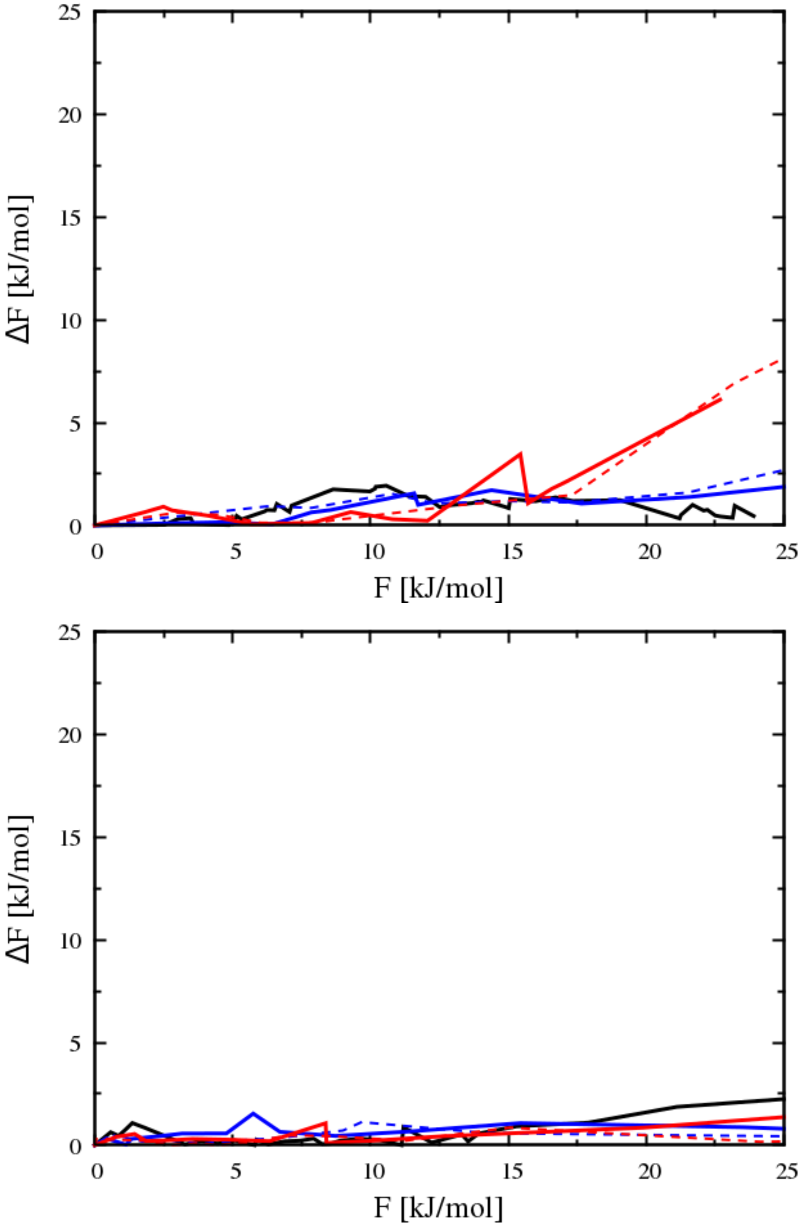


Figure 1.1: The convergence of simulations is showed by representing the difference in free energy between the landscapes $F(R_\alpha)$ and $F(R_g)$ (calculated at $t = 2\,\mu$s and $t' = 1.5\,\mu$s), as functions of F, for $R_\alpha$on the top panel and for $R_g$on the bottom panel. In both panels, the five simulations are displayed: water (black), urea 2M (solid blue), urea 5M (dashed blue), gnd 2M (solid red) and gnd 4M (dashed red).

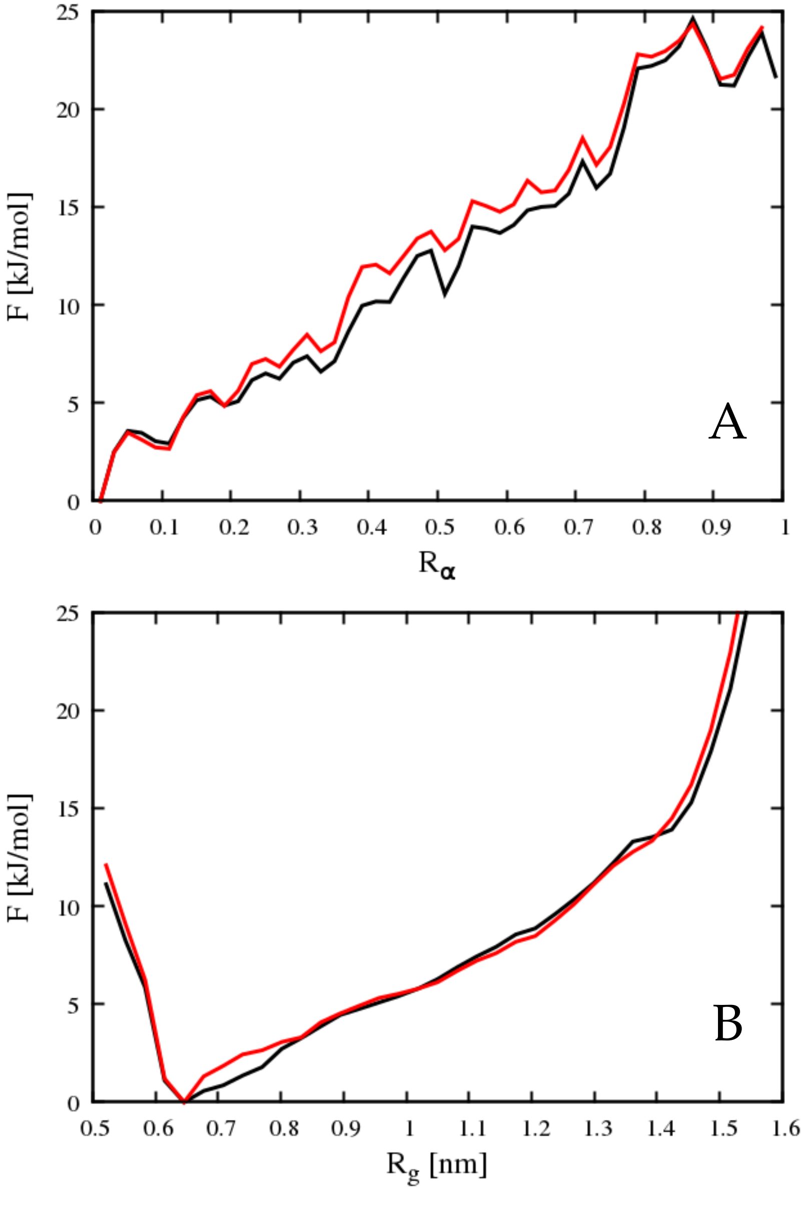


Figure 1.2: The monodimensional profiles of free energies $F(R_\alpha)$ (top panel) and $F(R_g)$ (bottom panel) are showed for the α–helix in water. In both panels, the profiles in red are calculated at $t = 1.5\,\mu$s, whereas those in black at $t = 2\,\mu$s.

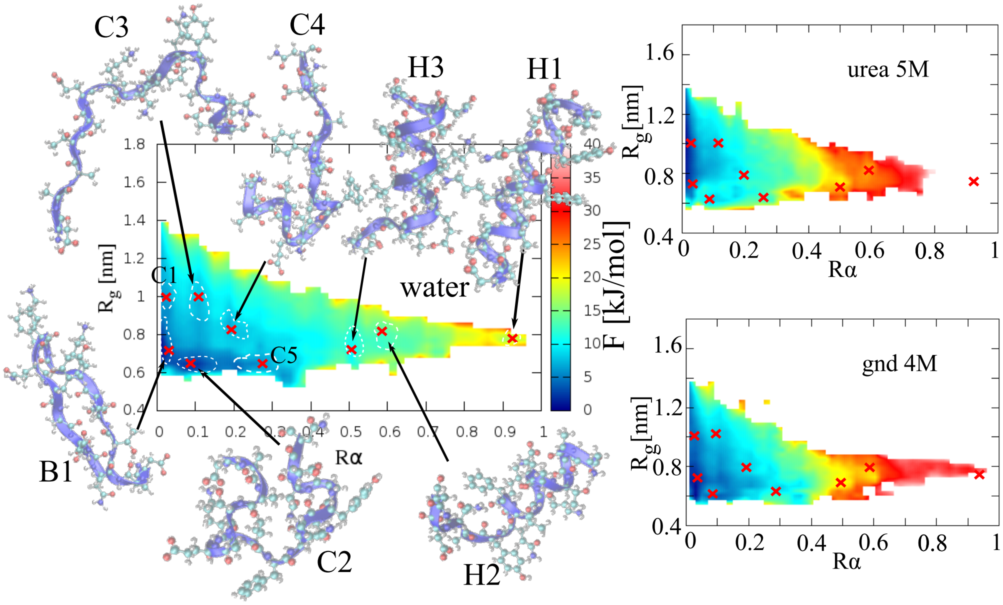


Figure 1.3: The free energy of the helix in water, 5M urea and 4M GndCl, as a function of the degree $R_\alpha$of helix formation and of the gyration radius R_g. Different states identified for the peptide in water are marked with a cross in the plot; they are labelled with H1, H2, H3 (different degree of formation of the helix), C1, C2, C3, C4, C5 (coils) and B1 (hairpin), and their mean structure is shown. The crosses are reported also in the free-energy plots of the urea and GndCl simulations, for comparison. The dashed lines indicate isoenergetic curves at $2.5\,\mathrm{kJ\,mol^{-1}}$ above each minimum.

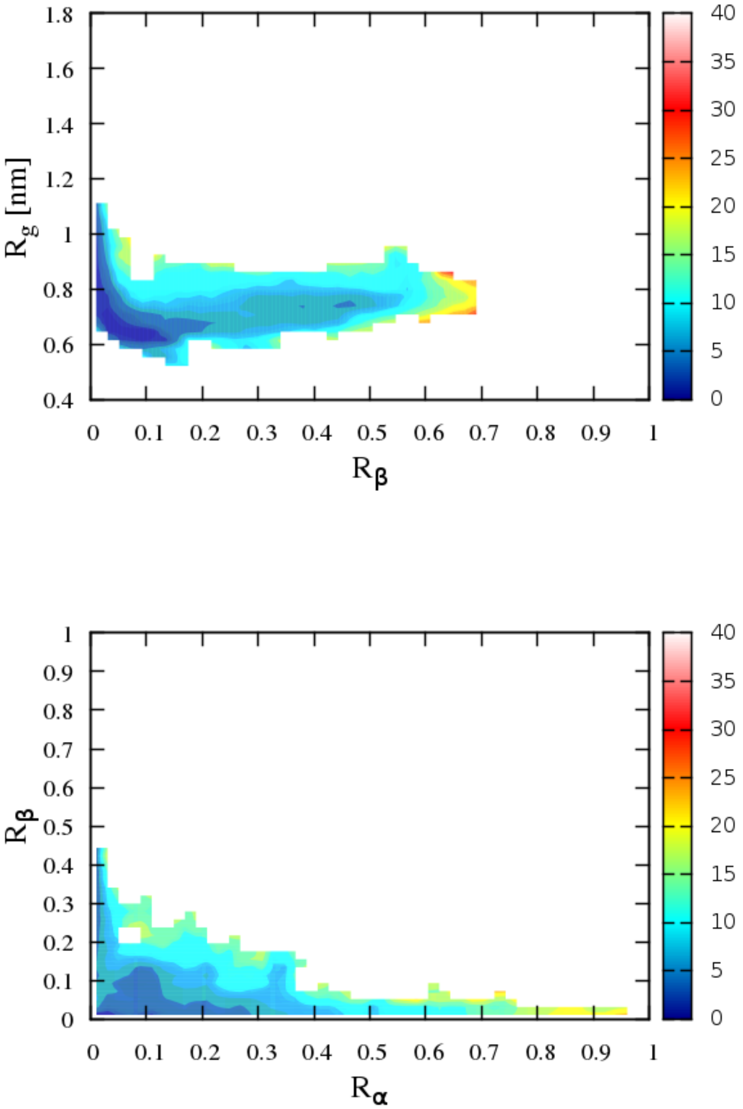
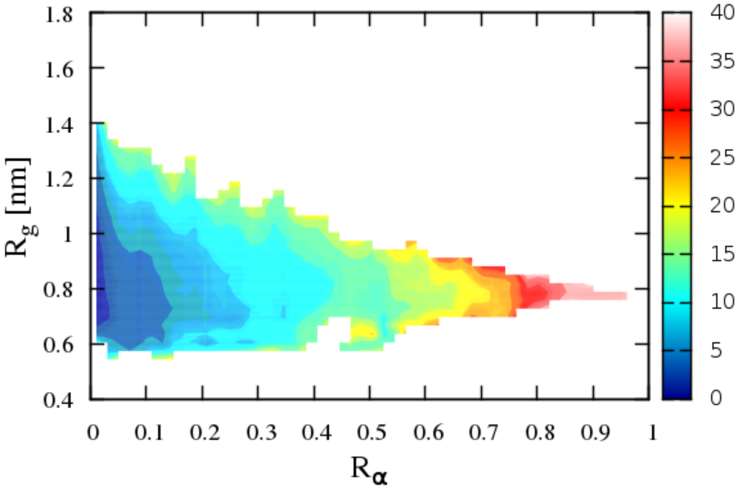


Figure 1.4: The free energy of the helix in water, as a function of the degree of β content R_β and of the gyration radius R_g (top panel) and of the degree of α content R_α and R_β (bottom panel).



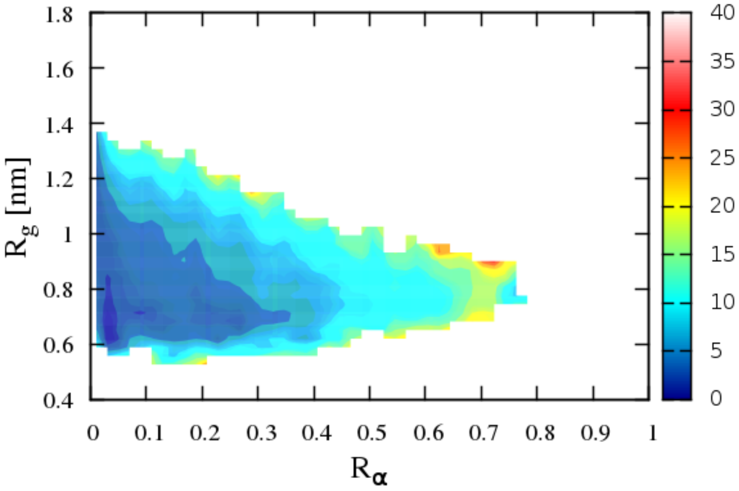


Figure 1.5: The free energies of the helix in 2M urea (top panel) and 2M GndCl (bottom panel) as functions of the degree of α content R_α and of the gyration radius R_g.

	K2D3		Contin		simulation	
	α	β	α	β	α	β
Water	5.7	9.6	11.1	4.1	14.1	0.6
GndCl 2M	8.1	8.7	15.4	3.8	17.1	2.5
GndCl 4M	3.5	9.4	4.1	1.8	0.8	0.4
Urea 2M	4.7	9.3	1.0	1.3	1.0	0.2
Urea 5M	4.7	9.6	0.1	0.0	0.3	0.0

Table 1.2: The percentage of α and β structure for the helix segment, as calculated by K2D3 [76], Contin [77] and from our simulations.

in Figs. 1.6 B, 1.7 A and B display a complex behavior that makes their interpretation uneasy. In pure water, the C_α chemical shifts are positive in the regions $23-26$ and $32-34$ and at residue 29, which usually indicates helical behavior, while is null or negative in the central region. This interpretation does not correspond to the actual population of secondary structures, displayed in Fig. 1.8 . In fact, the positive chemical shifts induced by the helical population, especially in the C-terminal region, is counterbalanced around residues 27 and 33 by the β–hairpin population, which contributes with negative chemical shifts. A similar trend is followed by the H_α chemical shifts, although here signs are reversed (for H_α it is negative chemical shifts to indicate helical behavior). The dependence of the chemical shifts on the type and concentration of denaturant is quite irregular, due to the erratic contribution of β content in compensating that of the helical population, which vice versa is quite regular (see. Fig. 1.8). The secondary chemical shifts associated with the C_β are markedly positive, indicating only β content, in contrast to what suggested by the signals from C_α and H_α. This is probably due to a failure of random-coil referencing [74] for our peptide. In fact, a downshift of $\approx 1\,\mathrm{ppm}$ of all C_β chemical shifts would result in data which are grossly consistent with C_α and H_α, although again not easy to interpret from a structural point of view.

1.3.3 Two-state approximation and m-values

Chemical denaturation is often described assuming a two state model and a linear dependence of the free energy difference between the two states on the concentration of denaturant [78], according to

$$\Delta F\left([D]\right) = \Delta F - m\,[D] \tag{1.1}$$

where ΔF is the free energy difference in water, $[D]$ is the concentration of denaturant and m is the proportionality constant. Usually, ΔF and m are calculated

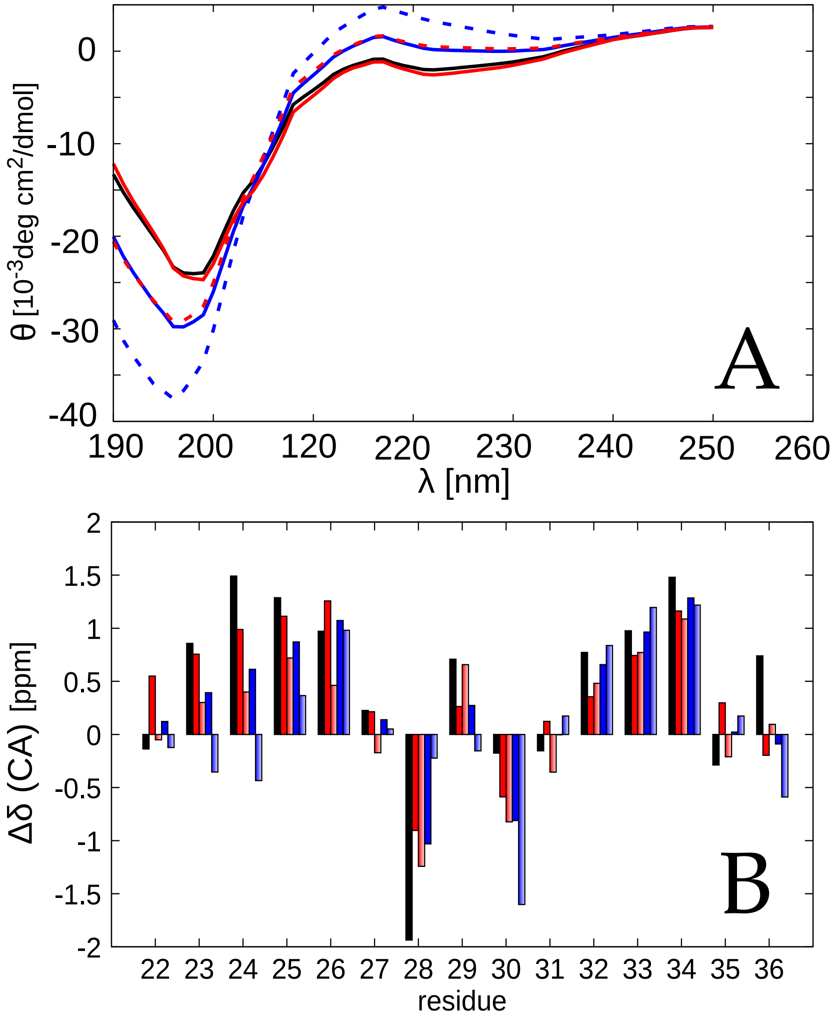


Figure 1.6: CD spectra (top panel) and chemical shifts of C_α atoms (bottom panel) predicted from the model in water (black curves), 2M urea (solid blue), 5M urea (blurred blue), 2M GdnCl (solid red) and 4M GndCl (blurred red).

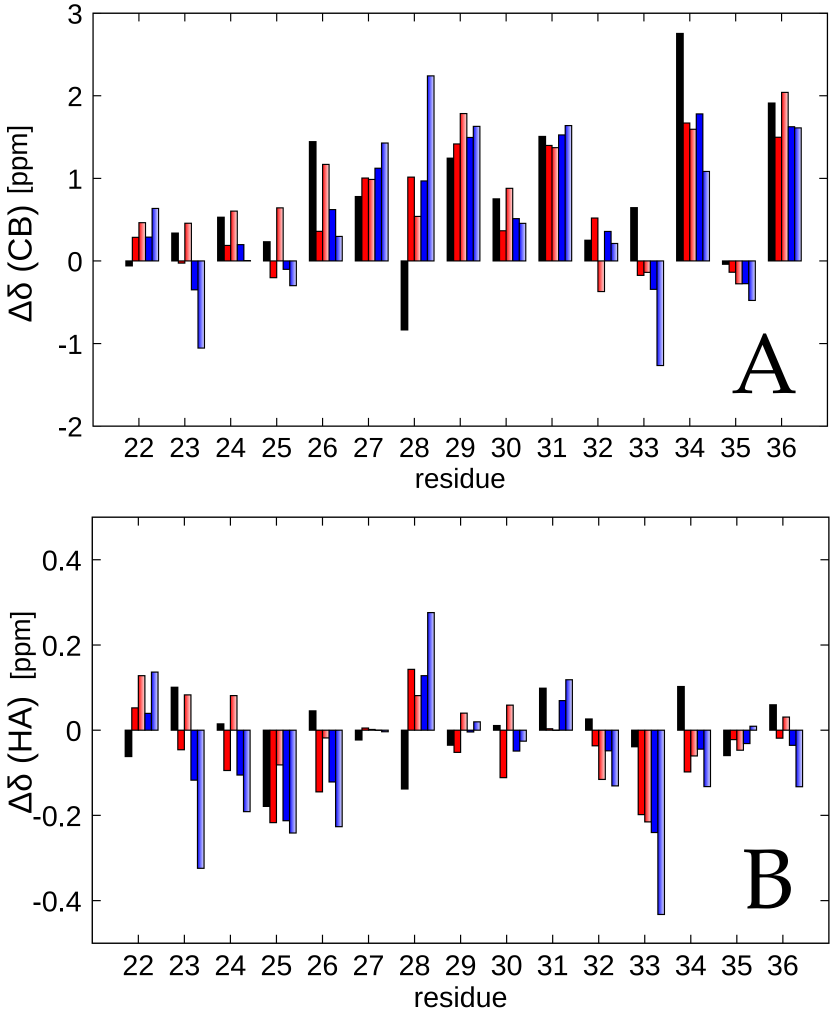


Figure 1.7: Chemical shifts of C_β (top panel) and H_α atoms (bottom panel) predicted from the model in water (black curves), 2M urea (solid blue), 5M urea (blurred blue), 2M GdnCl (solid red) and 4M GndCl (blurred red).

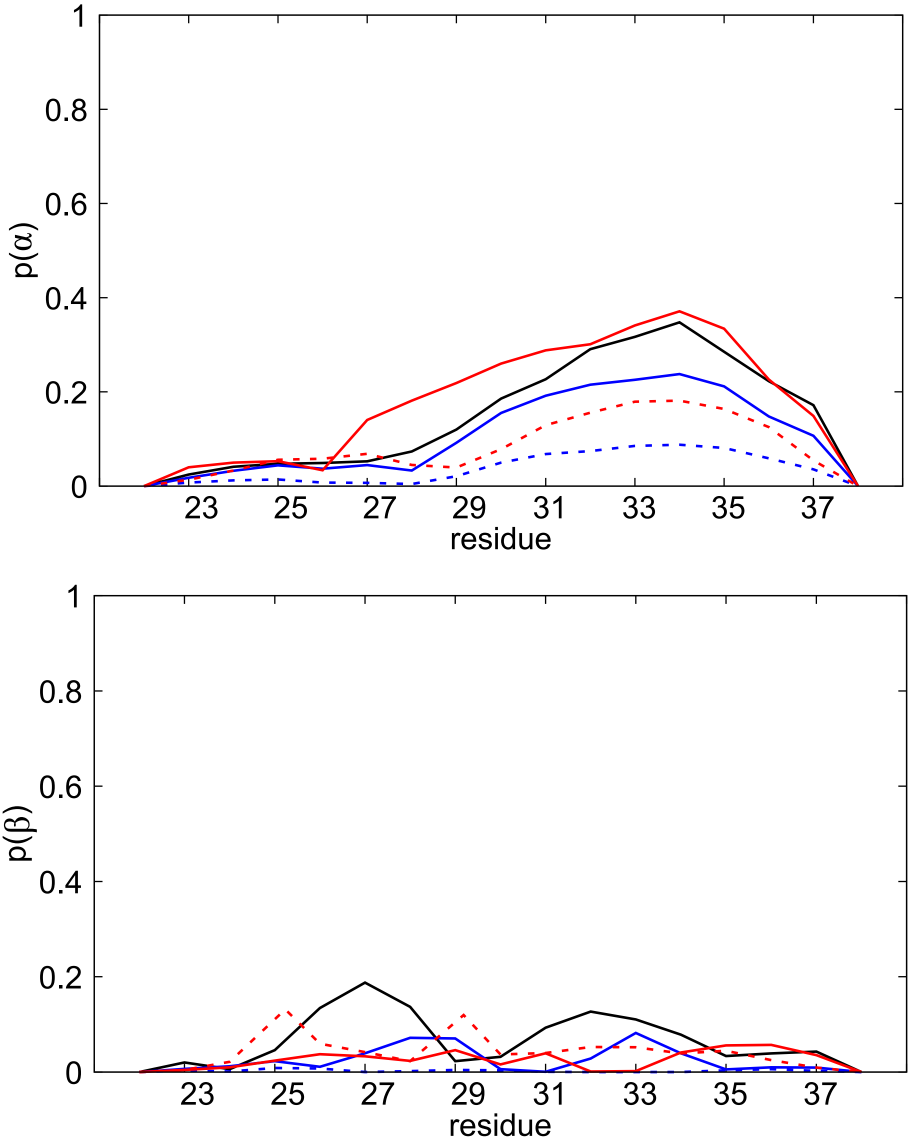


Figure 1.8: The probability of helical structures (above) and of extended beta structure (below) for each residues, calculated from the simulations in water (black curves), 2M urea (solid blue curves), 5M urea (blue dashed curves), 2M Gdn (solid red curves) and 4M Gnd (dashed red curves).

by means of a fit of the native probability p_N, obtained by fluorescence or CD, as a function of $[D]$, following

$$p_N = \frac{e^{\frac{-\Delta F([D])}{k_B T}}}{1 + e^{\frac{-\Delta F([D])}{k_B T}}} \tag{1.2}$$

However, in the present case we have only three points for each denaturant, and thus the non-linear fit is unfeasible. Consequently, we defined the native state on the basis of the free-energy profile of Fig. 1.3 as $R_\alpha > 0.5$, including states H1, H2 and H3. The corresponding free energies differences $\Delta F\left([D]\right)$ are displayed in Fig. 1.9 A. The curves are rather linear, and thus are in agreement with Eq. (1.1 with m-values $(3.0 \pm 0.1)\,\mathrm{kJmol^{-1}M^{-1}}$ for GndCl and $(2.9 \pm 0.3)\,\mathrm{kJmol^{-1}M^{-1}}$ for urea (the uncertainty being the standard error of the linear regression). The value in GndCl is essentially equal to that in urea, while usually the former is larger for globular proteins [78]. However, one should notice that the two-state picture gives only a partial picture of the effect of the denaturants. As shown in Figs. 1.3 and 1.5 , urea has a stronger effect in destabilizing the fully formed helix (state H1) than GndCl; vice versa, GndCl has a stronger effect in destabilizing the partially formed helices H2 and H3. This rearrangement of the probability distribution within the native state is lost in the two-state approximation. In the case of urea, the linear change in free energy corresponds to a linear change in the solvent accessible surface area (SASA), as displayed in Fig. 1.9 B. The free-energy gain per area in urea is approximately $7\,\mathrm{kJmol^{-1}nm^{-2}}$. This behavior is in agreement with the classical model of protein denaturation [79]. The effect of GndCl on the SASA is different, and the exposed area saturates at about $18\,\mathrm{nm^2}$, corresponding to conformations more compact than those denatured by urea (see also Fig. 1.3). A detailed comparison of the equilibrium SASA per residue under different conditions is displayed in Fig. 1.9 C. The SASA in urea are systematically larger for each residue than in water. On the other hand, in GndCl the SASA is more similar to that in water, except for GLU27, LYS28 and ASP36, which are charged residues.

1.3.4 Distribution of the solvent around fragment 22-38

The distribution of solvent around the peptide was investigated inspecting the radial distribution functions (rdf) of water and of denaturant molecules as a function of their minimum distance from each amino acid of the peptide. Two typical behaviors were found and illustrated in Fig. 1.10, where the rdf associated with ALA26 and GLU27 are displayed (those associated with the other amino acids are displayed in App. A). The shape of the rdf with water molecules is approximately the same for all residues (see Figs. 1.10 A and C). In presence of urea, the distribution of water molecules in the first shell around the residue (labeled as region S_1 in Fig. 1.10 E) is much depleted. Vice versa, in presence of GndCl the density of water in region S1 is weakly affected, and sometimes is even increased in 4M GndCl. In the second shell of water molecules (labeled as region S_2 in Fig. 1.10 E), the density is decreased in presence of urea and is unaffected in presence of GndCl at any concentration. The rdf of denaturant molecules, either urea or GndCl, is more residue-dependent. For most residues the rdf of denaturant is that displayed in Fig. 1.10 B. The density of urea is largely enriched in region S_1 with respect to water and there is a marked peak in region S_2 whose height is comparable with the density of water, but anyhow higher than the bulk density of urea. The rdf of Gnd displays peaks both in regions S_1 and S_2, in both cases lower than the corresponding density of water. Two residues display a different rdf for Gnd (see Fig. 1.10 D), displaying a much higher peak in region S_1 and essentially no peak in S_2. They are GLU27 and ASP36 which, not unexpectedly, are the two displaying a negative charge. The picture that emerges is that urea binds directly to the peptide, displacing water molecules, but GndCl does not. Thus, to investigate more the effect of GndCl on the stability of the peptide we focused on its electrostatic, long-range properties.

The fully formed helix has a dipole moment $\mu = 158$ D, shown in Fig. 1.11 A, which decreases to $\approx 90\,\mathrm{D}$ as the helix is disrupted into a coil displayed in Fig. 1.11 B. The dipole induces a separation of Gnd$^+$ and Cl$^-$ions which, in turn, produces an electric potential on the helix (see color distribution in Fig. 1.11 A). The separation of charges, besides being favorable from the point of view of the balance between Coulomb interaction and demixing entropy, gains Lennard-Jones energy between Gnd ions (see Fig. 1.12), presumably associated with hydrogen bonding between Gnd groups [80]. The result is a minimum in the free energy of the system when the dipole associated with the helix assumes minimum modulus (cfr. Fig. 1.11 B with Fig. 1.3 and Discussion section below).

1.3.5 Comparison with the denaturation of hairpin fragment

The free energy profiles of the second hairpin of GB1 (fragment $41 - 56$) in water, 4M GndCl and 5.5M urea were studied in a previous work [61] and are reported here in Figs. 1.13 and 1.14 . The result was that the peptide is stable in water at $T = 300\,\mathrm{K}$, displaying a partially-native intermediate shown as an inset in Fig. 1.13 . Urea disrupts completely the native region and stabilizes a state which resembles a random coil, while guanidine chloride has a milder effect, also maintaining the intermediate state (see Fig. 1.14). Here we analyze

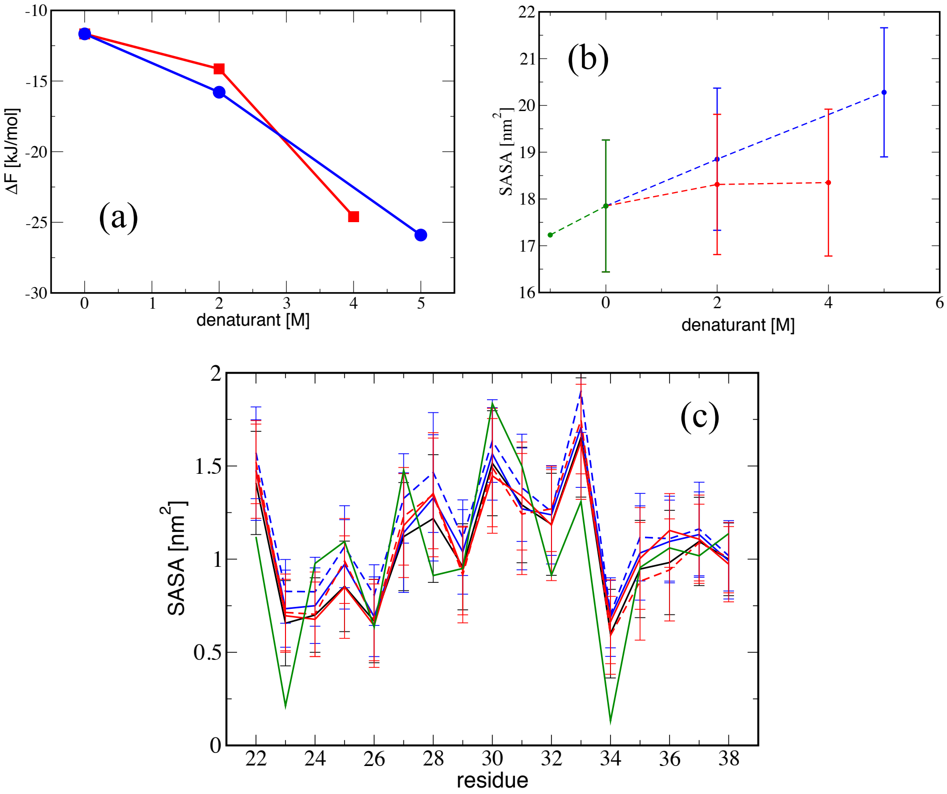


Figure 1.9: The free energy difference ΔF as a function of the concentration of urea (in blue) and Gdn (in red). (b) The average surface area exposed to the solvent; the error bars indicate the fluctuations around the average; the green point indicate the value associated with the crystallographic structure of the helix within the native protein. (c) the exposed surface area per residue; the color code is the same as in Fig. 2; the green curve refers to the crystallographic structure.

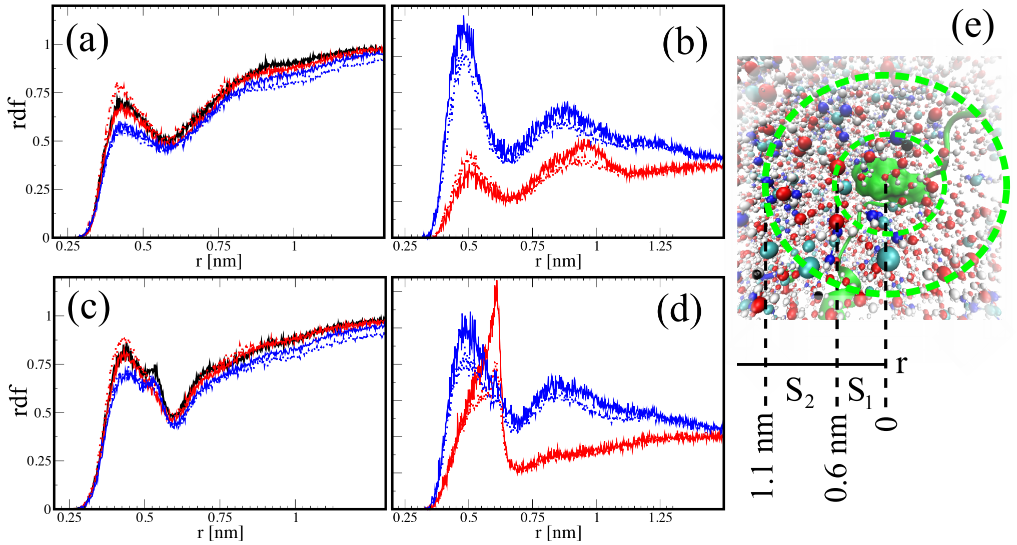


Figure 1.10: The radial distribution function of water (a,c) and of denaturant (b,d) arount A26 (a,b) and E27 (c,d). Black curves indicate the data in water, blue curves in urea (2M for the solid one, 5M for the dashed one), red curves in GndCl (2M and 4M for the solid and the dashed, respectively). A snapshot (e) of the helix in Gnd solution, where the first shell S_1 and a larger neighborhood S_2 are marked with dashed lines.

the properties of the solvent around the molecule to investigate the molecular mechanism of denaturation. In Fig. 1.15 A–D we display the rdf of water and denaturant around GLU42 and THR49 (the others are in Appendix A). Differently from the helix, water molecules in the simulation of the hairpin in GndCl experiences a modification both in the first and in the second shell, most notably around GLU42, ALA48, THR49, LYS50, THR55. The accumulation of Gnd is observed here not only around negatively-charged residues, but also around polar residues as ASN35 and ASN37. Also the native hairpin displays an electric dipole, but its modulus is 100 D, smaller than that of the helix, and consequently the charge separation in solution is also more limited, as showed in Fig. 1.15 E.

1.3.6 Experimental observables concerning the hairpin fragment

The relatively simpler structure of the free-energy profile of the hairpin fragment with respect to the helical fragment makes the interpretation of associated experimental observables potentially simpler. Fig. 1.16 A displays the CD spectrum and the secondary chemical shifts predicted by the simulations for the hairpin in water and in denaturants. The CD spectrum reports a clean β structure in water, which becomes more coil-like in GndCl and urea.

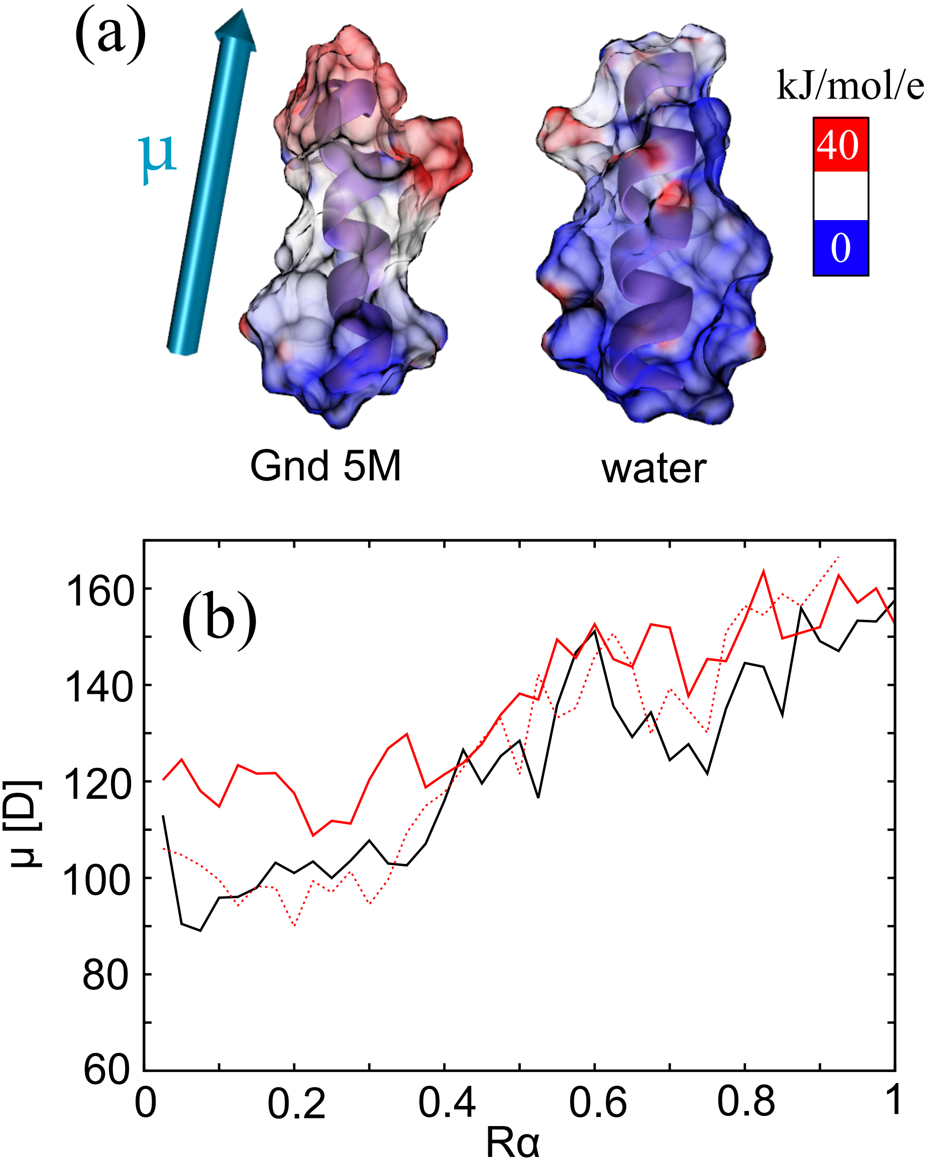


Figure 1.11: (a) The electric potential generated by the solvent on the surface of the native helix, obtained from the simulations in 5M Gnd (left panel) and in water (right panel). The colors range from blue, corresponding to a zero potential, to red, corresponding to $50\,\mathrm{k_B T/e}$. (b) The dipole moment of the helix as a function of its degree of formation R_α.

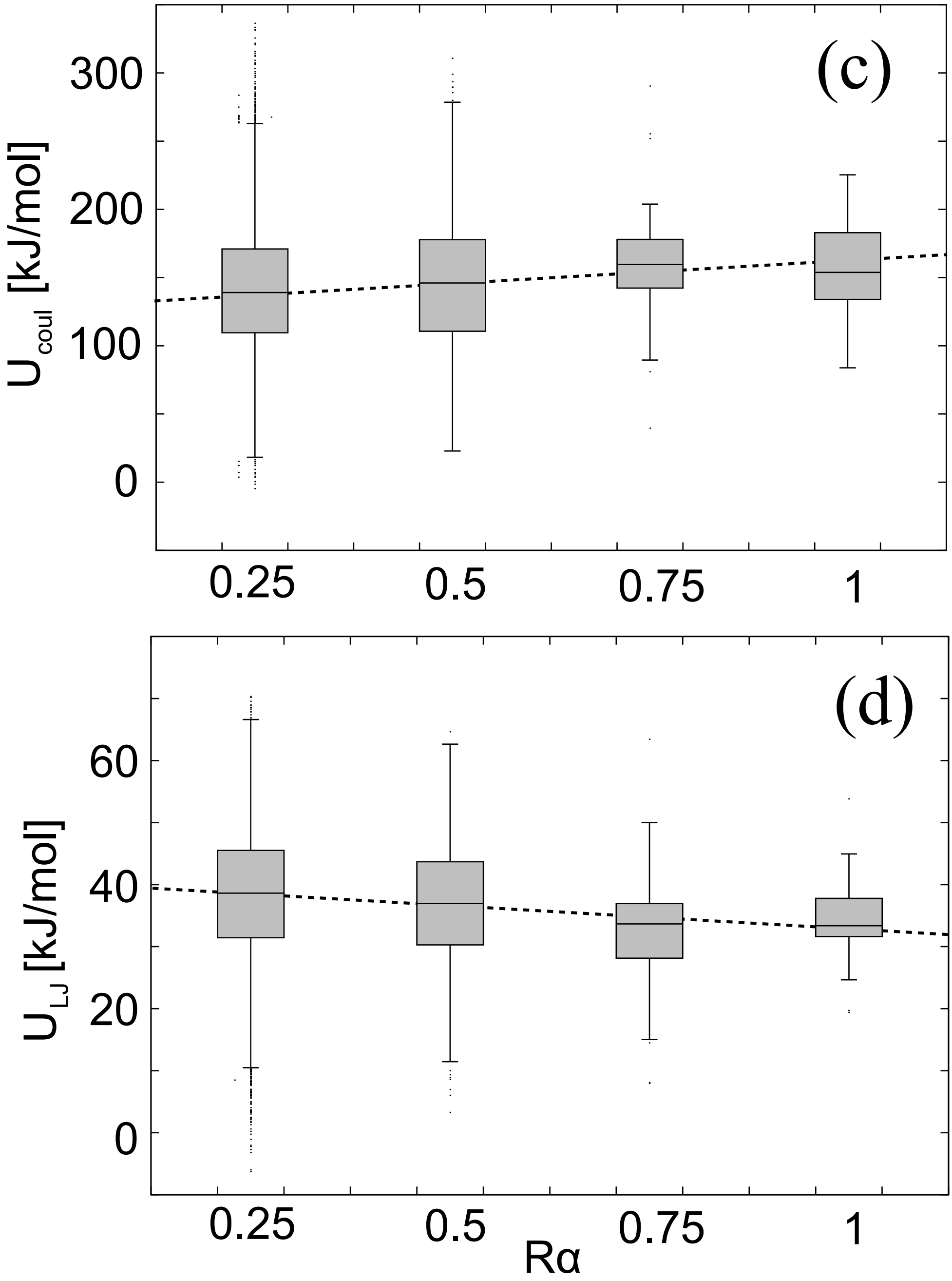


Figure 1.12: Boxplots of the Coulomb energy (above) and the Lennard Jones energy (below) between GndCl molecules as a function of the helicity of fragment $22 - -38$, calculated in the simulation in 4M GndCl.

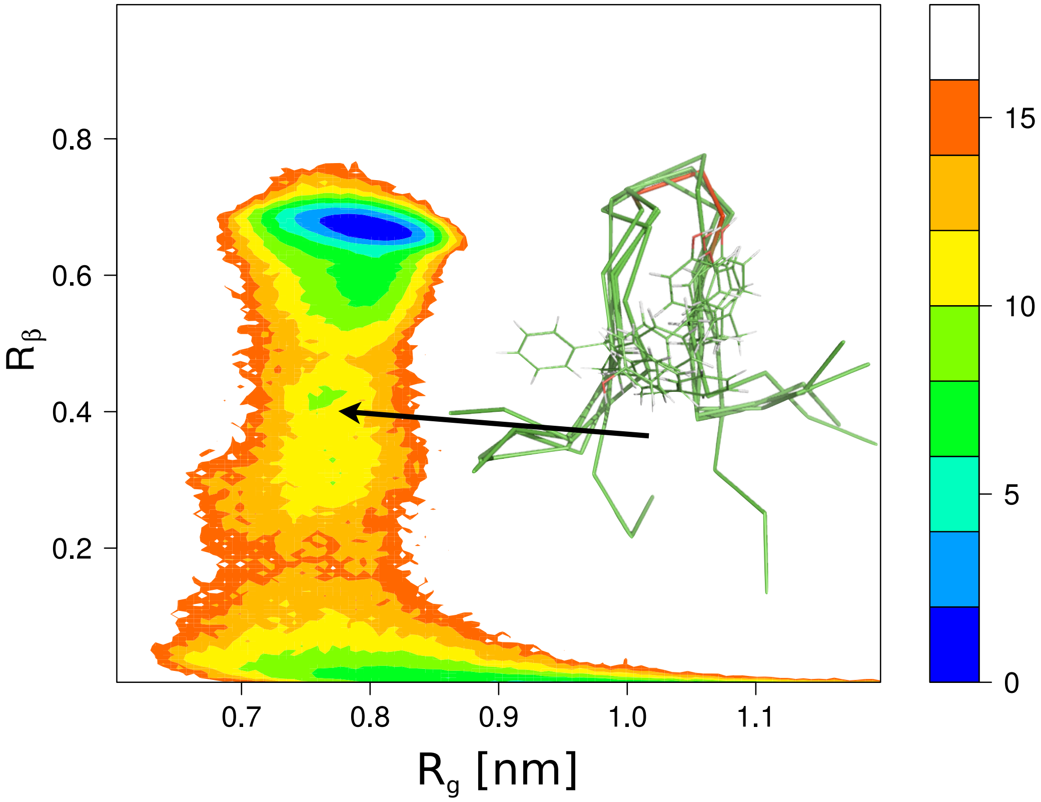


Figure 1.13: The free energy of the hairpin in water, as a function of the gyration radius R_g and of the degree of β–content R_β. In the inset, five representative conformations of the intermediate state are shown and TYR45 and PHE52 are highlighted. Adapted with permission from ref. [61]. Copyright ©2013, American Chemical Society.

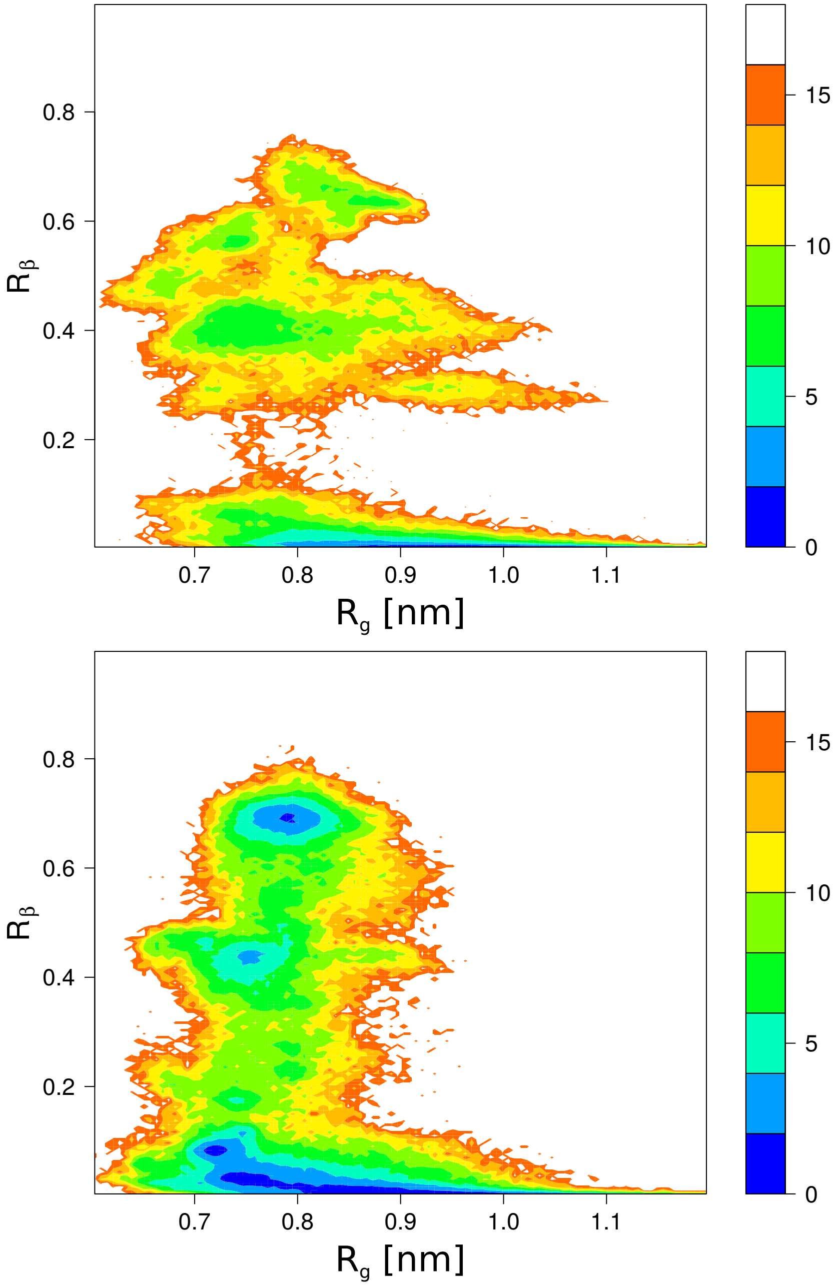


Figure 1.14: The free energy of the hairpin in 5.5M urea (top panel) and in 4M GndCl (bottom panel) as a function of the gyration radius R_g and of the degree of β–content R_β. Adapted with permission from ref. [61]. Copyright ©2013, American Chemical Society.

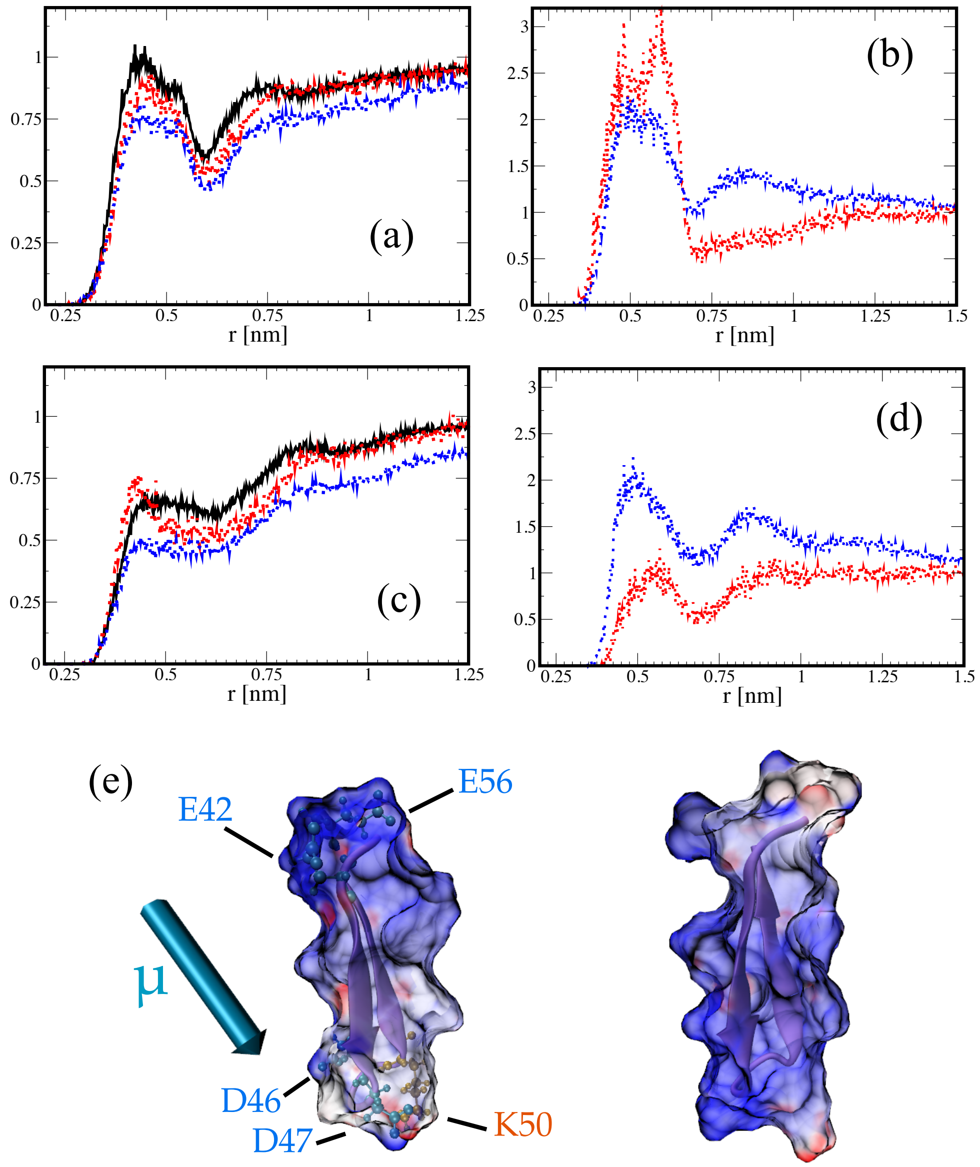


Figure 1.15: The rdf of water (a and c) and denaturant (b and d) around GLU42 (a and b) and THR49 (c and d). The black curves refer to the simulation in water, the blue curves to that in urea while the red curves to that in GndCl. (e) The electric potential generated by GndCl on the surface of the hairpin, as in Fig. 1.10 A.

The secondary chemical shifts (see Figs. 1.16 B and 1.17 A and B) in water are consistent with a β–hairpin, with negative values for C_α and positive values for C_β and H_α towards the termini. Under denaturing conditions, the picture becomes more involved. For example, in 5M urea the peptide is essentially coil, but the secondary chemical shifts of all atoms (blue striped bars in Figs. 1.16 B and 1.17) display an irregular behavior.
The m-value resulting from the simulation of the hairpin fragment is $4.0\,\mathrm{kJmol^{-1}M^{-1}}$ for urea and $1.88\,\mathrm{kJmol^{-1}M^{-1}}$ for GndCl. Although we have only two points and we cannot calculate the uncertainty of the m-values associated with the fig, if we assume an error comparable with that of the helix we can conclude that, in this case, the m-value is larger for urea. This behavior is in contrast with what observed for the helix and, in general, for globular proteins [78]. However, one should consider two facts. The first evidence is that, as in the case of the helix, the thermodynamics of the hairpin does not show only two states, while the definition of m-value is based on a two-state approximation. The second one is that the general trend reported in the literature reflects the behavior of stable proteins, while we are studying small peptides. Our results suggest that helices and hairpins display different weights in the destabilization free energies of globular proteins, in urea and in GndCl.

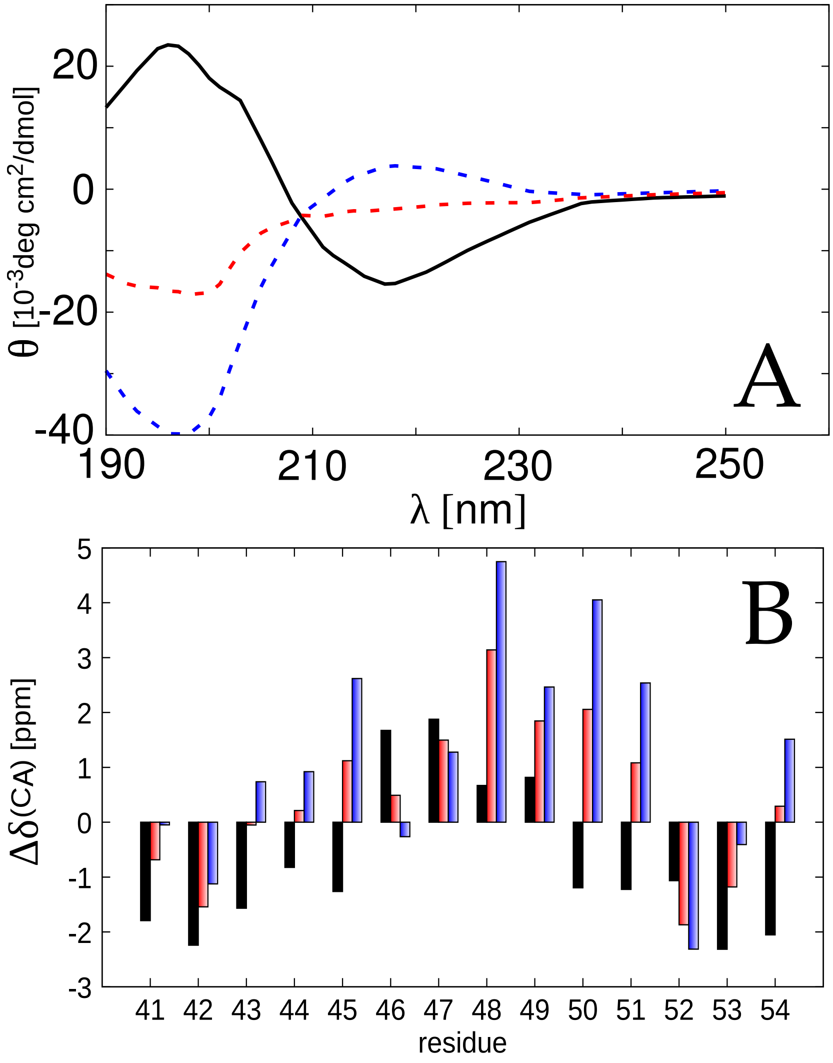


Figure 1.16: CD spectra (top panel) and chemical shifts of C_α atoms (bottom panel) predicted from the model in water (black curves), 5M urea (blurred blue) and 4M GndCl (blurred red).

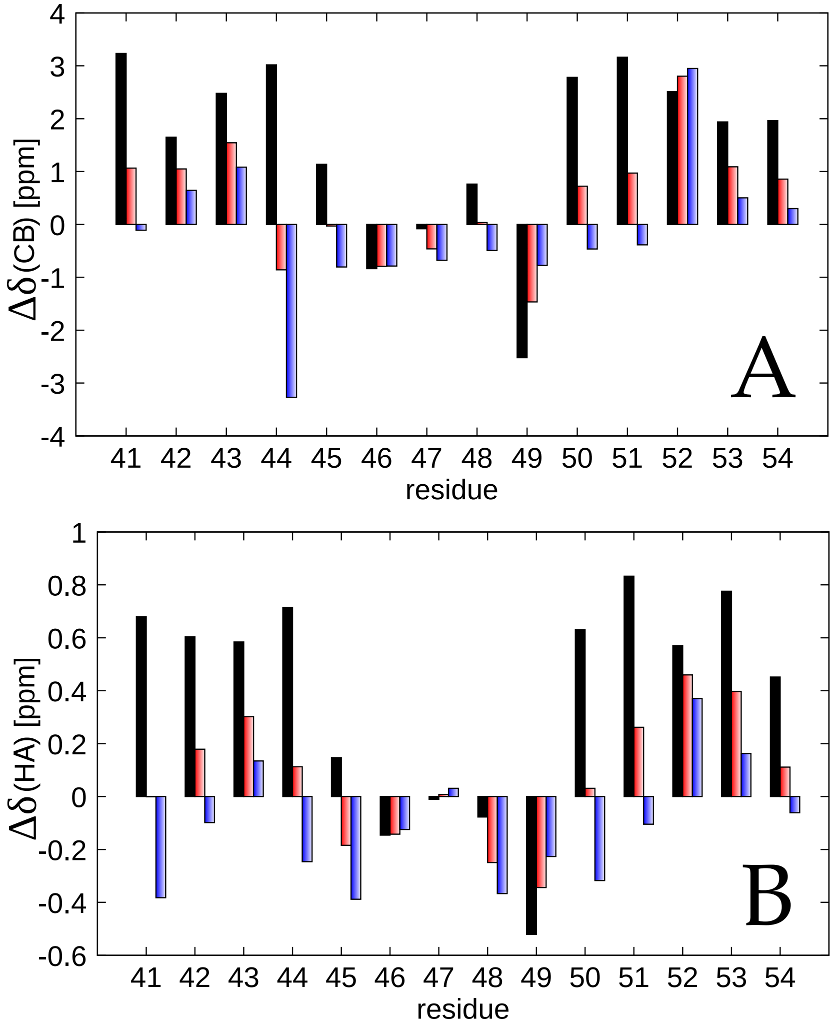


Figure 1.17: Chemical shifts of C_β (top panel) and H_α atoms (bottom panel) predicted from the model in water (black curves), 5M urea (blurred blue) and 4M GndCl (blurred red).

1.4 Discussion

The helix and the second hairpin of GB1 are among the most studied small peptides in the literature, both from an experimental and a computational point of view. As in the case of larger proteins, a standard experimental tool to probe the thermodynamic and conformational properties of these peptides is to denature them with urea or GndCl. However, the result of these experiments can be difficult to interpret, or even tricky. Simulations performed with advanced sampling techniques can be useful to interpret the raw experimental data and to monitor quantities that are difficult to access in experiments. In the specific case of the two fragments of GB1, the data on hairpin at $T = 300\,\mathrm{K}$ are rather simple to interpret. Its conformational space in water displays three well-defined states corresponding to the fully formed, to the half formed hairpin, and to a random coil. The effect of chemical denaturants is to decrease the population of the native state, without changing the structure of the conformational space. At $T = 300\,\mathrm{K}$ the population of helical states is negligibly low. Consequently, experimental observables like CD spectra and secondary chemical shifts just report the fraction of β–content of the peptide. Their main limitation is that they cannot distinguish between the fully formed hairpin populated with probability one half and the half-formed hairpin populated with probability one.

The situation for the helix is more complex. Several states, involving different degrees of structures α and β, compete with each other. The experimental data are the sum of the contributions of all these states. This is particularly problematic in the case of secondary chemical shifts, in which α and β give contributions with opposite sign, reporting a random-coil behavior when α and β states have comparable probabilities. The effect of urea appears rather simple from the simulations. It fills the first shell of solvent around the peptides in a residue-independent way, thus breaking the hydrogen bonds, which are the main interactions that stabilize either the helix or the hairpin. This is the reason why urea is so effective in stabilizing the denatured state of short peptides, containing only secondary structure, as compared with GndCl. In fact, in the case of full proteins, also stabilized by tertiary interactions, the denaturant power of urea is usually comparatively smaller [78]. This picture is in agreement with the results of hydrogen-exchange NMR experiments, which show that urea, but not GndCl, can form hydrogen bonds with peptides [18].

In fact, the denaturing mechanism of GndCl seems more complex. To some extent, Gnd accumulates in the neighborhood of the peptides, at various distances from its surface and differently for each type of aminoacid. As expected, the effect is very large for acid residues and negligible for basic residues. As a rule, it

does not seem to deplete the concentration of water in the first shell or to modify its radial distribution function appreciably.
Another effect of GndCl is to generate an electric field that interacts with the electric dipole of the peptide. This effect is very clear for the helix, which displays a large dipole moment due to the spatial alignment of the aminoacid, and is smaller for the hairpin, in which the (smaller) dipole moment is due to the specific sequence of acid and basic residues. The modulus of the electric dipole depends on the degree of formation of the secondary structure. In presence of GndCl, each of the effective charges that define the dipole is screened by ions of opposite charge. Thus, the two of them undergo a screened attractive Coulomb interaction that tends to decrease their (effective) distance and then to decrease the modulus of the dipole moment. The equilibrium state corresponds then to a small dipole, and thus to a denatured peptide.

CHAPTER 2

Dimensional reduction of Collective Variables

2.1 Introduction

Biomolecular models are usually very high–dimensional systems: proteins in solutions, as described in classical Molecular Dynamics simulations, are characterized by the three–dimensional positions and velocities of all the atoms they are constituted of. Consequently, they define a phase space whose number of dimensions is of order $\mathcal{O}(10^5 - 10^6)$. Studying the result of a Molecular Dynamics simulation in such a high–dimensional space is outrageously difficult, and consequently one usually follows the behavior of low–dimensional (often one–dimensional), Collective Variables $Y(\mathbf{r})$ which are function of the high–dimensional coordinates $\mathbf{r}$ of the system [35].

Also in dealing with experiments, one has usually the opportunity to monitor only low–dimensional Collective Variables. For example, the fluorescence intensity of tryptophanes, the nuclear chemical shifts in NMR spectra, the intensity of scattered X–rays in SAXS or the ellipticity in CD depend on the high–dimensional coordinates $\mathbf{r}$ of a system but can provide only a low–dimensional view of its conformational properties. While in the design of experiments the choice of the Collective Variables to monitor is determined by the technique itself, in simulations one can choose to compute and analyze essentially any function of the microscopic coordinates. The problem is how to make such a choice.

In the study of equilibrium properties of a system, the only feature that the Collective Variable must have is to be able to distinguish the relevant phases of the system, that is to be a good "order parameter". In the case of protein folding, for example, it should assume different values in the native state, in the denatured state and, when relevant, in intermediate states. The root mean square deviation of the atomic positions (RMSD) with respect to those of the native conformation, or the fraction q of native contacts usually can do that [31], although the latter is known as a "bad" Collective Variable when it comes to apply a bias on it [44].

The choice is less straightforward for the description of the time–dependent, dynamical properties of the system [32, 33]. For most purposes, it is useful that the Collective Variable Y obeys a Langevin equation of kind

$$\frac{dY}{dt} = D^{(1)}(Y) + \sqrt{2D^{(2)}(Y)} \cdot \eta(t) \tag{2.1}$$

Here, $\eta(t)$ is a stochastic variable with gaussian distribution whose moments are $\overline{\eta(t)} = 0$ and $\overline{\eta(t)\eta(t')} = \delta(t - t')$, while the bar indicates the average over the realizations of the stochastic variable. From a mathematical point of view, Eq. (2.1) is not completely defined, since a rule is needed for the timing of evaluation of $\sqrt{2D^{(2)}(Y)(t)}$ [81], namely for the interpretation of how to compute the term

$$\mathcal{I}(t, dt) = \int_t^{t+dt} \sqrt{2D^{(2)}(Y(\tau))} \cdot \eta(\tau) d\tau \tag{2.2}$$

in the integral form of Eq. (2.1). If both $\sqrt{2D^{(2)}(Y(t))}$ and $\eta(t)$ were continuous functions of t, then one would apply the first integral mean value theorem and Eq. (2.2) would become

$$\mathcal{I}(t, dt) = \sqrt{2D^{(2)}(Y(t^*))} \cdot \int_t^{t+dt} \eta(\tau) d\tau \tag{2.3}$$

where t^* is a specific point in the time interval $[t, t + dt]$, albeit a priori unknown. Unfortunately, $\eta(t)$ is far from being continuous, since it usually consists in a series of δ–function spikes of random sign [34] which yields the application of the mean value theorem unfeasible. The interpretation of $\mathcal{I}(t, dt)$ is then performed by evaluating the term $\sqrt{2D^{(2)}(Y)}$ in a constrainted linear combination of the values of Y at times t and $t + dt$

$$\mathcal{I}(t, dt) = \sqrt{2\left[\alpha Y(t) + (1 - \alpha)Y(t + dt)\right]} \cdot \int_t^{t+dt} \eta(\tau) d\tau \tag{2.4}$$

where α is a real number in the interval $[0, 1]$.

The two most common conventions for α are the Stratonovich interpretation ($\alpha = 1/2$) and the Itô interpretation ($\alpha = 0$), the former being a more appropriate representation in frameworks where the white noise is an approximation of a fluctuating noise with finite, short memory, whereas the latter is a natural starting point for numerical schemes [82]. Whatever the choice is made, $D^{(2)}(Y)$ is in both cases a position–dependent coefficient which controls the diffusion of Y within its low–dimensional space, while $D^{(1)}(Y)$ can be related to the equilibrium free energy $F(Y)$ through

$$D^{(1)}(Y) = -\frac{1}{\gamma(Y)}\frac{\partial F(Y)}{\partial Y} + \frac{1}{2}(1 - \alpha)\frac{\partial D^{(2)}(Y)}{\partial Y} \tag{2.5}$$

where the second term is necessary in order for the system to evolve towards the Boltzmann distribution. $D^{(1)}(Y)$ then can be regarded as an effective force acting on the Collective Variable, consisting in the classical, energetic term plus a diffusion–dependent correction.

Hence, both the coefficients are very rich of information on the system. If Eq. (2.1) holds, one can use Arrhenius equation to estimate reaction rates, can model the dynamics of the system as transitions between discrete states, and can use all the armoury of tools developed in the realm of Langevin equations [81, 35]. Besides the advantages in analysing and plotting a posteriori the relevant information, Collective Variables satisfying Eq. (2.1) can be used on–the–fly to bias the dynamics of the system and thus speed-up equilibrium sampling, like in the case of Umbrella Sampling, Metadynamics and Steered Molecular Dynamics [83, 84, 85], or even to obtain efficiently dynamical properties [32, 86].

An equation formally similar to Eq. (2.1) can always be written for any CV. However, a bad choice of the CV results in functions $D^{(1)}$ and $D^{(2)}$ which depend not only on Y, but on the whole history of the system; in other words, in this case Y undergoes a non–Markovian process [35]. This happens when Y does not really determine the properties of the relevant phase space where the dynamics of the system is likely to occur, but, on the contrary, to a given value of Y can correspond different, well–separated regions of the relevant microscopic phase space, in which the effective force $D^{(1)}$ and the effective diffusion coefficient $D^{(2)}$ are very different. In this case, having discarded the microscopic coordinates, only time can distinguish between the different phase–space regions at identical Y, resulting in a non–Markovian dynamics.

A necessary and sufficient condition for Eq. (2.1) to hold is that Y describes the slowest kinetic modes of the system [35]. This implies that the system can visit very quickly all the phase–space accessible regions for any fixed value of Y, thus equilibrating the coordinates perpendicular to it. In this context, fixed means that the relative change in value of Y is negligible with respect to that of the perpendicular coordinates. As a consequence, $D^{(1)}$ and $D^{(2)}$ are determined by the average contribution of all the phase–space regions displaying the same value Y, the time dependence is averaged out and only the dependence on Y remains. However, this is not a property which can be easily verified for any given function of the microscopic coordinates $\mathbf{r}$ of the system.

The study of the dynamics of a system by Eq. (2.1) requires the evaluation of $D^{(1)}$ and $D^{(2)}$ from a set of microscopic trajectories. Given a set of stochastic trajectories $\{\mathbf{r}(t)\}$, generated with some Molecular Dynamics algorithm from a point $\mathbf{r}(0)$ of conformational space, the effective force and diffusion coefficient defined by Eq. (2.1) in that point can be obtained by definition as the first two

Kramers–Moyal coefficients [81]. Defining $Y_t \equiv Y(\mathbf{r}(t))$, then

$$\begin{aligned} D^{(1)}(Y_0) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \overline{(Y_{\Delta t} - Y_0)} \\ D^{(2)}(Y_0) &= \frac{1}{2} \lim_{\Delta t \to 0} \frac{1}{\Delta t} \overline{(Y_{\Delta t} - Y_0)^2}. \end{aligned} \tag{2.6}$$

Physically, the limit $\Delta t \to 0$ means that $D^{(1)}$ and $D^{(2)}$ should be a property of Y only, independent on where the trajectories go to afterwards. In the case of MD simulations of biopolymers, the use of Eqs. (2.6) presents a serious problem, namely that any integrator that can be used to generate $\{\mathbf{r}(t)\}$ has a finite time step, and thus the limit $\Delta t \to 0$ cannot be evaluated.

Several works tried to estimate the drift and diffusion coefficients of Eqs. (2.6) in the limit of small Δt, by correction terms [36, 37], by iterative procedures [38], or by evaluating the adjoint Fokker–Planck operator [39, 40]. However one should stress that these works face the problem of evaluating $D^{(1)}$ and $D^{(2)}$ for generic time series of observables characterized by a low, uncontrollable, sampling rate. In the case of MD simulations, the minimum time period can be as small as an integration time step, which is smaller than any process involved in the microscopic dynamics. Assuming to know the reaction coordinate Y, an efficient way of extracting drift and diffusion coefficient from MD simulations was developed on the basis of a Bayesian approach [41, 42] and then applied to protein folding [43, 44]. The main result of these works is that the diffusion coefficient for variables that scale as the Euclidean distances between Cartesian coordinates depends strongly on Y, while the diffusion coefficient for variables that have a filtered dependence on such distances, like contact functions, depend weakly on Y. Using a maximum–likelihood principle [45], the drift and diffusion coefficients could be obtained as average of molecular dynamics trajectories, and a criterion for the choice of the sampling rate of the trajectories was introduced to minimize time correlations of noise.

An issue associated with the approach developed in ref. [41] is the following: the drift and diffusion coefficients $D^{(1)}$ and $D^{(2)}$ are a priori supposed to be functions only of the low–dimensional coordinate Y, and all the information regarding the microscopic conformation $\mathbf{r}$ is lost in the process of the evaluation of $D^{(1)}(Y)$ and $D^{(2)}(Y)$. In mathematical terms, the underlying hypothesis is that if two conformations $\mathbf{r}_0$ and $\mathbf{r}_1$ have the same value of Y, then also their coefficients must be equal.

$$\text{if } \exists \mathbf{r}_0, \mathbf{r}_1 : Y(\mathbf{r}_0) = Y(\mathbf{r}_1) = \overline{Y} \longrightarrow D^{(1-2)}(Y(\mathbf{r}_0)) = D^{(1-2)}(Y(\mathbf{r}_1)) = D^{(1-2)}(\overline{Y}) \tag{2.7}$$

Our goal here is rather different from those discussed above. Our primary task is to investigate the validity of the framework defined by Eq. (2.1) through the condition 2.7, and in particular whether it is possible to define the drift and diffusion coefficients $D^{(1)}$ and $D^{(2)}$ as a function only of the Collective Variable Y. In other words, we studied the validity of the hypothesis at the basis of refs. [41, 42, 43, 44, 45].

Another important difference is that, since we are interested in facing the problem from a computational perspective, we did not study directly the validity of true Langevin equation (2.1), but its finite–differences counterpart, defined within the scheme of a standard integrator at finite time step Δt. In fact, it is the finite–difference dynamic equation what one usually calculate in Molecular Dynamics simulations. The use of a finite–difference scheme allows us to override the issue of a choice between Itô, Stratonovich or other interpretation rules, although the former is the natural limit of our implementation for $\Delta t \to 0$, namely for the reversion from finite–differences to a stochastic differential equation.

In the next Sections we present an algorithm to obtain efficiently the drift and diffusion coefficient of the finite–time–step Langevin equation. Then we show to which extent $D^{(1)}$ and $D^{(2)}$ are function of Y only, and not depend on the detailed microscopy coordinates. We first applied this analysis to some test models and then to some popular CVs within simple alpha–helix and beta–hairpin models. Finally, we studied to which extent the calculated coefficients are in agreement with the equilibrium free energy of the system, and the dynamics in the reduced space is in agreement with the projection of the associated microscopic dynamics.

2.2 Drift and diffusion coefficients in finite–difference overdamped Langevin equation

2.2.1 Finite–difference equation in Euler–Maruyama approximation

We start by considering a set of M trajectories $\{\mathbf{r}(t)\}$ in the microscopic conformational space, which are the output of M Molecular Dynamics simulations of (unspecified) biomolecules. We also suppose that these trajectories are solution of a Langevin equation – be it complete or overdamped – which has been virtually integrated by a proper integrator making use of a *microscopic* time step Δt_{mic}, starting from the initial condition $\mathbf{r}(0)$ at fixed temperature T. For a correct integration of the equation of motion, Δt_{mic} needs to be shorter than the fastest physical process which can occur in the simulation. In practice, for the description of molecular processes a common choice consists in $\Delta t_{mic} \sim \mathcal{O}(10^{-15}\ s)$, as it allows the characterization of atomic bonds vibrations.
We further consider a Collective Variable Y, function of the microscopic coordinates $Y = Y(\mathbf{r})$, which we use to describe some relevant property of the system simulated. Here we have represented Y as a scalar function, though there is no prescription for its number of dimensions: the choice is dictated by common sense, as it is easier to visualize information on a monodimensional function, for instance in form of a time series. We want to investigate whether the projection of the microscopic dynamics on Y can be described as the output of the resolution of an overdamped Langevin equation, integrated by means of a virtual Euler–Maruyama integrator:

$$Y_{t+\Delta t} = Y_t + D^{(1)}(Y_t)\Delta t + \sqrt{2D^{(2)}(Y_t)\Delta t} \cdot \eta_t \tag{2.8}$$

where the dependence on the time ($Y(t)$) has been denoted with a subscript for ease of reading. Here, η_t is a white noise, namely an adimensional, Gaussian–distributed stochastic variable with zero average and $\overline{\eta_t \eta_{t+n\Delta t}} = \delta_{n,0}$, the delta being the Kronecker symbol. Within this finite–difference scheme, the drift and diffusion coefficients $D^{(1)}$ and $D^{(2)}$ can be interpreted as

$$D^{(1)}(Y) = -\frac{1}{\gamma}\frac{\partial F(Y)}{\partial Y} + \frac{1}{2}\frac{\partial D^{(2)}(Y)}{\partial Y} \tag{2.9}$$

$$D^{(2)}(Y) = D(Y) \tag{2.10}$$

where the second term on Eq. (2.9) can be usually neglected, at least in our framework when proper order of magnitudes are inserted (more details can be found in App. B). $F(Y)$ is the free energy associated to the Collective Variable Y, the first term of the definition of $D^{(1)}(Y)$ being the effective force acting in Y modulated by the inverse of the friction coefficient γ. On the other side, $D(Y)$ is simply the diffusion coefficient on Y, where the superscript has been dropped to highlight the fact that the left side of Eq. (2.10) comes from the application of Eq. (2.6), whereas $D(Y)$ is a physical property of the system under investigation. The purpose of this Section is to understand whether there exist two functions $D^{(1)}(Y)$ and $D^{(2)}(Y)$, along with an effective timestep Δt, for which the time evolution of Y – simulated with an Euler–Maruyama algorithm – displays the same moments $\overline{Y^k(t)}$, for any k, equal to $\overline{Y_t^k}$ at any time t, where $\overline{Y_t^k}$ are calculated from the trajectories generated by Eq. (2.8). More specifically, one would like that the difference between the true moments and those calculated by Eq. (2.8) goes to zero as $M \to \infty$.

In general, the timestep Δt of the effective equation (2.8) and the timestep Δt_{mic} of the underlying, full–dimensional simulation do not correspond, as Δt can be larger than Δt_{mic}. Consequently, the original dynamics could result as coarse–grained in time, when projected onto the effective dynamics. A relevant question would be then what is the most suitable timestep Δt for which the Eq. (2.8) best reproduces the microscopic dynamics, as it is not granted that the best choice is the smallest possible value, that is $\Delta t = \Delta t_{mic}$. For instance, a large value of Δt would have as nice effect that all the problems associated with the Markov–Einstein time scale would be avoided [40]. Moreover, since the non–Markovianity of the projected dynamics is a typical problem associated with dimensional reduction, the use of a large timestep Δt increases the probability of ending up into a Markovian process. If a set of $D^{(1)}(Y)$, $D^{(2)}(Y)$ and Δt exists such that the associated Eq. (2.8) is able to reproduce the dynamics of the system projected on Y (that is, its moments), then this dynamics is Markovian by definition.

In a purely theoretical scheme, where instead of Eq. (2.8) one deals with the true Langevin equation, the drift and the diffusion coefficients at a specific point Y_0 could be obtained by means of their mathematical definitions, as the first two coefficients of the Kramers–Moyal expansion of the distribution of Y [81].

$$D^{(n)}(Y) = \lim_{\tau \to 0} \frac{\overline{[Y(t+\tau) - Y(t)]^n}}{n!\tau} \tag{2.11}$$

and setting $t = 0$. It is worth to mention that all the higher Kramers–Moyal coefficients ($n \geq 3$) of the expansion should be identically zero, due to Pawula's theorem for the specific form of the Langevin equation we are using and its Euler–Maruyama counterpart [81]. In principle, the same procedure can be performed in a finite–differences framework: from Eq. (2.8), applying the Def. (2.11) one obtains

$$\begin{aligned} D^{(1)}(Y_0) &= \frac{\overline{Y_{\Delta t} - Y_0}}{\Delta t} - \sqrt{\frac{2D^{(2)}}{\Delta t}}\overline{\eta_0} \\ D^{(2)}(Y_0) &= \frac{\overline{[Y_{\Delta t} - Y_0]^2}}{2\Delta t} + \frac{1}{2}\left[D^{(1)}(Y_0)\right]^2 \Delta t - D^{(1)}\left(\overline{Y_{\Delta t} - Y_0}\right) = \\ &= \frac{\overline{[Y_{\Delta t} - Y_0]^2}}{2\Delta t} + \frac{1}{2}\left[D^{(1)}(Y_0)\right]^2 \Delta t - D^{(1)}(Y_0)\left(D^{(1)}(Y_0)\Delta t + \sqrt{2D^{(2)}(Y_0)\Delta t}\cdot\overline{\eta_0}\right) = \\ &= \frac{\overline{[Y_{\Delta t} - Y_0]^2}}{2\Delta t} - \frac{1}{2}\left[D^{(1)}(Y_0)\right]^2 \Delta t - D^{(1)}(Y_0)\sqrt{2D^{(2)}(Y_0)\Delta t}\cdot\overline{\eta_0}. \quad (2.12) \end{aligned}$$

In theory, the last terms of both quantities tend to zero, because $\overline{\eta_0}$ does as $M \to \infty$. Operatively, one prepares M systems which start from points $\mathbf{r}$ of the conformational space all displaying $Y(\mathbf{r}) = Y_0$. For every trajectory, the system is allowed to evolve for a time Δt, then the displacement $Y(\Delta t) - Y_0$ and its square are computed from the first and the last frame. Finally, their averages are substituted in Eqs (2.12) and eventually are corrected with the deterministic term (see $D^{(2)}(Y_0)$) considering the stochastic corrections (which one cannot compute, not knowing the values η_0) hopefully negligible, due to the nature of η. It is worthwhile to notice that, at variance with the expressions resulting from the standard Kramers–Moyal expansion (2.11), in Eqs (2.12) one does not have to compute a short–time limit, but the time increment Δt is the same which defines the dynamic equation (2.8).

2.2.2 Calculation of the drift coefficient in a test model

Unfortunately, to directly apply Eqs. (2.12) does not seem to be fruitful, because for finite values of M the average $\overline{\eta_0}$ introduces a non–negligible error. In fact, we tested the procedure above presented in a very simple test model, which consisted in a particle moving in a one–dimensional, harmonic potential

$$U(Y) = \frac{k}{2}(Y - Y_C)^2 \quad (2.13)$$

with the diffusion coefficient being constant over the space, as in the Einstein's formula: $D^{(2)}(Y) = k_B T\gamma \equiv D$, where $k_B T$ is the thermal energy. It is worth to notice that, in this case, the Collective Variable Ycorresponds to the microscopic coordinate r indicating the position of the particle. For this reason, the entropy associated to the Collective Variable is zero and its free energy is equal to the potential energy 2.13 which, following Eq. (2.9), leads to the drift coefficient

$$D^{(1)}(Y) = -\frac{k}{\gamma}(Y - Y_C) \quad (2.14)$$

We chose this example as it is probably the simplest system displaying non–zero drift nor diffusion coefficients. Moreover, it could be easily considered as the first–order approximation of any interaction – both attractive or repulsive – that can be found in biological systems, accordingly with the sign of k. We simulated this test model using the characteristic units of biomolecules: the length scale is expressed in nanometers, the energy scale is comparable to the thermal energy at room temperature ($k_BT \sim 2.5\,\mathrm{kJ\,mol^{-1}}$), which corresponds to forces of the order of tens of picoNewton. The numerical values used in the simulations were then chosen as typical orders of magnitude for biomolecules. In Fig. 2.1 the value of $D^{(1)}(Y)$ is shown, where $Y = 1.1\,\mathrm{nm}$. It was calculated through the application of the first of Eqs. (2.12) to different numbers M of trajectories, of length $0.1\,\mathrm{ns}$, which were generated by an Euler–Maruyama integrator using the numerical values of Tab. 2.1 with $\Delta t_{mic} = 10 \times 10^{-4}\,\mathrm{ns}$. For each n-th step, we considered $\Delta t = n\Delta t_{mic}$ and evaluated the average $(\overline{Y_{\Delta t} - Y_0})/\Delta t$ over all the M trajectories. For small values of Δt, comparable to Δt_{mic}, the resulting curves are too noisy and do not allow a solid determination of $D^{(1)}$. The reason lies in the evaluation of the stochastic number $\overline{\eta}$, which tends to zero as $\sim 1/M^{1/2}$. Inverting the first of Eqs. (2.12), one then finds that the condition for the noise to be negligible is

γ	$10 \times 10^{-11}\,\mathrm{kg\,s^{-1}}$ $(= 6 \times 10^{3}\,\mathrm{kDa\,ns^{-1}})$
k	$-1\,\mathrm{pN\,nm^{-1}}$
Y_C	$1\,\mathrm{nm}$
Y_0	$1.1\,\mathrm{nm}$
k_BT	$2.5\,\mathrm{kJ\,mol^{-1}}$
$D^{(1)}(Y)$	$-0.1\,\mathrm{nm\,ns^{-1}}$
$D^{(2)}(Y)$	$0.4\,\mathrm{nm^2\,ns^{-1}}$

Table 2.1: List of parameters for the simulations of the harmonic–potential model.

$$M \gg \frac{2D^{(2)}}{[D^{(1)}]^2\Delta t}, \tag{2.15}$$

which is stronger the smaller is Δt. As a consequence, the use of $\Delta t \approx \Delta t_{mic}$ is prevented. Indeed, the intrinsic time scale of a harmonic oscillator is $\tau_{int} = \gamma/k = Y_0/D^{(1)}$, and thus Δt_{mic} needs to be chosen several orders of magnitude smaller than this time scale: for instance, $h \cdot \tau_{int}$, with $h \sim 10^{-4}$. In such a scenario, the condition 2.15 on the noise becomes $M \gg 2D^{(2)}/hD^{(1)}Y_0$. Since in

typical systems the diffusion coefficient is comparable to the product of the drift times the value of Y, that is $D^{(2)} \sim D^{(1)}Y_0$, the final condition for the noise to be negligible is $M \gg 10^4$.
One could wonder whether the use of a larger Δt would reduce the stochastic noise, since doing so is equivalent to performing averages of an increased number of displacements. Indeed, a large time interval Δt can be considered as composed of m sub–intervals $\Delta t'$: the quantity $\overline{(Y_{\Delta t} - Y_0)}/\Delta t$ can be thus seen as an average over both the M trajectories and the m displacements in each trajectory, for a total of $M \cdot m$ terms. Unfortunately, the drawback relies in the fact that, in each trajectory, only the displacement in the first sub–interval starts exactly at Y_0, whereas the others are simply an approximation $Y_{n\Delta t'} \sim Y_0$, which of course becomes worse as m increases. Thus, for larger Δt, $D^{(1)}$ would be not a property of the specific point Y_0, but it would turn into the average of the drift coefficients over a number of points in the neighborhood of Y_0, number which can be higher as a greater value of Δt is chosen. As shown in Fig. 2.1, at large values of Δt the calculated drift coefficient, despite being less noisy than in the limit $\to 0$, departs from its true value $-0.1\,\mathrm{nm\,ns^{-1}}$. Practically speaking, the use of a larger Δt translates in trading a stochastic error for a systematic error and, as a consequence, the direct use of Eq. (2.12) is impractical.

2.2.3 The diffusion coefficient

If the drift coefficient cannot be easily recovered with the first of Eqs. (2.12), the situation is rather different for the diffusion coefficient. Indeed, the expression for $D^{(2)}$ includes three terms: the one in the middle is a deterministic correction depending on the square of $D^{(1)}$ which, as previously discussed, cannot be estimated. However, it could be neglected provided that

$$[D^{(1)}]^2 \ll \frac{2D^{(2)}}{\Delta t}. \tag{2.16}$$

Inserting in 2.16 the values listed above and using $\Delta t = 10 \times 10^{-4}\,\mathrm{ns}$, the condition reads $|D^{(1)}| \ll 90\,\mathrm{nm\,ns^{-1}}$, corresponding to forces approximately of the order of $10 \times 10^3\,\mathrm{pN}$, which are huge on a biological scale. Thus, the deterministic correction can be neglected in most cases we are interested in. The other source of error for the evaluation of $D^{(2)}$ is the random noise, which appears explicitly as the stochastic correction, third term of Eqs. (2.12). However, it goes to zero as $\Delta t M^{-1/2}$ and consequently it is negligible for large M even at finite Δt.
We have tested the calculation of $D^{(2)}$ from Eq. (2.12), assuming $\overline{\eta} = 0$, in two test cases: that of the harmonic spring, already considered above, and a more

complicated one which shall be discussed later. In the case of the harmonic potential, from simulations of length $0.2\,\mathrm{ns}$ and using the same parameters of Tab. 2.1, we calculated the quantity $\overline{[Y_{\Delta t} - Y_0]^2}/(2\Delta t)$ and the results are displayed in the upper panels of Fig. 2.2 . In Fig. 2.2 A is shown $D^{(2)}$, as a function of Δt, which is linear at $0.4\,\mathrm{nm}^2\,\mathrm{ns}^{-1}$ with an error smaller than 1% as expected from Eq. (2.12), provided that $D^{(1)}$ is not too large as suggested by Eq. (2.16). In fact, we display in Fig. 2.2 B the accuracy in the back–calculation of $D^{(2)}$ with the procedure described above. Indeed, results are good if the value of $D^{(1)}$ is less than $\sim 10\,\mathrm{nm}\,\mathrm{ns}^{-1}$, while the prediction becomes unreliable for larger drifts. Substantially, the condition in Eq. (2.16) provides a rule to estimate a priori whether the correction of the drift on the diffusion coefficient is important rather than not. Mathematically, it is a threshold curve which we displayed with an orange, dashed line in the same Fig. 2.2 B and which fits perfectly the separation line between the regions where the error is lower than $30 - 40\%$ and those where the error is higher. The maximum value of $D^{(1)}$ that allows the calculation of $D^{(2)}$ is for $\Delta t = \Delta t_{mic}$, and in this example assumes the value $D^{(1)} \approx 20\,\mathrm{nm}\,\mathrm{ns}^{-1}$. This drift corresponds to a force $\gamma D^{(1)} \sim 200\,\mathrm{pN}$, which is large with respect to the typical orders of magnitude of biomolecules. Consequently, we conclude that the use of $\Delta t = \Delta t_{mic}$ seems to be the best choice for the recovery of $D^{(2)}$, which is interesting because for $D^{(1)}$ the same choice entails very noisy results. We repeated the above procedure, starting from different initial points Y_0s and using $\Delta t = \Delta t_{mic}$ ($10 \times 10^{-4}\,\mathrm{ns}$) to back–calculate the values of $D^{(2)}(Y)$ along the Collective Variable Y. The results are shown as solid circles in Fig. 2.2 C, superposed to the curve which defines $D^{(2)}_{true} = 0.4\,\mathrm{nm}^2\,\mathrm{ns}^{-1}$. Once again, results are good only up to the value of Y_0 where $D^{(1)}(Y_0)_{true} \sim 10\,\mathrm{nm}\,\mathrm{ns}^{-1}$ (cfr. the purple curve in the same Fig.).

We carried out a second, more challenging one–dimensional test in order to better understand whether the ease of recovery of the diffusion coefficient was due to the simple mathematical form of $D^{(1)}(Y)$ and $D^{(2)}(Y)$ chosen for the previous model. The new test system was bistable, as can be seen in the purple curve in Fig. 2.2 F and had a sinusoidal diffusion coefficient, showed in the green curve in the same figure; the functional forms of the potential energy, of $D^{(1)}(Y)$ and of $D^{(2)}(Y)$ were

$$\begin{aligned}
U(Y) =& \frac{1}{4}aY^4 - \frac{1}{3}a(M + m_1 + m_2)Y^3 + \qquad (2.17)\\
&+ \frac{1}{2}a(Mm_1 + Mm_2 + m_1m_2)Y^2 + (c - aMm_1m_2)Y + cost\\
D^{(1)}(Y) =& aY^3 - a(M + m_1 + m_2)Y^2 + a(Mm_1 + Mm_2 + m_1m_2)Y + c - aMm_1m_2\\
D^{(2)}(Y) =& Asin(BY) + C
\end{aligned}$$

where the dependence on all the coefficients has been kept visible in order to highlight the meaning of each of them: m_1 and m_2 are the positions, in nanometers, of the two minima; M is the position of the barrier; a and c are two coefficients whose value has been found imposing the positions M, m_1 and m_2 and the heights of the barriers $H_1 = U(M) - U(m_1)$ and $H_2 = U(M) - U(m_2)$ (see Tab. 2.2). These shapes do not have direct relevance to describe any specific biological system, but they were chosen only because of their high complexity, following the idea that if the algorithm works for this case, it will work also for simpler, biologically–inspired force fields and diffusion coefficients. A time step $\Delta t_{mic} = 10 \times 10^{-6}\,\mathrm{ns}$ was used to generate the trajectories of length $2 \times 10^{-3}\,\mathrm{ns}$, for a collection of starting points Y_0s in the position range $0\,\mathrm{nm}$ to $5\,\mathrm{nm}$. The results are similar to the case of the harmonic spring, that is one can back–calculate the diffusion coefficient $D^{(2)}$ with a good accuracy (see Figs. 2.2 D, E and F), provided that the true drift coefficient is small enough that the system can diffuse within the time Δt. As a final note, we stress that here, in order to integrate the equation of motion and not to allow the simulation to explode, we needed to use a Δt_{mic} which is 2 orders of magnitude smaller than that used in the harmonic–potential system. This requirement, which a priori can be considered inconvenient as usually the bigger Δt_{mic} the better, has led to an improved recovery of $D^{(2)}(Y)$ with respect to the previous case because of the deterministic correction in 2.12 proportional to Δt. Indeed, for $\Delta t \sim \Delta t_{mic}$, the accuracy of the reconstruction is much higher than before (cfr. 2.2 E and B for comparison); also in the profiles in Figs 2.2 F and C one can notice that the black points start to depart from their true, expected values in different regimes: when $D^{(1)}(Y) \gtrsim 10\,\mathrm{nm\,ns^{-1}}$ in the harmonic–potential system and

a	$15\,873\,\mathrm{kJ\,nm^{-4}}$
c	$6.6666\,\mathrm{kJ\,nm^{-1}}$
m_1	$0.85\,\mathrm{nm}$
m_2	$1.15\,\mathrm{nm}$
M	$1\,\mathrm{nm}$
H_1	$3\,\mathrm{kj\,mol^{-1}}$
H_2	$1\,\mathrm{kj\,mol^{-1}}$
A	$2\,\mathrm{nm^2\,ns^{-1}}$
B	$5\,\mathrm{nm^{-1}}$
C	$1.6\,\mathrm{nm^2\,ns^{-1}}$
k_BT	$2.5\,\mathrm{kJ\,mol^{-1}}$

Table 2.2: List of parameters for the simulations of the bistable–potential model.

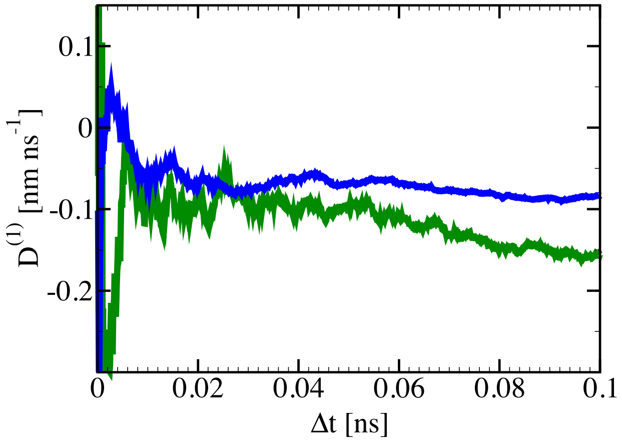


Figure 2.1: The value of $D^{(1)}$ as a function of Δt for the harmonic potential, calculated with Eq. (2.12) making use of different numbers M of trajectories. The green curve corresponds to $M = 10^4$, the blue curve to $M = 10^5$. The true value of $D^{(1)}$ is $-0.1\,\mathrm{nm\,ns}^{-1}$.

when $D^{(1)}(Y) \gtrsim 100\,\mathrm{nm\,ns}^{-1}$ in the bistable system.

2.2.4 Higher Kramers–Moyal coefficients

From the Langevin–like equation (2.8) one can also find the higher moments, like

$$\overline{(Y_{\Delta t} - Y_0)^3} - (D^{(1)}\Delta t)^3 - 6D^{(1)}D^{(2)}\Delta t^2 = z \cdot \overline{\eta} + w \cdot \overline{\eta^3}, \tag{2.18}$$

where $z = \sqrt{18[D^{(1)}]^4 D^{(2)} \Delta t^5}$ and $w = \sqrt{[2D^{(2)}\Delta t]^3}$. In the limit of small Δt, the left–hand side of this equation corresponds to the third Kramers–Moyal coefficient multiplied by Δt, so by extension we will label it as $D^{(3)}\Delta t$. In fact, combining Eq. (2.18) with the properties of $\overline{\eta}$ it follows that $D^{(3)}$ should vanish as $z/M^{1/2}$ for Eq (2.8) to hold. If this is the case, all higher moments should vanish by Pawula's theorem [81]. The validity of Eq. (2.18) is difficult to establish numerically, because of the problems already discussed in estimating $D^{(1)}$. Consequently, we only checked the condition

$$\overline{(Y_{\Delta t} - Y_0)^3} \ll \sqrt{[D^{(2)}\Delta t]^3} \quad (= w) \tag{2.19}$$

which means that the skewness of the displacements is negligible with respect to the diffusion in each run for small Δt. With the typical values of the diffusion coefficients we chose, $w \sim 10^{-7}\,\mathrm{nm}^3$. This condition is satisfied both in the case of the particle in the harmonic well and of the bistable system, except for extreme choices of the drift coefficient (cfr. Fig. 2.3).

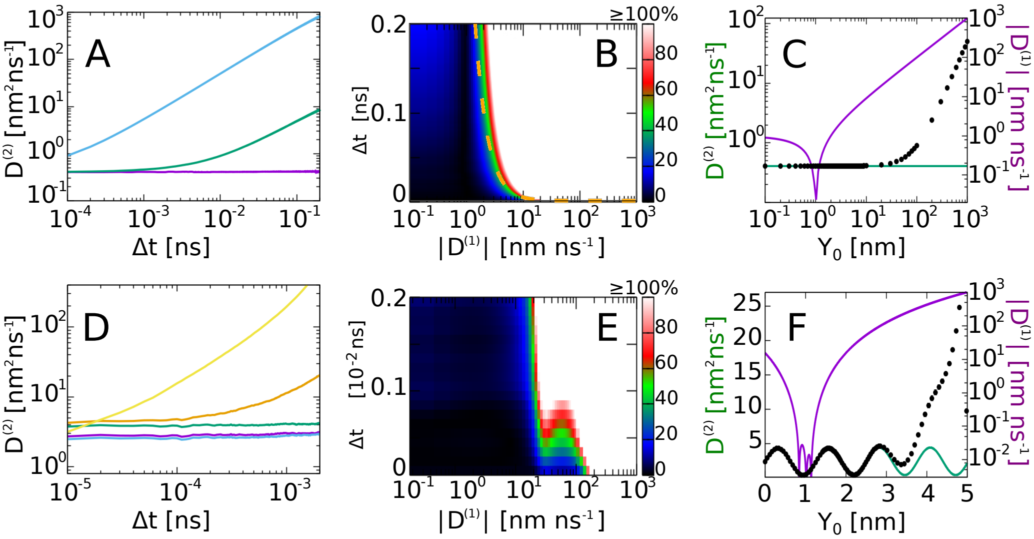
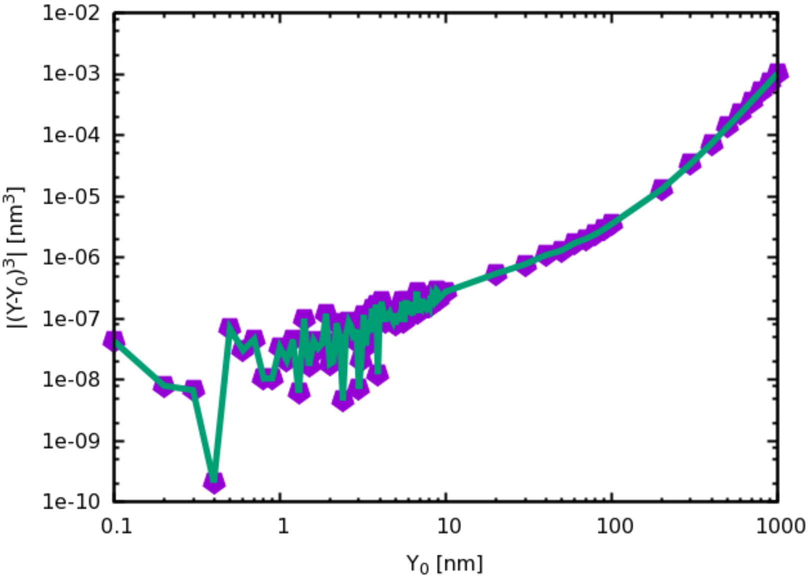


Figure 2.2: For the harmonic spring, (A) the back–calculated $D^{(2)}$ as a function of Δt in the cases $D^{(1)}_{true} = 1\,\mathrm{nm\,ns^{-1}}$ (purple line), $D^{(1)}_{true} = 10\,\mathrm{nm\,ns^{-1}}$ (green line) and $D^{(1)}_{true} = 10 \times 10^2\,\mathrm{nm\,ns^{-1}}$ (blue line). (B) The percentage error in the back–calculation with respect to the true value; the orange,dashed curve indicates the threshold given by Eq. (2.16). (C) The values of $D^{(1)}_{true}$ (purple curve) and $D^{(2)}_{true}$ (green curve) as a function of the elongation of the spring; the solid circles indicate the back–calculated profile of $D^{(1)}$ using $\Delta t = \Delta t_{mic}$. (D), (E) and (F) are the same as (A), (B) and (C), respectively, for the system displaying a two–state thermodynamics and a sinusoidal diffusion coefficient (see panel F).



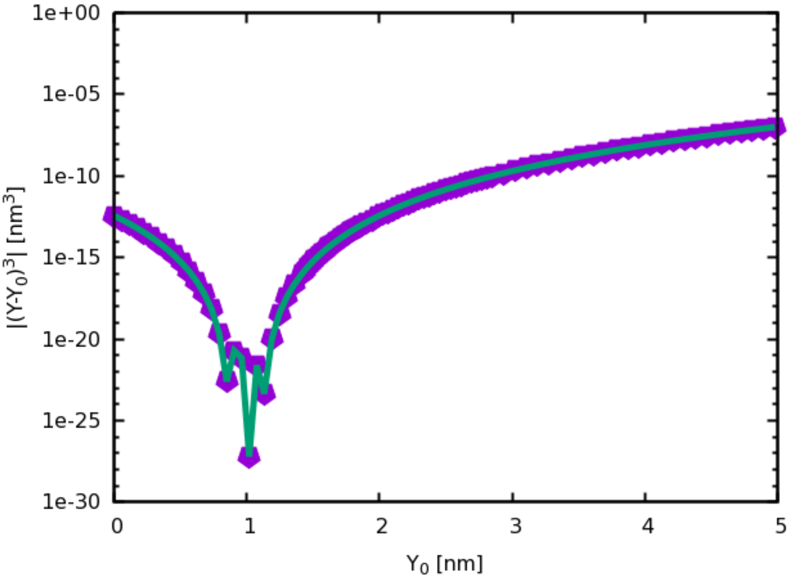


Figure 2.3: The value of the third moment of the displacement, calculated for the harmonic spring (above) and for the two–state system with sinusoidal diffusion coefficient (below).

2.3 Another strategy to calculate Drift and Diffusion coefficients

2.3.1 Drift and diffusion coefficients from an iterative approach

The strategy derived in the previous section did not provide a reliable way to recover both the drift and the diffusion coefficients at the same time. Instead of trying to evaluate more efficiently all the deterministic or stochastic corrections to the definitions of $D^{(1)}$ and $D^{(2)}$, we searched for an alternative way to calculate them. We then performed the following approximations: locally around each starting point Y_0 in the conformational space, the effective force acting on it is considered as depending linearly on Y, at least at the first order, while the diffusion coefficient is considered constant. Indeed, we expect that the underlying properties of the phase space do not really change much in the surroundings of each point Y_0, that is along the interval $(Y_0 - \varepsilon,\ Y_0 + \varepsilon)$. A constant diffusion coefficient reflects the hypothesis that the friction coefficient γ does not vary along the interval, namely that the average collisions occurring at the point $Y_0 - \varepsilon$ are the same occurring at the point $Y_0 + \varepsilon$, which sounds reasonable provided that ε is small enough and that Y is a smooth function of the cartesian coordinates $\mathbf{r}$. On the other side, regarding the drift coefficient, if we considered it constant the system would not migrate from the starting point Y_0, at least not in a statistically relevant way (there would be only free diffusion around that point). Then, a dependence on Y is needed to make the system evolve in time not only undergoing the effect of the diffusion, but with a real, effective force acting on it. The simplest, linear dependence on Y is what we chose as first approximation; this also corresponds to approximate the local, energetic landscape to a parabola, since the constant diffusion coefficient gives no contribution in the Eq. 2.9. Writing the force locally as $f(Y) = k(Y - Y_C)$, the drift coefficient becomes

$$D^{(1)}(Y) = \rho(Y - Y_C), \tag{2.20}$$

where $\rho = k/\gamma$. Defining $B = 1 + \rho\Delta t$, the Euler–Maruyama version of Langevin equation can thus be iterated to give

$$\begin{aligned}
Y_{\Delta t} &= BY_0 - (B-1)Y_C + (2D^{(2)}\Delta t)^{1/2}\eta_0; && (2.21)\\
Y_{2\Delta t} &= BY_{\Delta t} - (B-1)Y_C + (2D^{(2)}\Delta t)^{1/2}\eta_{\Delta t} = \\
&= B^2Y_0 - (B-1)(B+1)Y_C + (2D^{(2)}\Delta t)^{1/2}(B\eta_0 + \eta_{\Delta t}); \\
Y_{3\Delta t} &= B^3Y_0 - (B-1)(B^2+B+1)Y_C + && (2.22)\\
&\qquad + (2D^{(2)}\Delta t)^{1/2}(B^2\eta_0 + B\eta_{\Delta t} + \eta_{2\Delta t}); \\
&\dots && (2.23)
\end{aligned}$$

which can be generalized to n steps by performing the geometric sum as

$$\begin{aligned} Y_{n\Delta t} - Y_0 = (B^n - 1)(Y_0 - Y_C) + \\ + (2D^{(2)}\Delta t)^{1/2} \sum_i^{n-1} B^i \eta_{(n-i-1)\Delta t}. \end{aligned} \tag{2.24}$$

The drift coefficient at point Y_0, or better the two parameters ρ and Y_C which define it, can be found from the average displacement of Y, that is

$$\begin{aligned} \frac{\overline{Y_{n\Delta t} - Y_0}}{n\Delta t} = \frac{(1+\rho\Delta t)^n - 1}{n\Delta t}(Y_0 - Y_C) + \\ + \left(\frac{2D^{(2)}}{\Delta t}\right)^{1/2} \frac{1-(\rho\Delta t)^n}{1-\rho\Delta t}\frac{1}{n}\sum_{i=0}^{n-1}\overline{\eta}_i. \end{aligned} \tag{2.25}$$

The last term, containing the average $\overline{\eta}$, is expected to be negligible for essentially two reasons: initially, because its standard deviation goes to zero as $(M \times n)^{-1/2}$, in contrast with Eq. (2.12) where n was absent due to the different approach. Moreover, even if Δt is small, the prefactor $[D^{(2)}/\Delta t]^{1/2}$ is much smaller than in Eq. (2.12). For instance, let's consider $D^{(2)} \sim D^{(1)} Y_0$ and $\Delta t \sim 10^{-4} Y_0 / D^{(1)}$: the noise term results only of the order of $10^2 \cdot (M \times n)^{-1/2}$. Since both M and n can be tuned at will, by using more simulations or by making them last longer, one has the freedom to minimize the stochastic contribution up to the degree he or she wishes. We stress that, incidentally, the neglection of $\overline{\eta}$ allows Eq. (2.25) to converge, in the limit $n\Delta t \to 0$, to $\rho(Y_0 - Y_C)$, which is exactly what expected in a standard Langevin differential equation framework but whose validity was not granted in a finite–differences scheme.
Thus, neglecting $\overline{\eta}$ in Eq. (2.25) and defining $\tau = n\Delta t$ one can write

$$K(\tau) \equiv \overline{Y_\tau - Y_0} = [(1+\rho\Delta t)^{\tau/\Delta t} - 1](Y_0 - Y_C). \tag{2.26}$$

In this expression, in contrast with the previous approach we have a function of elapsed time τ at the left–hand side, whereas before we simply had a (unknown) number which was the result of the evaluation of a limit on the right–hand side, where conversely now there is an expression depending parametrically on ρ, Y_C and Δt. Similarly to the calculation of the drift coefficient in Eq. (2.25) it is possible to use the iterative procedure described above to obtain the diffusion coefficient. In fact from Eq. (2.24) one can obtain

$$\overline{(Y_\tau - Y_0)^2} = \left[(1+\rho\Delta t)^{\tau/\Delta t} - 1\right]^2 (Y_0 - Y_c)^2 - 2D^{(2)} \frac{1-(1+\rho\Delta t)^{2\tau/\Delta t}}{\rho(2+\rho\Delta t)}, \tag{2.27}$$

which can be combined with Eq. (2.26) to give

$$J(\tau) \equiv \overline{[Y_\tau - Y_0]^2} - \left(\overline{Y_\tau - Y_0}\right)^2 = 2D^{(2)} \frac{(1+\rho\Delta t)^{2\tau/\Delta t} - 1}{\rho(2+\rho\Delta t)}, \tag{2.28}$$

which depends parametrically on $D^{(2)}$, Δt and ρ. More details on the derivation of Eqs. (2.26) and (2.28) can be found in App. C .

2.3.2 Determination of the parameters

The idea is thus to fit the numerical values of $K(\tau)$ and $J(\tau)$, which can be extracted from the Molecular Dynamics simulations previously generated, with the expressions of Eqs. (2.26) and (2.28) in order to obtain the four parameters ρ, Y_C (which completely define $D^{(1)}$), $D^{(2)}$ and, in principle, also Δt. However, this quantity Δt sets the time scale of the time–dependent parameters. In fact, Eqs. (2.26) and (2.28) are invariant under the transformations $\rho \to \rho\Delta t$, $\tau \to \tau/\Delta t$ and $D^{(2)} \to D^{(2)}\Delta t$. As a consequence, the choice of Δt results in not being too critical, provided that it is small enough to allow a high–resolution determination of $J(\tau)$ and $K(\tau)$. Indeed, in Fig. (2.4) the accuracy in the reconstruction of $D^{(1)}$ and $D^{(2)}$ depends very poorly on Δt. Empirically, we find that the choice $\Delta t = 10\Delta t_{mic}$ provides a good balance between the resolution of $K(\tau)$ and $J(\tau)$ and the feasibility of the fit.
Since the shape of $J(\tau)$ is usually curved, as can be seen in Figs. 2.5 A and 2.6 A, we expect $J(\tau)$ to be completely specified by – at least – three different parameters. Because the theoretical expression for $J(\tau)$, Eq. (2.28), depends on exactly three parameters, we then expect to be able to fit all of them without running into overfitting. In particular, we expect to obtain ρ and $D^{(2)}$ from it. On the other hand, since Eq. (2.26) can be written as $\log K = \log(Y_0 - Y_C) + \log(...)$, we can then fit Y_C from the small–τ region of $\log K(\tau)$, which is generally rather flat (see dashed curves in Figs. 2.5 A and 2.6 A).

2.3.3 Results for the test models

In Fig. 2.5 we have reported the back–calculationf of both the drift coefficient $D^{(1)}$ and the diffusion coefficient $D^{(2)}$, for the test case of a one–dimensional particle in a harmonic potential. We stress that, in spite of the simplicity, $D^{(1)}$ could not be recovered directly by the definition (see Fig. 2.1), whereas $D^{(2)}$ was subject to a bias whose strength was a priori completely unknown (see green curves and solid circles in Fig. 2.2 C and F). In Fig. 2.5 B it is shown the percentage standard error in the determination of $D^{(1)}$; the error is lower than 10% in all cases and decreases for larger values of $D^{(1)}$, where the drift dominates over diffusion.

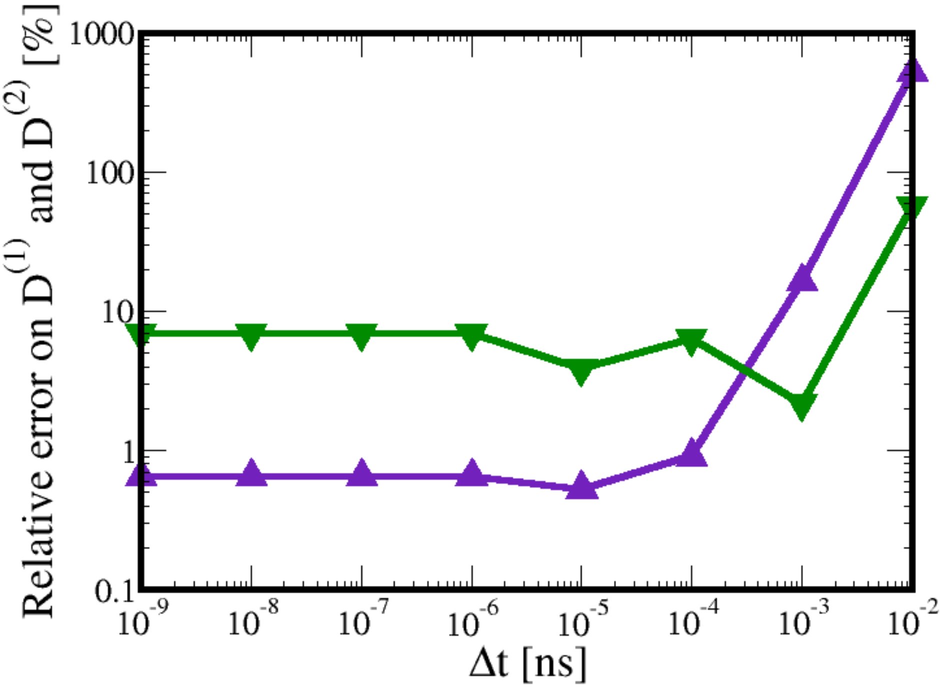


Figure 2.4: The quality of the reconstruction of $D^{(1)}$ (purple triangles) and $D^{(2)}$ (green triangles) poorly depends on Δt, provided that the two functions $J(\tau)$ and $K(\tau)$ have sufficient time resolution to be fitted.

Fig. 2.5 C shows the percentage standard error in the back–calculation of $D^{(2)}$: the error is less than 1%. Good results were obtained calculating $D^{(2)}$ by its definition (see Fig. 2.2) at small values of $D^{(1)}_{true}$; the present method extends those results to any biologically–relevant value of $D^{(1)}_{true}$. The overall reconstruction of the profile of $D^{(2)}$ and $D^{(2)}$ is displayed in Fig. 2.5 D.

Similar results were obtained for the more challenging system displaying a two–state thermodynamics and a sinusoidal diffusion coefficient (see lower panels of Fig. 2.2). In Fig. 2.6 we report the fits (panel A), the percentage standard error on $D^{(1)}$ (panel B) and $D^{(2)}$ (panel C), and the reconstruction of the profile of the two coefficients (panel D). Overall, $D^{(1)}$ can now be calculated with good accuracy, and $D^{(2)}$ with a better accuracy than using Eq. (2.12).

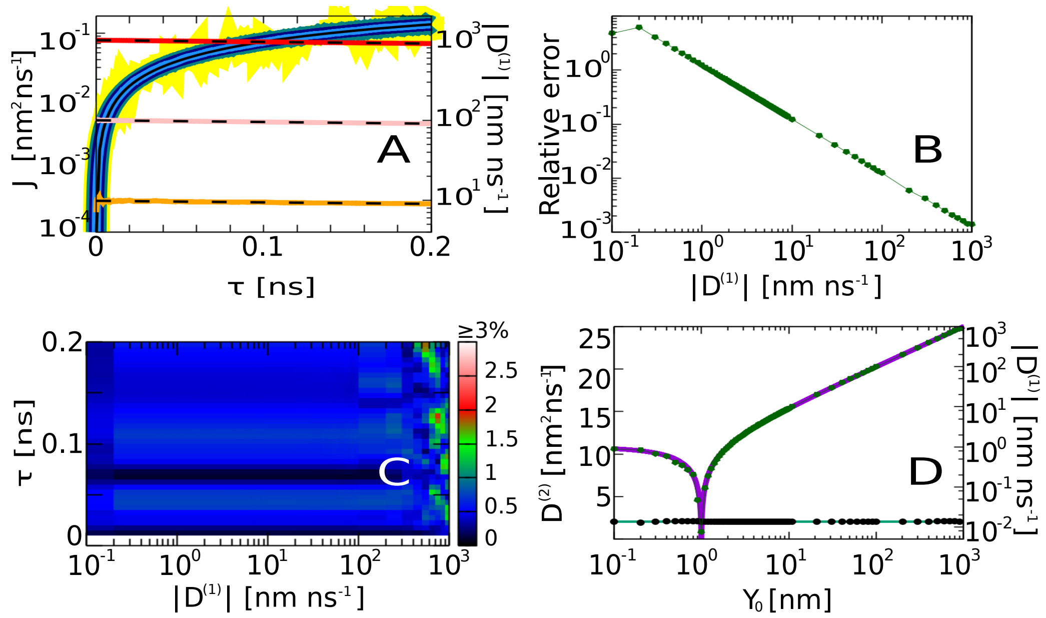


Figure 2.5: (A) The values J (solid black curves) and $|D^{(1)}|$ (dashed black curves) as a function of τ for the spring chain. The colored curves are the fits by Eqs. (2.26) and (2.28), respectively. (B) The percentage standard error in the determination of $D^{(1)}$. (C) The percentage standard error in the determination of $D^{(2)}$ as a function of τ and $D^{(1)}_{true}$. (D) the reconstruction (green and black dots, respectively) of the true behavior of $D^{(1)}$ and $D^{(2)}$ (solid curves).

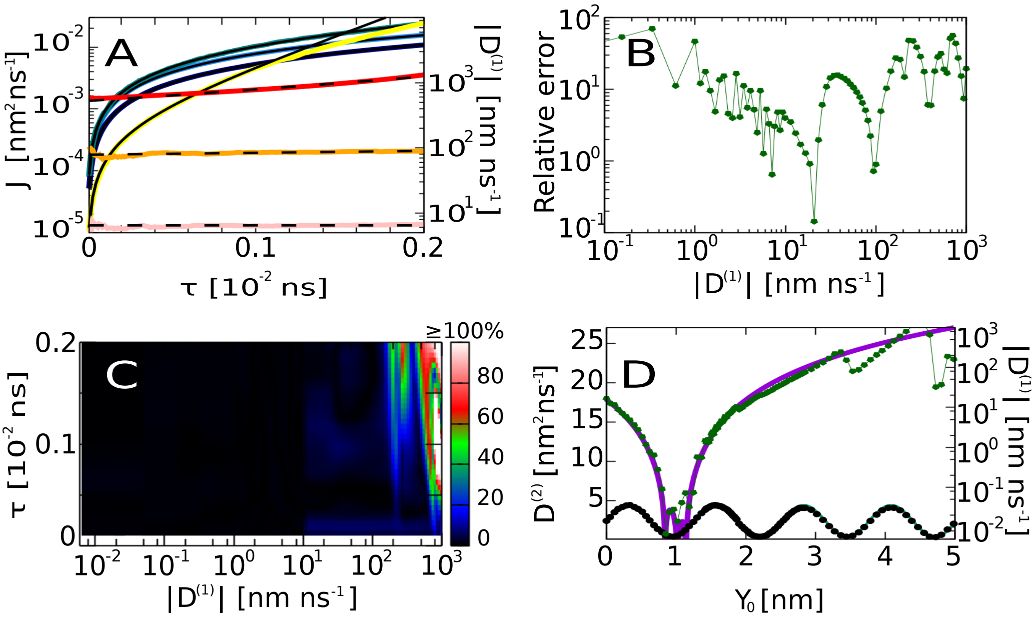


Figure 2.6: The same as Fig. 2.5 for a system displaying a two–state thermodynamics and a sinusoidal diffusion coefficient (cf. lower panels of Fig. 2.2).

2.4 Evaluation of the properties of Collective Variables

2.4.1 Application of the algorithm on the Collective Variables $dRMSD$ and q

Finally, the goal of the present section is an inspection on the properties of some Collective Variables commonly used to describe the dynamics of protein models. More specifically, we want to understand whether the dynamics of the system, projected from the high–dimensional phase space to the low–dimensional one, can be described by Eq. (2.8). A necessary requirement is that the Collective Variable needs to identify uniquely $D^{(1)}$ and $D^{(2)}$, which cannot depend on the microscopic coordinates (as stated in Eq. (2.7), which we report here for ease of reading)

$$\text{if } \exists \mathbf{r}_0, \mathbf{r}_1 : Y(\mathbf{r}_0) = Y(\mathbf{r}_1) = \overline{Y} \longrightarrow D^{(1-2)}(Y(\mathbf{r}_0)) = D^{(1-2)}(Y(\mathbf{r}_1)) = D^{(1-2)}(\overline{Y}) \tag{2.29}$$

Moreover, the Kramers–Moyal coefficients of the Collective Variable need to be zero for $n \geq\ 3$.

Operatively, to challenge a specific Collective Variable Y, we start by generating a set of microscopic conformations $\mathcal{Z} \equiv \{\mathbf{r}\}$. We then choose a value of Y we want to investigate, that is Y_0, and keep from the set $\mathcal{Z}$ only that set of conformations $\mathcal{S} \equiv \{\mathbf{r}'\}$ ($\mathcal{S} \subseteqq \mathcal{Z}$) satisfying $Y(\mathbf{r}') = Y_0$, or at least which are in the interval $[Y_0 - \varepsilon, Y_0 + \varepsilon]$ (we shall discuss more on ε later). From each conformation $\mathbf{r}'$ in $\mathcal{S}$, we extract $D^{(1)}(Y_0)$, $D^{(2)}(Y_0)$ and also $D^{(3)}(Y_0)$. The collection of all the resulting $D^{(1)}(Y_0)$, $D^{(2)}(Y_0)$ and $D^{(3)}(Y_0)$ will provide their distribution, associated with that set $\mathcal{S}$. If Y is a "good" reaction coordinate, namely if the condition 2.7 is met, such distributions should strongly peaked, identifying a single value for $D^{(1)}$, $D^{(2)}$ and should be identically zero for $D^{(3)}$. Therefore, the standard deviations σ_1 and σ_2 of the distributions of $D^{(1)}$ and $D^{(2)}$, respectively, and the root mean square difference σ_3 from zero of $D^{(3)}$ can be regarded as measures of the quality of the Collective Variable as reaction coordinate. An ideal reaction coordinate should display $\sigma_1 = \sigma_2 = \sigma_3 = 0$.

It is reasonable to assume that the quality of Collective Variables could be dependent on the kind of tridimensional structure possessed by the system under analysis. Consequently, we analyze separately the two basic structural units of proteins, that is α–helices and β–hairpins. In particular, we choose as study system the second hairpin of protein–G B1 domain and its helix, which were extracted from the Protein Data Bank reference structure 1PGB [87]. In addition, we use a structure–based, implicit–solvent potential [88] in order to make calculations particularly fast as we shall have to perform dozens of thousands of short simulations. The functional form of the potential is a sum of atom–atom interaction terms, all shaped in a Lennard–Jones fashion, where at the native distance of each pair of atoms the energy has its minimum -1 (in arbitrary energy units). We generated all the trajectories using Gromacs 5.1 ([64]), at a low temperature $T = 0.92$ (in energy units) at which the native conformations are stable; the post–processing of data has been performed with Plumed 2.3 ([63]).

We chose to focus our efforts on two Collective Variables, which are popular in the research field of protein folding and in general of the computational, structural biology. The first one is the distance–RMSD ($dRMSD$) of a generic conformation $\mathbf{r}^t$ to the native conformation $\mathbf{r}^N$

$$dRMSD(\mathbf{r}^t, \mathbf{r}^N) = \left[\frac{1}{N(N-1)} \sum_{ij, i \neq j} \left(|\mathbf{r}_i^t - \mathbf{r}_j^t| - |\mathbf{r}_i^N - \mathbf{r}_j^N| \right)^2 \right]^{1/2} \tag{2.30}$$

which is a sort of metric in the cartesian space, used to discriminate between conformations globally similar to the reference structure (where $dRMSD$ is lower, ideally zero in the reference structure itself) to those radically different (where $dRMSD$ is higher). In Eq. (2.30), the sum is carried out on all the pairs of atoms $\{i, j\}$ belonging to the structure $\mathbf{r}^t$ (or $\mathbf{r}^N$, as of course they must be compatible); N is the total number of atoms. The second Collective Variable is the fraction of native contacts q, defined as follows

$$q = \frac{1}{W} \sum_{ij \in C}^{W} s(d_{ij}) \qquad \text{with} \qquad s(d) = \begin{cases} 1 & \text{if } d \leq d_0 \\ 1 - \tanh\left(\frac{d - d_0}{r_0}\right) & \text{otherwise} \end{cases} \tag{2.31}$$

where C is the set of all the pairs of (non–hydrogen) atoms displaying a distance $d_{ij} < 5$ Å in the native conformation, provided that the distance in the sequence between the respective aminoacids was $|a_i - a_j| \geq 4$, in order to identify only the relevant, non–trivial native contacts. The sum is then performed on all the W pairs of atoms i and j belonging to this set C. All the parameters used in the definition of the function $s(d_{ij})$ are listed in Tab. 2.3 , along with the total number of native contacts found for the α–helix and the β–hairpin.

d_0	5 Å
r_0	1 Å
W_α	149
W_β	244

Table 2.3: List of parameters for the definition of the fraction of native contacts q.

To calculate $D^{(1)}$ and $D^{(2)}$, at first we generated a long trajectory, at $T = 0.92$ for both the α–helix and the β–hairpin, consisting in $2 \cdot 10^4$ frames each. We then labelled each frame with its own value of $dRMSD$ and of q. Eventually, from all the frames we extracted five subsets of 500 conformations each, whose $dRMSD$ lied in the ranges

- $[0.20, 0.21]\,\mathrm{nm}$
- $[0.25, 0.26]\,\mathrm{nm}$
- $[0.30, 0.31]\,\mathrm{nm}$
- $[0.35, 0.36]\,\mathrm{nm}$
- $[0.40, 0.41]\,\mathrm{nm}$

and other five subsets with q in the ranges

- $[0.10, 0.13]$
- $[0.30, 0.33]$
- $[0.50, 0.53]$
- $[0.70, 0.73]$
- $[0.90, 0.93]$

The width of intervals, $0.01\,\mathrm{nm}$ for $dRMSD$ and 0.03 for q, was chosen in the following way: we needed it to be small enough to let conformations to be rather similar, if belonging to the same interval, but large enough to have a statistically sound number of different frames, in order to build an histogram with the resulting data. Every frame was the starting conformation Y_0 for $M = 300$ independent simulations, in which the corresponding Collective Variable Y was monitored. From a single set of M simulations, the functions $K(\tau)$ and $J(\tau)$ could be drawn and fitted and the values of $D^{(1)}$ and $D^{(2)}$ were calculated, according to Eqs. (2.20), (2.26) and (2.28).

The resulting distributions of coefficients $D^{(1)}$ and $D^{(2)}$, calculated for the α–helix and the β–hairpin and then normalized over the 500 frames, are displayed in Figs. 2.7 and 2.8, respectively. The associated means and standard deviations are displayed in Figs. 2.9 A–D and 2.10 A–D, respectively. In the following lines we will refer essentially to latters, for ease of reading data. By looking to the case of the $dRMSD$, it seems that both the α–helix and the β–hairpin possess distributions of $D^{(1)}$ which are significantly wide, standard deviations being several times the value of the mean as can be seen in Figs. 2.9 A and 2.10 A. Whereas the mean is negative for all the values of the $dRMSD$, the monotonicity is different in the two systems, being decreasing for the α–helix and approximately flat in the case of the β–hairpin. The absence of a changement in the sign of $D^{(1)}$ is expected, as it suggests that all the chosen values Y_0s are at the same side (on the right in this case, because of the negative value of $D^{(1)}$) of an energy minimum, which can be common for all Y_0s or not. Indeed, the real, global minimum of both the α–helix and the β–hairpin lies roughly at $dRMSD \approx 0.1\,\mathrm{nm}$. For what concerns the diffusion coefficient instead, $D^{(2)}$ is better defined by the $dRMSD$ than $D^{(1)}$, as standard deviations are now only of the order of circa the half of the mean. Its value is lowest at low $dRMSD$, where the conformational space is narrower, and is always increasing in the case of the α–helix, while it slightly decreases at large $dRMSD$ in the case of the β–hairpin (see Figs. 2.9 C and 2.10 C) Interestingly, standard deviations of both $D^{(1)}$ and $D^{(2)}$ seem to be rather independent on the value of $dRMSD$.

On the other side, the fraction of native contacts q shows a drift coefficient $D^{(1)}$ whose standard deviation is of the order of the mean, in great contrast with the previous finding regarding the $dRMSD$. It is positive, except for at $q \approx 0.9$, indicating that the equilibrium state has possibly $0.7 < q < 0.9$ for both the α–helix and the β–hairpin (see Figs. 2.9 B and 2.10 B). As in the case of the $dRMSD$, the diffusion coefficient of the α–helix decreases monotonically to the native state, as shown in Fig. 2.9 D, while it has a bell–shaped behavior for the β–hairpin, as can be seen in Fig. 2.10 D. Overall, the standard deviation of $D^{(2)}$, relative to its mean, is comparable with that previously estimated of the $dRMSD$ but, in contrast with this, it shows a weak decrease when Y moves towards the native state. Finally, we stress that, following the criterion derived in Eq. (2.19), we find that the value of $D^{(3)}$ can be considered negligible, as shown in Figs. 2.11 and 2.12 .

2.4.2 Effective force and dynamics in reduced dimension

As previously reported in Eqs. (2.9) and (2.10), if a Collective Variable Y evolves in time according to the Eq. (2.1), then the effective force exerted on the system is composed by two terms: the one opposite to the gradient of the free energy F, calculated as a function of the Y, and the one proportional to the gradient of the diffusion coefficient, provided that $D^{(1)}$ effectively depends on Y [81].

Even though the approximation we performed considered $D^{(2)}$ as constant, providing then a theoretical zero contribution to the calculation of the drift $D^{(1)}$, in principle there is no reason to maintain the same approximation when it comes to analyze the results and looking at them from a global point of view. More specifically, in the approximation we considered as locally constant the value of $D^{(2)}(Y_0)$ in the surroundings of each conformation $\mathbf{r}$ whose value is Y_0. When we project the data orthogonally with respect to the Collective Variable Y, that is when we create the graphs such as those in Figs. 2.9 A–D and 2.10 A–D, all the information about the specific values Y_0s is collected at the same time, for each Y_0. This means that every point Y_0 on Figs. 2.9 A–D and 2.10 A–D gathers data generated with different values of $D^{(2)}$ and, for this reason, a priori one cannot consider $D^{(2)}$ as to be really constant over the space.

However, in the following we shall recover the effective force using Eq. (2.9) by neglecting the second term, that proportional to $\partial D^{(2)}(Y)/\partial Y$, for mainly two reasons. The first is due to the low order of magnitude of the gradient. Indeed, in all the panels 2.9 C and D, 2.10 C and D, the value of $D^{(2)}$ changes on a scale which is $\approx 10^{-1}$ times that of $D^{(1)}$. Let's take as reference example Fig. 2.10 D, which refers to the Collective Variable q: from the points 0.10 and 0.30 the value of $D^{(2)}(Y)$ increases from $\approx 3.5 \times 10^{-3}\,\mathrm{ns}^{-1}$ to $\approx 7 \times 10^{-3}\,\mathrm{ns}^{-1}$, resulting in $\Delta D^{(2)}(Y) \approx 3.5 \times 10^{-3}\,\mathrm{ns}^{-1}$. At the same time, $\Delta q = 0.2$. As a consequence, the value of the correction $1/2\,\partial D^{(2)}(Y)/\partial Y$, evaluated roughly as $\Delta D^{(2)}(Y)/\,(2\Delta q)$, is $8.75 \times 10^{-3}\,\mathrm{ns}^{-1}$ for the point $Y_0 = 0.10$ on the q axis. Simultaneously, the value of $D^{(1)}$ calculated with the algorithm is of the order $\approx 10^{-1}\,\mathrm{ns}^{-1}$ for the same point $Y_0 = 0.10$ (see Fig. 2.10 B). Basically, the correction accounts only for less than 10% of the total value of the drift $D^{(1)}$; consequently, it seems reasonable to neglect it (in App. B a similar argument was used to show that the time evolution of $p(Y)$ at a first approximation converges to the Boltzmann distribution). The second reason is more practical: the resolution of $D^{(2)}$ on the Y–axis is very poor in each of the Collective Variables under investigation, because we chose only five values of the Collective Variable to which apply the machinery, mainly to reduce the computational effort. Despite we do not expect $D^{(2)}$ to undergo huge variations in the points between those sampled by us, that is probably $D^{(2)}$ is smooth in each of the four intervals here defined, it seems a very crude approximation to evaluate the derivative on a so coarse–grained scale such that chosen in our analysis.

Therefore, in this context the drift $D^{(1)}$ exerted on the system at each point Y_0, is only proportional to the opposite of the gradient of the free energy F, calculated as a function of Y. In the formalism of Eq. (2.1), this term should be equivalent to an effective force $f = \gamma D^{(1)}$, where γ is the effective friction coefficient on the Collective Variable Y_0. We can further suppose that γ follows Einstein's relation $\gamma = T/D^{(2)}$, which was derived in the scenery of cartesian coordinates for what we should consider the Collective Variable "position", but which we extend also to this framework. A relevant question is then to which extent the effective force calculated with the algorithm is equal to the opposite gradient of the free energy F for the two variables under examination.

We then calculated the equilibrium free energies $F(dRMSD)$ and $F(q)$, for both the α–helix and the β–hairpin, by means of two replica–exchange simulations [28], using the same force–field adopted for the calculations of $D^{(1)}$ and $D^{(2)}$ [88]. They are displayed in the bottom panels of Figs. 2.9 and 2.10, for the α–helix and the β–hairpin, respectively. In Fig. 2.13 we compared the effective mean force $\gamma D^{(1)} = TD^{(1)}/D^{(2)}$ with the derivative $-\partial F(Y)\partial Y$ obtained from the numerical differentiation of the values displayed in Figs. 2.9 E and F, 2.10 E and F. In spite of the large standard deviation which affects the distributions of $D^{(1)}$ and $D^{(2)}$, as can be seen in Figs. 2.9 and 2.10, the two quantities are comparable, relative errors being in the worst case of the order of 50% of the force (in the case of the β–hairpin, see Fig. 2.13 C). The $dRMSD$ performs globally worse than q, especially in the case of the β–hairpin. Moreover, in three cases the two quantities are fairly correlated. Indeed, the Pearson's correlation coefficients are $r > 0.89$ for the α–helix and for q in the case of the β–hairpin, whereas only for the $dRMSD$ of this last case the correlation seems poor. One should stress that the correlation coefficients were calculated without considering the horizontal error bars, which can (actually, were) derived from the width of distributions in Figs. 2.7 and 2.8 , as they are so large that the whole computation of the correlation would have been pointless. Moreover, it is worth to point out that the two quantities compared in Fig. 2.13 have a completely different origin. In fact, while the effective force $TD^{(1)}/D^{(2)}$ is calculated point–wise from very short, dynamical simulations, the free energy F and its gradient were obtained from an equilibrium sampling. This feature is interesting as the agreement occurs in spite of the large width of the distributions of $D^{(1)}$ and $D^{(2)}$ mentioned above.

If the thermodynamic properties seem to be recovered by the machinery developed in previous Sections, we performed a further test for the drift and the diffusion coefficients on the recovery of the dynamics. To do so, we compared the dynamics simulated by Eq. (2.1) with the projection of the microscopic dynamics on the Collective Variable. Operatively, once we calculated $D^{(1)}(Y)$ and $D^{(2)}(Y)$ for the profile of an entire Collective Variable Y (only five points, actually), we could write an effective integrator where the functions $D^{(1)}(Y)$ and $D^{(2)}(Y)$ were coded in order to provide the same results found from the algorithm. The comparison has been carried out for the case of the variable q in the β–hairpin, as it was the variable with the best outcome in the determination of the effective energy (see Fig. 2.13 D). We display the average and the standard deviation of q as a function of the time in Fig. 2.14 . In green, data are taken from full–atom simulations, the same used to calculate $D^{(1)}$ and $D^{(2)}$; in red are shown data from the

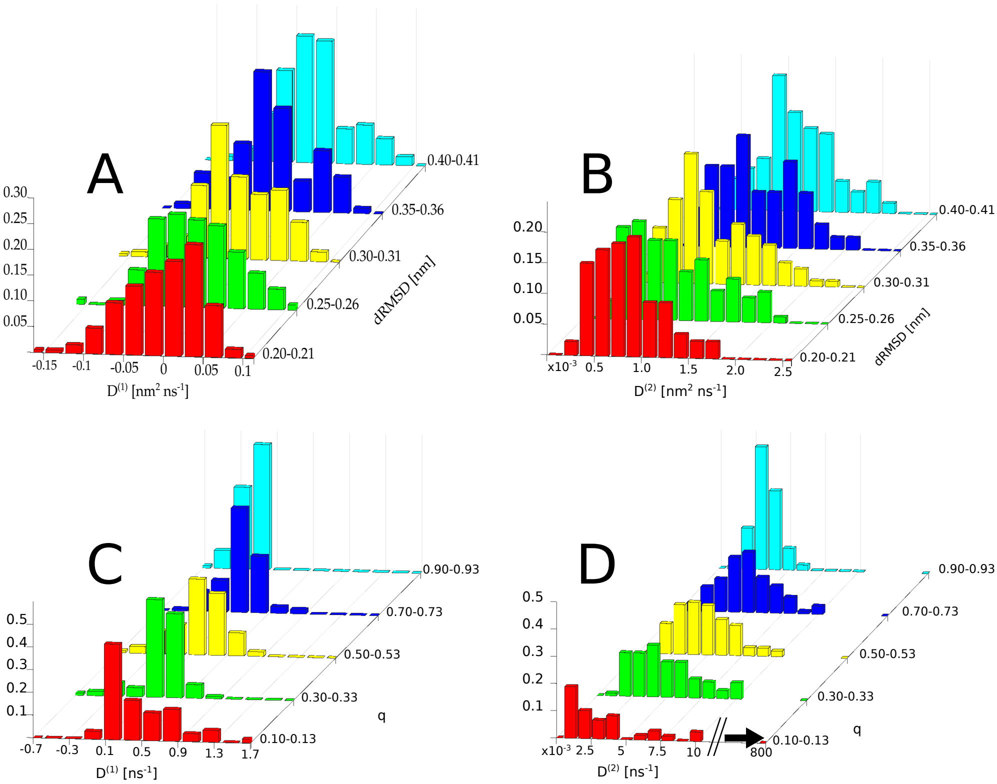


Figure 2.7: The distributions of drift (A–C) and diffusion (B–D) coefficients for the Collective Variables $dRMSD$ (upper panels) and q (lower panels), calculated in different intervals of the Collective Variable for the case of the α–helix.

resolution of the monodimensional Langevin equation with the $D^{(1)}$ and $D^{(2)}$ extracted from the green series. In the latter case, the average is performed over 100 simulations starting from the same unfolded conformation. As can be seen in Fig. 2.14 A, the overall folding time seems comparable in the two kind of simulations, being it of the order of few ps for both the cases, but the detailed dynamics is different. Indeed, the microscopic dynamics shows a slightly faster initial folding, followed by a further event at $\approx 18\,\mathrm{ps}$, which is not reported in the effective Langevin dynamics. However, the most important difference lies in the run–to–run variability, which is completely different in the two cases (see Fig. 2.14 B), as fluctuations of the value of q over time are many orders of magnitude smaller in the microscopic dynamics. The dynamical properties thus seem much more sensitive to the imperfections in the Collective Variable than to the equilibrium properties.

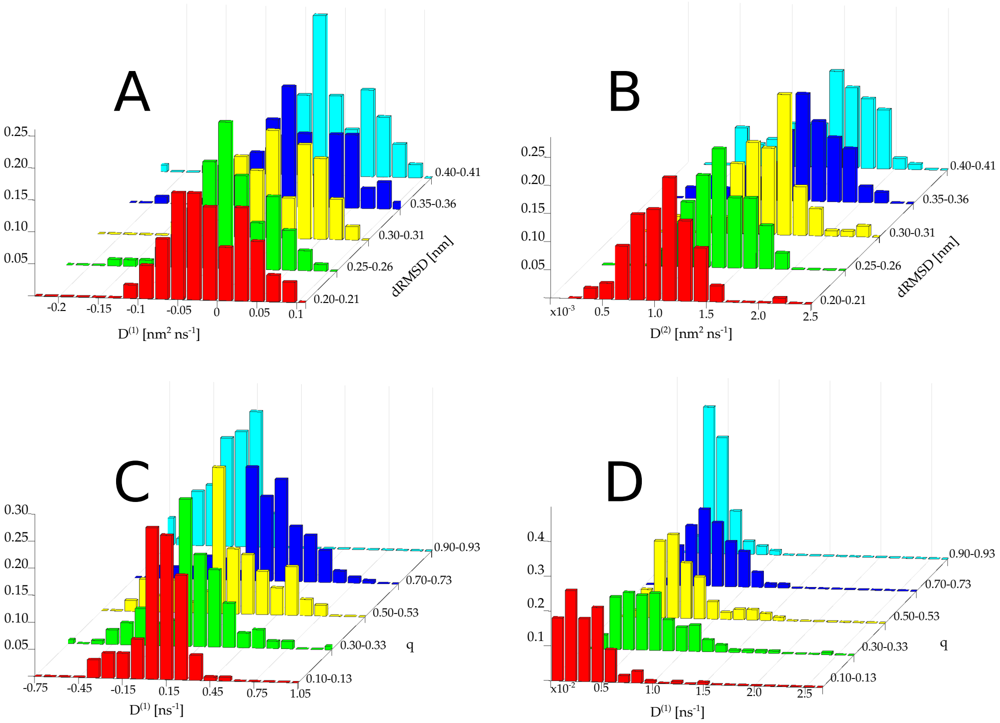


Figure 2.8: The distributions of drift (A–C) and diffusion (B–D) coefficients for the Collective Variables $dRMSD$ (upper panels) and q (lower panels), calculated in different intervals of the Collective Variablefor the case of the β–hairpin.

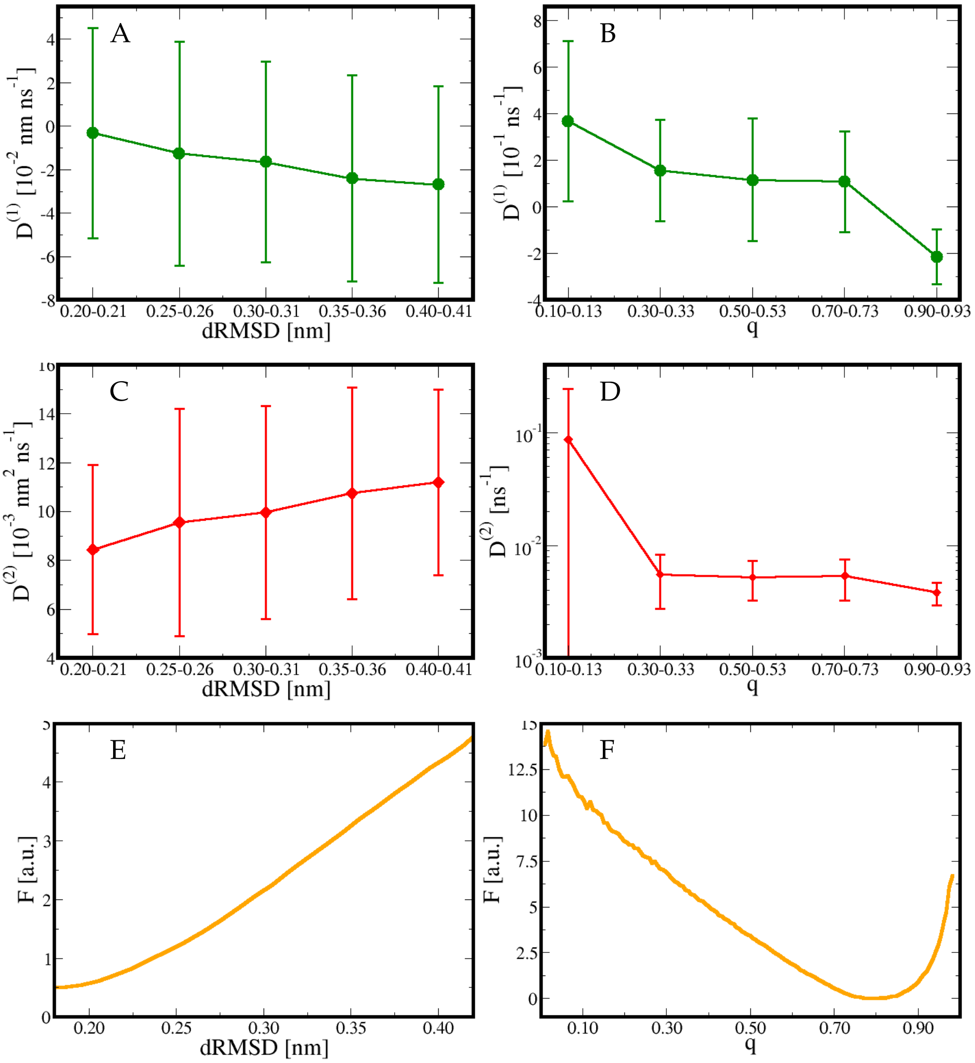


Figure 2.9: The average and the standard deviation, plotted as error bar, of the drift (A–B) and diffusion coefficient (C–D) for the α–helix, calculated in different intervals of $dRMSD$ (left panels) and q (right panels). Below, the free energies in the same region of interest for both $dRMSD$ (E) and q (F), at the simulation temperature $T = 0.92$ (in energy units).

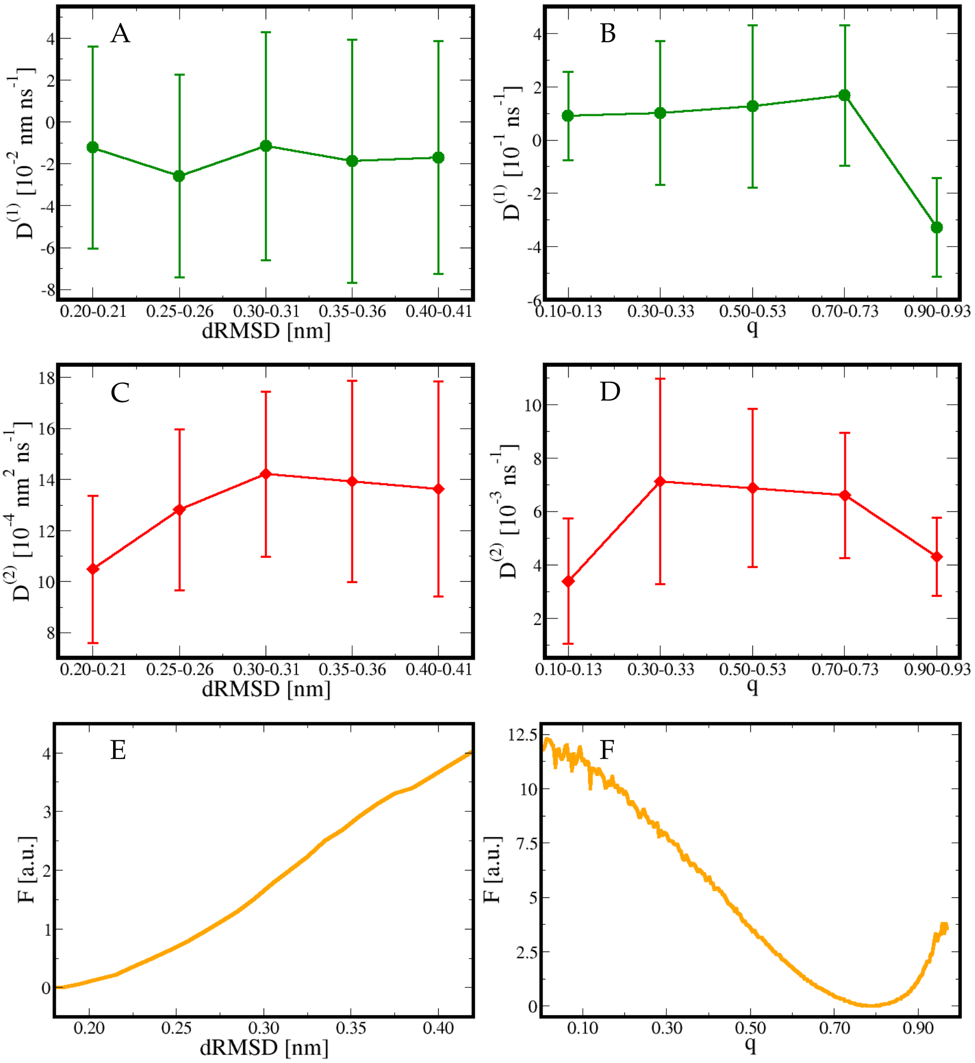
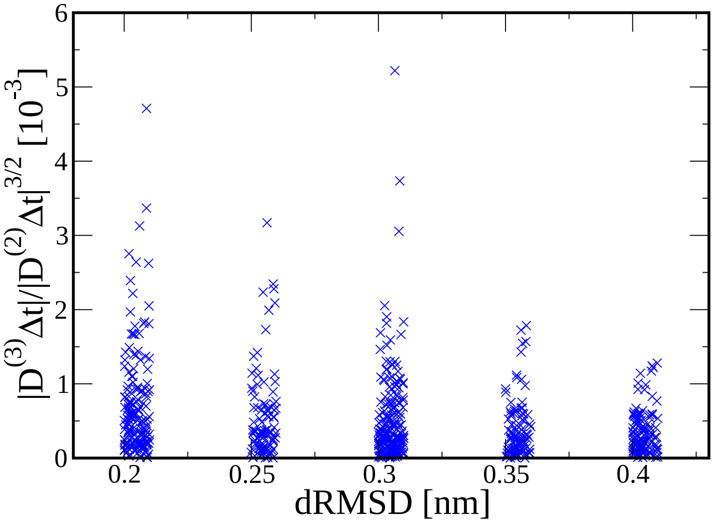


Figure 2.10: The average and the standard deviation, plotted as error bar, of the drift (A–B) and diffusion coefficient (C–D) for the β–hairpin, calculated in different intervals of $dRMSD$ (left panels) and q (right panels). Below, the free energies in the same region of interest for both $dRMSD$ (E) and q (F), at the simulation temperature $T = 0.92$ (in energy units).



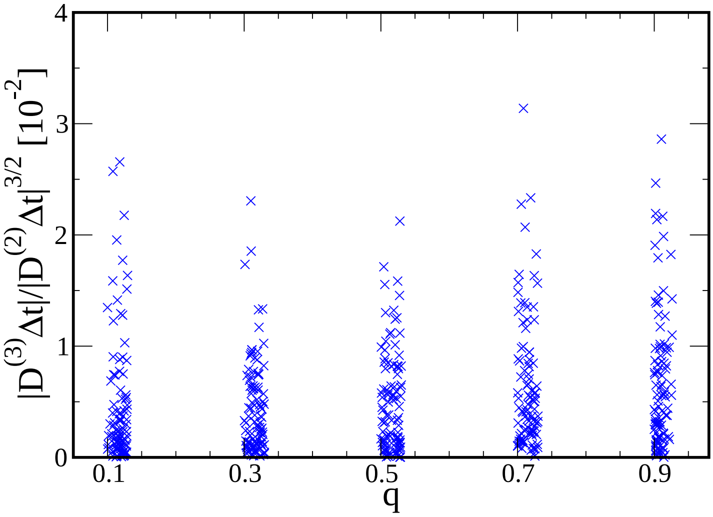
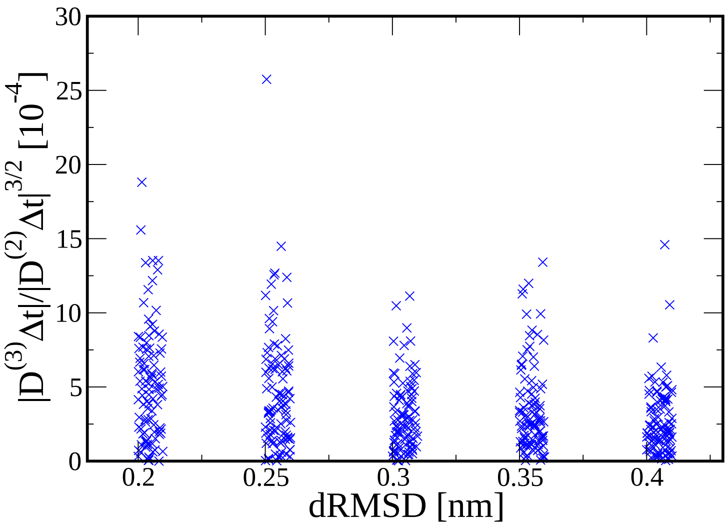


Figure 2.11: The ratio between $D^{(3)}\Delta t$ and $|D^{(2)}\Delta t|^{3/2}$, calculated for the Collective Variable $dRMSD$ (above) and q (below), for the α–helix.



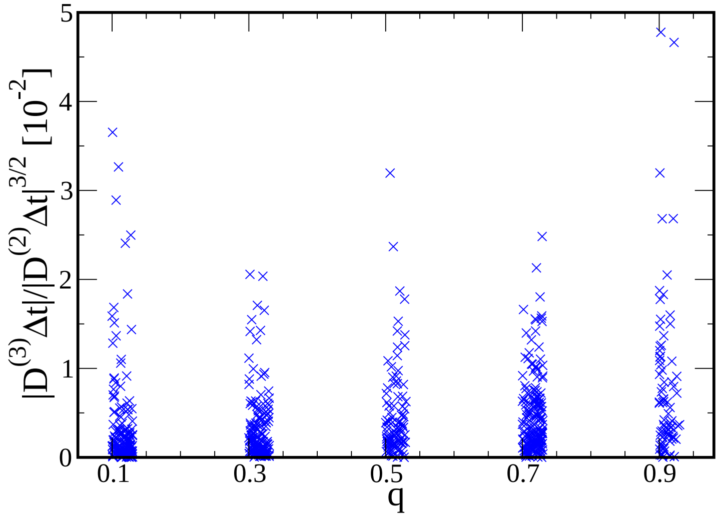


Figure 2.12: The ratio between $D^{(3)}\Delta t$ and $|D^{(2)}\Delta t|^{3/2}$, calculated for the Collective Variable $dRMSD$ (above) and q (below), for the β–hairpin.

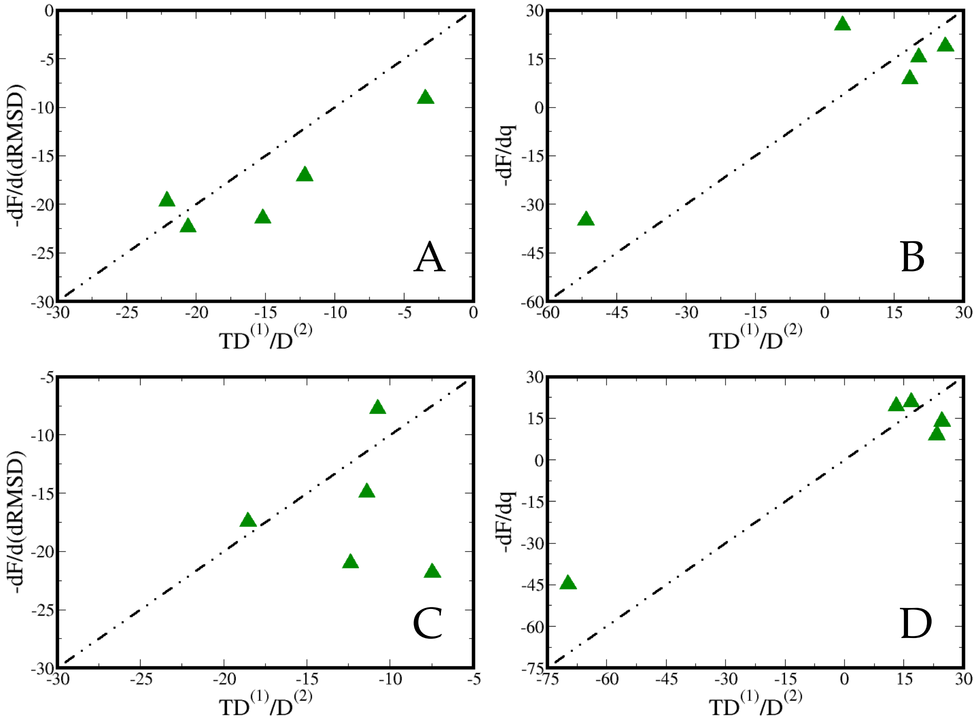


Figure 2.13: A comparison between effective force $TD^{(1)}/D^{(2)}$ obtained from the drift and diffusion coefficients and that obtained differentiating the equilibrium free energy (cf. Figs. 2.9 and 2.10) in the case of the α–helix using the $dRMSD$ (A) and q (B); in the case of the β–hairpin using the $dRMSD$ (C) and q (D).

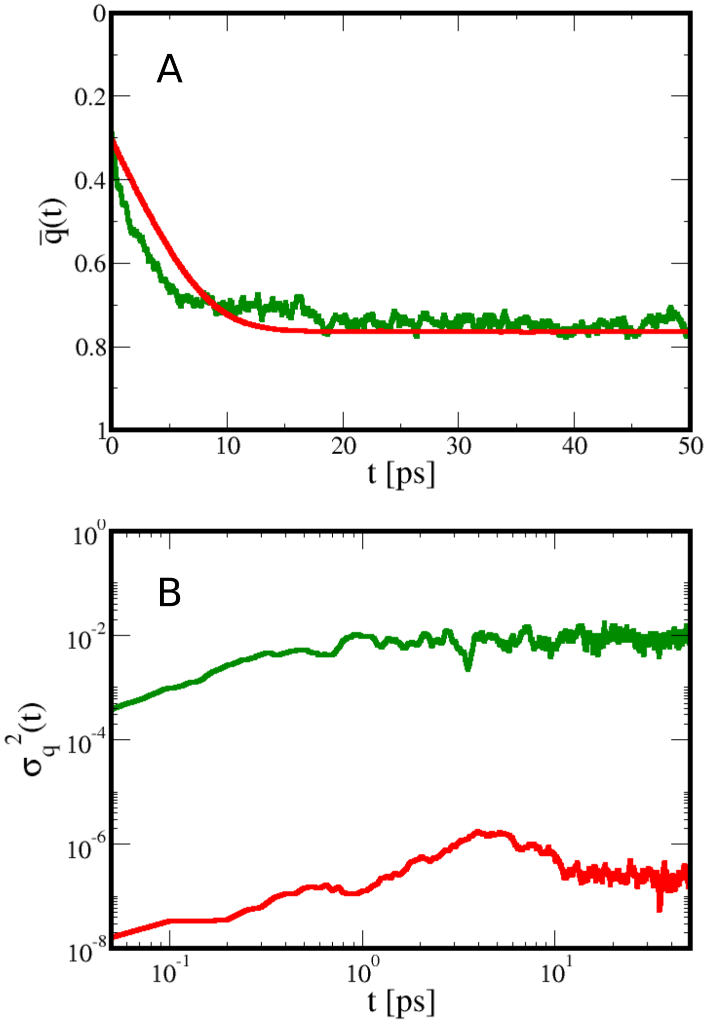


Figure 2.14: The average fraction of native contacts $\bar{q}$ (A) and its variance (B), as functions of the simulation time. Data in green are taken from full–atom simulations of the $\beta - hairpin$, while data in red are generated via the resolution of a monodimensional langevin equation. The relaxing times τ of $\bar{q}$ are $\tau_\beta = 3.6\,\mathrm{ps}$ for the full–atomistic simulations and $\tau_{1d} = 4.8\,\mathrm{ps}$ for the monodimensional trajectories.

2.5 Discussion

The results obtained in two one–dimensional test systems indicate that it is possible to back–calculate the diffusion coefficient with great precision (i.e., with an error lower than 1%) by direct application of its finite–difference definition (i.e., in Eq. (2.12)) if the drift coefficient is moderately small, that is if the drift does not induce a deterministic displacement within the time Δt used in Eq. (2.12). On the other hand, the calculation of the drift coefficient directly by definition would require an enormous ($\gg 10^5$) number of replicated simulations for each point of conformational space to obtain a stable result, and it is thus impractical. The strategy we developed based on an iterative solution of the finite–differences equations of motion allow one to back–calculate the drift coefficient with an error lower than 10% in the test cases, and to back–calculate correctly the diffusion coefficient also if the system undergoes strong drifts, up to those caused by nN forces, much stronger than typical biological forces.

In simulations carried out with simple models of α–helix and β–hairpin, the distribution of drift coefficients for different point of conformational space associated with the same value of the fraction of native contacts q has a spread which is of the order of half its average. This suggests that the Collective Variable q defines a drift coefficient, although with a non–negligible error bar. The same is not true for the $dRMSD$. In this case the drift coefficient is defined only in terms of its order of magnitude, the width of its distribution for fixed value of $dRMSD$ ranging from three to ten times the mean.

Anyway, for both variables the mean drift matches that obtained as derivative of the free energy F, calculated independently from a conformational sampling at equilibrium. This suggests that it is possible to build a one–dimensional approximated model of these peptides, in which the equilibrium and the dynamical properties are consistent with each other. Moreover, one could also exploit these results to calculate the equilibrium free energy of a system from short dynamical simulations. The diffusion coefficient is defined better than the drift for both the $dRMSD$ and the q, the width of the associated distribution being in both cases at worst half of the mean, and usually lower. This fact gives a sound basis to the calculations reported in refs. [41, 42, 43, 44] for the diffusion coefficient, and make it possible to exploit strategies in which the diffusion coefficient is artificially biased to enhance conformational sampling.

The large widths of the distributions of $D^{(1)}$ and $D^{(2)}$ are not really unexpected, since to a given value of q or $dRMSD$ correspond conformations which can be conformationally very different, and thus can display different energetic properties. Interestingly this is true for native–like conformations as well ($q = 0.90$ or

$dRMSD = 0.20\,\mathrm{nm}$), which are supposedly more homogeneous from the conformational point of view. Most likely, the steep dependence of the potential function which is used in the present force field (i.e., containing terms like $1/r^6$ or $1/r^{12}$) and which reflect the true interaction between the atoms of the system, plays an important role in defining the width of the observed distributions.

Although the agreement between the calculated mean drift coefficient and the equilibrium free energy of the peptides is reasonably good, the detailed dynamics is quite different in the dimensional–reduced model, especially in terms of run–to–run fluctuations of the Collective Variable. This suggests that the dynamical properties are more sensitive to approximations than the equilibrium properties. This asymmetry was already observed for their respective dependence on the force field [89].

Conclusions

The computational study of the denatured state of proteins can be an extremely challenging task, as it usually requires a thorough exploration of a huge phase space, whose dimension can rise up to 10^6 degrees of freedom for a biomolecule. However, it is necessary to provide an overview of the molecular mechanism through which common chaotropic agents, such as urea or guanidine chloride, act on the tridimensional structure of a protein and disrupt it, stabilizing the denatured phase and allowing it to be studied with standard experimental techniques. This topic has been addressed in the present thesis from a twofold perspective: one more practical, involving a set of simulations of polypeptides in solution, the other one more theoretical, concerning some properties of the Collective Variables used to describe the denatured state.

In the first part, we investigated the conformational properties of two peptides, the α–helix (residues $22 - 38$) and the second β–hairpin (residues $41 - 56$) of the protein G B1 domain, under the effect of distinct chemical environments: in water, in 2M or 5M urea and in 2M or 4M GndCl. The denaturation mechanism acting on the α–helix and the β–hairpin seems to be different, as the conformational properties of the two peptides in water are contrasting. Indeed, at physiological conditions the β–hairpin can be described as a three-state system, formed by the fully–formed native state, an half–formed intermediate state and a random coil, whereas the α–helix populates a plethora of states whose content in secondary structure can be rather heterogeneous. The addition of denaturants in solution reflects this dissimilarity, since the β–hairpin decreases the population of the native and the intermediate states, but overall the conformational space does not change its structure and back–calculated experimental values – CD spectra and chemical shifts – report only β content both in urea and in GndCl. On the other side, the mixed α and β content for the α–helix in denaturant solution results in

CD spectra and chemical shifts resembling those typical of a random coil. Moreover, the denaturing mechanisms of urea and GndCl appear to be very different from each other, the former competing with the molecules of water in forming hydrogen bonds with essentially all the residues, the latter favouring the interactions with the negative–charged amino acids, aside from generating an electric dipole whose effect is generally detrimental for secondary structures, especially the α–helices.

In the second part, we examined some mathematical properties of two Collective Variables (the fraction of native contacts q and the distance root mean square deviation $dRMSD$ from a reference structure), which are dimensional reductions commonly used to describe, among others, the denatured phase and its transition towards the native state. Indeed, a dimensional–reduction approach is extremely useful to analyze complex, biomolecular data and occasionally to bias Molecular Dynamics simulations in order to decrease their computational cost. However, the dynamics of a system in a reduced dimensional space can be as complicated as that in the original, full–dimensional one if the reduced coordinate Y is not chosen properly. Usually, a requirement is for the Collective Variable Y to describe the slowest motion of the system; this property makes Y to be controlled by an overdamped Langevin equation, where the dependence of the drift and the diffusion coefficients on the microscopic conformations is completely lost. Yet it is difficult to check directly this condition in a system as complex as a biopolymer. Here, we employed a different approach and developed an algorithm to calculate the drift and diffusion coefficients in the neighborhood of a microscopic point of the full–dimensional conformational space. This method is based on a finite–differences version of the Langevin equation and on an iterative evolution of the dynamics of the Collective Variable Y, under the approximation of locally linear force and locally constant diffusion coefficient. Such an approach allowed us to calculate the drift and diffusion coefficients for an ensemble of such points, corresponding to the same value of the reduced coordinate Y. We showed that the coefficients calculated for the length–invariant Collective Variables (such as a sum of contact functions like q) display a weaker dependence on the microscopic point than Euclidean distances (such as the $dRMSD$), but by any means this feature is a priori negligible. Nonetheless, the average drift and diffusion coefficients are compatible with the equilibrium properties of the system, whereas the dynamical properties are more sensitive to the lack of ideality of the Collective Variable.

Appendices

Appendix A

Aminoacid–specific radial distribution functions

A.1 α–helix

We report in Tab. A.1 the list of amino acids, along with the properties of their side chains concerning the electrostatic and the hydrophobicity. In the Figs. A.1 , A.2 , A.3 , A.4 and A.5 are shown the radial distribution functions for each amino acid with respect to each component of the solvent, except for the N-terminal ASP22 and the C-terminal GLY38, which are neglected since their increased degrees of freedom give rise to extremely noisy curves.

22	ASP	ch. negative
23	ALA	hydrophobic
24	ALA	hydrophobic
25	THR	polar
26	ALA	hydrophobic
27	GLU	ch. negative
28	LYS	ch. positive
29	VAL	hydrophobic
30	PHE	hydrophobic
31	LYS	ch. positive
32	GLN	polar
33	TYR	hydrophobic
34	ALA	hydrophobic
35	ASN	polar
36	ASP	ch. negative
37	ASN	polar
38	GLY	-

Table A.1: The whole list of amino acids and their electrostatic or hydrophobic properties are reported for the α–helix system.

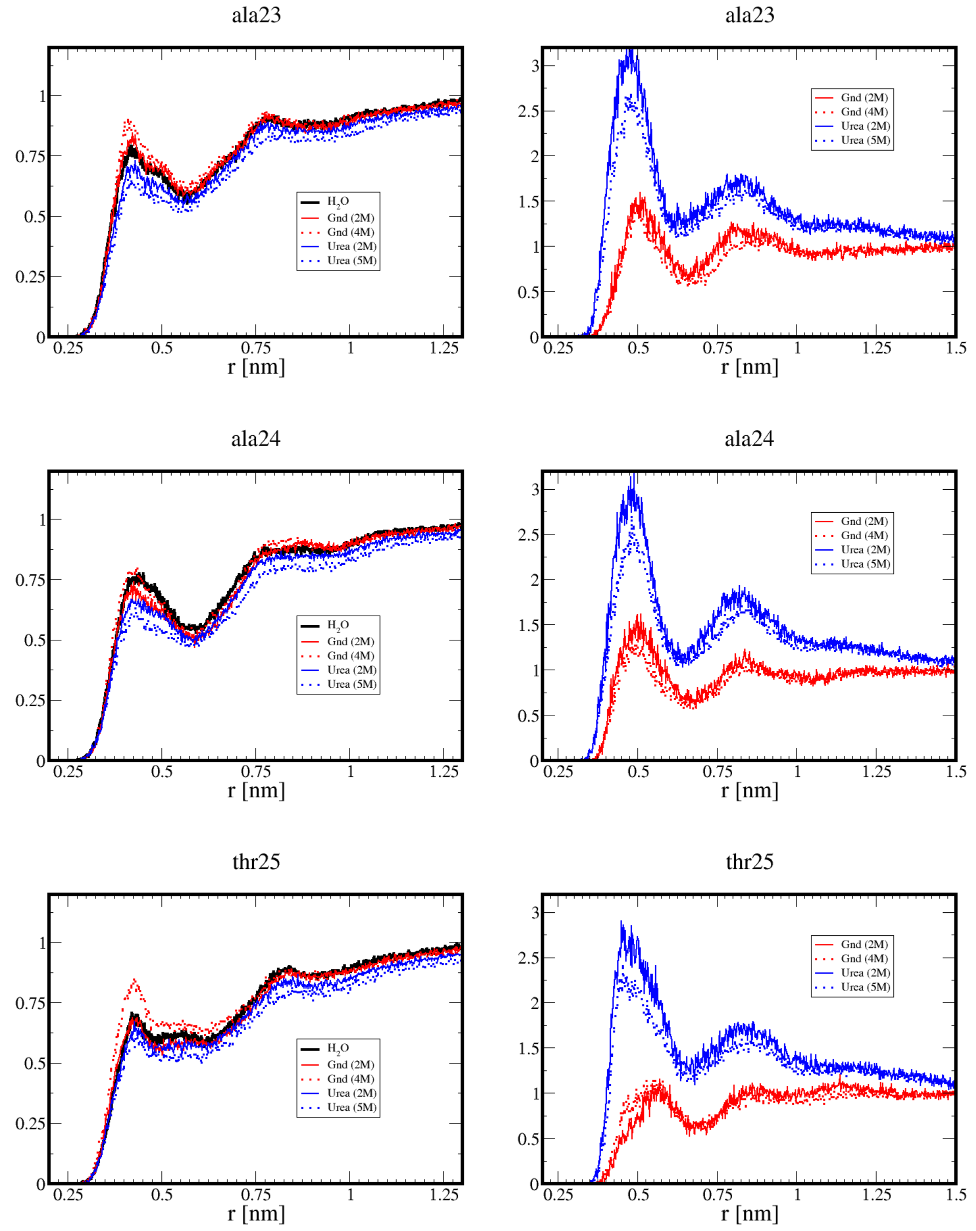


Figure A.1: Radial distribution functions for the amino acids listed in the title of each panel. On the left panels, the rdf is computed between the amino acid and the center of mass of the molecules of water, whereas on the right panels it is computed between the amino acid and the molecules of denaturant. Data refer to the simulation in water (black lines), 2M GndCl (solid red), 4M GndCl (dashed red), 2M urea (solid blue) and 5M urea (dashed blue).

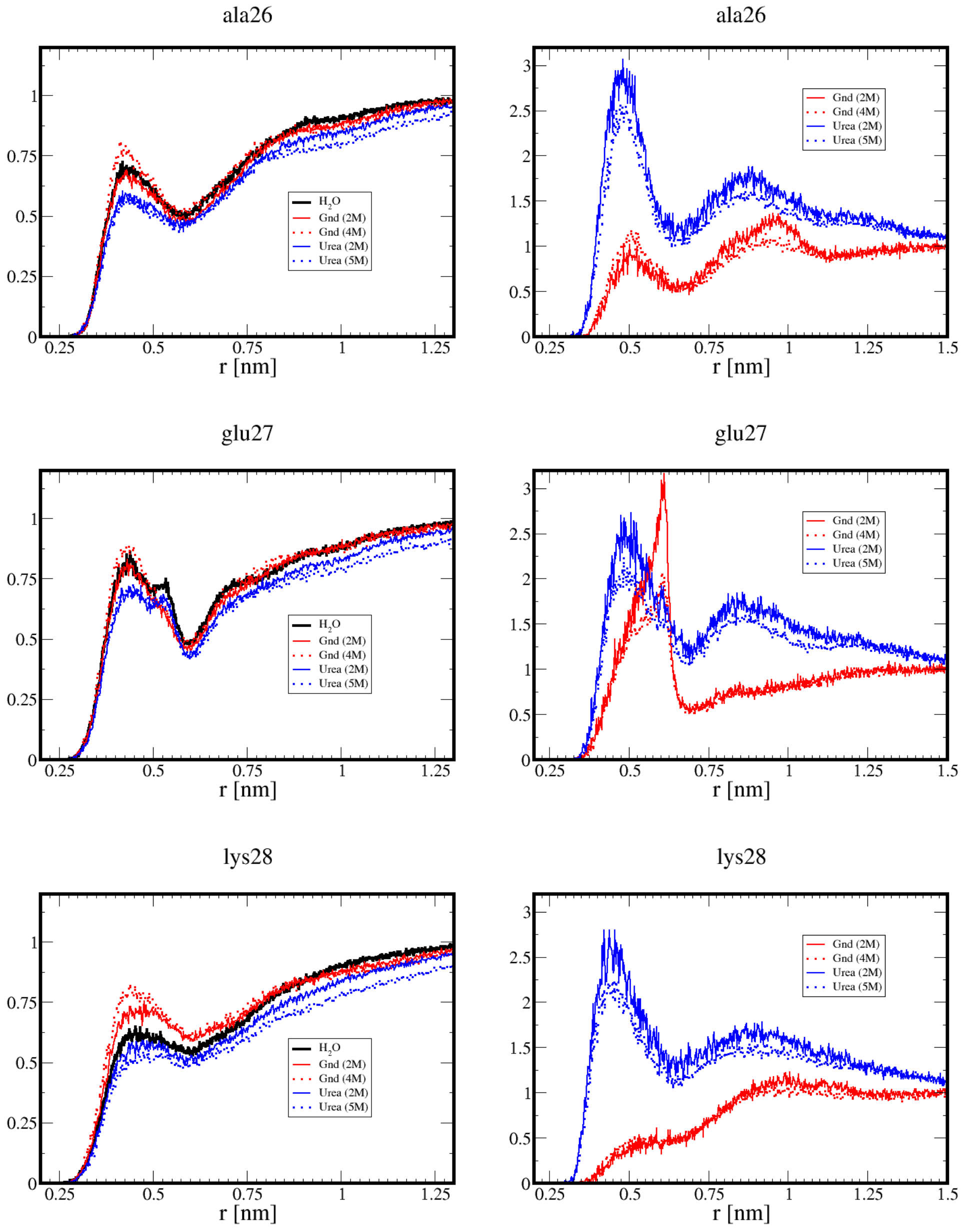


Figure A.2: See caption of Fig. A.1 .

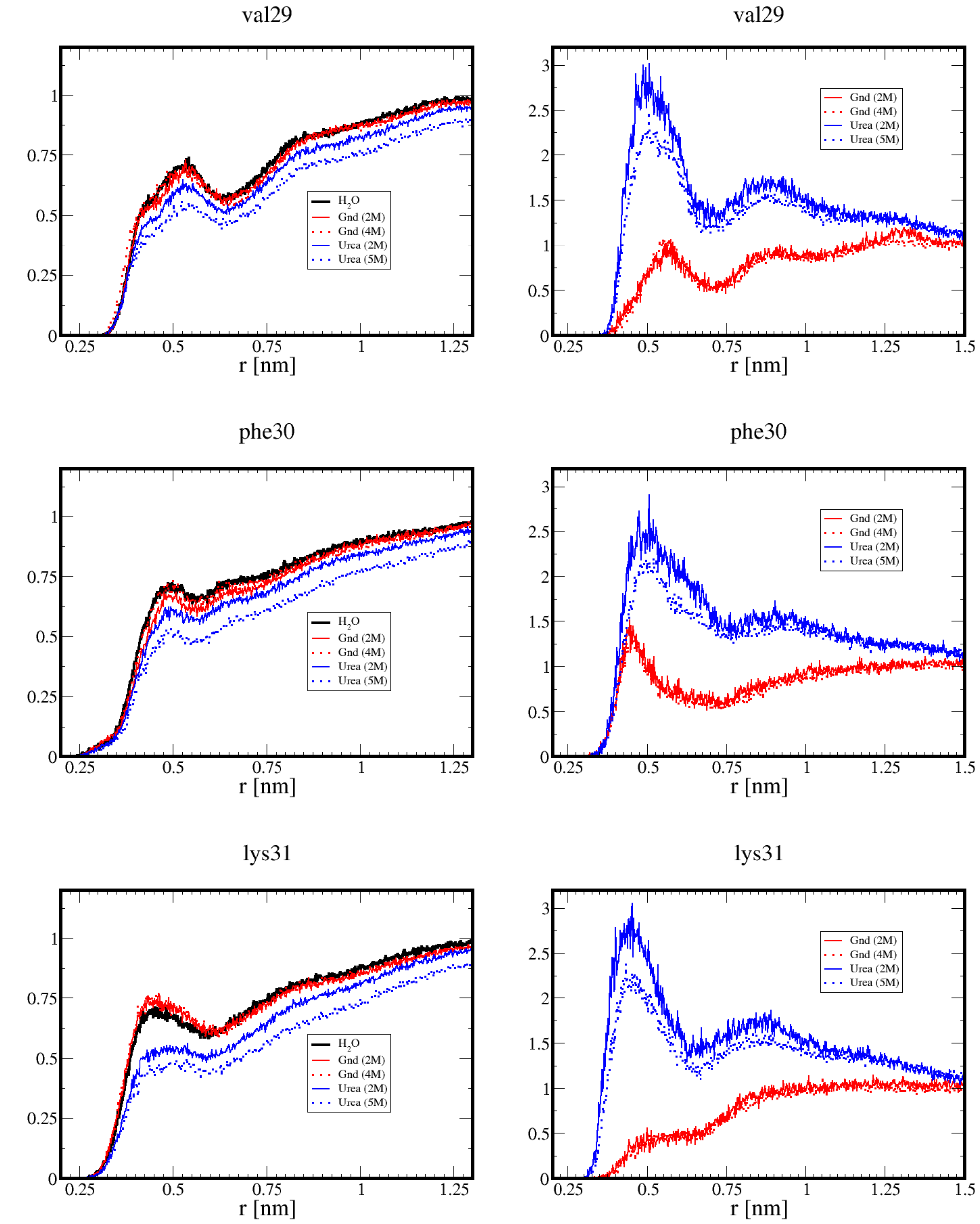


Figure A.3: See caption of Fig. A.1 .

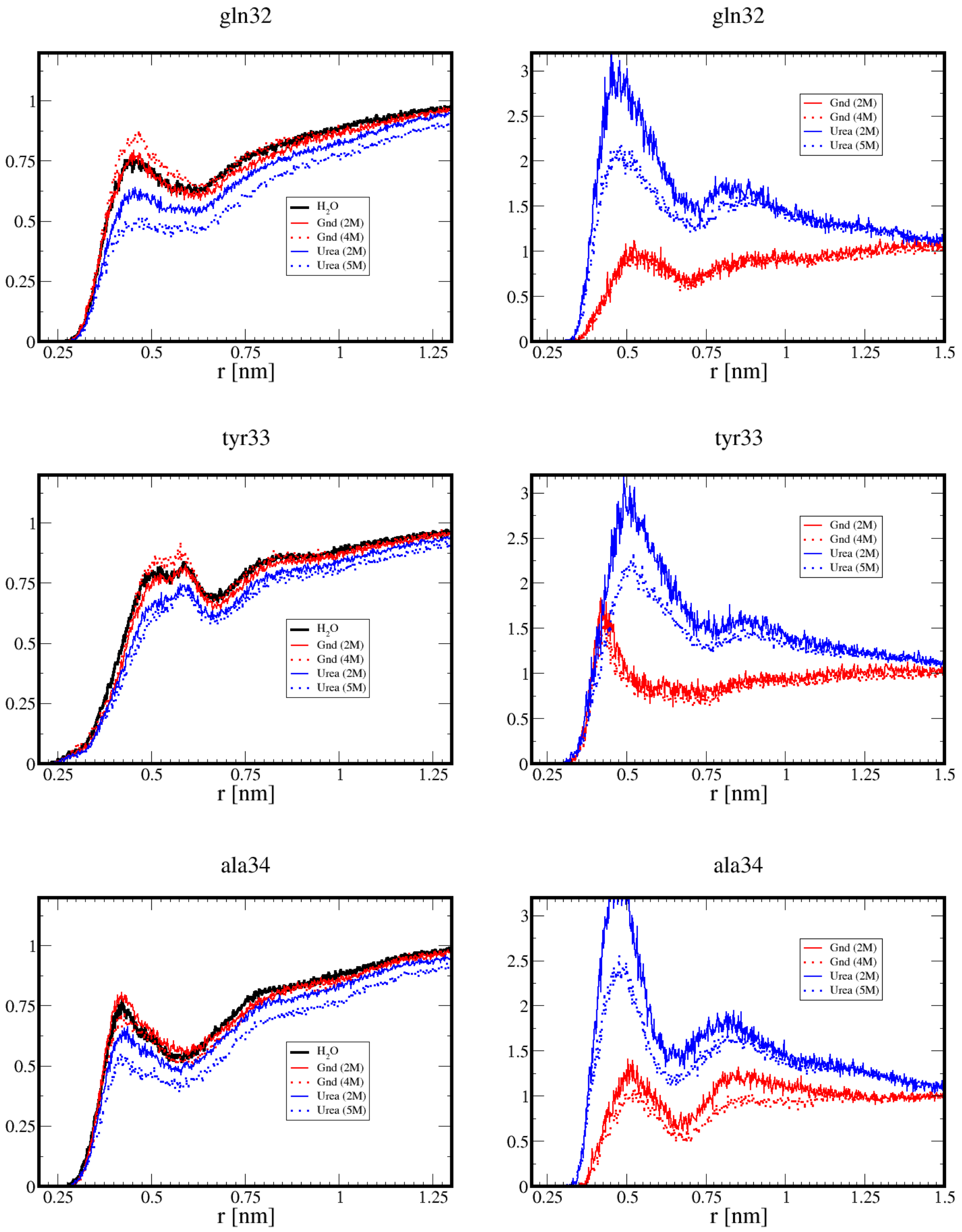


Figure A.4: See caption of Fig. A.1 .

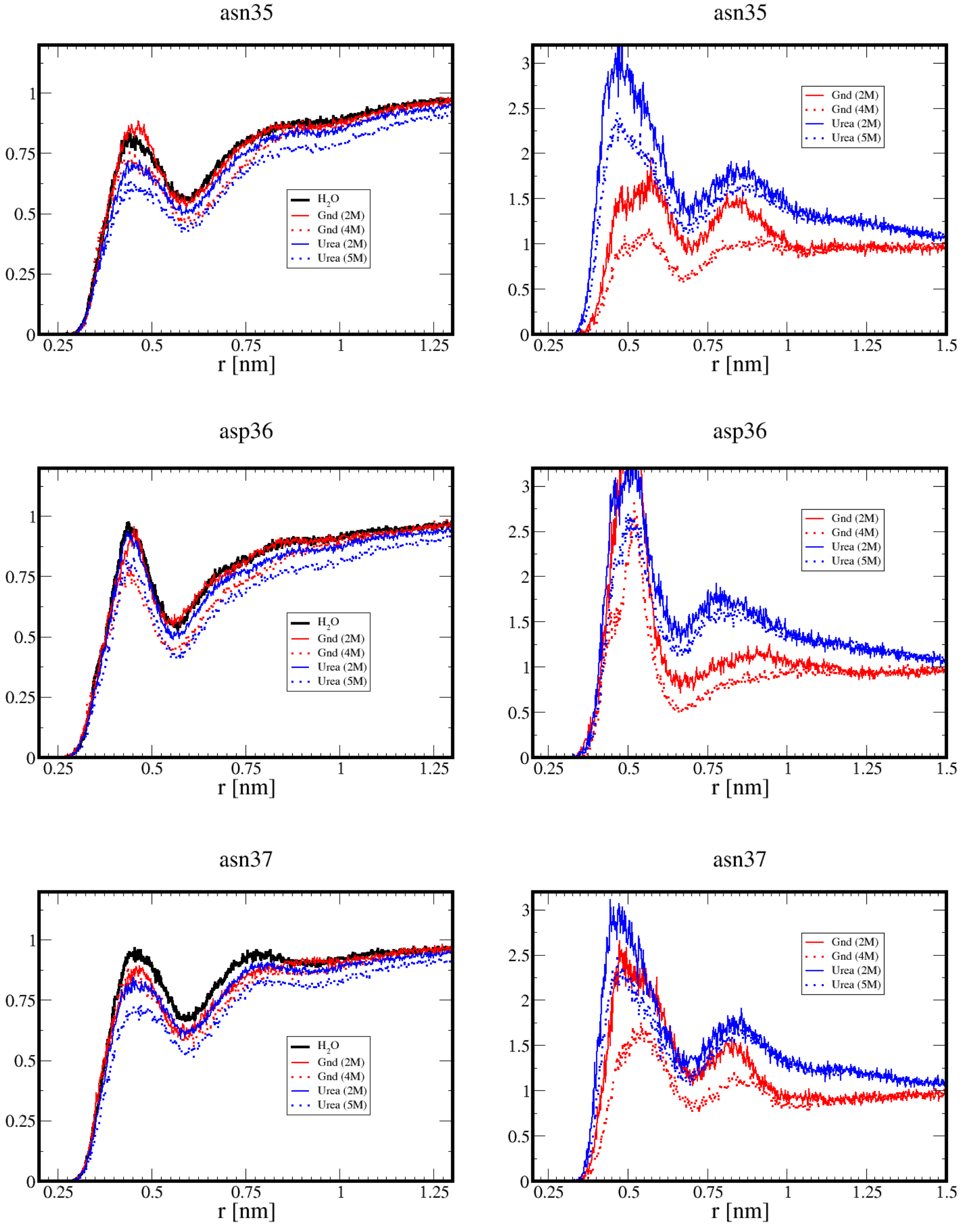


Figure A.5: See caption of Fig. A.1 .

A.2 β–hairpin

As in Tab. A.1 , in Tab. A.2 we report the list of amino acids and their properties, whereas in the Figs. A.6 , A.7 , A.8 , A.9 and A.10 are shown their radial distribution functions, except for those of the terminal amino acids.

41	GLY	-
42	GLU	ch. negative
43	TRP	hydrophobic
44	THR	polar
45	TYR	hydrophobic
46	ASP	ch. negative
47	ASP	ch. negative
48	ALA	hydrophobic
49	THR	polar
50	LYS	ch. positive
51	THR	polar
52	PHE	hydrophobic
53	THR	polar
54	VAL	hydrophobic
55	THR	polar
56	GLU	ch. negative

Table A.2: The whole list of amino acids and their electrostatic or hydrophobic properties are reported for the β–hairpin system.

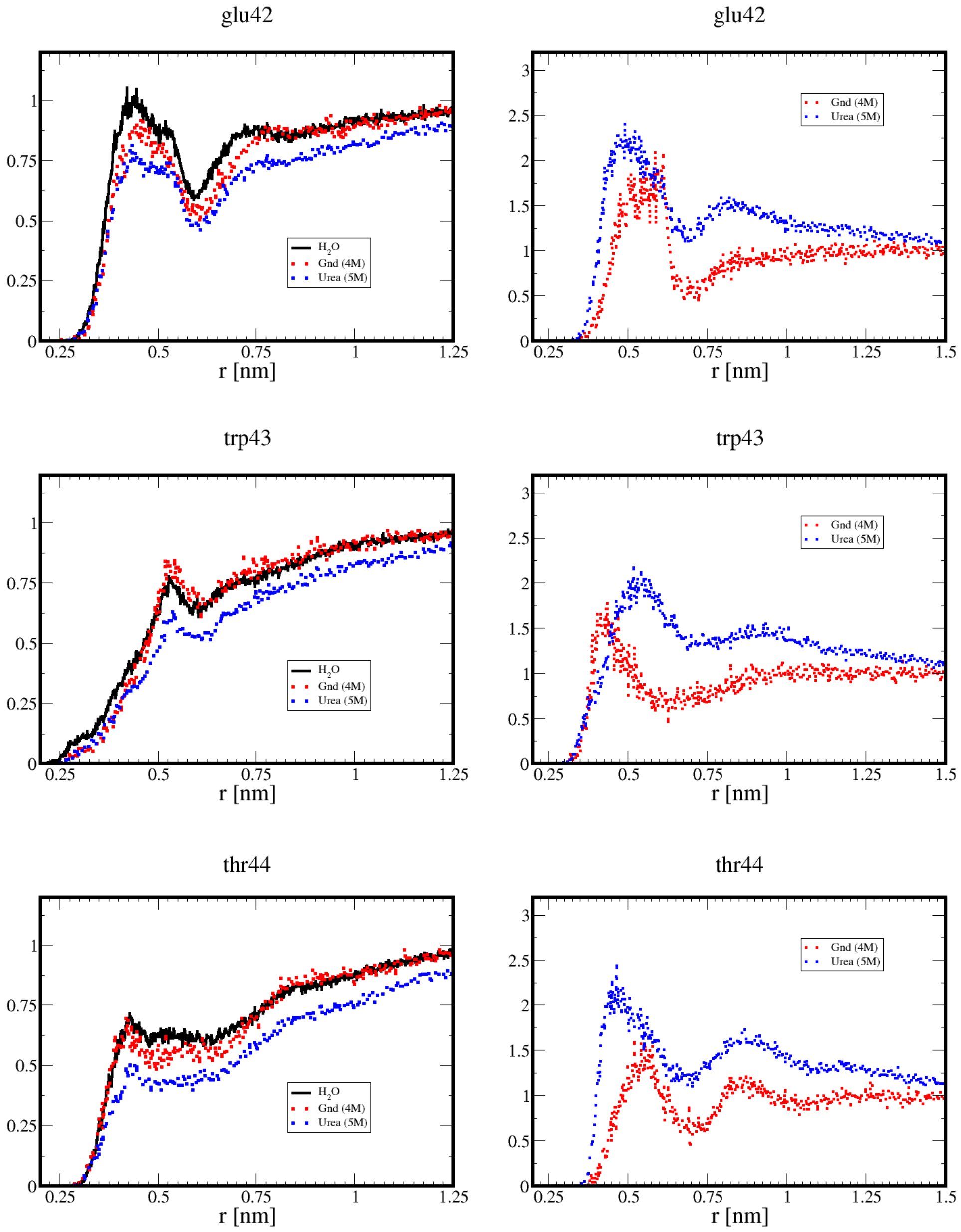


Figure A.6: Radial distribution functions for the amino acids listed in the title of each panel. On the left panels, the rdf is computed between the amino acid and the center of mass of the molecules of water, whereas on the right panels it is computed between the amino acid and the molecules of denaturant. Data refer to the simulation in water (black lines), 4M GndCl (dashed red) and 5M urea (dashed blue).

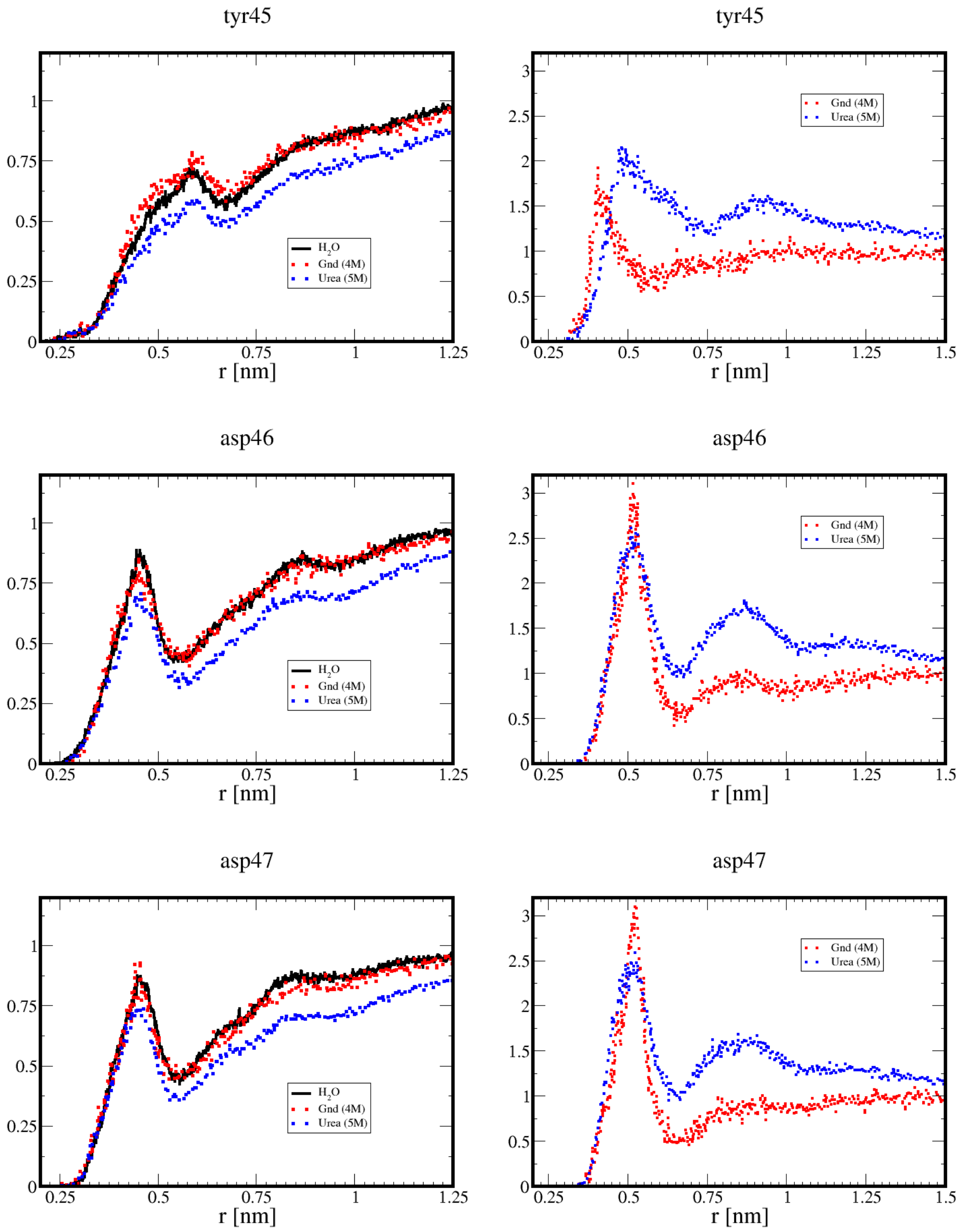


Figure A.7: See caption of Fig. A.6 .

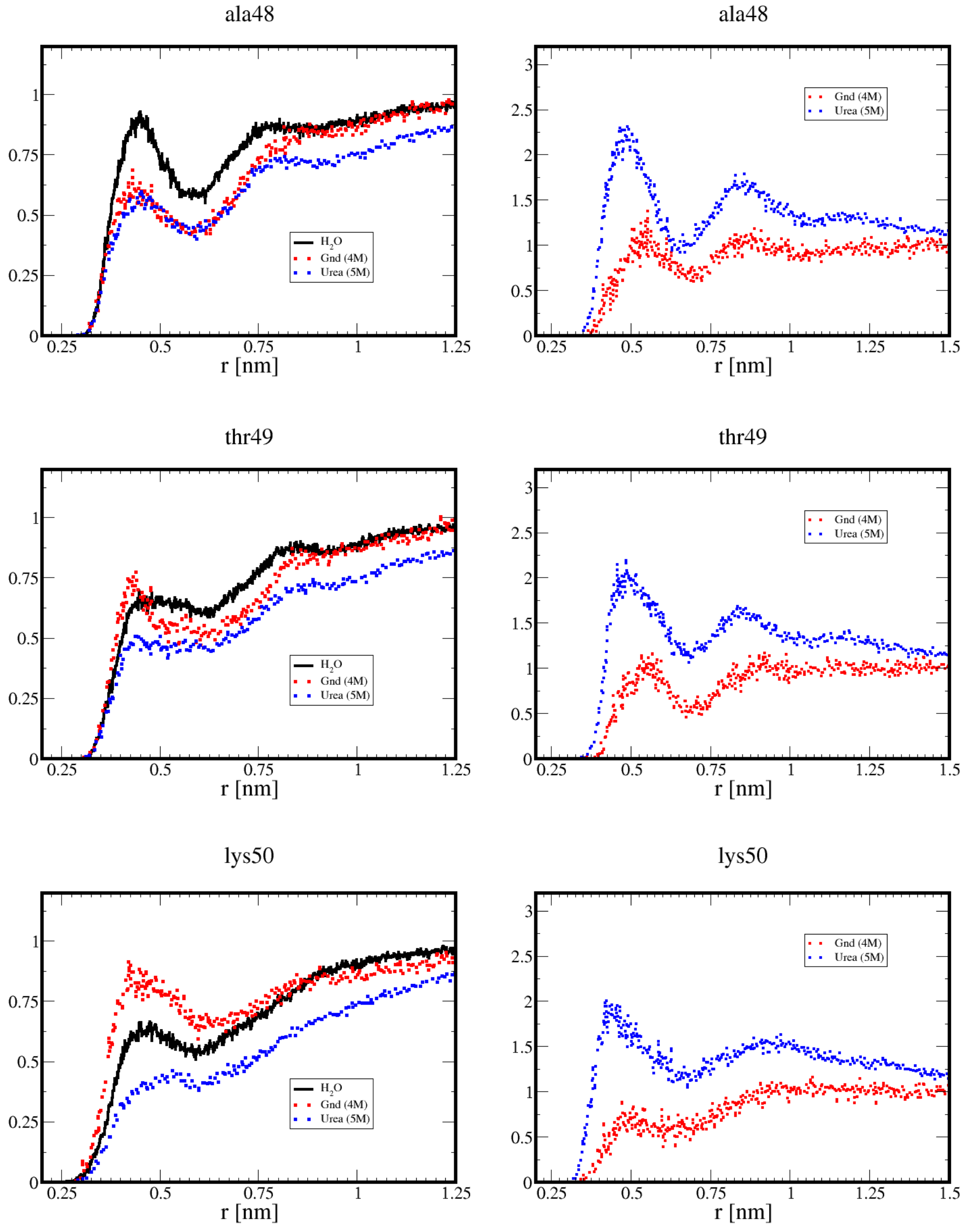


Figure A.8: See caption of Fig. A.6 .

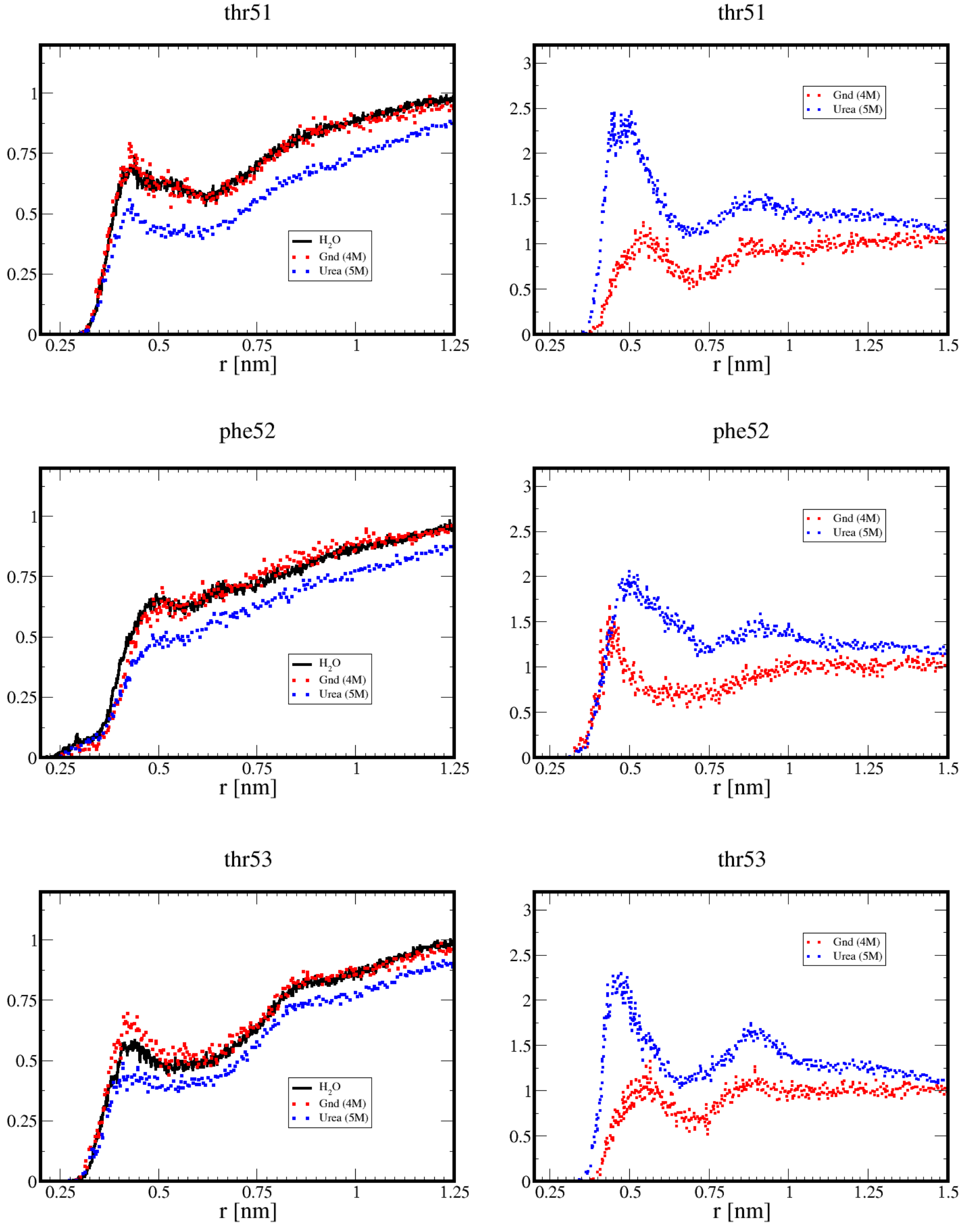


Figure A.9: See caption of Fig. A.6 .

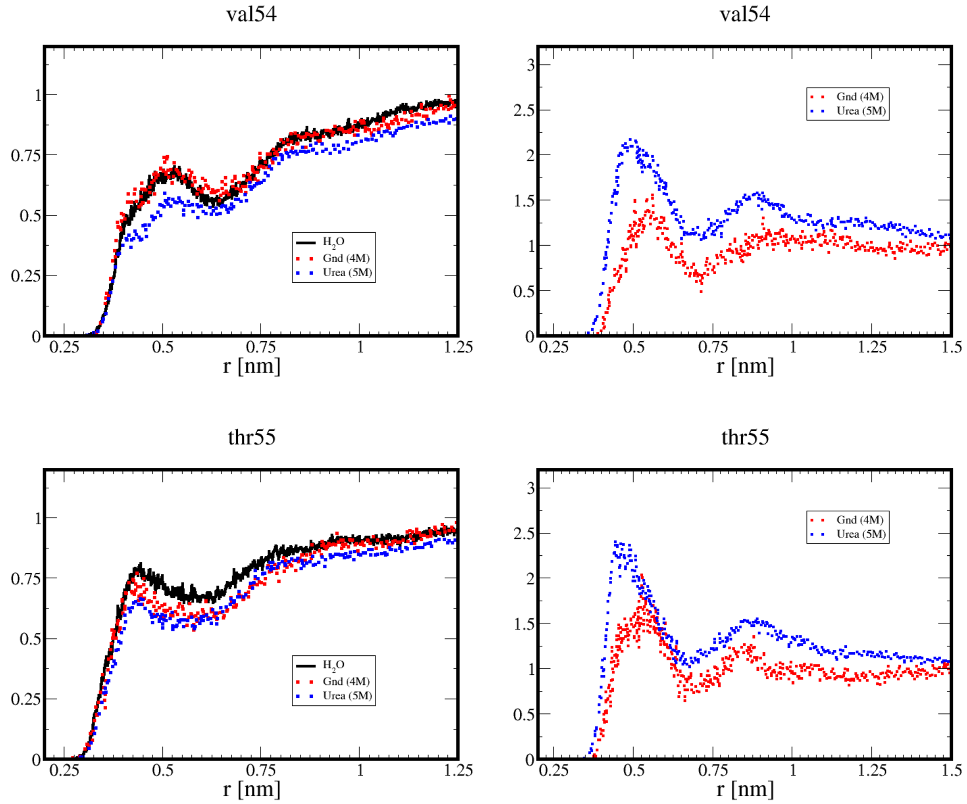


Figure A.10: See caption of Fig. A.6 .

APPENDIX B

On the convergence of $p(Y)$ to Boltzmann

We show here that $p(Y)$ converges in a first approximation to Boltzmann, when proper orders of magnitude for $D^{(1)}$ and $D^{(2)}$ are inserted in calculations. The integrator used has the form

$$Y(t+\Delta t) = Y(t) + D^{(1)}(Y(t))\Delta t + \sqrt{2D^{(2)}(Y(t))\Delta t}\cdot \eta_t \tag{B.1}$$

Inverting the integrator and introducing the notation $Y_t \equiv Y(t)$, we have an expression for η_t

$$\eta_t = \frac{Y_{t+\Delta t} - Y_t - D^{(1)}(Y_t)\Delta t}{\sqrt{2D^{(2)}(Y_t)\Delta t}} \tag{B.2}$$

We call $\Delta Y \equiv Y_{t+\Delta t} - Y_t$. Since the transition probability $W_{Y_t \to Y_{t+\Delta t}}$ (namely, the probability of performing a step ΔY) is the same of extracting the right η_t, then $p(\Delta Y) \equiv W_{Y_t \to Y_{t+\Delta t}} = p(\eta_t)$, where the latter is gaussian–distributed:

$$exp\bigg(-\frac{\eta_t^2}{2}\bigg) = exp\bigg[-\frac{1}{4D^{(2)}(Y_t)\Delta t}\bigg(\Delta Y - D^{(1)}(Y_t))\Delta t\bigg)^2\bigg] \tag{B.3}$$

On the other side, the transition probability $W_{Y_{t+\Delta t}\to Y_t}$ reads

$$exp\bigg(-\frac{\eta_{t+\Delta t}^2}{2}\bigg) = exp\bigg[-\frac{1}{4D^{(2)}(Y_{t+\Delta t})\Delta t}\bigg(-\Delta Y - D^{(1)}(Y_{t+\Delta t})\bigg)^2\bigg] \tag{B.4}$$

The ratio $W_{Y_t\to Y_{t+\Delta t}}/W_{Y_{t+\Delta t}\to Y_t}$ is then

$$\begin{aligned}\frac{W_{Y_t\to Y_{t+\Delta t}}}{W_{Y_{t+\Delta t}\to Y_t}} &= \frac{exp\bigg[-\frac{1}{4D^{(2)}(Y_t)\Delta t}\bigg(\Delta Y - D^{(1)}(Y_t)\Delta t\bigg)^2\bigg]}{exp\bigg[-\frac{1}{4D^{(2)}(Y_{t+\Delta t})\Delta t}\bigg(-\Delta Y - D^{(1)}(Y_{t+\Delta t})\Delta t\bigg)^2\bigg]} = \\ &= \frac{exp\bigg[-\frac{1}{2\alpha_t}\bigg(\Delta^2(Y) - 2\beta_t\Delta Y + \beta_t^2\bigg)\bigg]}{exp\bigg[-\frac{1}{2\alpha_{t+\Delta t}}\bigg(\Delta^2(Y) + 2\beta_{t+\Delta t}\Delta Y + \beta_{t+\Delta t}^2\bigg)\bigg]}\end{aligned} \tag{B.5}$$

where $\alpha_i = 2D^{(2)}(Y_i)\Delta t$ and $\beta_i = D^{(1)}(Y_i)\Delta t$. We rewrite Eq. (B.5) to highlight the differences between similar terms

$$\frac{W_{Y_t \to Y_{t+\Delta t}}}{W_{Y_{t+\Delta t} \to Y_t}} = exp\left[-\frac{1}{2}\Delta^2 Y \left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t+\Delta t}} \right) - \frac{\beta_t^2}{2\alpha_t} + \frac{\beta_{t+\Delta t}^2}{2\alpha_{t+\Delta t}} + \frac{\beta_t \Delta Y}{\alpha_t} + \frac{\beta_{t+\Delta t}\Delta Y}{\alpha_{t+\Delta t}} \right] \tag{B.6}$$

Eq. (B.6) is a product of three terms

$$\begin{aligned}(A) =& exp\left[-\frac{\Delta^2 Y}{2}\left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t+\Delta t}} \right)\right] \\ (B) =& exp\left[\frac{1}{2}\left(\frac{\beta_{t+\Delta t}^2}{\alpha_{t+\Delta t}} - \frac{\beta_t^2}{\alpha_t} \right)\right] \\ (C) =& exp\left[\frac{\beta_t \Delta Y}{\alpha_t} + \frac{\beta_{t+\Delta t}\Delta Y}{\alpha_{t+\Delta t}} \right]\end{aligned}$$

Re-inserting the values of β_i and α_i and making the assumptions that $D^{(1)}(Y_i) \equiv f_i/\gamma_i$ and $D^{(2)}(Y_i) \equiv k_B T/\gamma_i$, where f_i is a force at time i while γ_i is the friction coefficient at Y_i, it turns out that (C) is the classical Boltzmann term.

$$\begin{aligned} exp\left[\frac{\beta_t \Delta Y}{\alpha_t} + \frac{\beta_{t+\Delta t}\Delta Y}{\alpha_{t+\Delta t}} \right] =& exp\left[\frac{D^{(1)}(Y_t)\Delta Y}{2D^{(2)}(Y_t)} + \frac{D^{(1)}(Y_{t+\Delta t})\Delta Y}{2D^{(2)}(Y_{t+\Delta t})} \right] \\ =& exp\left[\frac{f_t \Delta Y}{2k_B T} + \frac{f_{t+\Delta t}\Delta Y}{2k_B T} \right] \\ =& exp\left[\frac{1}{k_B T}\frac{f_t + f_{t+\Delta t}}{2}\Delta Y \right] \\ =& exp\left[-\frac{\Delta U}{k_B T} \right] \end{aligned} \tag{B.7}$$

where we have used the definition of work

$$\begin{aligned}\mathcal{L} =& \int_{Y_t}^{Y_{t+\Delta t}} f(s)ds = -\Delta U \\ =& \overline{f} \int_{Y_t}^{Y_{t+\Delta t}} ds = \\ =& \frac{f_t + f_{t+\Delta t}}{2}\Delta Y\end{aligned}$$

On the other side, (B) is negligible being the argument of the exponent proportional to Δt

$$exp\left[\frac{1}{2}\left(\frac{\beta_{t+\Delta t}^2}{\alpha_{t+\Delta t}} - \frac{\beta_t^2}{\alpha_t} \right)\right] = exp\left[\frac{\Delta t}{2}\left(\frac{\left[D^{(1)}(Y_t)\right]^2}{2D^{(2)}(Y_t)} - \frac{\left[D^{(1)}(Y_{t+\Delta t})\right]^2}{2D^{(2)}(Y_{t+\Delta t})} \right)\right] \tag{B.8}$$

whereas (A) is

$$\begin{aligned} exp\Bigg[-\frac{\Delta^2 Y}{2}\bigg(\frac{1}{\alpha_t}-\frac{1}{\alpha_{t+\Delta t}}\bigg)\Bigg] =& exp\Bigg[-\frac{\Delta^2 Y}{4\Delta t}\bigg(\frac{D^{(2)}(Y_{t+\Delta t})-D^{(2)}(Y_t)}{D^{(2)}(Y_t)\cdot D^{(2)}(Y_{t+\Delta t})}\bigg)\Bigg] \\ \approx& exp\Bigg[-\frac{\Delta^2 Y}{4\Delta t}\cdot\frac{\Delta D^{(2)}(Y_t)}{\big[D^{(2)}(Y_t)\big]^2}\Bigg] \end{aligned} \tag{B.9}$$

where we approximated $D^{(2)}(Y_{t+\Delta t}) \approx D^{(2)}(Y_t)$ at the zeroth order. The order of magnitude of the exponent in Eq. (B.9) is $\mathcal{O}\left(\frac{\Delta D^{(2)}(Y)}{D^{(2)}(Y)}\right) \approx \mathcal{O}(10^{-1})$ (see for instance Figs 2.10). On the other side, the typical exponent in Eq. (B.7) is of the order of the unit $\mathcal{O}(1)$. Hence (A), *at least in this specific context*, can be considered a small correction to the Boltzmann term (C), and the system has the correct long–term thermal evolution.

APPENDIX C

Details on the derivation of $J(\tau)$ and $K(\tau)$

C.1 An expression for $Y(t + N\Delta t)$

We start by considering the integrator in the form

$$Y(t+\Delta t) = Y(t) + \frac{\Delta t}{\gamma} f(Y(t)) + \sqrt{\frac{2k_B T \Delta t}{\gamma}} \cdot \eta_t \tag{C.1}$$

where

$$\langle \eta_t \rangle = 0 \qquad ; \qquad \langle \eta_t \eta_{t'} \rangle = \delta_{t,t'} \tag{C.2}$$

and where the ratio $\frac{k_B T}{\gamma}$ is interpreted as the diffusion coefficient D for the variable Y; since γ is constant here, also it is D.

- k_B is the Boltzmann constant;
- T is the temperature;
- Y has the dimensions $[\mathrm{m}]$, for instance;
- γ has the dimensions $[\mathrm{u\,s^{-1}}]$
- Δt has the dimensions $[\mathrm{s}]$;
- f has the the dimensions $[\mathrm{kJ\,mol^{-1}\,m^{-1}}]$

The integrator C.1 is dimensionally coherent:

$$\left[\frac{\mathrm{s\,kJ\,mol^{-1}\,m^{-1}}}{\mathrm{u\,s^{-1}}}\right] = \left[\frac{\mathrm{s^2\,kg\,m^2\,s^{-2}\,mol^{-1}\,m^{-1}}}{\mathrm{u}}\right] = [\mathrm{m}] \tag{C.3}$$

$$\sqrt{\left[\frac{\mathrm{kJ\,mol^{-1}\,s}}{\mathrm{u\,s^{-1}}}\right]} = \sqrt{\left[\frac{\mathrm{kg\,mol^{-1}\,m^2\,s^{-2}\,s^2}}{\mathrm{u}}\right]} = [\mathrm{m}] \tag{C.4}$$

We further suppose that the force f is linear with respect to the Collective Variable Y:

$$f(Y(t)) = k(Y(t) - Y_C) \tag{C.5}$$

where k has the dimensions $\left[\mathrm{kJ\,mol^{-1}\,m^{-2}}\right]$ and where Y_C is the position of the basin of attraction by which the force is generated. Then, Eq. (C.1) becomes

$$\begin{aligned} Y(t+\Delta t) =& Y(t) + \frac{k\Delta t}{\gamma} Y(t) - \frac{k\Delta t}{\gamma} Y_C + \sqrt{\frac{2k_B T\Delta t}{\gamma}} \cdot \eta_t \\ =& \left(1 + \frac{k\Delta t}{\gamma}\right) Y(t) - \frac{k\Delta t}{\gamma} Y_C + \sqrt{2D\Delta t} \cdot \eta_t \\ =& \left(1 + \rho\Delta t\right) Y(t) - \rho\Delta t Y_C + \sqrt{2D\Delta t} \cdot \eta_t \\ =& BY(t) - (B-1)Y_C + \alpha\eta_t \end{aligned} \tag{C.6}$$

where we inserted the definition of D, along with $\rho \equiv \frac{k}{\gamma}$, $B \equiv (1+\rho\Delta t)$ and $\alpha \equiv \sqrt{2D\Delta t}$. Eq. (C.6) is only the first iteration of the algorithm. We can apply the same machinery to $Y(t+\Delta t)$ to find $Y(t+2\Delta t)$

$$\begin{aligned} Y(t+2\Delta t) =& BY(t+\Delta t) - (B-1)Y_C + \alpha\eta_{t+\Delta t} \\ =& B\left[BY(t) - (B-1)Y_C + \alpha\eta_t\right] - (B-1)Y_C + \alpha\eta_{t+\Delta t} \\ =& B^2 Y(t) - (B-1)(B+1)Y_C + \alpha\left(B\eta_t + \eta_{t+\Delta t}\right) \end{aligned} \tag{C.7}$$

$$\begin{aligned} Y(t+3\Delta t) =& BY(t+2\Delta t) - (B-1)Y_C + \alpha\eta_{t+2\Delta t} \\ =& B\left[B^2 Y(t) - (B+1)(B-1)Y_C + \alpha\left(B\eta_t + \eta_{t+\Delta t}\right)\right] - (B-1)Y_C + \alpha\eta_{t+2\Delta t} \\ =& B^3 Y(t) - (B-1)(B^2+B+1)Y_C + \alpha\left(B^2\eta_t + B\eta_{t+\Delta t} + \eta_{t+2\Delta t}\right) \end{aligned} \tag{C.8}$$

and so on; the generic expression reads

$$Y(t+N\Delta t) = B^N Y(t) - (B-1)Y_C \sum_{i=0}^{N-1} B^i + \alpha \sum_{i=0}^{N-1} B^i \eta_{t+(N-i-1)\Delta t} \tag{C.9}$$

Using the property

$$\sum_{i=0}^{n} x^i = \frac{1-x^{n+1}}{1-x} \rightarrow \sum_{i=0}^{n-1} x^i = \frac{1-x^n}{1-x} \tag{C.10}$$

in Eq. (C.9), we find the final expression for $Y(t+\Delta t)$

$$\begin{aligned} Y(t+N\Delta t) =& B^N Y(t) - (B-1)Y_C \frac{1-B^N}{1-B} + \alpha \sum_{i=0}^{N-1} B^i \eta_{t+(N-i-1)\Delta t} \\ =& B^N (Y(t) - Y_C) + Y_C + \alpha \sum_{i=0}^{N-1} B^i \eta_{t+(N-i-1)\Delta t} \end{aligned} \tag{C.11}$$

C.2 Average displacement $\langle Y(t+N\Delta t)-Y(t)\rangle$ and fluctuations

We shall now compute the average values of the following expressions

$$\langle Y(t+N\Delta t)-Y(t)\rangle \qquad ; \qquad \langle [Y(t+N\Delta t)-Y(t)]^2\rangle$$

The first one reads

$$\begin{aligned}
\langle Y(t+N\Delta t)-Y(t)\rangle =&\langle \left[B^N(Y(t)-Y_C)+Y_C-Y(t)+\alpha\sum_{i=0}^{N-1}B^i\eta_{t+(N-i-1)\Delta t}\right]\rangle \\
=&\langle (B^N-1)(Y(t)-Y_C)\rangle + \langle \alpha\sum_{i=0}^{N-1}B^i\eta_{t+(N-i-1)\Delta t}\rangle \\
=&(B^N-1)(Y(t)-Y_C)+\alpha\sum_{i=0}^{N-1}B^i\langle\eta_{t+(N-i-1)\Delta t}\rangle \\
=&(B^N-1)(Y(t)-Y_C)+\alpha\langle\eta\rangle\sum_{i=0}^{N-1}B^i \\
=&(B^N-1)(Y(t)-Y_C) \\
=&\left((1+\rho\Delta t)^N-1\right)(Y(t)-Y_C) \\
=&\left((1+\rho\Delta t)^{\tau/\Delta t}-1\right)(Y(t)-Y_C)
\end{aligned} \tag{C.12}$$

where $\tau \equiv N\Delta t$ and where $\langle\eta_{t+(N-i-1)\Delta t}\rangle$ exits the sum, since the i in the subscript is only a label with no effect on the values η can assume at each time. We call $K(\tau)$ the average displacement occurring after a time interval τ

$$\begin{aligned}
K(\tau) \equiv&\langle Y(t+\tau)-Y(t)\rangle \\
=&\left((1+\rho\Delta t)^{\tau/\Delta t}-1\right)(Y(t)-Y_C)
\end{aligned} \tag{C.13}$$

which is precisely Eq. (2.26) for a generic starting point $Y(t)$. On the other side, the second expression reads

$$\begin{aligned}
\langle[Y(t+N\Delta t)-Y(t)]^2\rangle =&\langle\left[(B^N-1)(Y(t)-Y_C)+\alpha\sum_{i=0}^{N-1}B^i\eta_{t+(N-i-1)\Delta t}\right]^2\rangle\\
=&\langle(B^N-1)^2(Y(t)-Y_C)^2\rangle+\langle\left[\alpha\sum_{i=0}^{N-1}B^i\eta_{t+(N-i-1)\Delta t}\right]^2\rangle\\
=&K^2(\tau)+\alpha^2\langle\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}B^iB^j\eta_{t+(N-i-1)\Delta t}\ \eta_{t+(N-j-1)\Delta t}\rangle\\
=&K^2(\tau)+\alpha^2\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}B^iB^j\delta_{i,j}\\
=&K^2(\tau)+\alpha^2\sum_{i=0}^{N-1}B^{2i}\\
=&K^2(\tau)+\alpha^2\frac{1-B^{2N}}{1-B^2}\\
=&K^2(\tau)+2D\Delta t\ \frac{1-(1+\rho\Delta t)^{2\tau/\Delta t}}{1-(1+\rho\Delta t)^2}\\
=&K^2(\tau)+2D\frac{(1+\rho\Delta t)^{2\tau/\Delta t}-1}{\rho(2+\rho\Delta t)}
\end{aligned}\tag{C.14}$$

Then, we define the quantity $J(\tau)$ as

$$\begin{aligned}
J(\tau)\equiv&\langle[Y(t+\tau)-Y(t)]^2\rangle-[\langle Y(t+\tau)-Y(t)\rangle]^2\\
=&2D\frac{(1+\rho\Delta t)^{2\tau/\Delta t}-1}{\rho(2+\rho\Delta t)}
\end{aligned}\tag{C.15}$$

which is exactly Eq. (2.28); the dependence on the starting point $Y(t)$ is lost as $J(\tau)$ represents the fluctuations of the displacement $Y(t+\tau)-Y(t)$, a quantity which cannot depend on the initial condition if the diffusion coefficient D is constant.

Bibliography

[1] E. Fermi, Pasta J. and S. M. Ulam (1955). Studies of Nonlinear Problems. *Los Alamos Report LA–1940*.

[2] W. F. van Gunsteren and H. J. C. Berendsen (1977). Algorithms for Macromolecular Dynamics and Constraint Dynamics. *Mol. Phys.*, 34(5):1311–1327.

[3] W. F. van Gunsteren and M. Karplus (1982). Effect of Constraints on the Dynamics of Macromolecules. *Macromolecules*, 15(6):1528–1544.

[4] McCammon J. A., Gelin B. R. and Karplus M. (1972). Dynamics of Folded Proteins. *Nature*, 267(5612):585–590.

[5] Levitt M. and Warshel A. (1975). Computer Simulation of Protein Folding. *Nature*, 253(5494):694–702.

[6] J. D. Bryngelson and P. G. Wolynes (1987). Spin Glasses and the Statistical Mechanics of Protein Folding. *Proc. Natl. Acad. Sci.*, 84(21):7524–7528.

[7] J. D. Bryngelson and P. G. Wolynes (1989). Intermediates and Barrier Crossing in a Random Energy Model (with Applications to Protein Folding). *J. Phys. Chem.*, 93(19):6902–6915.

[8] E. I. Shakhnovich and A. V. Finkelstein (1989). Theory of cooperative transitions in protein molecules. i. why denaturation of globular protein is a first-order phase transition. *Biopolymers*, 28(10):1667–1680.

[9] D. Shortle (1996). The Denatured State (the Other Half of the Folding Equation) and its Role in Protein Stability. *FASEB J.*, 10(1):27–34.

[10] K. A. Dill and D. Shortle (1991). Denatured States of Proteins. *Annu. Rev. Biochem.*, 60:795–825.

[11] J. R. Gillespie and D. Shortle (1997). Characterization of Long-Range Structure in the Denatured State of Staphylococcal Nuclease. I. Paramagnetic Relaxation Enhancement by Nitroxide Spin Labels. *J. Mol. Biol.*, 268(1):158–169.

[12] J. R. Gillespie and D. Shortle (1997). Characterization of Long-Range Structure in the Denatured State of Staphylococcal Nuclease. II. Distance Restraints from Paramagnetic Relaxation and Calculation of an Ensemble of Structures. *J. Mol. Biol.*, 268(1):170–184.

[13] D. Shortle and M. S. Ackerman (2001). Persistence of Native-Like Topology in a Denatured Protein in 8 M Urea. *Science*, 293(5529):487–489.

[14] Y. Tang, M. J. Goger and D. P. Raleigh (2006). NMR Characterization of a Peptide Model Provides Evidence for Significant Structure in the Unfolded State of the Villin Headpiece Helical Subdomain. *Biochemistry*, 45(22):6940–6946.

[15] A. Morrone, M. E. McCully, P. N. Bryan, M. Brunori, V. Daggett, S. Gianni and C. Travaglini-Allocatelli (2011). The Denatured State Dictates the Topology of Two Proteins with almost Identical Sequence but Different Native Structure and Function. *J. Biol. Chem.*, 286(5):3863–3872.

[16] A. T. Alexandrescu, C. Abeygunawardana and D. Shortle (1994). Structure and Dynamics of a Denatured 13 1 -Residue Fragment of Staphylococcal Nuclease:. *Biochem.*, 5(33):1063–1072.

[17] E. P. O'Brien, R. I. Dima, B. Brooks and D. Thirumalai (2007). Interactions between Hydrophobic and Ionic Solutes in Aqueous Guanidinium Chloride and Urea Solutions: Lessons for Protein Denaturation Mechanism. *J. Am. Chem. Soc.*, 129(23):7346–7353.

[18] W. K. Lim, J. Rösgen and S. W. Englander (2009). Urea, but not Guanidinium, Destabilizes Proteins by Forming Hydrogen Bonds to the Peptide Group. *Proc. Natl. Acad. Sci. U. S. A.*, 106(8):2595–2600.

[19] A. Berteotti, A. Barducci and M. Parrinello (2011). Effect of Urea on the β-hairpin Conformational Ensemble and Protein Denaturation Mechanism. *J. Am. Chem. Soc.*, 133(43):17200–17206.

[20] S. J. Whittington, B. W. Chellgren, V. M. Hermann and T. P. Creamer (2005). Urea Promotes Polyproline II Helix Formation: Implications for Protein Denatured States. *Biochem.*, 44(16):6269–6275.

[21] A. Möglich, F. Krieger and T. Kiefhaber (2005). Molecular Basis for the Effect of Urea and Guanidinium Chloride on the Dynamics of Unfolded Polypeptide Chains. *J. Mol. Biol.*, 345(1):153–162.

[22] M. C. Stumpe and H. Grubmüller (2007). Interaction of Urea with Amino Acids: Implications for Urea-Induced Protein Denaturation. *J. Am. Chem. Soc.*, 129(51):16126–16131.

[23] M. D. Baer and C. J. Mundy. An ab initio Approach to Understanding the

Specific Ion Effect. In *Faraday Discuss.*, volume 160, pages 89–101, (2013).

[24] R. J. Cooper, S. Heiles, M. J. Ditucci and E. R. Williams (2014). Hydration of Guanidinium: Second Shell Formation at Small Cluster Size. *J. Phys. Chem. A*, 118(30):5657–5666.

[25] D. Cui, S.-C. Ou and S. Patel (2015). Protein Denaturants at Aqueous–Hydrophobic Interfaces: Self-Consistent Correlation between Induced Interfacial Fluctuations and Denaturant Stability at the Interface. *J. Phys. Chem. B*, 119(1):164–178.

[26] G. M. Torrie and J. P. Valleau (1977). Non-Physical Sampling Distributions in Monte-Carlo Free-Energy Estimation - Umbrella Sampling. *J. Comp. Phys.*, 23(2):187–199.

[27] A. Khachaturyan, S. Semenovskaya and B. Vainshtein (1979). Statistical-Thermodynamic Approach to Determination of Structure Amplitude Phases. *Sov. Phys. Crystallography*, 24(5):519–524.

[28] Y. Sugita and Y. Okamoto (1999). Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chemical Physics Letters*, 314(1-2):141–151.

[29] S. Piana and A. Laio (2007). A Bias-Exchange Approach to Protein Folding. *J. Phys. Chem. B*, 111(17):4553–4559.

[30] A. Barducci, M. Bonomi and M. Parrinello (2011). Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(October):826–843.

[31] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka and E. S. Shakhnovich (1998). On the Transition Coordinate for Protein Folding. *J Chem Phys*, 108(1):334–350.

[32] A. Berezhkovskii and A. Szabo (2005). One-Dimensional Reaction Coordinates for Diffusive Activated Rate Processes in Many Dimensions. *Journal of Chemical Physics*, 122(1):014503–014506.

[33] P. V. Banushkina and S. V. Krivov (2016). Optimal Reaction Coordinates. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 6(6):748–763.

[34] A. W. C. Lau and T. C. Lubensky (2007). State-Dependent Diffusion: Thermodynamic Consistency and its Path Integral Formulation. *Physical Review E*, 76(1):011123–011139.

[35] R. Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, Oxford UK, (2001).

[36] M. Ragwitz and H. Kantz (2001). Indispensable Finite Time Corrections for Fokker-Planck Equations from Time Series Data. *Physical Review Letters*, 87 (25):254501–254504.

[37] C. Anteneodo and R. Riera (2009). Arbitrary-Order Corrections for Finite-Time Drift and Diffusion Coefficients. *Physical Review E*, 80(3):031103–031110.

[38] D. Kleinhans, R. Friedrich, A. Nawroth and J. Peinke (2005). An Iterative Procedure for the Estimation of Drift and Diffusion Coefficients of Langevin Processes. *Physics Letters A*, 346(1-3):42–46.

[39] Steven J. Lade (2009). Finite Sampling Interval Effects in Kramers-Moyal Analysis. *Physics Letters A*, 373(41):3705–3709.

[40] C. Honisch and R. Friedrich (2011). Estimation of Kramers-Moyal Coefficients at Low Sampling Rates. *Physical Review E*, 83(6 pt. 2):066701–066706.

[41] G. Hummer (2005). Position-Dependent Diffusion Coefficients and Free Energies from Bayesian Analysis of Equilibrium and Replica Molecular Dynamics Simulations. *New Journal of Physics*, 7(34):1–14.

[42] S. Sriraman, I. G. Kevrekidis and G. Hummer (2005). Coarse Master Equation from Bayesian Analysis of Replica Molecular Dynamics Simulations. *J. Phys. Chem. B*, 109(14):6479–6484.

[43] R. B. Best and G. Hummer (2006). Diffusive Model of Protein Folding Dynamics with Kramers Turnover in Rate. *Physical Review Letters*, 96(22): 228104–228107.

[44] R. B. Best and G. Hummer (2010). Coordinate-Dependent Diffusion in Protein Folding. *Proceedings of the National Academy of Sciences of the United States of America*, 107(3):1088–93.

[45] C. Micheletti, G. Bussi and A. Laio (2008). Optimal Langevin Modeling of Out-of-Equilibrium Molecular Dynamics Simulations. *Journal of Chemical Physics*, 129(7):074105–074112.

[46] V. N. Uversky (2002). Natively Unfolded Proteins: a Point where Biology Waits for Physics. *Protein Sci.*, 11(4):739–756.

[47] J. A. Rupley (1964). The Hydrolysis of Chitin by Concentrated Hydrochloric Acid, and the Preparation of Low-Molecular-Weight Substrate for Lysozyme. *Biochim. Biophys. Acta*, 83(3):245–255.

[48] G. I. Makhatadze and P. L. Privalov (1992). Protein Interactions with Urea and Guanidinium Chloride. A Calorimetric Study. *J. Mol. Biol.*, 226(2):491–505.

[49] C. Camilloni, A. Guerini Rocco, I. Eberini, E. Gianazza, R. A. Broglia and G. Tiana (2008). Urea and Guanidinium Chloride Denature Protein L in Different Ways in Molecular Dynamics Simulations. *Biophys. J.*, 94(12):4654–

61.

[50] M. K. Frank, G. M. Clore and A. M. Gronenborn (1995). Structural and Dynamic Characterization of the Urea Denatured State of the Immunoglobulin Binding Domain of Streptococcal Protein G by Multidimensional Heteronuclear NMR Spectroscopy. *Prot. Sci.*, 4(12):2605–2615.

[51] H. S. Chung, J. M. Louis and W. A. Eaton (2010). Distinguishing between Protein Dynamics and Dye Photophysics in Single-Molecule FRET Experiments. *Biophys. J.*, 98(4):696–706.

[52] J. Kuszewski, G. M. Clore and A. M. Gronenborn (1994). Fast Folding of a Prototypic Polypeptide: the Immunoglobulin Binding Domain of Streptococcal Protein G. *Protein Sci.*, 3(11):1945–1952.

[53] O. Tcherkasskaya, J. R. Knutson, S. A. Bowley, M. K. Frank and A. M. Gronenborn (2000). Nanosecond Dynamics of the Single Tryptophan Reveals Multi-State Equilibrium Unfolding of Protein GB1. *Biochemistry*, 39(37): 11216–11226.

[54] F. J. Blanco, G. Rivas and L. Serrano (1994). A Short Linear Peptide that Folds into a Native Stable β–hairpin in Aqueous Solution. *Nat. Struct. Biol.*, 1(9):584–590.

[55] F. J. Blanco and L. Serrano (1995). Folding of Protein G B1 Domain Studied by the Conformational Characterization of Fragments Comprising Its Secondary Structure Elements. *Eur. J. Biochem.*, 230(2):634–649.

[56] F. J. Blanco, A. R. Ortiz and L. Serrano (1997). Role of a Nonnative Interaction in the Folding of the Protein G B1 Domain as Inferred from the Conformational Analysis of the α-helix Fragment. *Fold. Des.*, 2(2):123–133.

[57] G. Bussi, F. L. Gervasio, A. Laio and M. Parrinello (2006). Free-Energy Landscape for β-hairpin Folding from Combined Parallel Tempering and Metadynamics. *J. Am. Chem. Soc.*, 128(41):13435–13441.

[58] M. Bonomi, D. Branduardi, F. L. Gervasio and M. Parrinello (2008). The Unfolded Ensemble and Folding Mechanism of the C-terminal GB1 β-hairpin. *J. Am. Chem. Soc.*, 130(42):13938–13944.

[59] R. B. Best and J. Mittal (2011). Microscopic Events in β-hairpin Folding from Alternative Unfolded Ensembles. *Proc. Natl. Acad. Sci. U. S. A.*, 108 (27):11087–11092.

[60] R. Capelli, F. Villemot, E. Moroni, G. Tiana, A. van der Vaart and G. Colombo (2016). Assessment of Mutational Effects on Peptide Stability through Confinement Simulations. *J. Phys. Chem. Lett.*, 7(1):126–130.

[61] R. Meloni, C. Camilloni and G. Tiana (2013). Sampling the Denatured State of Polypeptides in Water, Urea, and Guanidine Chloride to Strict Equilibrium Conditions with the Help of Massively Parallel Computers. *J. Chem. Theory Comput.*

[62] R. B. Best, D. de Sancho and J. Mittal (2012). Residue-Specific α-helix Propensities from Molecular Simulation. *Biophys. J.*, 102(6):1462–1467.

[63] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni and G. Bussi (2014). PLUMED 2: New Feathers for an Old Bird. *Comp. Phys. Comm.*, 185(2):604–613.

[64] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and Lindahl E. (2015). GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX*, 1–2:19–25.

[65] F. Pietrucci and A. Laio (2009). A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *J. Chem. Theory Comput.*, 5(9):2197–2201.

[66] X. Biarnés, F. Pietrucci, F. Marinelli and A. Laio (2012). METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. *Comput. Phys. Commun.*, 183(1):203–211.

[67] N. A. Baker, D. Sept, S. Joseph, M. J. Holst and J. A. McCammon (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.*, 98(18):10037–10041.

[68] Humphrey W., Dalke A. and Schulten K. (1996). VMD – Visual Molecular Dynamics. *J. Mol. Graph.*, 14(1):33–38.

[69] J. Stone. *An Efficient Library for Parallel Ray Tracing and Animation*. Master's thesis, Computer Science Department, University of Missouri-Rolla, (1998).

[70] `http://gnuplot.sourceforge.net/`.

[71] `http://plasma-gate.weizmann.ac.il/Grace/`.

[72] N. Greenfield and G. D. Fasman (1969). Computed Circular Dichroism Spectra for the Evaluation of Protein Conformation. *Biochemistry*, 8(10):4108–4116.

[73] D. Frishman and Argos P. (1995). Knowledge-Based Protein Secondary Structure Assignment. *Proteins*, 23(4):566–579.

[74] Y. Shen and A. Bax (2007). Protein Backbone Chemical Shifts Predicted from Searching a Database for Torsion Angle and Sequence Homology. *J. Biomol. NMR*, 38(4):289–302.

[75] B. H. Zimm and J. K. Bragg (1959). Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. *J. Chem. Phys.*, 31(2):526–535.

[76] C. Louis-Jeune, M. A. Andrade-Navarro and C. Perez-Iratxeta (2012). Prediction of Protein Secondary Structure from Circular Dichroism using Theoretically Derived Spectra. *Proteins Struct. Funct. Bioinforma.*, 80(2):374–381.

[77] S. W. Provencher and J. Glöckner (1981). Estimation of Globular Protein Secondary Structure from Circular Dichroism. *Biochemistry*, 20(1):33–39.

[78] A. R. Fersht and V. Daggett (2002). Protein Folding and Unfolding at Atomic Resolution. *Cell*, 108(4):573–582.

[79] C. Tanford (1970). Protein Denaturation. C. Theoretical Models for the Mechanism of Denaturation. *Adv. Protein Chem.*, 24:1–95.

[80] V. Hoepfner, V. L. Deringer and R. Dronskowski (2012). Hydrogen-Bonding Networks from First-Principles: Exploring the Guanidine Crystal. *J. Phys. Chem. A*, 116(18):4551–4559.

[81] H. Risken. *The Fokker-Planck Equation - Methods of Solutions and Applications*. Springer–Verlag, Berlin GER, (1989).

[82] R. Friedrich, J. Peinke, M. Sahimi and M. Reza Rahimi Tabar (2011). Approaching Complexity by Stochastic Methods: from Biological Systems to Turbulence. *Phys. Rep.*, 506(5):87–162.

[83] A. Laio and M. Parrinello (2002). Escaping Free-Energy Minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12562–12566.

[84] W. Zheng, M. A. Rohrdanz and C. Clementi (2013). Rapid Exploration of Configuration Space with Diffusion-Map-Directed Molecular Dynamics. *J. Phys. Chem. B*, 117(42):12769–12776.

[85] C. Abrams and G. Bussi (2014). Enhanced Sampling in Molecular Dynamics using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy*, 16(1):163–199.

[86] P. Tiwary and B. J. Berne (2016). Kramers Turnover: from Energy Diffusion to Spatial Diffusion using Metadynamics. *Journal of Chemical Physics*, 144 (13):20–23.

[87] P. J. Kraulis (1991). Similarity of Protein G and Ubiquitin. *Science*, 254(5031): 581–582.

[88] J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu and J. N. Onuchic (2010). SMOG@ctbp: Simplified Deployment of Structure-Based Models in GROMACS. *Nucleic Acids Research*, 38(SUPPL. 2):657–661.

[89] S. Piana, K. Lindorff-Larsen and D. E. Shaw (2011). How Robust are Protein Folding Simulations with respect to Force Field Parameterization? *Biophysical journal*, 100(9):L47–L49.

Ringraziamenti

"Giunto al termine del giorno,
cerco fra le coltri un poco di speranza."
– Elio, *Nubi di ieri sul nostro domani odierno (abitudinario)*

"All'inizio ho pensato di aver capito.
Poi di non aver capito davvero.
Poi, ancora, di aver capito."
– Il sottoscritto, ne *Gli Scafisti Anonimi*

Gennaio 2017, termine di un percorso che è iniziato nel lontano settembre del 2006. Fresco di diploma di maturità (e di ~~sbronza~~ gioia post-festeggiamenti per i mondiali, *of course*), non avevo la minima idea di cosa mi sarebbe capitato negli anni a venire. Devo essere onesto: mi ero iscritto alla facoltà di Fisica con l'idea - molto velleitaria - di diventare uno di quei professoroni che dissertano di fisica teorica e massimi sistemi, vincono i Nobel e/o finiscono nei libri di storia. Avevo letto troppi testi di divulgazione scientifica, temo. Un giorno (non di pioggia) di qualche anno dopo, per caso, mi capita l'occasione di sentir parlare di questo corso - *si chiama "Fisica delle Proteine"*, mi dicono - e del suo professore, che si è messo a spiegare il Random Walk mimando la camminata di un ubriaco in mezzo all'aula. L'argomento è potenzialmente interessante, il prof. sembra simpatico, perché non approfondire? Ed è così che l'anno successivo mi sono ritrovato immerso in un contesto fatto di *palline e stecche*, catene fluttuanti e simulazioni che crashavano ogni due per tre.. e tutto ciò mi piaceva, mi piaceva veramente un sacco. Mi piaceva a tal punto che ho deciso di trascorrere quattro anni e mezzo a simulare *palline e stecche* (ehi, mica solo quello. A volte c'erano pure delle molle!).

Storielle a parte, è per questo che voglio ringraziare quel professore, Guido Tiana: per essere stato in grado di trasmettermi quella passione per le scienze della vita analizzate con l'occhio del fisico, quell'interesse nei confronti dei meccanismi più sottili e quell'abilità nel saper osservare un problema da differenti punti di vista per ricavarne sempre un messaggio stimolante che ora permeano - o che vorrei permeassero - il mio modo di affrontare la vita. Il tutto, guidandomi già dai tempi della Laurea Magistrale fino al completamento di questa Tesi e dandomi sempre una grande libertà nel definire i miei metodi ed obiettivi.

Il secondo ringraziamento è senza dubbio per Carlo Camilloni, il quale in più di un'occasione è stato da me definito *Deus ex Machina* per la naturalezza con cui riesce a trovare il modo giusto per affrontare i problemi e che ha generosamente speso molte ore del suo tempo con me, in utili consigli su come affrontare il mio lavoro in maniera efficace ed in stimolanti discussioni, di carattere scientifico ma spesso anche personale.

Durante i miei anni trascorsi in università, ho avuto modo di conoscere decine di persone con cui sono orgoglioso di aver condiviso i momenti migliori - e pure quelli peggiori - della mia carriera accademica. Temo che un elenco esaustivo sia fuori dalle mie possibilità, ma tra di questi meritano una menzione particolare Riccardo Capelli (la causa prima del mio percorso di studi, nonostante le pessime premesse con cui abbiamo iniziato a conoscerci) e Francesco Villa, oltre naturalmente a tutti gli Scafisti Anonimi.

Sarà forse una frase fatta, ma è immensamente vero: sapere che c'è sempre qualcuno, a casa, che farà sempre il tifo per me è quello che, in certe occasioni, mi ha convinto a perseverare in quel che facevo. Prima erano i miei genitori, Silvia e Gianni, ad insegnarmi che talvolta basta una parola o un sorriso per svoltare una giornata storta, ed ora è Alessandra, con cui ho trascorso dieci anni meravigliosi e che ben presto diventerà mia moglie. Mio fratello Gianluca, il quale mi ritiene l'artefice di quello che è diventato: ora spesso è lui la mia fonte di ispirazione e ne sono immensamente orgoglioso; mia zia Paola, sempre disponibile e prodiga di consigli; Giovanna, Enrico e Marco, che mi hanno fin da subito accolto come un figlio e fratello.

Ringrazio infine l'Università degli Studi di Milano e il Ministero dell'Istruzione, Università e Ricerca per il supporto finanziario attraverso la borsa di studio, senza la quale questa Tesi non avrebbe mai visto la luce.