Unconstrained Structure Formation in Coarse-Grained Protein Simulations

Tristan Bereau

April 2011

Department of Physics
Carnegie Mellon University
Pittsburgh, PA 15213

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Thesis committee:
Markus Deserno, Chair
Maria Kurnikova
Mathias Lösche
Robert H. Swendsen
Daniel M. Zuckerman

Abstract

The ability of proteins to fold into well-defined structures forms the basis of a wide variety of biochemical functions in and out of the cell membrane. Many of these processes, however, operate at time- and length-scales that are currently unattainable by all-atom computer simulations. To cope with this difficulty, increasingly more accurate and sophisticated coarse-grained models are currently being developed.

In the present thesis, we introduce a solvent-free coarse-grained model for proteins. Proteins are modeled by four beads per amino acid, providing enough backbone resolution to allow for accurate sampling of local conformations. It relies on simple interactions that emphasize structure, such as hydrogen bonds and hydrophobicity. Realistic α/β content is achieved by including an effective nearest-neighbor dipolar interaction. Parameters are tuned to reproduce both local conformations and tertiary structures. By studying both helical and extended conformations we make sure the force field is not biased towards any particular secondary structure. Without any further adjustments or bias a realistic oligopeptide aggregation scenario is observed.

The model is subsequently applied to various biophysical problems: (i) kinetics of folding of two model peptides, (ii) large-scale amyloid-β oligomerization, and (iii) protein folding cooperativity. The last topic—defined by the nature of the finite-size thermodynamic transition exhibited upon folding—was investigated from a microcanonical perspective: the accurate evaluation of the density of states can unambiguously characterize the nature of the transition, unlike its corresponding canonical analysis. Extending the results of lattice simulations and theoretical models, we find that it is the interplay between secondary structure and the *loss* of non-native tertiary contacts which determines the nature of the transition.

Finally, we combine the peptide model with a high-resolution, solvent-free, lipid model. The lipid force field was systematically tuned to reproduce the structural and mechanical properties of phosphatidylcholine bilayers. The two models were cross-parametrized against atomistic potential of mean force curves for the insertion of single amino acid side chains into a bilayer. Coarse-grained transmembrane protein simulations were then compared with experiments and atomistic simulations to validate the force field. The transferability of the two models across amino acid sequences and lipid species permits the investigation of a wide variety of scenarios, while the absence of explicit solvent allows for studies of large-scale phenomena.

Acknowledgments

There are a number of people who have contributed to this thesis in one way or another—whether they taught me something new, shared their scientific interests and ideas, or simply enjoyed some company for a drink or a (preferably good) meal after a day's work. They have made my time in graduate school both enriching and enjoyable.

First and foremost, I would like to thank my Ph.D. advisor, Markus Deserno—working with him has been an incredible experience. I am grateful to him for suggesting the field of protein coarse-graining, providing critical advice and insight, and allowing much freedom in my research. Not only have I learned a great deal from his good judgment, care for detail, and enthusiasm for research, Markus has also taught me many non-science related skills, such as eating noodles with chopsticks, folding T-shirts very quickly, and tying my shoes the *right* way (chirality will come back in chapter 2 in the context of amino acids, rather than shoelaces). His spending time with his students and post-docs outside of school has forged a strong cohesion among us (also known as the "Awesome Deserno Group").

I would like to thank the other members of my thesis committee: Maria Kurnikova, Mathias Lösche, Bob Swendsen, and Dan Zuckerman. Joint group meetings with Mathias' group has provided lively discussions and useful feedback, often followed by enjoyable dinners. I am indebted to Bob for teaching me the basics of computational statistical mechanics and efficient ways of calculating thermodynamics (Appendix A) during a semester-long research project.

Thanks to Bob, I also had the chance to attend David Landau's CSP workshop in Athens, GA (2009), during which I met Michael Bachmann, who was then a young investigator at the Forschungszentrum Jülich, Germany. I have had the chance to collaborate with Michael and learned a great deal from his experience on finite-size thermodynamics. His expertise, patience, and devotion to research have been a valuable example to me. I thank him and his group for welcoming me to Jülich twice.

I was very fortunate to visit the Max Planck Institute for Polymer Research in Mainz, Germany, several times during my Ph.D. Working in Kurt Kremer's theory group—and interacting with its members—has been a memorable experience, both stimulating and enjoyable. I hope all of them know they are missed.

David C. Stone and Senthil Kumar Muthiah made important contributions to chapters 3 and 5, respectively. I am grateful to Christine Peter, Luca Monticelli, Will Noid, and Thomas Weikl for providing critical help and advice in my research. Many thanks to Mingyang Hu for proofreading parts of this thesis. I acknowledge support from an Astrid and Bruce McWilliams Fellowship.

Zun-Jing Wang, Mingyang Hu, and Cem Yolcu have been incredible office mates. I thank

Zun-Jing for sharing her experience on membrane simulations, and Mingyang and Cem for endless discussions about physics, computer-related topics, food, and everything else. On the personal side, I am especially grateful to Irene Cooke, Michelle Hicks Ntampaka, Mary Jane Hutchison, Agnieszka Kalinowski, Xiaofei Li, Luxmi, Susan Santa-Cruz, Nishtha Srivastava, Donna Thomas, Ryan Booth, Matteo Broccio, Patrick Diggins, David Frank, Jason Giulieri, Frank Heinrich, Venky Krishnamani, Oscar Marchat, Radu Moldovan, Pedro, Pierre-Louis Pernet, Rémy Praz, Prabhanshu Shekhar, Sidd Shenoy, and Karpur Shukla for making my stay in Pittsburgh so enjoyable. I also thank David Cesaro from the Shadyside-based, french restaurant *Brasserie 33* for several incredibly fun evenings.

Je suis reconnaissant à Murat Kunt et Rémi André pour m'avoir tous deux convaincu d'étudier la physique. Je termine en remerciant ma famille, en particulier mes parents et mon frère, pour leurs encouragements et leur amour, bien loin de la maison.

List of Publications

Chapter 2:

T. Bereau and M. Deserno. *Generic coarse-grained model for protein folding and aggregation*, J. Chem. Phys. **130** (2009), no. 23, 235106–235120.

Chapter 4:

T. Bereau, M. Bachmann, and M. Deserno. *Interplay between secondary and tertiary structure formation in protein folding cooperativity*, J. Am. Chem. Soc. **132** (2010), no. 38, 13129–13131.

T. Bereau, M. Deserno, and M. Bachmann. *Structural basis of folding cooperativity in model proteins: Insights from a microcanonical perspective*, Biophys. J. (2011), in press (doi:10.1016/j.bpj.2011.03.056).

Chapter 6:

T. Bereau, Z.-J. Wang, and M. Deserno. *High-resolution, solvent-free coarse-grained model for protein-lipid interactions*, in preparation.

Appendix A:

T. Bereau and R. H. Swendsen. *Optimized convergence for multiple histogram analysis*, J. Comput. Phys. **228** (2009), no. 17, 6119–6129.

Contents

1 Background and Motivation

1.1 Proteins

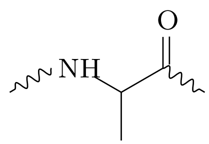
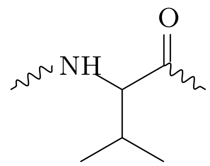
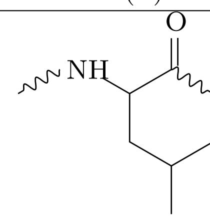
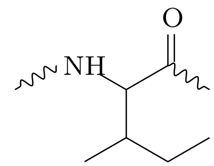
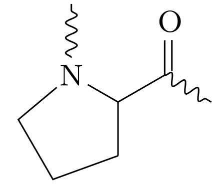
Proteins and peptides are polymeric organic compounds made of amino acids, which constitute one of the four major building blocks of molecular biology.[1] They are evolutionarily optimized heteropolymers, whose physical and material properties more often than not exceed what can be readily understood from conventional polymer physics reasoning, which derives much of its strength from uniformity, randomness, and the law of large numbers. In contrast, the complexity of proteins rests on the different physical and chemical properties of their monomers, the twenty physiological amino acids, and their intricate combination into what at cursory inspection only *seems* to be a random heteropolymer sequence. Moreover, the main interactions that drive their folding into intricate secondary, tertiary and quaternary functional structures are weak, comparable to thermal energy. The overall stability of a protein is perilously marginal [Pac90, PSMG96], so proteins very often rely on cooperative effects to keep them in their native structure—one appealing reason for why they might be so much bigger than what their comparatively small active centers would make one suspect [SHHB09]. Their ability to fold provides a basis for the many biochemical functions they provide within organisms, e.g., catalysis, cell signaling, immune responses, motors, channels, structural and mechanical building blocks [GG08, Cre92, Fer98, BTS10].

1.1.1 Amino acids

Proteins are linear polymers built from amino acids. The generic chemical structure of all amino acids is characterized by a central (so-called α-)carbon which is connected to an amino group, a carboxyl group, a side chain, and one more hydrogen.[2] The asymmetry of the α-carbon atom results in *chiral* molecules, which feature a non-superposable mirror image. These two optical isomers are denoted L- and D-. Only the side chain varies between different residues. Table 1.1 and Table 1.2 provide the name, structure, and certain chemical characteristics of the twenty standard (naturally occurring) amino acids [Hay10]. This collection provides a wide library of possible primary structures, i.e., amino acid sequences. The chemical properties involved are key to understand the large-scale characteristics of proteins.

[1] The other three being lipids, nucleic acids, and polysaccharides.

[2] The amino and carboxyl groups become amide and carbonyl, respectively, once a peptide bond forms between two amino acids through a condensation reaction.

	Alanine Ala (A)	Valine Val (V)	Leucine Leu (L)	Isoleucine Ile (I)	Proline Pro (P)
Structure					
Type	hydrophobic	hydrophobic	hydrophobic	hydrophobic	hydrophobic
Side-chain pK_a	–	–	–	–	–

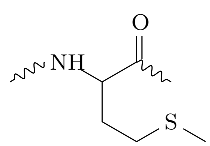
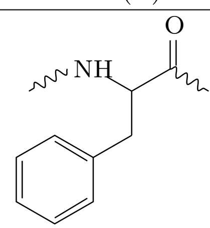
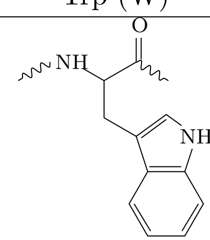
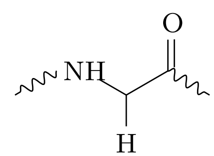
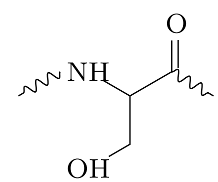
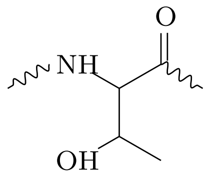
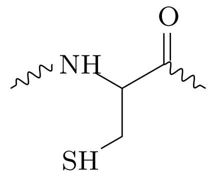
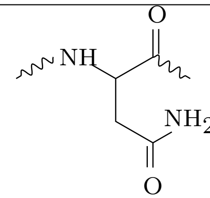
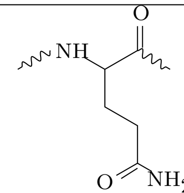
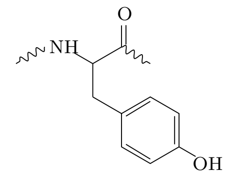
	Methionine Met (M)	Phenylalanine Phe (F)	Tryptophan Trp (W)	Glycine Gly (G)	Serine Ser (S)
Structure					
Type	hydrophobic	hydrophobic	hydrophobic	hydrophilic	hydrophilic
Side-chain pK_a	–	–	–	–	–

Table 1.1: Amino acid names, structures, and chemical characteristics (part 1/2). Squiggly lines represent the peptide bonds connecting neighboring amino acids.

	Threonine Thr (T)	Cysteine Cys (C)	Asparagine Asn (N)	Glutamine Gln (Q)	Tyrosine Tyr (Y)
Structure					
Type	hydrophilic	hydrophilic	hydrophilic	hydrophilic	hydrophilic
Side-chain pK_a	–	8.1	–	–	10.1

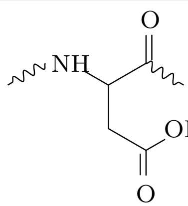
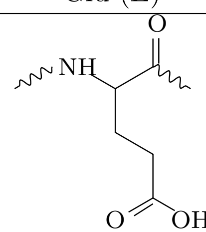
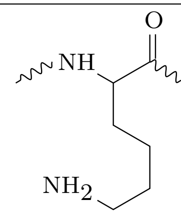
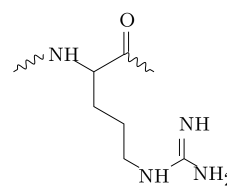
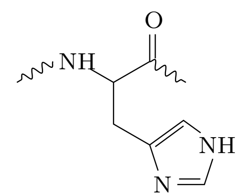
	Aspartic acid Asp (D)	Glutamic acid Glu (E)	Lysine Lys (K)	Arginine Arg (R)	Histidine His (H)
Structure					
Type	acidic	acidic	basic	basic	basic
Side-chain pK_a	3.7	4.1	10.7	12.1	6.0

Table 1.2: Amino acid names, structures, and chemical characteristics (part 2/2). Squiggly lines represent the peptide bonds connecting neighboring amino acids. Note that the neutral form of histidine exists in two forms because the hydrogen atom shown on the imidazole ring can be located on either of the two nitrogen atoms.

The interactions taking place between amino acids arise mainly from electrostatics:

Covalent bond ($\sim 100\, k_{\mathrm{B}}T$) chemical bond due to the sharing of pairs of electrons between atoms.

Salt bridge ($\sim 10\, k_{\mathrm{B}}T$) attraction between two oppositely charged residues.

Coulomb ($\sim 1 - 10\, k_{\mathrm{B}}T$) interaction between charged atoms; the presence of the surrounding solvent involves large gradients in the dielectric constant ϵ (≈ 80 in water compared to ≈ 5 in the protein interior), such that the strength of the interaction varies greatly.

Hydrogen bond ($3 - 10\, k_{\mathrm{B}}T$) attractive, highly-directional, non-bonded interaction between a polar hydrogen atom ("donor") and an electronegative atom (e.g., nitrogen, oxygen) with a nonbonding orbital ("acceptor"); the hydrogen atom must be covalently bonded to another electronegative atom to leave it with a partial positive charge [Mar07].

van der Waals ($\lesssim 1\, k_{\mathrm{B}}T$) the van der Waals interactions between atoms describe an attraction due to (*i*) fluctuating multipoles (including neutral atoms), i.e., London dispersion, and (*ii*) permanent-multipole–induced-multipole forces, i.e., Debye induction.

Hydrophobic effect ($\lesssim 1\, k_{\mathrm{B}}T$) entropic driving force for self-association of non-polar groups in water; the hydrophobic effect describes the loss of water entropy due to a hydrophobic group which enforces constraints on the hydrogen-bond network in its vicinity [Tan80].

Apart from covalent bonds, all interactions are weak (i.e., comparable to thermal energy at room temperature). Thus (*i*) thermal fluctuations will have a predominant role in the structures of proteins, (*ii*) the conformations adopted by polypeptides tend to be marginally stable, and (*iii*) entropic effects will largely dominate, associating proteins to the class of soft-matter systems.

The chemical structure of each amino acid is shown in Table 1.1 and Table 1.2. Four main amino acid types can be distinguished: hydrophobic, hydrophilic, acidic, and basic. The first two refer to non-polar and polar side chains, respectively, a distinction which impacts their (self-)association in water. Acidic and basic amino acids may be negatively or positively charged, respectively, depending on their pK_a (i.e., logarithmic measure of the chemical equilibrium between an acid and its conjugate base) and the surrounding pH (i.e., logarithmic measure of the hydrogen ion activity in a solution). Given the side-chain pK_a of each amino acid in Table 1.1 and Table 1.2, a solution at neutral pH (≈ 7) will contain negatively charged Asp and Glu, positively charged Lys and Arg, while the protonation of His will depend sensitively on the pH (because pH and pK_a have similar values) [GG08].

1.1.2 Folding and thermodynamics

Most proteins can only perform their biochemical function while in their native—folded—state.[3] For instance, the establishment and control of a voltage gradient across the cell membrane requires ion channels and pumps to properly assemble across the bilayer. The ability of proteins to reliably stabilize a given structure out of many possible conformations is a remarkable feature that distinguish them from most polymers. This property stems from the heterogeneity and chemical variety of amino acids: some interactions will be favored over others, and this bias creates the ensemble of stable conformations. Moreover, the pioneering work of Anfinsen showed that, provided suitable environmental conditions (e.g., temperature and pH) hold, an unfolded protein will *spontaneously* fold into its native structure [Anf72].[4] This suggests that protein folding, in its simplest form, is a self-assembly process and that the thermodynamics of the system is contained within the amino acid sequence of the chain. For recent reviews on protein folding, see [BVP11, DOW$^+$07].

The existence of a unique native structure (including its thermal fluctuations) implies that no other state exhibits a lower free energy. Hence, the resulting "free energy landscape" of a protein exhibits a single global minimum.

In order to cope with small changes in the surrounding environment, the free energy minimum must be stable. Stability is quantified by the difference between the free energy of the native structure and the free energy of the next lowest metastable state(s). Lattice simulations showed that the most stable single-ground-state heteropolymer sequences exhibit a funnel-like energy landscape, in which large variations in energy and entropy compete and result in small free energies of only $\approx 1 - 10\, k_{\mathrm{B}}T$ [BOSW95, OW04]. Indeed, the energetics involved in folding an extended chain into its native conformation require, among other things, the formation of many hydrogen bonds and hydrophobic contacts—leading to a large *gain* in energy and a strong *loss* of entropy (see Figure 1.1). A steep energy surface around the native state optimizes thermodynamic stability.

Another consequence of the funnel-like energy landscape theory is kinetic accessibility between the native state and *any* unfolded conformation. This resolves Levinthal's paradox [Lev69], which argues that sequentially sampling all possible protein conformations in order to attain the correct folded state would require astronomically long folding times.[5] The study of kinetic pathways, which describes *how* proteins fold in time, has lately been the subject of intense research and has greatly benefited from computer simulations (e.g., [Caf06, VBBP10]).

---

[3] A state, as defined in statistical physics, is a probability density in phase space. It does not map to a *single* conformation, which would have no statistical weight. Instead, it is really understood as an ensemble.

[4] However, more complex proteins may require chaperones in order to fold [Fin99].

[5] One wonders, though, whether this can really be considered a paradox. The very idea of a sequential search through all possible conformations seems to negate fundamental concepts from thermodynamics and statistical mechanics.

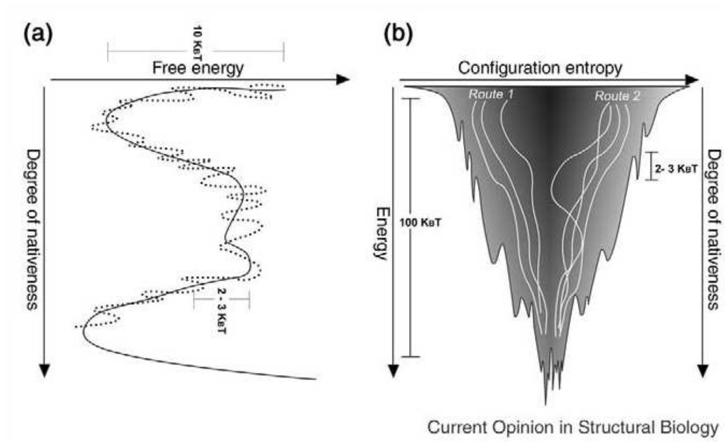


Figure 1.1: Cartoon of the folding energy landscape is shown in (b). This illustrates the competition between few low-energy, low-entropy folds (bottom; folded) with a large number of high-energy, high-entropy conformations (top; unfolded). The ruggedness of the landscape suggests the existence of kinetic traps. On the left (a), the corresponding free energy as a function of a valid order parameter (i.e., one that can describe the evolution of the system) shows two minima: one at high energy, high entropy (unfolded) and the other at low energy, low entropy (folded). Reprinted from Current Opinion in Structural Biology, **14**, J. N. Onuchic and P. G. Wolynes, *Theory of protein folding*, 70–75 [OW04], Copyright (2004), with permission from Elsevier.

1.2 Computer simulations

The use of computer simulations to predict and understand the behavior of proteins (as well as many other physical systems, ranging from quarks to the universe) is an evergrowing field. They provide a highly resolved picture that is complementary both to experimental measurements and theoretical (i.e., analytical) calculations. By defining interaction potentials between neighboring atoms (i.e., bonded interactions) and atom pairs (i.e., nonbonded interactions), one can sample a time or ensemble average of a system by means of molecular dynamics or Monte Carlo simulations, respectively.[6] While Monte Carlo simulations sample states by selecting random configurations according to a predefined distribution function (e.g., the Boltzmann distribution for canonical sampling), molecular dynamics numerically integrates the (classical) equations of motion of the system, usually subject to extra noise and friction terms (called "thermostats") which create the ensemble of interest (e.g., [AT93, FS01]).

[6]For more details on the equivalence between time and ensemble averages, see the introduction of Appendix A.

1.2.1 Analysis

The collection of conformations sampled from a simulation are then analyzed in order to answer a specific question, compare data with experiment, or gain insight into a system. The overwhelming amount of information contained in the $3N$ microscopic degrees of freedom of a N-particle system calls for appropriate, low-dimensional *observables* that characterize the macroscopic state of system. These observables are evaluated at suitably chosen time intervals during the simulation for subsequent analysis. Such analysis is often straightforward: in a canonical simulation,[7] for instance, the canonical average of any observable $\mathcal{O}$ can simply be obtained from the arithmetic mean of all (hopefully equilibrated) data points $\mathcal{O}_i$ (see subsection A.1.1 on page 117). For example, the total energy of a system, E, can be recorded at any time during the simulation. The histogram of equilibrated values will converge towards the probability distribution function $p(E) \propto \Omega(E)\exp\left(-E/k_\mathrm{B}T\right)$, where $\Omega(E)$ is the density of states.

Such histograms readily provide the means to characterize the stability of a structure through the evaluation of *free energies*.[8] Projected along an observable $\mathcal{O}$, the free energy $F(\mathcal{O})$ describes the stability of the system as a function of this observable. Canonically, this is expressed as

$$F(\mathcal{O}) = -k_\mathrm{B}T \ln\left(\frac{p(\mathcal{O})}{p_\mathrm{max}}\right), \tag{1.1}$$

at a temperature T and using an arbitrary reference probability p_max. By definition, it is impossible to evaluate an "instantaneous" free energy $F_i(\mathcal{O}_i)$ at any given time i in the simulation, because the free energy is an ensemble property. Practically, calculating Equation 1.1 requires a bit more sophistication, because exhaustively sampling $p(\mathcal{O})$ by brute force counting in a canonical simulation fails for all but very small systems. Smarter techniques to calculate free energies are presented in Appendix A and used throughout the present thesis.

1.2.2 Atomistic simulations

Atomistic simulations follow the motion of every single atom and describe interactions between them using a classical force field. One of the earliest attempts to study atomistically the time-evolution of proteins from molecular dynamics was presented by McCammon, Gelin, and Karplus [MGK77]. They presented a 9 ps-long simulation of the bovine pancreatic trypsin inhibitor protein in vacuo at an all-atom resolution. This study is illustrative in a number of ways:

- While (breathtakingly) short, the simulation nevertheless showed the predominance

[7] i.e., a simulation in which suitable thermostats ensure that the time average of the simulated trajectory coincides with the canonical state.

[8] Recall that in a soft-matter system, the most stable configuration is given by the minimum in the *free energy*, rather than the energy.

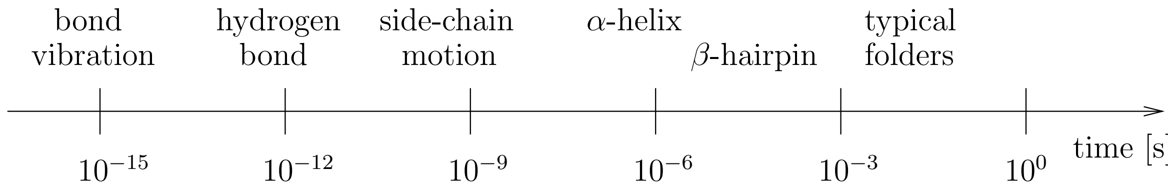


Figure 1.2: Several timescales involved in protein folding.

of fluctuations and contributed in dismantling the old belief that proteins are static bricks of matter.

- The time-averaged structure deviates from the X-ray configuration. While expected (the structure in the crystal is constrained; the simulation is run in vacuo), it is not clear how much of the difference is simply due to force-field inaccuracies. The impact of such systematic errors on the resulting structures is often difficult to probe. In more than three decades, atomistic force-fields have become substantially more accurate, and yet they can still show secondary structure bias in small protein folding simulations [FPRS09].

- The time-scales involved in protein simulations span many orders of magnitude, ranging between bond vibration ($\sim 10\,\text{fs}$) and protein folding (from μs to more than a second), as shown in Figure 1.2. From a simulation point of view, the proper integration of the equations of motion requires a time step that is roughly 10 times smaller than the fastest degree of freedom in the system (in the range $1-2\,\text{fs}$ for atomistic resolution). The computational power available thus limits the accessible timescale. As shown in Figure 1.2, the simulation time of McCammon *et al.* is about 10 orders of magnitude away from the folding time of a typical protein. Nowadays, technological advancements allow for simulations of small proteins (i.e., $10-50$ residues) in the μs timescale (e.g., [MC07, FPRS09, VBBP10]). Simulations on specific hardware have recently reached 1 millisecond [SMLL+10].

While corresponding quantum simulations would yield much more accurate results, their use is severely limited to very small systems (i.e., $\sim 10-100$ atoms) because of obvious computational limitations. Their use in biomolecular simulations is thus restricted to the study of active sites or localized chemical reactions.

1.2.3 Coarse-graining

As mentioned above, the advancement of protein simulations is strongly limited by force-field development and computational power. While the invention of sophisticated simulation methods (e.g., generalized-ensemble techniques, distributed computing, specific hardware architecture) has helped investigating problems which arise at longer time- and length-scales, the difficulty in attaining thermodynamic equilibrium has called for alternative techniques. One of them is called *coarse-graining*.

Coarse-graining relies on the concept of separation of length- and time-scales in physical systems. Figure 1.2 illustrates the correlation between timescales and typical processes involved in protein folding, i.e., larger processes happen over longer times. These longer timescale movements depend, to a large extent, on the *average behavior* of the faster ones—rather than their detailed dynamics. By lowering the level of resolution, coarse-grained (CG) simulations average over fast degrees of freedom to focus on larger (time- and length-) scales [Toz05, Vot08].

Computationally, coarse-graining offers enticing features: a smaller number of quasi-atoms—or *beads*—decreases the computational requirements and thus accelerates the speed of Monte Carlo or molecular dynamics simulations. In addition, the coarse-grained potentials tend to be softer than their atomistic counterparts so that larger integration time steps can be used. Coarse-graining also smoothens out the free energy landscape by reducing molecular friction, which artificially accelerates the dynamics even more and makes phase space both smaller and more navigable.

Yet, the development of a new coarse-grained model is not without effort, as it requires proper mapping (between atoms and CG beads), parametrization (i.e., potentials of interaction), and testing for legitimate validation. Similar to atomistic force-fields, the interaction potentials are tuned to reproduce data from higher resolution simulations (e.g., atomistic) and/or experimental measurements. Several important caveats associated with coarse-graining are essential to keep in mind:

- Some systems (such as proteins) may heavily depend on small local interactions to stabilize a given conformation, which makes the process of "throwing away detail" so much more challenging.

- While there exist different systematic coarse-graining procedures (e.g., Iterative Boltzmann Inversion, Inverse Monte Carlo, Force Matching; see [RJL$^+$09]) which optimize interaction potentials to best reproduce a reference system (e.g., pair correlation functions, average forces), there is no sure-fire way of producing a reliable and robust model. Top-down parametrizations offer an alternative approach to coarse-graining, including physics and knowledge-based models, where meso/macroscopic information of the system is used to construct a simplified model. For instance, the peptide model presented in the present thesis was *not* derived from a systematic parametrization mainly due to the large body of residue-residue pair interactions (20×20 residues: 210 interactions),[9] but rather constructed as a physics/knowledge based model. While interactions in physics-based models are constructed using physical arguments (e.g., "beads should include excluded volume"), knowledge-based potentials are derived from a statistical analysis of protein structures in the Protein Data Bank [WWWa].

[9] Pairwise residue potentials were recently derived from atomistic simulations for the 210 pairs of amino acids by Betancourt and Omovie [BO09].

- Gauging the applicability, strengths, and shortcomings of a coarse-grained model is key to avoid misusing and misinterpreting results predicted from it.

The construction of a coarse-grained model also requires agreeing on a set of units with which dimensional physical quantities can be measured. In the absence of electrostatics, all units can be constructed from the explicit definition of length $\mathcal{L}$, energy $\mathcal{E}$, and mass $\mathcal{M}$. It is necessary to map these fundamental units to "real" (e.g., SI[10]) units in order to relate the coarse-grained simulations to atomistic or experimental data. Note that when studying thermodynamic ("static") properties, masses drop out of any measurable quantity (see Technical Point 1.1 for a derivation)—the proper calibration of bead masses is, in this context, irrelevant. It is only in the dynamics that masses have a significant effect (e.g., the heavier the bead, the larger its inertia). This naturally raises the question: "Why not reproduce the dynamics as well?" Because matching *static* properties alone provides no constraint on the dynamics. The model displays *some* coarse-grained dynamics, as quantified by the unit of time $\tau = \mathcal{L}\sqrt{\mathcal{M}/\mathcal{E}}$, but this unit does not provide a correct measure for the long time dynamics of the real system; it only describes the "instantaneous" dynamics of the coarse-grained system. For instance, it implies velocities $\boldsymbol{v}_i$ that lead to kinetic energies which satisfy the equipartition theorem

$$\frac{1}{2}m_i\langle \boldsymbol{v}_i^2\rangle = \frac{3}{2}k_{\mathrm{B}}T. \tag{1.2}$$

The main reason why the dynamics are not automatically recovered is because the reduction of molecular friction—fewer beads give rise to a smoother energy landscape—accelerates the true dynamics. In fact, this is generally seen as a *good* thing, because it enables more efficient sampling. One thus refers to how much faster the coarse-grained model is by means of a *speed-up factor*. An attempt to calculate this quantity for the model introduced in the present thesis is presented in chapter 3.

Overall, the field of coarse-graining has greatly evolved and become increasingly sophisticated, such that it is now recognized as a complementary tool to experiments and atomistic simulations in various fields (e.g., the MARTINI force field in the context of transmembrane protein simulations [MRY+07, MKP+08]). Apart from computational speedup, one of the appealing features of coarse-graining is the amount of *insight* that can be gained:[11] one hopes that the main structural mechanisms involved in a complex system, such as a protein, need not be described by *all* of its degrees of freedom but rather a small subset. Coarse-graining consists of judiciously identifying the important degrees of freedom to better understand the problem at hand.

In this respect, the coarse-grained peptide model presented in the next chapter is an attempt to produce a generic (i.e., transferrable) model that includes enough biochemical

[10] The *Système International d'unités* [WWWb] is the modern form of the metric system. It relies on the following units: metre, kilogram, second, ampere, kelvin, candela, and mole. While overwhelmingly used in the commerce and scientific communities throughout the world, certain countries still resist its invasion.

[11] "The purpose of computing is insight, not numbers." R. W. Hamming [Ham87].

Technical Point 1.1 Masses do not affect the thermodynamics

Consider a system of N particles expressed as a function of their coordinates $\boldsymbol{r}_i$ and momenta $\boldsymbol{p}_i$. Assuming the potential energy only depends on the coordinates, the Hamiltonian can be written

$$\mathcal{H} = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + V(\boldsymbol{r}_1, \dots, \boldsymbol{r}_N), \tag{1.3}$$

where $V(\boldsymbol{r}_1, \dots, \boldsymbol{r}_N)$ describes the interaction between all particles. The corresponding classical partition function is

$$Z = \int \frac{\mathrm{d}\boldsymbol{p}^N \mathrm{d}\boldsymbol{r}^N}{N! h^{3N}} \mathrm{e}^{-\beta\left[\sum_{i=1}^{N} \frac{p_i^2}{2m_i} + V(\boldsymbol{r}_1,\dots,\boldsymbol{r}_N)\right]}, \tag{1.4}$$

where h is Planck's constant and $\beta = 1/k_\mathrm{B}T$. Because the kinetic term only involves momenta and the potential only depends on coordinates, the integral can be split into two terms

$$Z = \underbrace{\int \frac{\mathrm{d}\boldsymbol{p}^N}{N! h^{3N/2}} \mathrm{e}^{-\beta \sum_{i=1}^{N} \frac{p_i^2}{2m_i}}}_{\text{Ideal gas contribution}} \underbrace{\int \frac{\mathrm{d}\boldsymbol{r}^N}{h^{3N/2}} \mathrm{e}^{-\beta V(\boldsymbol{r}_1,\dots,\boldsymbol{r}_N)}}_{\text{Interactions}}. \tag{1.5}$$

The second term in Equation 1.5 does not show any dependence on particle masses (this is only true if V does not depend on the set of momenta $\boldsymbol{p}_i$). The first term is the ideal gas contribution. Its integral can be solved analytically

$$\int \frac{\mathrm{d}\boldsymbol{p}^N}{N! h^{3N/2}} \mathrm{e}^{-\beta \sum_{i=1}^{N} \frac{p_i^2}{2m_i}} = \frac{1}{N!} \left(\frac{2\pi k_\mathrm{B} T}{h^2}\right)^{3N/2} \prod_{i=1}^{N} m_i^{3/2}. \tag{1.6}$$

It can readily be seen that the mass dependent term contributes a constant prefactor, thus merely shifting the free energy.

This shows that, as far as *static* properties are concerned, the proper choice of particle masses is irrelevant. Previous studies have used this feature to increase the integration time step in molecular simulations (e.g., [WW10]).

Incidentally, this factorization of the partition function only works *classically*. The situation is very different for the quantum partition function $\mathrm{Tr}\left(e^{-\beta\hat{\mathcal{H}}}\right)$. Since positions and momenta do not commute, a factorization of the exponential à la $\exp\left[f(\hat{P}) + g(\hat{Q})\right] = \exp\left[f(\hat{P})\right] \cdot \exp\left[g(\hat{Q})\right]$ is not possible. This indeed gives rise to thermostatically observable effects, e.g., the strength of hydrogen bonds which involve deuteriums is slightly bigger [Kat65].

details to choose secondary/tertiary structure on its own while cutting down significantly on the overall resolution ([BD09]). It is later applied to several biophysical problems: (*i*) the folding kinetics of α-helix and β-hairpin peptides (chapter 3), (*ii*) the thermodynamics of protein folding cooperativity from a microcanonical perspective (chapter 4; [BBD10]), (*iii*) the aggregation of β-amyloid peptides (chapter 5), and (*iv*) the cross-parametrization of the force-field with a CG lipid model (chapter 6).

2 Coarse-Grained Peptide Model

An intermediate-resolution, implicit-solvent, coarse-grained peptide model is introduced. The high level of resolution devoted to the backbone allows for unconstrained secondary structure formation (unlike Gō models). The model is shown capable of folding simple peptides and reproduce a realistic oligopeptide aggregation scenario using a single force field.

The field of coarse-grained protein modeling is very diverse and has a rich history owing to a wide variety of problems to tackle, as well as length- and time-scales to look at (e.g., [Gō83, SO94, BOSW95, SHG00, SH01, CGO02, FIW02, DBB+03, HGB03, FTLSW04, PDU+04, Toz05, AFS06, ANV07, DM07, AYS08, Cle08, HWW08, HJBI08, MKP+08, TZV08, YFHG08, Bet09, ACCDP10, SEB10]). Various levels of resolution have been designed to study many different problems. On the coarser-side of particle-based simulations, conformational effects of hydrophobic interactions were studied using lattice simulations [LD89]. This is a very powerful tool that is still widely used when looking at large-scale cooperativity effects. Soon, off-lattice simulations were developed using one bead per amino acid with implicit solvent, famous examples are Gō models [Gō83]. This level of resolution allows for much more conformational freedom, which is key to structural studies. One underlying constraint in Gō models is that structure is biased towards the native configuration of the protein because the remaining degrees of freedom don't suffice to accurately represent the system's phase space, including secondary structure motifs. Intermediate resolution models (more than one bead per amino acid) have been designed to investigate structural properties of proteins while emphasizing certain aspects. For instance, the recently introduced MARTINI force field [MKP+08] opts for a high resolution on the protein's side chains, while the backbone is represented by only one bead per amino acid. The force field was parametrized using partitioning coefficients between water and a (similarly coarse-grained) lipid membrane. By doing so, protein-lipid systems, such as transmembrane proteins, can be accurately investigated (e.g., peptide aggregation and pore formation in a lipid bilayer [TSV+08]). Other models with a comparable overall resolution shift the emphasis (in terms of modeling detail) on the backbone instead of the side chain in order to look at structure and conformational properties without biasing the force field to the native configuration. Several force fields (see e.g., [TLSW99, ISW00, FIW02]) have been reported to fold *de novo* helical proteins. These models incorporate only a subset of amino acids, emphasizing their chemical effects (e.g., hydrophobic, polar, glycine residue).

Intermediate level resolution models have shown promising results in capturing local conformations, and reproducing basic aspects of secondary structure formation while gaining much computational efficiency compared to atomistic models. This is partly due to the removal of solvent, which allows for significant speedup, as water typically represents the bulk of a simulation in such systems. As a result, it is necessary to treat important solvent effects implicitly, as they are determining factors in a protein's conformation. This, of course, is also one of the main limiting factors of such approaches.

While α-helices are comparatively easy to obtain in such models, β-sheets and structures are more difficult to stabilize. There are several reasons for this. First, the enthalpic gain per amino acid is weaker compared to α-helices [YH95]. Second, Yang and Honig [YH95] have shown that side-chain–side-chain interactions have a decisive role in sheet formation. And third, the stabilization energy contains a contribution from interactions between dipoles of successive peptide bonds that is usually neglected in simple models, yet it favors the β- over the α-structure [CSM06]. Apart from these local effects, the stability of extended conformations also depends greatly on cooperativity. Other than stabilizing folds, this can also lead to peptide aggregation. Besides being an interesting physical problem, peptide aggregation is associated with countless biological processes. It also plays a crucial role in many diseases, ranging from sickle cell anemia [LBZ+00] to Alzheimer's [LL06].

In this chapter, we present a CG model of a four-bead-per-amino-acid model in implicit solvent. It differs from previously mentioned intermediate level force fields [TLSW99, ISW00, FIW02] in several ways. First, by improving on amino acid specificity[1] it provides a more detailed free energy landscape. Second, protein folding is quantitatively probed by comparing our molecular dynamics (MD) simulations with experimental data, instead of the lowest energy structure that is sampled. Third, after tuning our force field with respect to one protein (in terms of tertiary structure reproduction), it is tested on other proteins to understand how reliable this procedure is. Fourth, an important design criterion for our model is its ability to produce a realistic balance between α-helical and β-extended conformations, thereby avoiding a bias toward any particular secondary structure.[2] Finally, we monitor the aggregation of small peptides (into β-sheets) to test whether a realistic aggregation scenario in the long-time and large length-scale regime can be achieved.

In order to parametrize and test our force field as finely as possible, we systematically compare the performance of our CG model with experimental data. We hasten to add, though, that refining CG models is no attempt to compete with atomistic force fields. Such an endeavor strikes us as neither likely to succeed, nor to be in line with the reasons one pursues coarse-graining in the first place, namely to gain a physical understanding of fundamental mechanisms and universals of complex molecular structures. However, in systems as delicate as marginally stable proteins a subtle local interaction can have a substantial global impact, and uncovering causations of this type is well within the scope of CG studies.

This chapter is divided into several parts: the mapping scheme will explain how atomistic

[1] A full spectrum of amino acid hydrophobicities is used rather than, say, a smaller subset which represents types of amino acids (e.g., hydrophobic, polar, charged; [DBB+03]). See below for details.

[2] This is in contrast to other models that can only fold helical proteins (e.g., [TLSW99, ISW00, FIW02]) or tune secondary structure propensity via the temperature (e.g., [DBB+03]).

details were coarse-grained out, the different interactions as well as parameter tuning and simulation methods will be described, and finally several applications will show to what extent the model can reproduce structural properties.

2.1 Mapping scheme

2.1.1 Overall geometry

An amino acid is modeled by three or four beads (Figure 2.1). These beads represent the amide group N, central carbon C_α, carbonyl group C', and (for non-glycine residues) a side chain C_β. The first three beads belong to the backbone of the protein chain, whereas the last one represents the side chain and is responsible for amino acid specificity. This high level of backbone resolution is necessary to account for the characteristic conformational properties underlying secondary protein structure. As far as reducing the number of degrees of freedom is concerned, this high resolution is regrettable, as the backbone is represented almost atomistically. Indeed, models that do not require the CG protein to represent local structure generally do away with most (if not all) backbone beads (e.g., the MARTINI model for proteins [MKP+08]). However, here we explicitly aim at a model that is at least in principle capable of finding secondary structure by itself. This is for instance necessary in applications where this structure is known to change (e.g., misfolding, spontaneous aggregation) or not known at all.

2.1.2 Parameter values

Geometric parameters were taken from existing peptide models [TLSW99, ISW00, DBB+03] and are reported in Table 2.1.[3] Even though the spatial arrangement of the beads was fixed beforehand, the van-der-Waals radii were left as free parameters. Following the abovementioned references, C_β was set at the location of the first carbon of the side chain (hence our nomenclature), directly connected to the backbone. Its location will generally not coincide with the center of mass of the atomistic side chain (which for larger and flexible side chains has no fixed position with respect to the backbone), but the concomitant substantial reduction of tuning parameters is necessary for our parametrization scheme, as we will see below.

All side chain beads have been given the same van-der-Waals radius, except for glycine, which is modeled without a side chain. This accounts for the biggest difference in the Ramachandran plot of amino acids, namely the large flexibility of an achiral glycine residue, as opposed to the substantial chiral sterical clashes between all the others [FP02]. On the other hand, it does not represent the size differences between non-glycine residues and will thus likely cause problems if packing issues are important, e.g., inside globular proteins.

[3] Ref. [BD09] incorrectly expressed k_{angle} in units of $\mathcal{E}/\mathrm{deg}^2$ instead of $\mathcal{E}/\mathrm{rad}^2$.

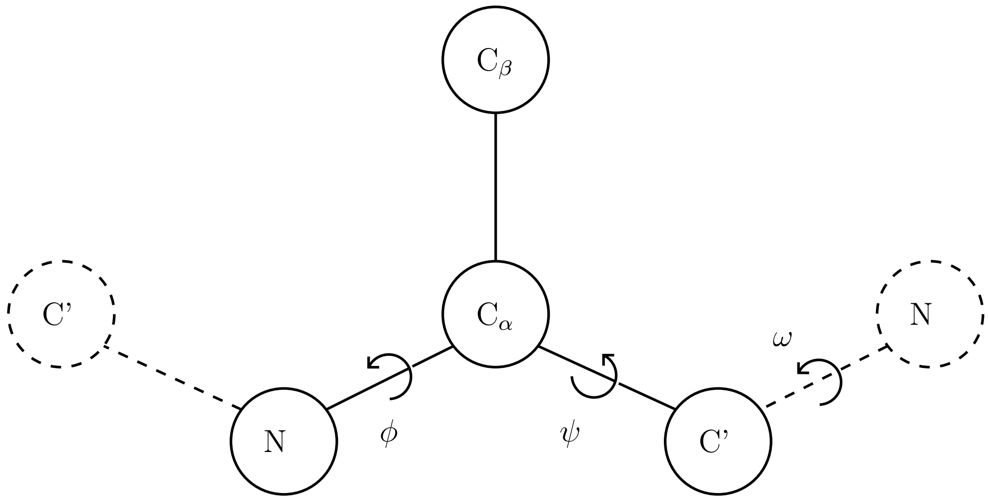


Figure 2.1: Schematic figure of the local geometry of the protein chain. The solid beads comprise one amino acid. Neighboring amino acid beads are represented in dashed lines. Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

Both the location and the size of the side chain is thus modeled in an approximate and highly simplified way. Why not be more sophisticated? Since these degrees of freedom are accounted for, one might as well give them the best possible parameter values. Ideally this is indeed what one would like to do, but the catch is that the necessary tuning is very difficult. Having 20 different amino acids gives—in the worst case—$20^3 = 8000$ local Ramachandran plots for the (ϕ, ψ) angles between three consecutive amino acids. These would first need to be determined atomistically and then—via some suitable matching procedure—translated into CG side chain properties. Clearly, many obvious simplifications would be possible and the task is not nearly as daunting. The number of free parameters would nevertheless be substantially increased and their tuning would require both automated techniques and enormous computing resources. In contrast, in the present model we aim to keep the number of free parameters as low as possible, such that judicious tuning by hand is still a viable option. We will see below that it is also successful. While optimization of side chain parameters will remain a long term goal, this is certainly not the point where to start.

Finally, amino acids that are in the middle of a protein chain form peptide bonds with their neighbors. This is not so at the ends of the chain, and the structure is slightly different. Nonetheless, we model the end beads identically.

	Bond lengths			
	NC_α	$C_\alpha C'$	$C'N$	$C_\alpha C_\beta$
r_0 [Å]	1.455	1.510	1.325	1.530
k_{bond} [$\mathcal{E}$/Å^2]	300	300	300	300

	Bond angles				
	$NC_\alpha C_\beta$	$C_\beta C_\alpha C'$	$NC_\alpha C$	$C_\alpha C'N$	$C'NC_\alpha$
θ_0 [deg]	108	113	111	116	122
k_{angle} [$\mathcal{E}/\text{rad}^2$]	300	300	300	300	300

	Dihedrals				
	ϕ^*	ψ^*	ω	ω_{Pro}	improper
k [$\mathcal{E}$]	-0.3	-0.3	67.0	3.0	17.0
n	1	1	1	2	1
φ_0 [deg]	0	0	180	0	$\mp$120

Table 2.1: Bonded interaction parameters used in the model. The dihedrals denoted with an asterisk were determined during parameter tuning (see section 2.4). All parameters are expressed in terms of the intrinsic units of the system (see subsection 2.1.3). k represents the interaction strength of Fourier mode n (see main text), with equilibrium value φ_0. ω_{Pro} refers to the ω dihedral around the peptide bond for a proline residue. The sign of the improper dihedral angle φ_0 is linked to the chirality of the isomer; the L-form requires a negative sign. For each angular potential, only a single mode n was used.

2.1.3 Units

All lengths are measured in units of $\mathcal{L}$, which we choose to be 1 Ångström. For the energies we found it convenient to relate them to the thermal energy, since it is this balance which determines the overall protein conformation. We thus define the energy unit $\mathcal{E} = k_{\text{B}}T_{\text{r}} = 1.38 \times 10^{-23}\text{JK}^{-1} \times 300\,\text{K} \approx 4.1 \times 10^{-21}\text{J} \approx 0.6\,\text{kcal}\,\text{mol}^{-1}$ as the thermal energy at room temperature.

Masses will be measured in the unit $\mathcal{M}$, which is the mass of a single CG bead. We will assume all beads to have the same mass.[4] An amino acid weighs on average 110 Da. By distributing mass equally among the four beads N, C_α, C_β, and C', this gives an average mass of $\mathcal{M} \simeq 4.6 \times 10^{-26}$ kg.

The natural time-unit in our simulation is $\tau = \mathcal{L}\sqrt{\mathcal{M}/\mathcal{E}}$. Using the length, energy,

[4] The precise parametrization of masses only matters for dynamical issues—it does not affect equilibrium properties (see Technical point 1.1 on page 11).

and mass-mappings from above, we find $\tau \sim 0.1\,\mathrm{ps}$. As explained in the previous chapter (subsection 1.2.3 on page 8), this unit of time correctly describes the instantaneous dynamics of a fictitious CG bead-spring system. However, it does *not* measure the time which the real protein system requires to undergo the same conformational change as observed in the simulation. An attempt to quantify *how much faster* the coarse-grained model operates is presented in chapter 3. It should be recalled that as far as equilibrium questions are concerned the precise time mapping is, of course, irrelevant.

2.2 Interactions

2.2.1 Bonded interactions

The local structure is constrained by bonded interactions. Bonds and angle potentials are chosen to be harmonic:

$$V_{\mathrm{bond}}(r) = \frac{1}{2}k_{\mathrm{bond}}(r-r_0)^2 \, , \qquad (2.1a)$$

$$V_{\mathrm{angle}}(\theta) = \frac{1}{2}k_{\mathrm{angle}}(\theta-\theta_0)^2 \, . \qquad (2.1b)$$

The spring constants k_{bond} and k_{angle} are set high enough to keep these coordinates close to their minimum (within $\sim 5\%$). Table 2.1 reports these parameters.

Up to thermal fluctuations bonds and angles are thus fixed. Flexibility of the overall structure enters through the dihedrals, the possibility to rotate around a chemical bond. In the case of proteins, two out of three backbone dihedrals are very flexible and are responsible for the diverse set of local conformations. These dihedrals are the ϕ and ψ coordinates, defined by the sets of beads C'NC_αC' and NC_αC'N, respectively (see Figure 2.1). They describe the angle between two planes (e.g., ϕ is the angle between the planes C'NC_α and NC_αC') and obey the following convention: taking any four beads #1,2,3,4 and looking along the vector from bead #2 to bead #3, the angle "0" will correspond to the conformation in which beads #1 and #4 point into the same direction (i.e., when they visually overlap). The rotation of plane #1,2,3 with respect to plane #2,3,4 away from this state defines the angle; the counterclockwise sense counts positive (Figure 2.2). Because the potential of rotation around the bond between sp^3- and sp^2-hybridized atoms has a rather low barrier compared to thermal energy at room temperature, we let the beads rotate freely. However, we will later include a contribution to the coordinates ϕ and ψ accounting for an effective non-bonded dipolar interaction (see below).

The third dihedral along the backbone chain, ω, defined by C_αC'NC_α, is located at the peptide bond (see Figure 2.1). This bond corresponds to the rotation around two sp^2-hybridized atoms, which involves a symmetric potential with two minima, separated by a rather high barrier. The two conformations, *cis* and *trans*, have an angle of 0° and 180°, respectively. The *cis* conformation tends to be sterically unfavored for most amino acids, except for proline, where there is no specific preference due to its special side chain linkage.

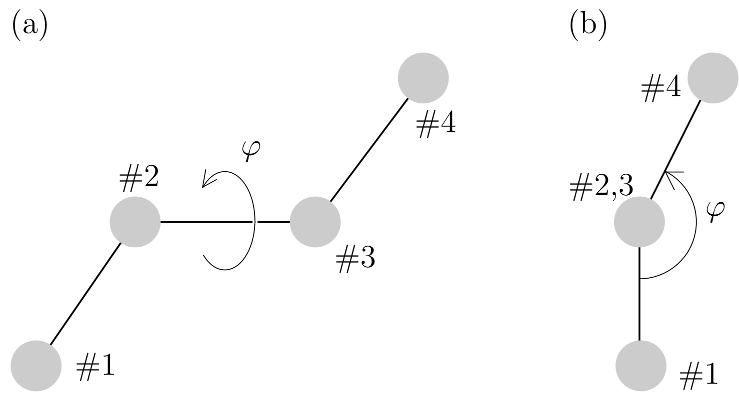


Figure 2.2: Schematic figure of the convention used when measuring a dihedral φ from the four beads #1,2,3,4. Side view (a) and view along the vector from bead #2 to bead #3 (b).

Generally, dihedrals can be written as a Fourier series in the rotation angle. Here we will restrict to a single mode and describe the interaction as

$$V_{\mathrm{dih}}(\varphi) = k_n \Big[1 - \cos(n\varphi - \varphi_{n,0})\Big] \tag{2.2}$$

with coefficient k_n and phase $\varphi_{n,0}$. In this model we represent the peptide bond using only one minimum ($n = 1$) centered around the *trans* conformation. In this case $\varphi_0 \equiv \varphi_{1,0}$ is the equilibrium orientation of the dihedral and $k \equiv k_1$ is the stiffness describing deviations from the equilibrium angle. For a peptide bond located right before (i.e., on the N-terminal side) a proline residue, we model the isomerization by a dihedral potential with two minima ($n = 2, k \equiv k_2$), one at the *cis* conformation, and the other one at *trans*. This allows for a more natural representation of the different conformations proline can take. Depending on the problem one is interested in (and the time scales which matter), the energy barrier can be tuned to either freeze the isomerization, or set to a low value to allow efficient sampling. We chose the latter in this work. This choice will of course affect the kinetics of the system.

The central carbon C_α not only links the backbone to the side chain, its sp^3 hybridization imposes a tilted orientation of the $C_\alpha C_\beta$ vector compared to the $NC_\alpha C'$ plane. Its four bonds are located at the vertices of a tetrahedron, linking the backbone atoms N and C', as well as the C_β side chain and an extra hydrogen (not modeled by us). This has an important consequence, because a carbon atom with four different substituents is *chiral* and hence optically active. All amino acids except glycine exist as two different stereoisomers. The L-form is realized in native amino acids: looking at the central carbon C_α, with the hydrogen atom pointing away, the isomer has L-form if the three other substituents C', C_β, and N are arranged in a counterclockwise fashion ("CORN-rule"). This amino acid chirality is a central feature in proteins and their secondary structure, and we account for it by including an "improper dihedral" between the beads $NC_\alpha C'C_\beta$. This keeps a tilt between the backbone plane, $NC_\alpha C'$, and the plane intersecting the side chain with two backbone beads, $C_\alpha C'C_\beta$, such that all angles are correct and the CORN-rule is satisfied. The interaction has the same form as other dihedrals, given by Equation 2.2. The two stereo-isomers only differ in the sign of the dihedral equilibrium angle φ_0 and can thus both be modeled.

2.2.2 Non-bonded interactions

Probably the biggest challenge in any coarse graining scheme is determining the non-bonded interactions. Unlike bonded interactions, their form is not intrinsically obvious and the system behavior depends very sensitively on them. In the following section every interaction introduced will require at least one free parameter that has to be determined by tuning. The key technical difficulty of this enterprise is that all parameters are typically highly correlated. Optimization is thus an intrinsically multidimensional problem and we therefore intend to limit the number of free parameters as much as possible. While one might envision "hands-off" tuning schemes in which optimization occurs in an automated fashion [MBF⁺00], for the present problem we found this difficult to implement for two reasons: first, parameter variations often have a rather inconspicuous impact on target observables and the determination of the right gradient in parameter space thus can require very substantial computer time. And second, some optimization aims are hard to quantify in numbers and rather require judgment and choice—e.g., the question how one balances the quality of a local Ramachandran plot against global folding characteristics.

Backbone

Steric interactions are closely linked to secondary and tertiary structures for two reasons: first, local interactions along the protein chain will shape the Ramachandran plot; second, contact between distant parts of the amino acid chain will determine protein packing on larger scale. In order to model a local excluded volume, we use a purely repulsive Weeks-Chandler-Andersen (WCA) potential

$$V_{\mathrm{bb}}(r) = \begin{cases} 4\epsilon_{\mathrm{bb}}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6} + \frac{1}{4}\right], & r \le r_{\mathrm{c}} \\ 0, & r > r_{\mathrm{c}} \end{cases} \tag{2.3}$$

where $r_{\mathrm{c}} = 2^{1/6}\sigma_{ij}$ and σ_{ij} is the arithmetic mean between the two bead sizes involved, following the Lorentz-Berthelot mixing rule. Just like the bead sizes, the energy ϵ_{bb} is a free parameter, though we use only one parameter for all backbone-backbone and backbone-side chain interactions, since for the WCA potential the energy scale is largely immaterial. Following the practice in atomistic simulations, we do not calculate excluded volume interaction between beads that are less than three bonds apart, since their distance is largely fixed through the bonded interactions.

Side chain interactions

Amino acids differ in their water solubility. This can be quantified experimentally by measuring the partitioning of residues between water and a hydrophobic environment (e.g., [FP83]). The ratio of densities (or strictly speaking: *activities*) of a residue in the two environments can be translated into a free energy of transfer from one medium to another [MS97]. Hydrophobicity is one prominent cause for certain amino acids to attract. However, there are other reasons why residues interact (e.g., charges or hydrogen bonds between side chains) and this combination can be probed by statistical analyses of residue-residue contacts in proteins [MJ96, SJKG97, MJ99, BT99, WL00]. One then arrives at a phenomenological interactions energy between any two residues A and B that depend on the number of close AB-contacts that are found in a pool of protein structures.[5] This mean-field approach (it averages over all neighboring contacts) not only contains information on the relative hydrophobicity of amino acids, but also partially incorporates effects coming from additional interactions (e.g., salt bridges or side-chain hydrogen bonds). In the absence of explicit solvent we represent this phenomenological cohesion by introducing an effective attraction (of standard Lennard-Jones 12-6 type) between C_β side chain beads, whose strength is mapped to such a statistical analysis of residue-residue contacts. Specifically, we used Miyazawa and Jernigan's (MJ) statistical analyses [MJ96] to extract a *relative* attraction strength between residues. To translate this into an *absolute* scale, *one* additional free parameter ϵ_{hp} is needed.

Miyazawa and Jernigan analyzed residue-residue contacts in crystallized proteins. By modeling interactions via square-well potentials, they obtained interaction strengths $\epsilon_{ij}^{\text{MJ}}$ for every i-j pair of residues. We reduced the resulting 20×20 interaction matrix further by deconvolving it (see below) into 20 interaction parameters ϵ_i (one for each amino acid), which approximately recreate all interactions as the geometric mean of the two amino acids involved, $\epsilon_{ij}^{\text{MJ}} \approx \epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$, following the Lorentz-Berthelot mixing rule. Each term is then normalized

$$\epsilon_i' = \frac{\epsilon_i - \min_k \epsilon_k}{\max_k \epsilon_k - \min_k \epsilon_k} \tag{2.4}$$

such that the most hydrophilic residue has a weight of 0 and the most hydrophobic a weight of 1, and the normalized interaction contact is denoted $\epsilon_{ij}' = \sqrt{\epsilon_i' \epsilon_j'}$. Finally, we multiply this term by the overall interaction scale ϵ_{hp}. One limitation in varying the interaction strength of a Lennard-Jones potential is that a low ϵ_{ij}' will tend to flatten out the repulsive part of the interaction. This will, as a result, fade the excluded volume effect for certain side chain beads, which is likely to exacerbate packing problems in dense regions. To overcome this issue and keep the same excluded volume for all side chain beads, we model the overall interaction by using a Lennard-Jones potential for the attractive part linked to a purely repulsive WCA potential for smaller distances. We join the two potentials at the minimum value of the interaction in such a way that both the potential and its first

[5] This is commonly referred to as a "knowledge-based approach."

derivative are continuous. Overall, the interaction will have the following form

$$V_{\text{hp}}(r) = \begin{cases} 4\epsilon_{\text{hp}} \left[\left(\frac{\sigma_{C_\beta}}{r} \right)^{12} - \left(\frac{\sigma_{C_\beta}}{r} \right)^{6} + \frac{1}{4} \right] - \epsilon_{\text{hp}}\epsilon'_{ij}, & r \leq r_{\text{c}} \\ 4\epsilon_{\text{hp}}\epsilon'_{ij} \left[\left(\frac{\sigma_{C_\beta}}{r} \right)^{12} - \left(\frac{\sigma_{C_\beta}}{r} \right)^{6} \right], & r_{\text{c}} \leq r \leq r_{\text{hp,cut}} \\ 0, & r > r_{\text{hp,cut}} \end{cases} . \tag{2.5}$$

Relative (unnormalized) coefficients ϵ_i were calculated by minimizing the expression

$$\chi^2 = \frac{1}{N} \sum_{i,j \geq i=1}^{N} \chi_{ij}^2 \ , \tag{2.6}$$

where $\chi_{ij} = \epsilon_{ij}^{\text{MJ}} - \sqrt{\epsilon_i \epsilon_j}$, N is the number of matrix coefficients (210 independent elements in a 20×20 symmetric matrix), and the sum goes over all such elements. The normalized coefficients ϵ'_i that were obtained by simulated annealing followed by proper scaling (Equation 2.4) are reported in Table 2.2.

Let us quantify the quality of this deconvolution and the suitability of the amino-acid specific hydrophobic strength ϵ'_i. Recall that the correlation coefficient c between two data sets $\{X_i\}$ and $\{Y_i\}$ is defined as

$$c = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma_X} \right) \left(\frac{Y_i - \overline{Y}}{\sigma_Y} \right) \ , \tag{2.7}$$

where n is the number of data points in each set, $\overline{X}$ and $\overline{Y}$ are their averages, and σ_X and σ_Y are their standard deviations, respectively. Our inferred 210 ϵ_{ij} values and their original $\epsilon_{ij}^{\text{MJ}}$ counterparts have a correlation coefficient of 98%, which decreases by only 3 points when comparing the MJ matrix to the normalized interaction contacts ϵ'_{ij}. Moreover, the 20 individual values ϵ_i as well as the ϵ'_i have a 87% correlation with the experimental hydrophobicity scale measured by Fauchere and Pliska [FP83]. Since the MJ matrix accounts for more than hydrophobicity, this further drop in the correlation coefficient is expected. However, its still relatively large value suggests that the hydrophobic effect is the dominant contribution to the MJ energies. This is the reason why we refer to the interactions (Equation 2.5) summarily as "hydrophobicity." The fitting procedure gave a χ^2 value of 0.064, which translates into an average relative error $\overline{\Delta\epsilon} = 0.25$ between coefficients along the diagonal of the MJ matrix, where this deviation is defined by $\Delta\epsilon_{ii} = \chi_{ii}/\epsilon_{ii}^{\text{MJ}}$. Even though most coefficients did not deviate more than 15% from the MJ matrix, Lysine, the most hydrophilic residue, is off by a factor of 4. Various sets of parameters with a comparable χ^2 value showed equivalent correlation properties, even though deviations were located on different amino acids. This rules out the hypothesis of a systematic failure of our $\mathcal{N}^2 \rightarrow \mathcal{N}$ deconvolution.

A comparison between the normalized hydrophobicity scale derived from the MJ matrix (Table 2.2) and experimental hydrophobicity scales (e.g., [FP83] and Table 1.1 and Table 1.2) shows discrepancies for several amino acids:

	Lys	Glu	Asp	Asn	Ser	Arg	Gln	Pro	Thr	Gly*
	K	E	D	N	S	R	Q	P	T	G*
ϵ'_i	0.00	0.05	0.06	0.10	0.11	0.13	0.13	0.14	0.16	0.17
$\Delta\epsilon_{ii}$	4.00	0.50	0.16	0.01	0.05	0.20	0.20	0.10	-0.01	-0.05
χ_{ij}	-0.48	-0.45	-0.19	-0.02	-0.08	-0.31	-0.31	-0.17	0.02	0.11

	His	Ala	Tyr	Cys	Trp	Val	Met	Ile	Phe	Leu
	H	A	Y	C	W	V	M	I	F	L
ϵ'_i	0.25	0.26	0.49	0.54	0.64	0.65	0.67	0.84	0.97	1.00
$\Delta\epsilon_{ii}$	-0.11	0.00	0.03	-0.14	0.05	-0.02	0.01	0.02	0.04	0.05
χ_{ij}	0.35	0.01	-0.14	0.76	-0.24	0.12	-0.05	-0.12	-0.32	-0.38

Table 2.2: Normalized scale of amino acid hydrophobicities ϵ'_i using the Lorentz-Berthelot mixing rule for the cross terms, as well as relative and absolute error, $\Delta\epsilon_i$ and χ_{ij}, from the diagonal elements of the MJ matrix (see text for definition). Note that the side chain of glycine (marked with an asterisk in the table) is not modeled.

Pro Proline is categorized as hydrophobic from experimental hydrophobicity scales, while it is given a low normalized relative hydrophobicity $\epsilon'_i = 0.14$ from the MJ matrix deconvolution. The reason is straightforward: proline tends to play a role in turns and loops due to its unique chemical structure.[6] This creates an anomaly between its water/oil partitioning free energy and its average distance to other amino acids.

Cys Cysteine exhibits the opposite behavior: while it consists of a polar uncharged group, the relative hydrophobicity is rather high ($\epsilon'_i = 0.54$—close to tryptophan, a weakly hydrophobic residue). This is explained by the tendency of cysteine to form disulfide bridges which might become buried inside hydrophobic cores.

These two amino acids are likely to be the main sources of discrepancy between the two descriptions of hydrophobicity presented here.

It is possible to account for solvent effects in even further detail, for instance by including the layering of water molecules around the solute into the effective potentials [CGO02]. In our attempt to develop a simple force field and only keep a few important aspects of protein interactions, and in view of the approximation already made, we decided against such local details.

Hydrogen bonds

Since our model does not contain any electrostatics, it is necessary to model hydrogen bonds implicitly as well. The interaction depends on the relative distance and orientation of an amide and a carbonyl group. A real amide group is composed of a nitrogen with a hydrogen, whereas the carbonyl group has a carbon double-bonded to an oxygen. The hydrogen bond is favored when the N, H, and O atoms are aligned. Several interaction potentials for hydrogen bonding have been proposed in the literature [TLSW99, ISW00, SH01, GCLK02, MG07, YFHG08]. For its simplicity and corresponding CG mapping, we follow Irbäck *et al.* [ISW00] by using a radial 12-10 Lennard-Jones potential combined with an angular term

$$V_{\mathrm{hb}}(r,\theta_{\mathrm{N}},\theta_{\mathrm{C}}) = \epsilon_{\mathrm{hb}}\left[5\left(\frac{\sigma_{\mathrm{hb}}}{r}\right)^{12} - 6\left(\frac{\sigma_{\mathrm{hb}}}{r}\right)^{10}\right] \times \begin{cases} \cos^2\theta_{\mathrm{N}}\cos^2\theta_{\mathrm{C}}, & |\theta_{\mathrm{N}}|,|\theta_{\mathrm{C}}| < 90^\circ \\ 0, & \text{otherwise} \end{cases} \tag{2.8}$$

where r is the distance between the two beads N and C', σ_{hb} is the equilibrium distance (Table 2.3), θ_{N} is the angle formed by the atoms HNC' and θ_{C} corresponds to the angle NC'O (Figure 2.3). The main motivation for using a power of 10 instead of 6 in the Lennard-Jones potential is a narrower confinement of the hydrogen bond length. Since our model does not represent hydrogens and oxygens, these particle positions were calculated via the local geometry of the backbone. Any NC' pair can form a hydrogen bond, except if N belongs to proline, since its side chain connects to the preceding amide on the backbone.

[6] Recall that the side chain of proline bonds to the amide group (see Table 1.1 on page 2).

Backbone excluded volume			
σ_{N} [Å]	$\sigma_{\mathrm{C}_\alpha}$ [Å]	$\sigma_{\mathrm{C}'}$ [Å]	ϵ_{bb} [$\mathcal{E}$]
2.9	3.7	3.5	0.02

Hydrophobicity		
$\sigma_{\mathrm{C}_\beta}$ [Å]	ϵ_{hp} [$\mathcal{E}$]	$r_{\mathrm{hp,cut}}$ [Å]
5.0	4.5	10*

Hydrogen bonding		
σ_{hb} [Å]	ϵ_{hb} [$\mathcal{E}$]	$r_{\mathrm{hb,cut}}$ [Å]
4.11*	6	8*

Table 2.3: Non-bonded interactions. The length σ represents the diameter of a bead. Most parameters were determined after parameter tuning, except the ones denoted by an asterisk. See section 2.4.

The hydrogen bond leads to one more free parameter, its interaction strength ϵ_{hb}. Technical point 2.1 derives the force associated with the potential described in Equation 2.8.

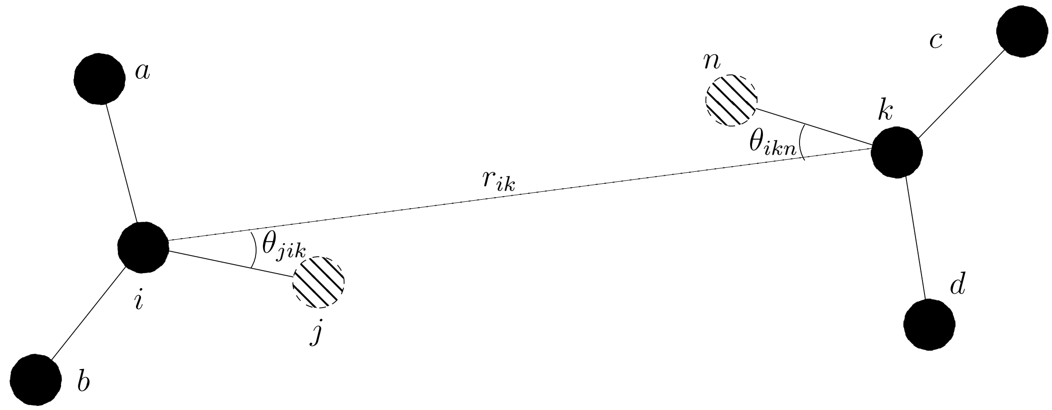
The main drawback of such a multibody potential is the necessity to implement its functional form in a MD simulation package. While this interaction was implemented in the ESPResSo package [LAMH06, WWWc], not all simulation engines easily allow for customized interaction potentials. Alternatively, others have successfully modeled effective hydrogen bonds using sets of *pair* interactions, in which directionality is recreated by an attractive interaction along the axis of the hydrogen bond (e.g., N-C' atoms) with a set of repulsive interactions in its neighborhood [DBB$^+$03, HWJW10]. An early attempt to convert Equation 2.8 into pair potentials suggests that small discrepancies between the energetics of the multibody potential and a pair-only analog drastically affect the conformations sampled. Further investigations will be required to successfully replace Equation 2.8.

Electrostatics

There is no explicit treatment of side chain charges in the force field. Specifically, we do not model the interaction between charged residues. However, this piece of information is partially included in the MJ matrix, as the method is based on statistical analysis of residue-residue distances.[7] The electrostatic interaction involved between two charged residues will be implicitly sampled, and its effect reflected in the interaction coefficient. Nevertheless,

[7] Unfortunately the $^+/_-$ charge asymmetry gets lost after the deconvolution of the MJ interaction matrix.

Technical Point 2.1 Force derivation of the hydrogen bond interaction



We consider the following multi-body potential

$$V(r_{ik}, \theta_{jik}, \theta_{ikn}) = \epsilon \left[5 \left(\frac{\sigma}{r_{ik}} \right)^{12} - 6 \left(\frac{\sigma}{r_{ik}} \right)^{10} \right] \cos^2 \theta_{jik} \cos^2 \theta_{ikn} \tag{2.9}$$

where the indices represent particles in a geometry described in the figure (bonded partners represented by full lines). The component α of the force exerted on particle l is given by $\boldsymbol{f}_l^\alpha = -\partial V / \partial \boldsymbol{r}_l^\alpha$. The angular dependence of the potential is calculated using the relative positions of *virtual* particles j and n (dashed beads) such that (i) their position is constructed by simple vector additions $\boldsymbol{r}_{ij} = \boldsymbol{r}_{ai} + \boldsymbol{r}_{bi}$ and $\boldsymbol{r}_{kn} = \boldsymbol{r}_{ck} + \boldsymbol{r}_{dk}$, where $\boldsymbol{r}_{ij} = \boldsymbol{r}_j - \boldsymbol{r}_i$, and ($ii$) the forces acting on them is redistributed among all others.
Denoting $S(r_{ik})$ as the radial part of the potential and its derivative

$$S'(r_{ik}) = 60 \left(\frac{\sigma^{10}}{r_{ik}^{11}} - \frac{\sigma^{12}}{r_{ik}^{13}} \right), \tag{2.10}$$

the force is given by

$$\begin{aligned}
\boldsymbol{f}_l^\alpha = {} & \cos^2 \theta_{jik} \cos^2 \theta_{ikn} \frac{r_{ik}^\alpha}{r_{ik}} S'(r_{ik}) (\delta_{li} - \delta_{lk}) \\
& - 2 \cos^2 \theta_{ikn} \cos \theta_{jik} S(r_{ik}) \left\{ (2\delta_{li} - \delta_{la} - \delta_{lb}) \left(\frac{r_{ik}^\alpha}{r_{ij} r_{ik}} - \cos \theta_{jik} \frac{r_{ij}^\alpha}{r_{ij}^2} \right) \right. \\
& \qquad \left. + (\delta_{lk} - \delta_{li}) \left(\frac{r_{ij}^\alpha}{r_{ij} r_{ik}} - \cos \theta_{jik} \frac{r_{ik}^\alpha}{r_{ik}^2} \right) \right\} \\
& - 2 \cos^2 \theta_{jik} \cos \theta_{ikn} S(r_{ik}) \left\{ (2\delta_{lk} - \delta_{lc} - \delta_{ld}) \left(\frac{-r_{ik}^\alpha}{r_{ik} r_{kn}} - \cos \theta_{ikn} \frac{r_{kn}^\alpha}{r_{kn}^2} \right) \right. \\
& \qquad \left. + (\delta_{li} - \delta_{lk}) \left(\frac{r_{kn}^\alpha}{r_{ik} r_{kn}} + \cos \theta_{ikn} \frac{r_{ik}^\alpha}{r_{ik}^2} \right) \right\},
\end{aligned} \tag{2.11}$$

δ_{ij} corresponds to the Kronecker delta.

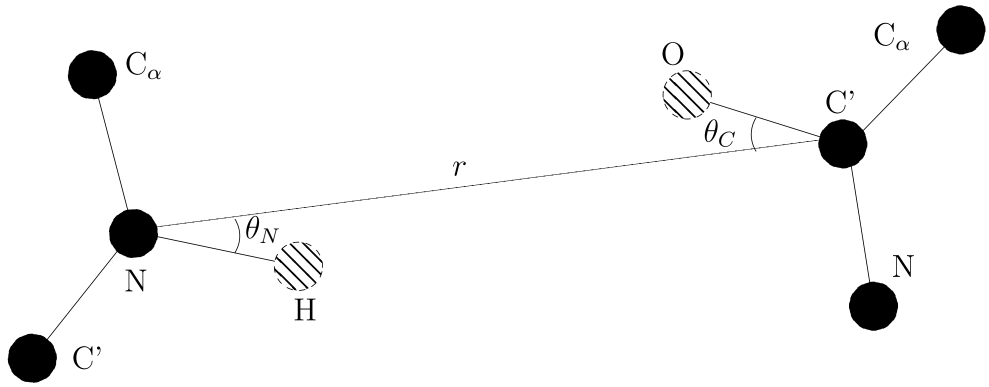


Figure 2.3: Schematic figure of the hydrogen-bond interaction. The light beads (H and O) are not explicitly modeled in the simulation; their positions are inferred from their bonded neighbors. Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

an explicit treatment of charged residues would allow one to look into properties that depend on the environment's pH or ionic strength. For a solution that has a high salt concentration (e.g., under physiological conditions), ions are able to screen most of the electrostatics, such that a Debye-Hückel potential would be appropriate to model this interaction. By compensating for the difference in binding energy for all the coefficients involved, one could disentangle charge effects from the MJ matrix. This, however, has not been done in the present model.

Dipole interaction

The interactions described above were sufficient to fold and stabilize α-helices, but not β-sheets. Chen *et al.* [CSM06] have pointed out that there is an important contribution usually neglected in generic models: carbonyl and amide groups at the peptide bond form dipoles that interact with each other. Mu and Gao [MG07] showed that the nearest-neighbor interaction is enough to sufficiently raise the occurence of β-conformations. Effectively, all dipoles along a helix are parallel compared to more favorable antiparallel neighboring dipoles on a β-sheet.

From a computational stand point, a dipole-dipole interaction

$$V_{\mathrm{dd}}(\boldsymbol{p}_i, \boldsymbol{p}_j) = \frac{\epsilon_{\mathrm{dd}}}{r^3}\Big[\boldsymbol{p}_i \cdot \boldsymbol{p}_j - 3\,(\boldsymbol{p}_i \cdot \hat{\boldsymbol{r}})(\boldsymbol{p}_j \cdot \hat{\boldsymbol{r}})\Big] \tag{2.12}$$

between two dipoles $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ at a distance $\boldsymbol{r}$ from each other is inconvenient because it is long-ranged. However, nearest-neighbor dipoles are all separated roughly by the same distance, as all amino acids have the same backbone geometry. All dipoles also

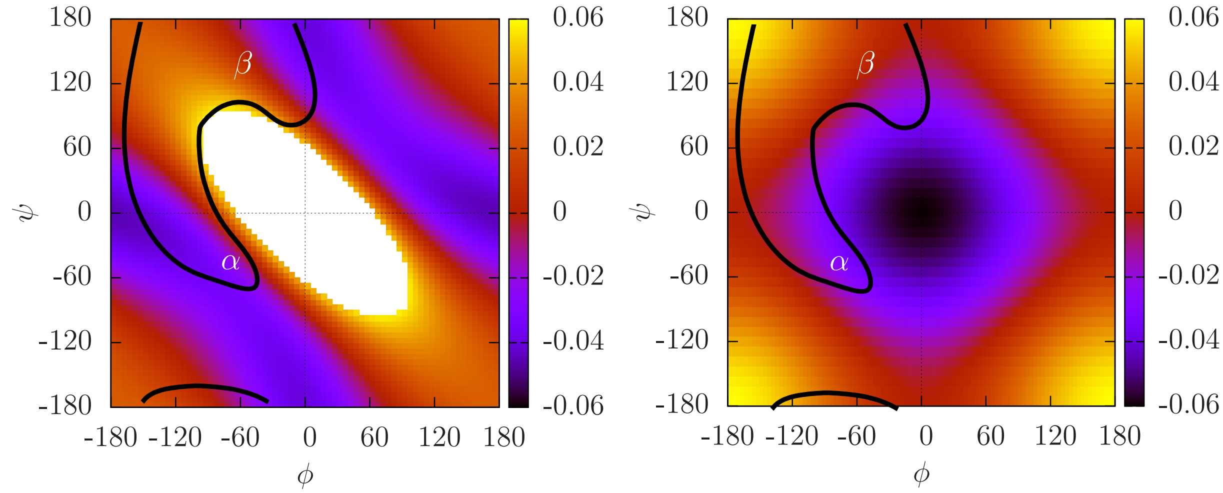


Figure 2.4: Map of the nearest-neighbor dipole-dipole interaction for all sets of dihedral angles ϕ and ψ (left), and the decoupled Fourier series approximation (right). The central part of the left plot was not reproduced in order to emphasize local difference in other regions of the plot (as can be seen in Figure 2.5, this anyways is a sterically hindered region). Sterically favored regions of the plot are circumscribed by a thick line, in addition to labels of α and β regions. The two graphs were shifted and scaled for comparison. Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

have the same magnitude, as they are formed from the same atoms. Therefore, the key component of the interaction lies in the relative orientation between dipoles, and not in their magnitude or relative distance. Successive dipoles therefore capture the orientation of the local backbone geometry. Two neighboring dipoles will effectively measure the angle difference between the two planes $\mathrm{C'}_{i-1}\mathrm{N}_i\mathrm{C}_{\alpha,i}$ and $\mathrm{C}_{\alpha,i}$ $\mathrm{C'}_i\mathrm{N}_{i+1}$, where the index keeps track of the amino acid involved (see Figure 2.1). As the effect is completely localized and only affects the conformation of the amino acid backbone, we treat this interaction as a *bonded* one, by effectively biasing the dihedral potentials of ϕ and ψ. To do so, we first calculated Equation 2.12 for all combinations of dihedral angles with a 5° resolution. The result is plotted on Figure 2.4 (left). The (sterically forbidden) central part of the plot was removed to emphasize local differences in allowed regions.

In order to be efficient, the potential should decouple along the two coordinates, i.e., it must be expressible as a sum $U(\phi, \psi) \simeq U(\phi) + U(\psi)$. We use a single cosine function centered around $\phi, \psi = 0$ with identical amplitude along both coordinates to approximate the neighboring dipole potential (Figure 2.4 (right)). Higher modes in the series have shown to be negligible:

$$V_{\rm dip}(\phi,\psi) = k_{\rm dip}\Big[(1-\cos\phi) + (1-\cos\psi)\Big]. \tag{2.13}$$

The value of the optimally tuned free parameter $k_{\rm dip}$ is reported in Table 2.1. The discrepancy between the plots is due to the enforced decoupling of the two coordinates ϕ and ψ.[8] Even though the final result looks rather inaccurate on the whole domain of the function, it nevertheless recreates the one important effect of the interaction: the β region is more favored than the α region (see labels in Figure 2.4 (left)). Moreover, the quality of the fit should only be tested along the physically relevant domains of the Ramachandran plot, most notably the α and β regions. In this sense, Equation 2.13 makes for a good approximation of the dipole interaction, and is enough to recreate the physics that favors β regions.

2.3 Simulations

MD simulations were performed with the ESPResSo package [LAMH06]. Simulations in the canonical ensemble (NVT) were achieved by using a Langevin thermostat with friction constant $\Gamma = \tau^{-1}$. The temperature was expressed in terms of the intrinsic unit of energy, $\mathcal{E}$. The force field is parametrized in order to reproduce a temperature of $T = 300\,\mathrm{K}$. The integration time step used for all simulations is $\delta t = 0.01\,\tau$.

2.3.1 Initial conformations

Initial peptide conformations were generated using either: (*i*) a random structure, as described in Technical Point 2.2, or (*ii*) atomistic or coarse-grained structure from a PDB file. A PDB file, which contains the position of all atoms in a protein structure, can easily be coarse-grained into the representation of this model by simply keeping the positions of amide nitrogen N, central carbon C_α, carbonyl group C', and side chain atom C_β to their atomistic coordinates while ignoring all others.

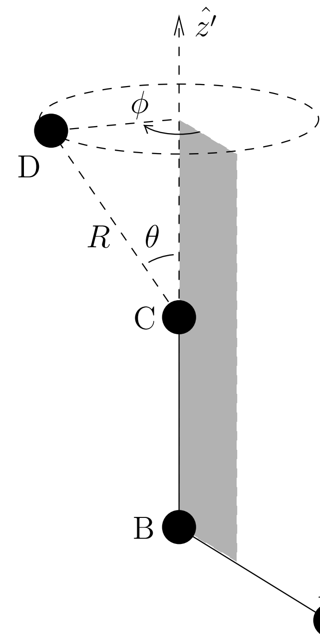
2.3.2 Warmup

Initial conformations often contain partially overlapping beads which produce high-energy steric clashes. The stability of the discretized integrator for the equations of motion depends on the steepness of the forces involved (the stiffest interaction determines the time-step). Steric clashes will result in numerical instabilities due to the integration of steep forces (e.g., large contribution from the r^{-12} term in the Lennard-Jones interaction).

[8] Using different amplitudes, $k_{\rm dip}$, for the two coordinates, or phase shifts (e.g., $\phi-\phi_0$), did not improve the agreement between the two plots significantly.

Technical Point 2.2 Generation of a random chain

Upon building a chain, most bonds (i.e., bond lengths, angles) severely constrain its geometry. Only the dihedral interactions are weak enough to allow large-scale flexibility. A random structure therefore corresponds to assigning random dihedrals along the chain.



The chain was built sequentially starting from the N-terminus (see Figure 2.1 for chain topology). The very first N bead is randomly placed inside the simulation box. The next C_α bead is placed at the surface of a sphere with radius corresponding to the bond length NC_α (see Table 2.1) with random orientation [AT93]. Next, the C_β bead is constrained by the bond length $C_\alpha C_\beta$ and the bond angle $NC_\alpha C_\beta$, leaving out one degree of rotational freedom. (Rotational freedom in the beads N, C_α, and C_β only applies for the first residue. All other beads are fully constrained by an additional dihedral angle; see Figure 2.2.) For this particle, as well as all subsequent ones, it has shown easiest to calculate the coordinates of the new bead using the previous ones in a *local* coordinate system, as described in [PHR+05] and briefly summarized below.

Let A, B, C, and D, be four neighboring beads where all particles but D have been placed. The unit vector between beads B and C, $\widehat{bc} = \mathbf{BC}/|\mathbf{BC}|$, defines the local z'-axis, whereas the previous bond vector $\mathbf{AB}$ will be oriented along both the z' and x' axes. The local coordinate system can then be defined from the cross product of $\mathbf{AB}$ and $\widehat{bc}$

$$\hat{n} = \frac{\mathbf{AB} \times \widehat{bc}}{|\mathbf{AB} \times \widehat{bc}|}. \tag{2.14}$$

The new atom D is expressed in the local coordinate system in terms of spherical coordinates

$$\mathbf{D'} = \begin{pmatrix} R\sin\theta\cos\phi \\ R\sin\theta\sin\phi \\ R\cos\theta \end{pmatrix}, \tag{2.15}$$

where R is the bond length between C and D, θ is the angle between B, C, and D, and ϕ is the (random) dihedral angle between A, B, C, and D.

The transformation matrix $M = [\widehat{bc}, \hat{n} \times \widehat{bc}, \hat{n}]$ allows to calculate the position of D in the original coordinate system from the position of C and D': $\mathbf{D} = M\mathbf{D'} + \mathbf{C}$.

Free parameters	Tuning method
$\sigma_{\mathrm{N}}, \sigma_{\mathrm{C}_\alpha}, \sigma_{\mathrm{C}'}, \sigma_{\mathrm{C}_\beta}, \epsilon_{\mathrm{bb}}$	Ramachandran plot
$\epsilon_{\mathrm{hp}}, \epsilon_{\mathrm{hb}}, k_{\mathrm{dip}}$	Folding characteristics

Table 2.4: Table of free parameters in this CG model. The main test that was used to determine a given parameter is denoted in the second column.

To avoid such behavior, it is necessary to first relax such overlaps before running the simulation. Atomistic simulations typically rely on finding a local potential energy minimum near the starting structure using, e.g., steepest descent or conjugate gradient methods [HKvdSL08]. Here we warmup the system by running a simulation with all forces capped at a specific value. The force value at which capping is performed is slowly increased until it can be removed altogether.

2.4 Parameter tuning

There are various ways coarse-grained force fields can be parametrized. For instance: only allowing non-bonded interactions between native contacts (Gō-type models) [Gō83]; partitioning measurements of amino acids between water and a hydrophobic medium [MKP+08]; structure-based coarse-graining based on all-atom simulations [ANV07]; knowledge-based potentials which intend to optimize parameters by using large pools of existing structures [FTLSW04].

Parameter tuning in top-down CG models aims at reproducing a selected subset of structural or energetic system properties. Since these parameters tend to be correlated, a given set needs to be tested at all scales. In our force field, *local* conformations are tuned to reproduce probability distribution functions of dihedral angles, which by a slight extension of standard terminology we also called Ramachandran plots (see subsection 2.4.1). Large scale (*global*) properties are targeted by studying folding events of a helical peptide (see subsection 2.4.2). The final set of parameters was identified as the one we felt most capable in reproducing properties on both levels. The physical conditions (temperature, density, etc.) of the force field will be set by the systems we try to match.

Note that on the global level we tune our parameters using only one protein. Of course, adding more proteins into the "training set" would incorporate more information, presumably leading to a better founded force field. There exist various successful parametrization schemes that rest on large ensembles of data [MS96, FTLSW04, MC06]. This, however, needs to be balanced against the need to test, how reliable a given force field handles proteins that were not part of its training set—a point we deemed more relevant.

Table 2.4 lists the eight free parameters that need to be determined. Because of time constraints, and to obtain some intuition and feeling of each interaction involved, we made a point of having our model tunable by hand, which is why we required the number of free parameters to be as low as possible. As explained above, this is the main reason why we decided against amino acid specific bead sizes. Adding even more free parameters, on top of being time consuming during tuning, would make it difficult to obtain a consistent set of parameters that would correctly describe both local and global conformations. Different bead sizes would involve different Ramachandran plots, and all backbone parameters would need to be consistent throughout.

The free parameters were tuned by trying to constrain parameter space as much as possible, for instance by eliminating unphysical behavior (e.g., sterically hindered β region in the Ramachandran plot, or too much helicity in secondary structures). Combining both local and global tests was enough to settle for a satisfying set of parameters, using the constraint that the dipole interaction strength was maximized. Even though this may sound arbitrary when looking for a realistic α/β content ratio, it turned out to be very difficult to use β structures as tests because they are so weakly stabilized. Indeed, we have found that the final set point is still not strong enough to fully stabilize β-sheets during folding events (see below). This shows that maximizing the dipole interaction strength in this model does not lead to oversampling of β content, but merely sampling as much extended conformations as possible before the force field cannot stabilize helical structures anymore. Other simple tests can be used to exclude regions of parameter space. For example, a hydrogen bond interaction that is too strong will lead to proteins that fold into one long helix. Too strong hydrophobic interactions will collapse proteins into globules, even native elongated helical structures. Bead size parameters were initially taken from other CG models (e.g., [DBB+03, ISW00, TLSW99]) and tuned as little as possible to recreate enough sampling of α/β content, while suppressing sterically hindered regions.

As for any physical system, the representative sampling of its phase space is prerequisite to obtaining accurate thermodynamic information. Different schemes have been developed to characterize and estimate the population of thermodynamic states [MLP04, TV77, PS04, DMS+06]. In the present case, thermodynamic calculations were performed by combining parallel tempering [SW86] with the Weighted Histogram Analysis Method (WHAM) [FS88, KRB+92, KRB+95] (more details can be found in Appendix A). The main idea is to combine energy histograms from canonical simulations at various temperatures in order to reconstruct the density of states of the system. The information contained in these histograms is used to calculate a consistent set of free energy differences between each simulation. Converging these free energies was done by using a recently developed highly efficient algorithm (see [BS09] and Appendix A). Once the density of states is reconstructed, one can obtain continuous approximations to all thermodynamic observables. By combining WHAM with parallel tempering, we effectively improve sampling by reducing correlations between data points.

2.4.1 Local conformations: Ramachandran plot

The Ramachandran plot [RRS63] records the occurrence and frequency of successive (ϕ, ψ) angles in a protein. Since backbone flexibility is almost exclusively due to these two

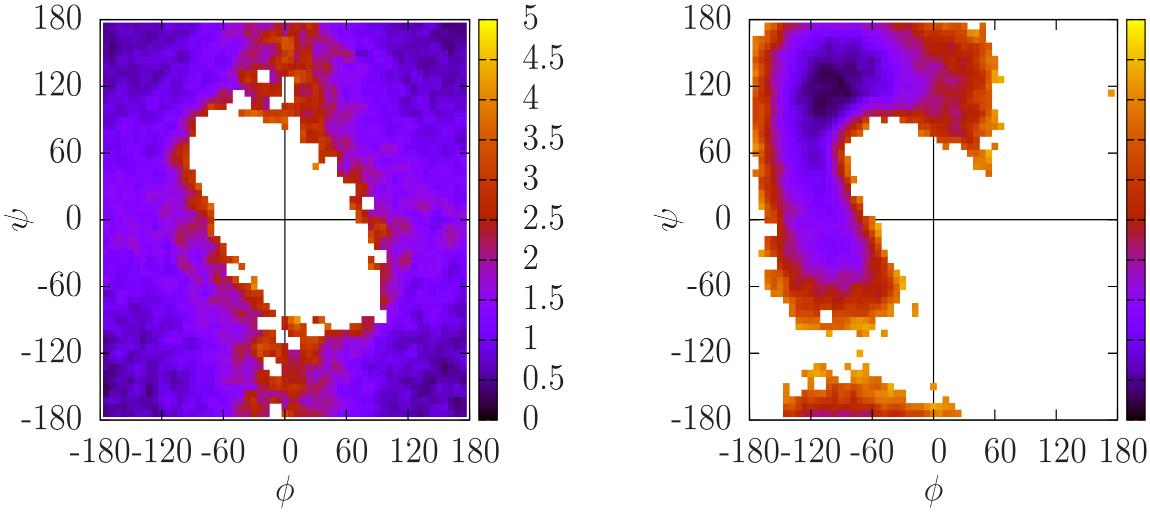


Figure 2.5: Free energy plots of tripeptides Gly-Gly-Gly (left) and Gly-Ala-Gly (right), as a function of successive dihedrals ϕ and ψ, calculated at our reference temperature $T = 1\,\mathcal{E}/k_{\mathrm{B}}$. The coloring represents the free energy difference with the lowest conformation, in units of $k_{\mathrm{B}}T$. Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

coordinates, the Ramachandran plot is an ideal reporter of local (secondary) structure: α-helices and β-sheets belong to peaks in different regions of the plot. And since proteins are highly constrained systems, low energy points on the Ramachandran plot are rather well localized. Their accurate sampling is therefore prerequisite to the formation and stabilization of reliable structures on larger scales. In the following we will be concerned with the (thermal) distribution of the (ϕ, ψ) angles surrounding some particular amino acid and, in a slight stretch of standard terminology, also refer to this probability density as a Ramachandran plot.

The free parameters that most directly constrain the Ramachandran plot are the different bead sizes $\{\sigma_{\mathrm{N}}, \sigma_{\mathrm{C}_\alpha}, \sigma_{\mathrm{C}'}, \sigma_{\mathrm{C}_\beta}\}$ and, to a lesser extent, the excluded volume energy prefactor ϵ_{bb}. We disentangled hydrogen bond and hydrophobicity effects from the Ramachandran plot by studying systems made of only three amino acids. From a steric point of view we only distinguish between glycine and non-glycine amino acids,[9] by either not having a side chain bead at all (Gly) or by using a generic bead representing the 19 other amino acids (Ala, for the sake of concreteness). It is then sufficient to study the two Ramachandran plots of Gly-Gly-Gly and Gly-Ala-Gly tripeptides, the smallest systems that contain relevant information on successive dihedral angles ϕ and ψ. The reason why we sur-

[9] Proline and pre-proline are modeled like any other non-glycine residues.

round the amino acid of interest with two Gly is to avoid hydrophobic interactions between neighboring side chains. As a result, we solely probe steric effects. The Ramachandran plots derived from the final set of parameters are shown in Figure 2.5 as free energy plots obtained from using parallel tempering at temperatures $k_{\mathrm{B}}T/\mathcal{E} \in \{0.5, 0.7, 1.0, 1.3, 1.6, 1.9, 2.2, 2.5\}$ and reconstructing the density of states with WHAM. The free energy plot is calculated at our reference temperature $k_{\mathrm{B}}T/\mathcal{E} = 1$. The shading represents the free energy difference with respect to the lowest conformation, in units of $k_{\mathrm{B}}T$. Notice the inherent asymmetry in the Gly-Ala-Gly system, which reflects the chirality of the α-carbon. Both α-helix $(-60°, -60°)$ and β-sheet $(-60°, 130°)$ regions are well populated, in agreement with Ho *et al.* [HTB03]. Proper balance and connectivity between the two regions is crucial for protein folding. This is tuned by the bead sizes and excluded volume energy, but also depends on the dipole interaction k_{dip} (see below). The achiral Glycine, on the other hand, has no side chain, and permits many more conformations. One therefore often finds glycine residues at the ends of helices.

A particular challenge was the fact that we model neither the amide-hydrogen nor the carbonyl-oxygen explicitly, yet their steric effects strongly shape the Ramachandran plot [HTB03]. This required subtle adjustments of the bead sizes of the N and C′ atoms compared to their conventional van der Waals radii.

A poor sampling of local conformations can thwart the formation of realistic secondary structure. Moreover, the *relative* weight of characteristic regions of the Ramachandran plot determines to a large extent the α/β content. Even though the analysis of above-mentioned tripeptides accounts for steric effects and the dipole interaction, it does not consider hydrogen bonds and side chain interactions which are also important to stabilize secondary structure. For this reason it is difficult to ascertain the quality of conformational distributions without studying larger structures.

2.4.2 Folding of a three-helix bundle

In this section we study full size proteins to parametrize large scale interactions. We used proteins found in the Protein Data Bank [WWWa] that were resolved experimentally in aqueous solvent.

Our choice of reference protein is constrained by the limitations of our model. For instance, salt- or disulfide-bridges cannot yet be represented and should thus play no role in the reference protein either. Also, it was important to start with a simple structure rather than a globular protein for which packing and cooperativity are more important. Following Irbäck *et al.* [ISW00] and Takada *et al.* [TLSW99], we also tuned our force field on a *three-helix bundle*. Direct comparisons with their models is difficult, though. First, these authors do not incorporate specificity on every amino acid and only represent a few amino acid types (e.g., hydrophobic, polar, glycine residue). Second, they only compared their simulations to the lowest-energy structure found during the simulation, rather than experimental data. In contrast, we use the *de novo* protein α3D (73 residues) and systematically compare our results with the real structure resolved experimentally (using NMR) [WCB+99]. The

Name	PDB ID	Structure	Sequence
α3D	2A3D	Three-helix bundle	MGSWA EFKQR LAAIK TRLQA LGGSE AELAA··· FEKEI AAFES ELQAY KGKGN PEVEA LRKEA··· AAIRD ELQAY RHN
GS-α_3W	1LQ7	Three-helix bundle	GSRVK ALEEK VKALE EKVKA LGGGG RIEEL··· KKKWE ELKKK IEELG GGGEV KKVEE EVKKL··· EEEIK KL
S-824	1P68	Four-helix bundle	MYGKL NDLLE DLQEV LKNLH KNWHG GKDNL··· HDVDN HLQNV IEDIH DFMQG GGSGG KLQEM··· MKEFQ QVLDE LNNHL QGGKH TVHHI EQNIK··· EIFHH LEELV HR
S-836	2JUA	Four-helix bundle	MYGKL NDLLE DLQEV LKHVN QHWQG GQKNM··· NKVDH HLQNV IEDIH DFMQG GGSGG KLQEM··· MKEFQ QVLDE IKQQL QGGDN SLHNV HENIK··· EIFHH LEELV HR
R1-69	1R69	Five short helices	SISSR VKSKR IQLGL NQAEL AQKVG TTQQS··· IEQLE NGKTK RPRFL PELAS ALGVS VDWLL··· NGTSD SNVR
aIF2β	1K8B	Two helices and a four stranded β-sheet	EILIE GNRTI IRNFR ELAKA VNRDE EFFAK··· YLLKE TGSAG NLEGG RLILQ RR
MBH12	1K43	β-hairpin	RGKWT YNGIT YEGR
		β-hairpin	VVVVV$^{\mathrm{D}}$PGVVV VV

Table 2.5: Structure and amino acid sequence of all proteins studied in this chapter.

amino acid sequence is given in Table 2.5. A similar protocol was followed by Favrin *et al.* [FIW02] in order to study a different three-helix bundle (PDB: 1BDD).

A first attempt in tuning parameters consisted of simulating proteins starting from their native structure. Testing for stability is a rapid means to constrain parameter space, but not sufficiently so as to actually determine their values. This is consistent with the picture of a deep funnel-like free energy landscape [BOSW95]: the free energy minimum of a native state is sufficiently deep compared to unfolded states that a folded protein is very stable against force field parameter variations. Further tuning was therefore mainly achieved by studying folding events using a set of trial runs with different parameters. Observation of three-dimensional structures with VMD [HDS96] was well suited to characterize simulations. The software was also used to render protein images throughout the present thesis.

Folding was studied in the following way: the only input into our simulations was the sequence of amino acids and the temperature. The initial conformation (determined by the collection of dihedral angles ϕ and ψ) was chosen randomly, and the integration started by

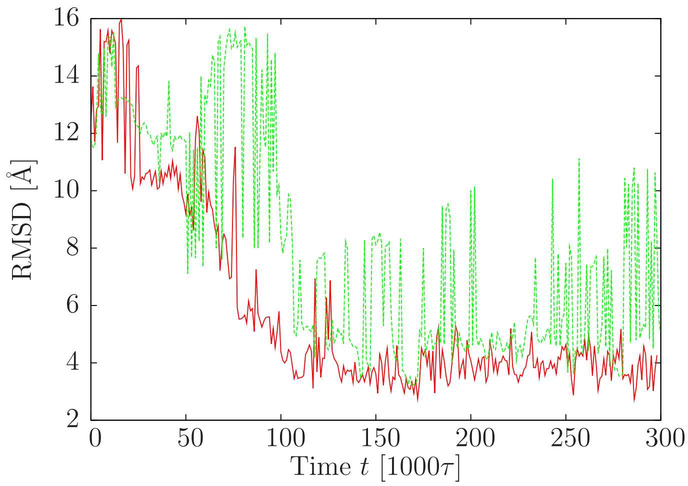


Figure 2.6: RMSD of the CG protein α3D (full line) and S-824 (dashed line) compared with experimentally resolved structures. Both simulations were run at $T = 1\,\mathcal{E}/k_\mathrm{B}$. Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

warming up non-bonded interactions to relax high energy steric clashes. We used parallel tempering for all simulations to avoid kinetic traps. Structural observables were measured at $k_\mathrm{B}T = 1\,\mathcal{E}$, the temperature at which the force field was tuned. Simulations were set at eight different temperatures: $k_\mathrm{B}T/\mathcal{E} \in \{1.0, 1.1, \ldots, 1.4, 1.6, 1.9, 2.2\}$). MC swaps between different temperatures were attempted every $10\,\tau$, the average acceptance rate was around 10%. We tested convergence to a global minimum by checking that different initial conditions consistently equilibrate to the same structure. A combination of thermodynamic and kinetic studies (see below) will allow us to show two important features. First, the temperature used for parameter tuning, $k_\mathrm{B}T = 1\,\mathcal{E}$, is below the folding temperature T_f of α3D, above which the unfolded conformation becomes the most stable state. Second, $k_\mathrm{B}T = 1\,\mathcal{E}$ is above the glass transition temperature T_g, below which the energy landscape becomes very rugged and creates severe kinetic traps. It was indeed possible to observe folding events in conventional (i.e., not using parallel tempering) simulations within this range of temperature.

Quantitative comparison between the CG and the experimental structures can be made by calculating the root-mean-square-deviation (RMSD) between corresponding α-carbons on the two chains (after optimal mutual alignment). Figure 2.6 reports the RMSD of a protein in the lowest ($k_\mathrm{B}T = 1\,\mathcal{E}$) replica of a parallel tempering MD run as a function of time, using the RMSD Trajectory Tool within the VMD package [HDS96]. These results were obtained with the parameters reported in Table 2.1, Table 2.2, and Table 2.3. The average error between the equilibrated simulation and the NMR structure is around 4 Å,

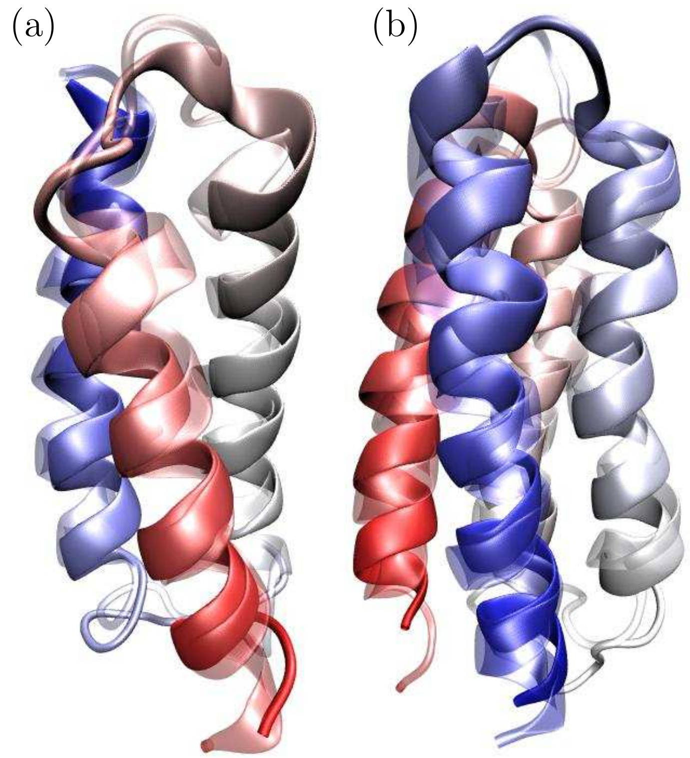


Figure 2.7: Equilibrated structures of three-helix bundle α3D (a) and four-helix bundle S-824 (b) sampled at $T = 1\,\mathcal{E}/k_\mathrm{B}$. Superposition of simulated structure (opaque) with experimental data (transparent) is displayed. The STRIDE algorithm [FA95] was used for secondary structure assignment (thick ribbons represent α-helices on the figure). Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

after about $100\,000\,\tau$ and at $k_\mathrm{B}T/\mathcal{E} = 1$, the temperature at which the native conformation represents the free energy minimum. A superposition of the simulated structure with the experimental one is shown in Figure 2.7. The STRIDE algorithm [FA95] was used to assign secondary structure. Overall the conformation is very well reproduced considering that we have a resolution of only 4 beads per amino acid, and that no *a priori* knowledge of secondary/tertiary structure was provided to the force field. Helix regions had formed at the right place, and amino acids were arranged in order to bury hydrophobic beads between the three helices, away from the implicit solvent.

To characterize the stability of this protein, we also performed thermodynamic calculations using WHAM and parallel tempering at the temperatures $k_\mathrm{B}T/\mathcal{E} \in \{0.8, 0.9, \ldots, 1.4, 1.6, 1.9, 2.2\}$. By reconstructing the density of states, we can estimate the folding temperature $k_\mathrm{B}T_\mathrm{f} \simeq 1.2\,\mathcal{E}$, the point at which the folded and unfolded states are equally

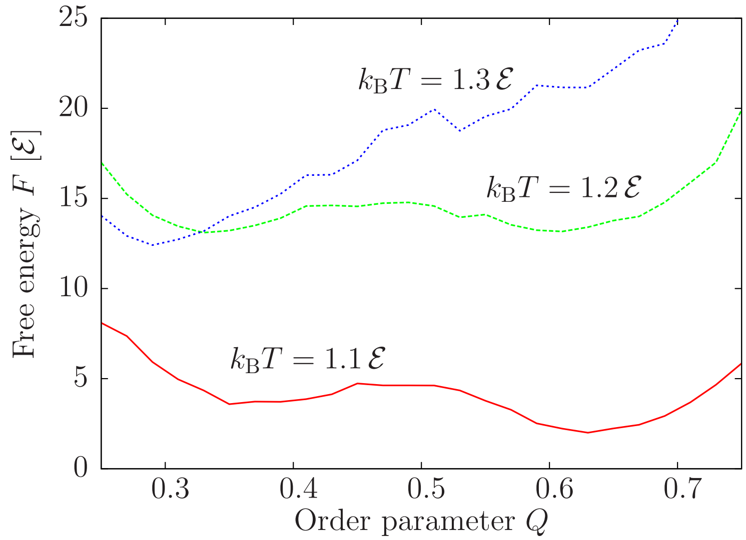


Figure 2.8: Free energy profile as a function of nativeness order parameter Q (Equation 2.16) below ($T = 1.1\,\mathcal{E}/k_B$), at, and above ($T = 1.3\,\mathcal{E}/k_B$) the folding temperature $T_f = 1.2\,\mathcal{E}/k_B$ for α3D. Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

populated. This gives a measure of the stability of the system: below T_f the native state is the most likely conformation. In Figure 2.8, we plot the free energy below, at, and above the folding temperature as a function of the nativeness order parameter Q as introduced by Takada *et al.* [TLSW99]. It measures the distance r_{ij} between pairs i and j of C_α beads between the NMR data and CG simulations:

$$Q = \left\langle \exp\left[-\frac{1}{9\sigma^2}\left(r_{ij}^{\mathrm{NMR}} - r_{ij}^{\mathrm{CG}} \right)^2 \right] \right\rangle_{ij} \tag{2.16}$$

where the average goes over all pairs ij and $\sigma = 1\,\text{Å}$. The folded conformation lies in the basin $Q \gtrsim 0.6$ whereas all unfolded conformations (in which not all three helices have properly formed) occur for $Q \lesssim 0.5$. It should be noted that all three curves in the graph have been calculated by using the same reference point, meaning that the vertical shift between curves accounts for the free energy difference in going from one temperature to another. The folding temperature is close to $1.2\,\mathcal{E}/k_B$. To make sure the model is also able to sample this important part of phase space in conventional simulations, we provide a stability run at the folding temperature starting from a random conformation. It can be seen that the system repeatedly switches between folded and unfolded states and roughly spends as much time in either one (Figure 2.9).

In 13 out of 15 independent parallel tempering simulations the protein folded to the native state at a temperature $T = 1\,\mathcal{E}/k_B$. However, the folding time varied substantially

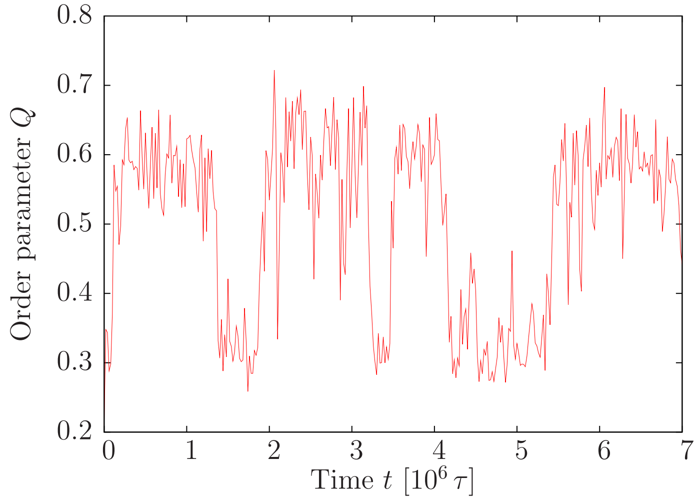


Figure 2.9: Conventional (i.e., not using parallel tempering) simulation of α3D at $T = 1.2\,\mathcal{E}\,/k_{\mathrm{B}}$. The nativity parameter Q is plotted against time. The protein alternates between folded ($Q \gtrsim 0.6$) and unfolded conformations ($Q \lesssim 0.5$). Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

between different simulations. The kinetics of folding of this protein was studied by running conventional simulations at various temperatures. For each temperature $k_{\mathrm{B}}T/\mathcal{E} \in \{0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$ we ran 10 simulations and measured the average time it took to fold the protein to its native conformation, if it ever did in the time scale of the simulation ($2{\times}10^6\,\tau$). The results are reported in Figure 2.10. Temperatures $0.7\,\mathcal{E}/k_{\mathrm{B}}$ and $0.8\,\mathcal{E}/k_{\mathrm{B}}$ did not yield a single folding event, suggesting the onset of glassy behavior [SO94, BOSW95]. The glass transition temperature T_{g} can be estimated following a simple pragmatic scheme suggested by Socci and Onuchic [SO94]: it is the temperature where the mean folding time is the average of the minimum folding time τ_{min} (lowest point in the graph) and the largest time scale one is willing to invest in the simulation, τ_{max} (highest boundary in the graph): $\tau_{\mathrm{g}} = (\tau_{\mathrm{min}} + \tau_{\mathrm{max}})/2$. This average time is plotted as a horizontal line in the graph. One can then estimate what temperature this folding time corresponds to (Figure 2.10). In our case, we can safely assume that $T_{\mathrm{g}} < 0.9\,\mathcal{E}/k_{\mathrm{B}}$, meaning that the protein does not experience glassy behavior when simulating at our reference temperature $T = 1\,\mathcal{E}/k_{\mathrm{B}}$. Moreover, combining results from thermodynamic calculations and kinetic studies shows that there is a range of temperatures $T_{\mathrm{g}} < T < T_{\mathrm{f}}$ in which the system is not experiencing glassy behavior, but is still “cold” enough such that the native state is the most stable conformation.

Irbäck *et al.* [ISW00] as well as Takada *et al.* [TLSW99] have reported a degeneracy

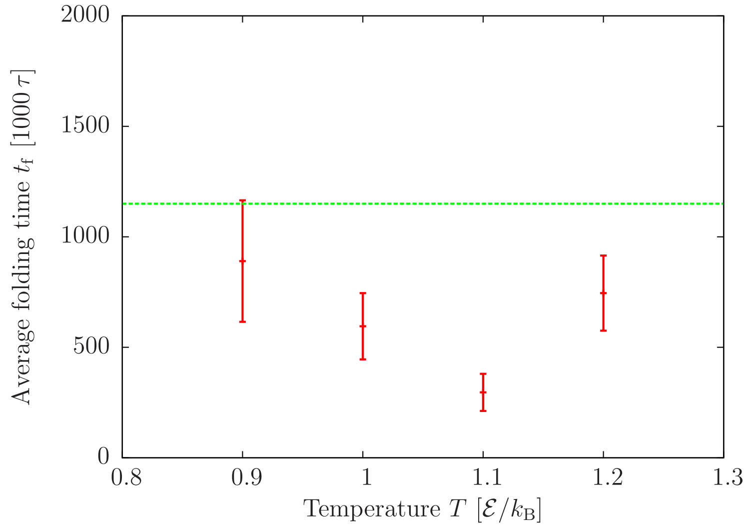


Figure 2.10: Kinetic studies of the α3D three-helix bundle CG protein. The average folding time t_f is plotted against temperature. For temperatures ranging from $T = 0.7\,\mathcal{E}/k_B$ to $T = 1.2\,\mathcal{E}/k_B$, about 10 simulations were run and we measured the first passage time to the native state. The line represents the average between the minimum folding time and the time scale of the simulation. This can be used to estimate the glass transition temperature (see text). Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

in the CG structures of their helix bundles: there are two ways three helices can pack (see Figure 2.11), and their models were not able to discriminate the two different tertiary structures. NMR experiments on α3D found a ratio between clockwise and counterclockwise topologies of several percents, leading to a free energy difference of a few k_BT at room temperature.[10] From 15 independent simulations we ran, one of them did not fold within $300\,000\,\tau$, and 13 converged to the NMR structure—a counterclockwise topology (Figure 2.11 (a)); only one had the other topology (illustrated in Figure 2.11 (b)). While it is encouraging to see that our model is able to distinguish these topologies, it is not guaranteed that this will work equally well for other proteins.

[10] W. F. DeGrado, personal communication.

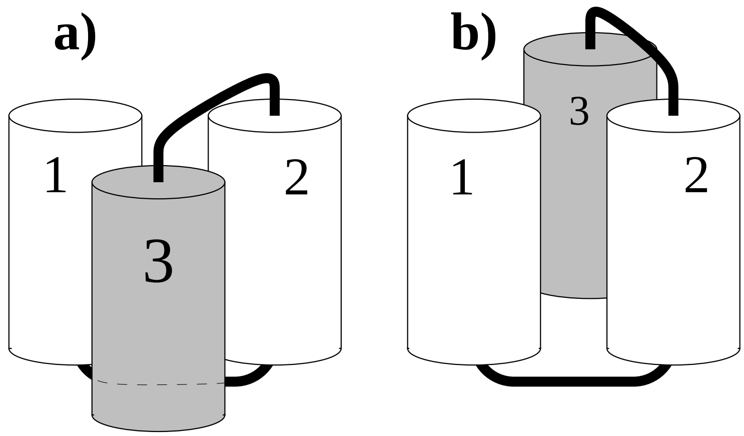


Figure 2.11: Schematic figure of the two possible topologies in forming a three-helix bundle. The native fold of protein α3D corresponds to a counterclockwise topology (a), that of GS-α_3W is clockwise (b). Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

2.5 Applications and tests

2.5.1 Folding

All simulations mentioned from this point onward have not been part of the parameter tuning training set. They come out as independent checks and features of the force field. Thermodynamic and kinetic studies were not performed for the different proteins of this section. Here, we study the equilibrium conformations of various sequences at a temperature of $k_{\mathrm{B}}T/\mathcal{E} = 1$, which lies between T_{g} and T_{f} for our reference protein, α3D. In this respect, we expect to avoid glassy behavior for similarly complex proteins whose native state is folded.

In order to test the folding features of the model, we first studied another *de novo* three-helix bundle, GS-α_3W. Even though the fold is very similar to α3D, it has 67 amino acids and a completely different primary sequence. Also, the native structure, obtained from NMR [DTF$^+$02], has the *opposite* topology (clockwise) compared to α3D. From 10 independent parallel tempering runs, $300\,000\,\tau$ long each, one of them did not fold within this amount of time (helices formed, but did not arrange properly). Out of the 9 remaining structures, 5 folded consistently to the native clockwise topology (Figure 2.11 (b)), and 4 to the other one (Figure 2.11 (a)). It should be noted that this sequence had been designed such that its native structure leads to favorable salt-bridge interactions [DTF$^+$02]. As we do not incorporate electrostatics (and thus salt bridges) explicitly, we expect the CG model to have difficulties in discriminating between the two tertiary structures.

In order to further probe the folding features of different α-helical rich folds, we studied a four-helix bundle, S-824, consisting of 102 amino acids [WKF$^+$03]. Even though the secondary structure is overall rather similar to the abovementioned three-helix bundle, the tertiary structure and amino acid sequence is completely different. Again, the reference structure is taken from experimental data [WKF$^+$03]. From 6 independent parallel tempering runs, each $600\,000\,\tau$ long, our force field successfully folded the protein into a four-helix bundle for every simulation except one, which did not have time to properly align its fourth helix. The RMSD is shown in Figure 2.6 for a simulation which converged to the right topology. As can be seen, the RMSD went below 4 Å, which, again, is very satisfactory considering the level of resolution and the complete absence of structure bias in the force field. It should be noted that what appears as large fluctuations on the graph are actually frequent MC swaps between replicas of the parallel tempering ladder. Fairly different structures from neighboring replicas are energetically comparable (which is the reason why they swap temperatures[11]), as can be seen on the RMSD plot. Just as in the three-helix bundle case, this protein can fold into several different topologies. Out of the five simulations which converged to a four-helix bundle tertiary structure, two of them represented the NMR topology. RMSD values for other topologies ranged between 5 Å and 8 Å. A snapshot of the equilibrated structure is shown in Figure 2.7 (b).

Also a second *de novo* four-helix bundle was used to test the force field. Even though the tertiary structure resembles the abovementioned S-824, the amino acid sequence of S-836 is completely independent (though it also has 102 amino acids) and the topology is different. Out of 3 independent runs, all of them successfully folded in a four-helix bundle structure within $600\,000\,\tau$, by comparing qualitatively the CG protein with the NMR structure [GKBH08]. However, none of them converged to the right topology.

Our model has proven very efficient in finding the equilibrium conformation of various helical structures, up to small deviations, and independently of their tertiary structure (i.e., number of helices) or sequence of amino acids. The fact that none of these proteins were part of the parameter tuning strongly indicates that our CG model captures important aspects of protein physics.

The limits of the model were reached when simulating globular proteins, such as R1-69 [MSA$^+$89], and aIF2β [CH02]. The chain collapsed into a molten globule, but the arrangement of secondary structures (collections of α-helices and β-sheets) was not accurately reproduced, leading to an incorrect tertiary structure. This suggests a missing sufficiently deep free energy minimum, most likely due to the limitations of the CG model in terms of cooperativity and realistic packing (recall that all side chains have the same bead size). The RMSD values did not drop below 10 Å.

Stabilizing a single β-hairpin in small proteins is difficult because this relies on very weak

[11] Recall the parallel tempering update rule between replicas i and j:

$$p = \min\left(1, \frac{\exp\left(-\beta_i E_j - \beta_j E_i\right)}{\exp\left(-\beta_i E_i - \beta_j E_j\right)}\right) = \min\left(1, \mathrm{e}^{(E_i - E_j)(\beta_i - \beta_j)}\right) \,,$$

where $\beta_i^{-1} = k_\mathrm{B} T_i$ [SW86, NB99].

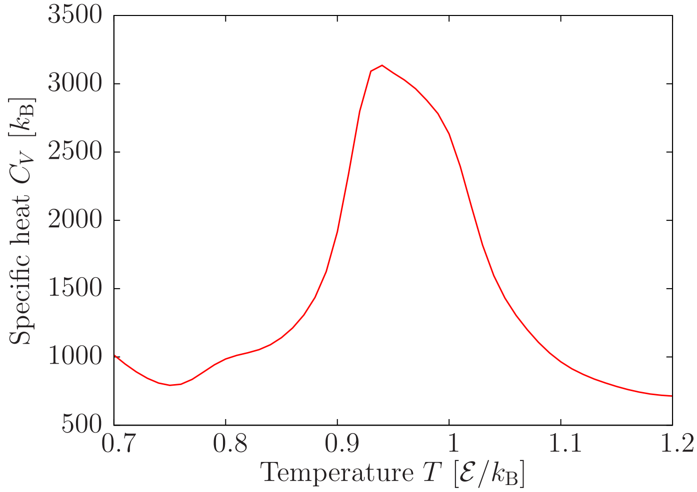


Figure 2.12: Specific heat of 15 GNNQQNY peptides in a cubic box of size 55 Å. The peak around $T = 0.95\,\mathcal{E}/k_B$ separates a low-temperature phase, rich in high-β content aggregates, from a high temperature phase where no aggregates form. Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

interactions. We simulated the *de novo* MBH12 peptide for $300\,000\,\tau$. It consists of 14 residues and forms a β-hairpin in water [PdlPL+02]. Our model is not able to stabilize it. The simulation shows a high tendency to form an α-helix, where 40 % of all conformations are helical, whereas only 2 % are extended (β-sheet like). However, the CG model can successfully fold a designed β-hairpin, sequence $V_5{}^{D}PGV_5$, which contains a D-proline in order to sterically favor hairpin formation [Gel98]. This peptide has been recently characterized using atomistic [FAC00] and structure-based coarse-grained simulations [TZV08].

2.5.2 Aggregation

Gsponer *et al.* [GHC03] recently reported atomistic simulations of small aggregation events in water. Heptapeptides GNNQQNY from the yeast prion protein Sup35 were shown to form β-sheet aggregates. These authors did a quantitative analysis of the number of 2- and 3-aggregates in the system at room temperature.

We studied the abovementioned scenario by simulating fifteen identical peptides in a box of size 55 Å,[12] without matching density with the atomistic run. Indeed, while Gsponer *et al.* simulated their system in a restricted sphere of 150 Å diameter and applying forces to constrain the system in the center, we set periodic boundary conditions in a cubic

[12]Ref. [BD09] incorrectly states a box of size 40 Å.

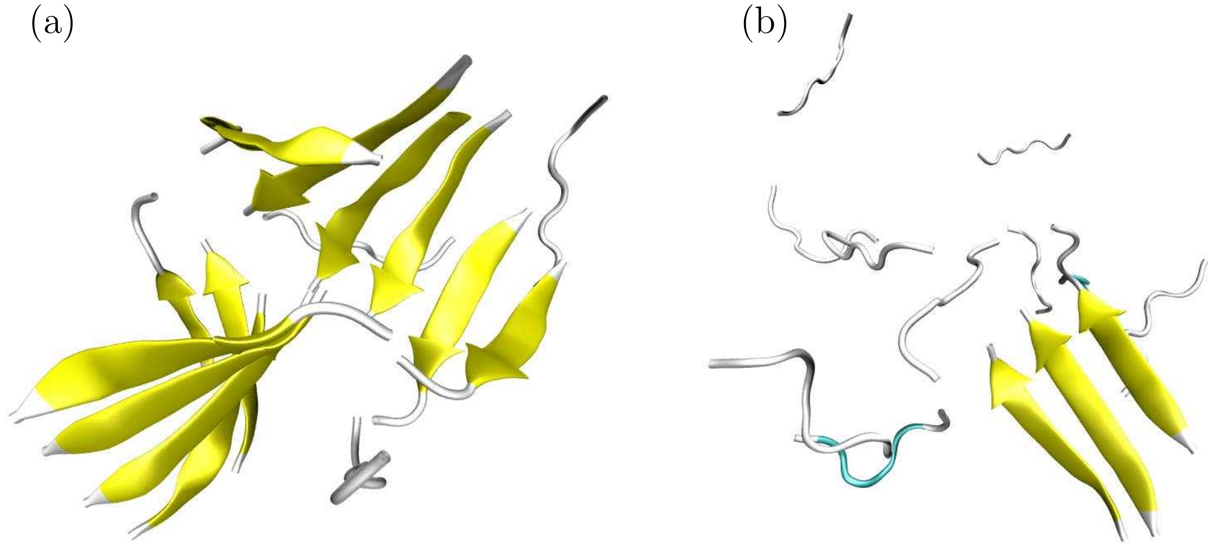


Figure 2.13: Snapshot of representative conformations at: (a) $T = 0.8\,\mathcal{E}/k_\mathrm{B}$ where peptides aggregate into a condensed phase and (b) $T = 1.0\,\mathcal{E}/k_\mathrm{B}$ which mostly samples disordered monomers. These two temperatures are located on both sides of the system's transition, as shown in Figure 2.12. Reprinted with permission from Bereau, T. and Deserno, M. *J. Chem. Phys.* **130** (2009), 235106 [BD09]. Copyright 2009, American Institute of Physics.

box. Even though this represents a rather dense system in order to drive aggregation, we checked that similar structures were sampled when simulating more dilute systems. Initial configurations were chosen randomly, and we ran parallel tempering simulations at temperatures $k_\mathrm{B}T/\mathcal{E} \in \{0.7, 0.8, 0.85, 0.9, 0.95, 1.0, 1.1, 1.2\}$ for $500\,000\,\tau$ each. We used WHAM to calculate the specific heat of the system (Figure 2.12). A clear peak occurs between lower temperatures, with formation of long-range fibrillar structures (Figure 2.13 (a)), and higher temperatures where the system mostly samples random coil monomers (Figure 2.13 (b)). The temperature dependence of the system's β-propensity is illustrated in Figure 2.14, where the free energy was calculated as a function of the ratio of residues in a β conformation for temperatures $k_\mathrm{B}T = \{0.9, 1.0, 1.1\}\,\mathcal{E}$. The results show that while the lowest temperature shows a free energy minimum at a non-zero ratio (i.e., $\approx 17\%$), the two temperatures above the transition exhibit a monotonically increasing curve, indicating that the formation of β structures (i.e., extended, β-sheets) is unfavorable. The data points are horizontally shifted due to the binning of the order parameter.

Table 2.6 shows a detailed analysis of the amount of β propensity sampled from the different temperature replicas. The top-most part of the table compares the average number of residues which exhibited antiparallel and parallel sheet conformations. The column on the right shows the average number of residues in a β-conformation for each temperature.

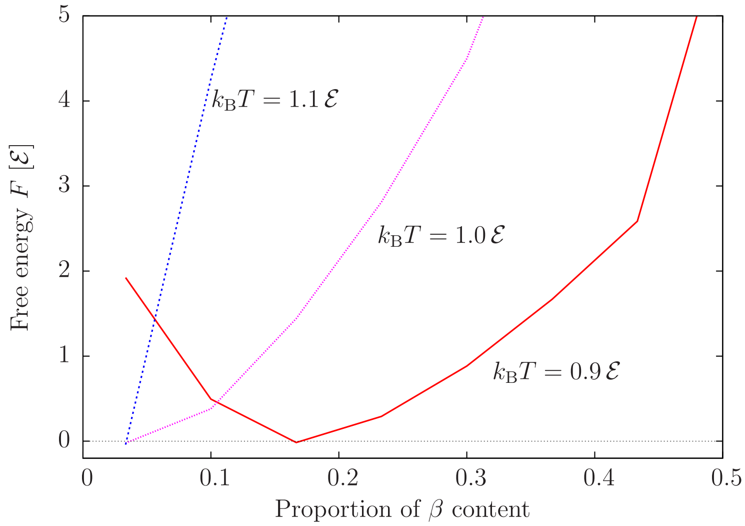


Figure 2.14: Free energy as a function of the (normalized) amount of β content for 15 GNNQQNY peptides at temperatures $T = \{0.9,\ 1.0,\ 1.1\}\,\mathcal{E}/k_{\mathrm{B}}$. All curves were shifted vertically such that $F = 0$ corresponds to the lowest free energy. The horizontal shift (i.e., the first data points are away from zero) is due to the binning of the order parameter upon calculating F.

There is a sharp drop in β-propensity between $T = 0.95\,\mathcal{E}/k_{\mathrm{B}}$ and $T = 1.0\,\mathcal{E}/k_{\mathrm{B}}$, which corresponds to the specific heat peak position (Figure 2.12). At lower temperatures, where aggregation occurs, we mostly observe parallel sheets over antiparallel.[13] Interestingly, this is in agreement with the study of Gsponer *et al.* and could be due to the hydrophobic interactions of the C-terminal tyrosine. To test this, we performed single point mutations in order to create a symmetric sequence. In this case parallel β-sheets also turned out to be more stable than antiparallel ones, which is unexpected since antiparallel β-sheets are generally believed to be lower in free energy [FP02]. One possible explanation is that the model is lacking electrostatic interactions at the N- and C-termini of the chains, which will favor antiparallel sheets, as the two ends have opposite charges.

These parallel GNNQQNY β-sheets also have the tendency to align within a plane, with the C-termini facing each other. This evidently results from the attraction between the C-terminal tyrosines, the most hydrophobic amino acid in this peptide.

To show that their force field was not biased towards aggregation, the authors also simulated a water-soluble control peptide SQNGNQQRG and found a difference in the amount of β-sheets formed. We compared the phase behavior of GNNQQNY and SQNGNQQRG by using WHAM on both sequences, but did not find statistically significant differences

[13]Note that β-propensity at $T = 0.7\,\mathcal{E}/k_{\mathrm{B}}$ is *lower* than around the transition. This is most-likely a sampling artifact, suggesting that this replica hasn't quite reached equilibrium. Lower-temperature replicas produce larger autocorrelations and thus require longer simulation times.

T $[\mathcal{E}/k_\mathrm{B}]$	Aggregation amount antiparallel	parallel	total
0.7	0.01	0.15	0.16
0.8	0.03	0.14	0.17
0.85	0.01	0.19	0.20
0.9	0.01	0.19	0.20
0.95	0.03	0.16	0.19
1.0	0.01	0.06	0.07
1.1	0.00	0.01	0.01
1.2	0.00	0.00	0.00

Table 2.6: Amount of aggregation sampled for 15 GNNQQNY at different temperatures. Percentage of antiparallel and parallel sheet formation, as well as the sum of both contributions.

over the studied temperature range. This suggests that some of the details necessary to distinguish the thermodynamics of these two peptides are too subtle for our force field to represent. Since the simulation temperature of Gsponer *et al.* in our case maps to $T = 1\,\mathcal{E}/k_\mathrm{B}$, which is where we essentially find the phase transition (Figure 2.12), effects only captured by the atomistic force field can indeed be expected to lead to substantial differences. Previous studies have shown how differences in CG force field parameters affect structure, β-sheet propensity, and aggregation behavior of different sequences [BS07].

All of these aggregation results were obtained using the same force field with no additional parameter adjustment. Other CG models have previously demonstrated aggregation events on a larger scale [PDU$^+$04, FOYHG07]. Here our goal was to show that we can study aggregation events using a force field that is tuned to reproduce simple folding features without biasing secondary or tertiary structure. This is important when looking at spontaneous aggregation or misfolding pathways, where one aims to reproduce general behavior without constraining the protein's structure towards a certain state that might not even be known or well-defined.

2.6 Summary

We have presented a new CG implicit solvent peptide model. Its intermediate resolution of four beads per amino acid permits accurate sampling of local conformations and thus secondary structure. Following cautious parameter tuning, the CG model is able to fold simple proteins such as helix bundles. Folding of a three-helix bundle was used to incorporate large-scale aspects of the force field, whereas the successful folding event of other helical bundles provided independent checks of reliability. Thermodynamic and kinetic studies of the three-helix bundle were carried out to verify that the folding temperature T_f was above the glass transition temperature T_g for this protein. The model was systematically compared to NMR data in order to optimize parameter tuning and precisely determine how much fine-scale information this CG model still contains. Of course, our model is not intended to compete with atomistic simulations, which is not the point of CG models; yet, carefully balancing several key contributions to the force field is a prerequisite to perform meaningful studies involving secondary and tertiary structure formation. Globular shaped proteins have proven more difficult to stabilize, presumably because accurate packing and strong cooperativity are not well enough captured. We also observe aggregation events of small β-sheets without retuning the force field. A realistic α/β balance, coupled with basic folding features, make the CG model very suitable for the large-scale and long-term regime that many biological processes require. Indeed, a force field that is not biased toward the protein's native conformation will likely give rise to insightful thermodynamic and kinetic studies when the structure is not known, not well defined, strongly perturbed from the native state, or adjusts during aggregation events.

3 Folding Kinetics of Two Model Peptides: α-helix and β-hairpin

This chapter attempts to evaluate the speedup factor f_t involved in the coarse-grained dynamics of the model. The folding kinetics of two model peptides—an α-helix and a β-hairpin—are calculated and compared to experimental estimates.

Coarse-graining is an attempt to reproduce key properties of a system at a reduced level of resolution. As mentioned in subsection 1.2.3, matching *static* properties alone provide no constraint on the dynamics of a coarse-grained system. Also, the reduction of molecular friction naturally leads to faster dynamics. One thus often resorts to estimating *how much faster* the model system is compared to reality. Assuming all dynamic processes were to happen homogeneously faster it should be possible to estimate the value of the speedup factor f_t that links coarse-graining with the real dynamics. Measuring the same dynamic process both numerically and experimentally would yield f_t. Examples include lateral diffusion of lipids [MKP+08] and folding simulations of peptides [TZV08]. While one cannot generally expect this assumption to hold, it offers, if nothing else, a rough estimate of the coarse-grained time-scale.

In the following we will evaluate the value of f_t by monitoring the folding kinetics of two model peptides: the α-helix forming $(AAQAA)_3$ and the β-hairpin $V_5{}^DPGV_5$ (see Figure 3.1). While only one peptide is necessary for the evaluation of the speedup factor, the second peptide provides a check for the transferability of f_t since these two dynamic processes are independent.

$(AAQAA)_3$ has been studied extensively, both experimentally [SYSB91, SDS94, ZJM+05] and computationally [FAC00, ZJM+05, ZTIV07, TZV08, CDL+09], and was shown to strongly stabilize a helical conformation. $V_5{}^DPGV_5$, composed of a D-proline followed by a glycine residue, was designed to enhance the conformational stability of β-hairpins as the geometry of D-Pro matches the right-handed twist of a β-sheet structure [KAB96, Gel98]. It can therefore reduce the entropic cost associated with turn formation—the rate limiting event in β-hairpin folding—and thus stabilize the folded state by increasing the folding rate [XPG06]. The hydrophobic tail is composed of valine residues.

3.1 Thermodynamics

In order to characterize the folding kinetics of a peptide, one first needs a suitable order parameter which describes the evolution of the system (essentially answering the question:

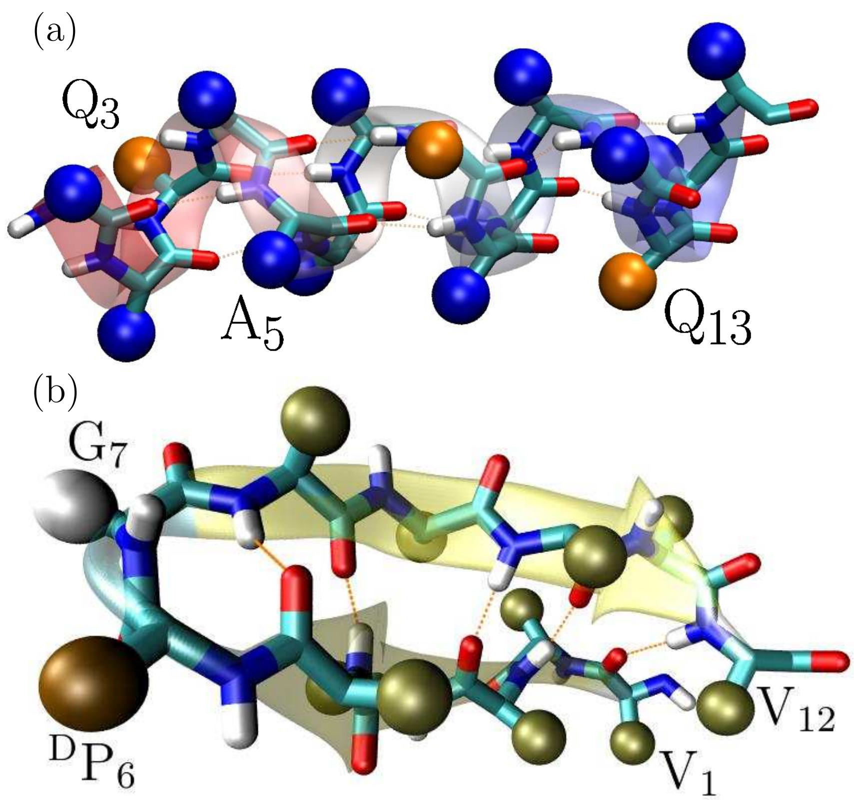


Figure 3.1: CG snapshots of the folded conformations of (a) α-helix $(\mathrm{AAQAA})_3$ sampled at $T = 1.0\,\mathcal{E}/k_\mathrm{B}$ and (b) β-hairpin $\mathrm{V}_5^\mathrm{D}\mathrm{PGV}_5$ sampled at $T = 0.7\,\mathcal{E}/k_\mathrm{B}$. Labels indicate amino acid names with their sequence position. The larger beads represent side chains and the light dashed lines show hydrogen bonds. Rendering was done with VMD [HDS96].

"is it folded?"). Following a similar (implicit-solvent atomistic) study of the same peptides [FAC00], we use the C_α root-mean-square deviation (RMSD) relative to the sampled coarse-grained structure with lowest potential energy (after optimal mutual alignment).

Thermodynamic information of the two peptides was obtained by combining parallel tempering [SW86] with the weighted histogram analysis method (WHAM) [KRB+95, KRB+92, FS88] (see Appendix A). The free energy as a function of RMSD was calculated according to Equation A.17 on page 122. Simulations were run at temperatures $k_\mathrm{B}T/\mathcal{E} \in \{0.7, 0.8, \ldots, 1.4\}$ for both $(\mathrm{AAQAA})_3$ and $\mathrm{V}_5{}^\mathrm{D}\mathrm{PGV}_5$ with a total simulation time per replica of $5 \times 10^6\,\tau$.

The results for the α-helix are shown in Figure 3.2 (a) for temperatures $T = \{1.0, 1.2, 1.3\}\,\mathcal{E}/k_\mathrm{B}$. All three curves have been calculated using the same reference point, such that the vertical shift between curves accounts for the free energy difference in going from one

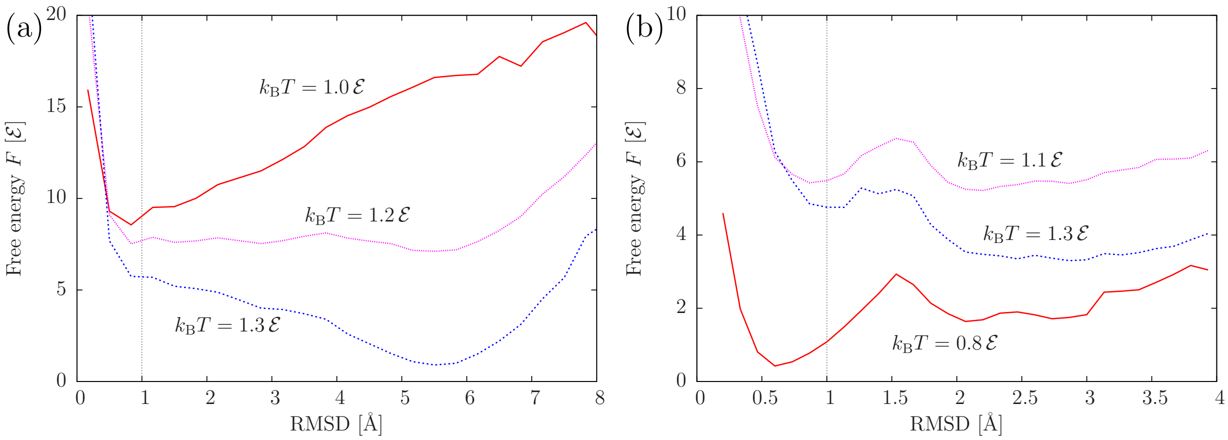


Figure 3.2: Free energy as a function of the RMSD from the sampled structure with lowest potential energy for (a) $(AAQAA)_3$ and (b) $V_5{}^{D}PGV_5$. The three curves represent the free energy at temperature $T = \{1.0,\ 1.2,\ 1.3\}\,\mathcal{E}/k_B$ and $T = \{0.8,\ 1.1,\ 1.3\}\,\mathcal{E}/k_B$, respectively. The vertical line indicates a 1 Å RMSD used as the folding threshold in section 3.2.

temperature to another. From analyzing the evolution of the free energy as a function of temperature, it is straightforward to identify folded conformations (corresponding to lower values of the RMSD ≈ 1 Å) from the unfolded ensemble (RMSD $\approx 5 - 6$ Å). Note that $k_BT = 1.2\,\mathcal{E}$ samples both ensembles with roughly equal free energies but does not exhibit any statistically significant free energy barrier in the middle.[1]

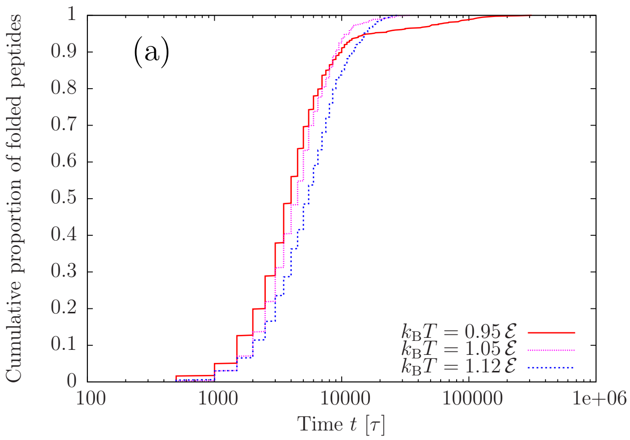
The corresponding results for the β-hairpin are shown in Figure 3.2 (b) for temperatures $T = \{0.8,\ 1.1,\ 1.3\}\,\mathcal{E}/k_B$. Compared to $(AAQAA)_3$, both the variations in RMSD and free energy difference are much lower, indicating that

- The change in chain extension (between compact and unfolded) is reduced for the hairpin $V_5{}^{D}PGV_5$.

- Energetics between folded and unfolded ensembles are comparable.

The second point is best illustrated by the absence of any peak in the canonical specific heat curve (data not shown). Still, the free energy curves in Figure 3.2 (b) clearly distinguish the folded (≈ 1 Å) from the unfolded ensembles ($\approx 2 - 3$ Å). The presence of a rather large free energy barrier should be interpreted with care.[1]

[1] The transition temperature T_c cannot be identified uniquely in the canonical ensemble for finite-size systems. Different observables will yield different values of T_c. Likewise, the presence of a free energy barrier is specific to the order parameter used in finite-size systems. See chapter 4 for details.



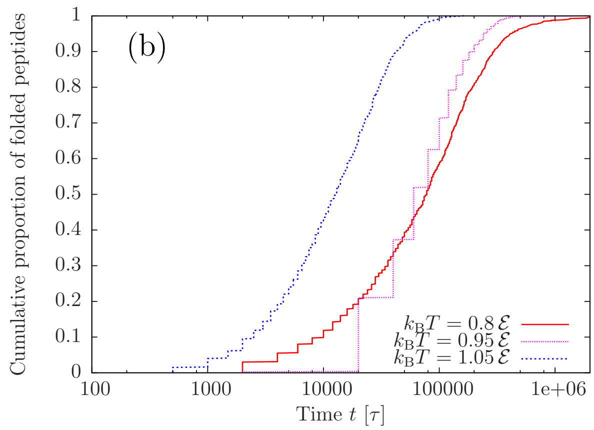


Figure 3.3: Cumulative proportion of folded peptides as a function of time for (a) $(AAQAA)_3$ and (b) $V_5{}^DPGV_5$ at different temperatures. Right boundaries correspond to the maximum simulation time.

3.2 Folding kinetics

Here, we evaluate folding times at a given temperature by measuring the average time taken to fold each peptide from a random conformation to the folded state. Various temperatures were considered. Following Ferrara *et al.*, we used the C_α-RMSD to follow the evolution of each system and set the folding threshold to 1.0 Å [FAC00]. As shown in Figure 3.2, such a value is well within the folded state of each peptide. Each folding time was evaluated from 800 independent simulations. Random conformations were first equilibrated for $2\,\tau$ to avoid steric clashes. Simulations were then run for up to $2 \times 10^6\,\tau$, or until the RMSD had reached the folding threshold.

The (cumulative) distribution functions of folding times are shown in Figure 3.3. The right boundaries correspond to the maximum simulation time. All simulations folded within the maximum simulation time, avoiding artifacts when calculating averages. The shoulder in the distribution function of $(AAQAA)_3$ at $k_BT = 0.95\,\mathcal{E}$ (around 95 %) is not due to statistical noise: an identical set of 800 simulations was run under the same conditions and yielded a nearly-identical curve (data not shown). Its origin is unclear.

The average[2] folding times extracted from the cumulative distribution functions shown in Figure 3.3 are presented in Table 3.1. The kinetics of $(AAQAA)_3$ yield an average folding time in the range $5\,000 - 7\,000\,\tau$ at $k_BT \approx \mathcal{E}$. The results indicate the presence of a *minimum* folding time between $k_BT = 0.95\,\mathcal{E}$ and $k_BT = 1.12\,\mathcal{E}$. The thermodynamic analysis of Figure 3.2 suggest that this range of temperature is somewhat below the folding transition of the peptide, indicating that folding is most efficient in the presence of

[2]The *median* may be more appropriate when the cumulative distribution functions of folding times show significant long-time tails.

	Average folding time [τ]	
T [$\mathcal{E}/k_B$]	$(AAQAA)_3$	$V_5{}^DPGV_5$
0.8		140 000
0.95	7 900	92 000
1.0		19 000
1.05	5 200	
1.12	6 400	

Table 3.1: Average folding times for $(AAQAA)_3$ and $V_5{}^DPGV_5$ at different temperatures.

weaker fluctuations. This illustrates the predominance of enthalpic contributions—mainly through hydrogen bonds. Interestingly, Kaya and Chan suggested that kinetic two-state cooperativity can be characterized by larger $T_{\min}/T_f$ ratios, where $T_{\min}$ is the temperature of minimum folding time and T_f the folding temperature [KC00b]. We will show in chapter 4 that $(AAQAA)_3$ indeed shows strong features of two-state cooperativity from accurate calculations of its density of states.

$V_5{}^DPGV_5$ shows a much longer folding time scale: $19\,000\,\tau$ at $k_BT = \mathcal{E}$ and up to $140\,000\,\tau$ at $k_BT = 0.8\,\mathcal{E}$. The increasing folding time upon lowering temperature is evident: even though weakened thermal fluctuations stabilize the folded ensemble, an unfolded chain is more likely to experience kinetic traps due to the ruggedness of the free-energy landscape. This forms the basis of the statistical mechanics of glasses in proteins and peptides [SO94, BOSW95]. The fact that the hairpin seems to be more sensitive to a decrease in temperature illustrates the importance of thermal fluctuations upon folding: entropic contributions are predominant in the hydrophobic collapse of the Val residues. The turn region, on the other hand, is sterically constrained.

3.3 Time-scale mapping

In light of the results obtained in the last section, we can now evaluate the speedup factor f_t that links the coarse-grained simulations with the "true" dynamics (e.g., as measured experimentally). Despite the existence of experimental studies on $(AAQAA)_3$ [SYSB91, SDS94], there is little data on the kinetics of folding. We will therefore estimate the average folding time of this peptide from similar α-helix forming sequences with comparable chain length. Experimental studies have shown that such peptides fold in $100 - 500$ ns at room temperature [WZG+04, HGZ+02, EMT+98, XPG06]. The approximate evaluation of the experimental time scale will only allow us to roughly evaluate the speedup factor, i.e., characterize its order of magnitude. Assuming coarse-grained simulations at $k_BT \approx \mathcal{E}$ describes the room-temperature behavior of $(AAQAA)_3$, the speedup factor is evaluated by simply dividing experimental by coarse-grained folding times $f_t = t_{\text{real}}/t_{\text{cg}}$. Recall from subsection 2.1.3 on page 17 that the coarse-grained unit of time was evaluated as

$\tau \sim 0.1\,\mathrm{ps} = 100\,\mathrm{fs}$. This yields

$$f_\mathrm{t} = \frac{t_\mathrm{real}}{t_\mathrm{cg}} \sim \frac{500\,\mathrm{ns}}{5\,000 \times 100\,\mathrm{fs}} = 10^3. \tag{3.1}$$

Equation 3.1 provides an estimate of the speedup factor from folding kinetics of the α-helix alone. We can now compare this value with folding time scales of β-hairpin folding. While, again, no experimental kinetic study is available for the sequence $\mathrm{V_5{}^DPGV_5}$, it is generally believed that most β-hairpins fold at least 10 times slower than helices (e.g., [XPG06] and references therein). The same factor was reported from implicit-solvent atomistic simulations of $\mathrm{(AAQAA)_3}$ and $\mathrm{V_5{}^DPGV_5}$ [FAC00].[3] Although it is difficult to calibrate the coarse-grained temperature in order to reproduce the room-temperature behavior of $\mathrm{V_5{}^DPGV_5}$, the set of average folding times in Table 3.1 suggest a ratio of folding times $t_\mathrm{hairpin}/t_\mathrm{helix}$ in the range $5 - 20$. While approximate, the results are consistent with a ratio of 10 between helix and hairpin folding times, suggesting an overall speedup factor $f_\mathrm{t} = 10^3$. How much can we trust these results? Zhou *et al.* showed that their coarse-grained model folded $\mathrm{V_5{}^DPGV_5}$ *faster* than a 15-residue polyalanine [ZTIV07]. This could represent a genuine characteristic of their model, but might also reflect an unfortunate choice in the folding threshold. There is no reason to believe that the threshold selected here is any more accurate than theirs, and its effect on the overall kinetics is unclear.

[3] Even though the absolute time scales of folding were too fast due to the lack of solvent friction, the ratio between helix and hairpin yielded a value of 10.

4 Protein Folding Cooperativity from a Microcanonical Perspective

Two-state protein folding cooperativity, defined by a depletion of states lying energetically between folded and unfolded conformations, is unambiguously identified using accurate density of states calculations. The link between thermodynamics and structure highlights the interplay between secondary structure formation and the concomitant loss of non-native tertiary contacts in helical peptides.

4.1 Protein folding cooperativity

Two-state protein folding is defined by a single free-energy barrier between folded and unfolded conformations at the transition temperature T_c, whereas downhill folders do not exhibit folding barriers—they rather show a significant population of intermediates [DŠK98]. The determination of this property conveys important information on both the thermodynamics as well as the kinetic pathways of proteins [BOSW95, Jac98]. Folding cooperativity is characterized by the nature of the underlying finite-size transition: two-state proteins correspond to a finite-size first-order transition (with an associated free energy barrier and nonzero latent heat), whereas downhill folders exhibit a continuous transition (i.e., no barrier) [Pri79, Pri82, CBD95]. The existence of such a free energy barrier implies a depletion of the population of intermediate conformations.[1] This characterization, based on the energetics of the system, is often difficult to extract—especially experimentally. Fortunately, various proxies have been developed to characterize protein folding cooperativity [Jac98]:

- The characterization of two-state folding from sharp transitions in order parameters (e.g., circular dichroism signal at 222 nm as a measure of helicity [HB94]). This, however, is seldom reliable because both two-state and downhill folders may display strong sigmoidal features [CBD95].
- A widely used test for a two-state transition is a calorimetric criterion which probes features in the canonical specific heat curve [Pti95, KC00b]. The calorimetric criterion attempts to extract the nature of the thermodynamic transition from the "sharpness" of the canonical specific heat curve (see Technical point 4.1 for a derivation of

[1] More accurately, it gives rise to a region where the number of states grows slower than $\mathrm{e}^{E/k_\mathrm{B}T_\mathrm{c}}$, where T_c is the transition temperature.

the calorimetric criterion). However, this criterion does not hold as a *sufficient* condition to identify two-state transitions [ZHK99] and does not offer a clear distinction between weakly two-state and downhill folders.

- So-called "chevron plots" characterize the relationship between folding/unfolding *rates* and concentration of chemical denaturant. Two-state folders feature two linear arms (thus the term "chevron") that intersect when folded and unfolded conformations are equally probable. The folding region (located between no denaturant—native conditions—and chevron mid-point) will show a *negative* slope: the more denaturant is added, the slower proteins fold. Conversely, the unfolding region (denaturant higher than the chevron mid-point) will show a *positive* slope: the more denaturant is added, the faster proteins *unfold*. Downhill folding proteins exhibit additional "rollovers" in the folding region which correspond to a slow-down in folding time in near-native conditions [FP02]. Kaya and Chan showed that such rollovers are due to slow, kinetically trapped intermediates—representative of downhill (i.e., non-two-state) folders [KC03]. This method thus probes the *kinetics* rather than the thermodynamics of protein folding.

- Förster resonance energy transfer (FRET) measurements allow to monitor the distribution of end-to-end distances in proteins [Lak99]. By studying the change in this distribution upon denaturation (e.g., increase in chemical denaturant), one may distinguish two-state from downhill folders: two-state distributions will show a clear dip between folded and unfolded distribution whereas downhill folders will show a shift in populations from folded to intermediates to unfolded [SE08]. FRET has shown extremely valuable in characterizing protein folding cooperativity (i.e., distinguishing between two-state and downhill folding), provided the relaxation rate between native and denatured ensembles is smaller than the single-molecule detection rate of the apparatus [HSS$^+$07].

From a simulation point of view, energy distributions can, in principle,[2] be sampled. The following work is based on an accurate determination of the density of states to unambiguously characterize the nature of the finite-size transition.

4.2 Microcanonical analysis

It is possible to determine the density of states in a standard canonical computer simulation at temperature T^* of interest: sample the probability density $p(E)$ of finding an energy E. The density of states $\Omega(E)$ is then proportional to $p(E)\,\mathrm{e}^{E/k_\mathrm{B}T^*}$, and hence the entropy is (up to a constant) given by $S(E) = k_\mathrm{B} \ln \Omega(E) = \text{const.} + k_\mathrm{B} \ln p(E) + E/T^*$. One may proceed to analyze the system *microcanonically*, i.e., to study the thermodynamics of $S(E)$, in the neighborhood of $\langle E \rangle_{T^*}$. The advantage is that we essentially directly analyze

[2] If one waits long enough.

Technical Point 4.1 Derivation of the calorimetric ratio

This derivation follows Ptitsyn [Pti95], and Kaya and Chan [KC00b]. The calorimetric criterion consists of assuming two-state folding: the entire protein folds as a whole. Energetically, this criterion consists in assuming that the van 't Hoff enthalpy ΔH_{vH}, which corresponds to the enthalpy of a *cooperative unit*, is equal to the enthalpy measured by calorimetry, ΔH_{cal}. A ratio $\Delta H_{\mathrm{vH}}/\Delta H_{\mathrm{cal}} \approx 1$ quantifies this two-state folding assumption. These two quantities can simply be measured from calorimetric experiments or simulations.

Standard calorimetric measurements give access to the excess enthalpy from heat capacity scans

$$\langle \Delta H \rangle_T = \langle H \rangle_T - H_{\mathrm{N}} \ , \tag{4.1}$$

where H is the enthalpy of the excess system, H_{N} is the enthalpy of the native state, and $\langle \cdot \rangle_T$ denotes a canonical average at temperature T. It is assumed that the native state is made of a single conformation with a unique temperature-independent enthalpy H_{N}. The corresponding specific heat yields

$$C_P = \frac{\partial \langle \Delta H \rangle}{\partial T} = \frac{\langle H^2 \rangle_T - \langle H \rangle_T^2}{k_{\mathrm{B}} T^2} \ . \tag{4.2}$$

The calorimetric enthalpy ΔH_{cal} corresponds to the excess enthalpy at a sufficiently high temperature T_∞ such that heat denaturation is complete: $\Delta H_{\mathrm{cal}} = \langle \Delta H \rangle_{T_\infty}$.

The van 't Hoff enthalpy can be obtained from the van 't Hoff equation [MS97] associated with an equilibrium constant K

$$\Delta H_{\mathrm{vH}} = k_{\mathrm{B}} T^2 \frac{\mathrm{d} \ln K}{\mathrm{d}T} = k_{\mathrm{B}} T^2 \frac{1}{\theta(1-\theta)} \frac{\mathrm{d}\theta}{\mathrm{d}T} \ , \tag{4.3}$$

where $\theta(T)$ is a suitable, normalized order parameter that describes the evolution of the system: $K = [\text{unfolded}]/[\text{folded}] = \theta/(1-\theta)$. The van 't Hoff enthalpy is commonly evaluated at the mid-point temperature T_{mid} of the parameter θ, such that $\theta(T_{\mathrm{mid}}) = 1/2$ and

$$\Delta H_{\mathrm{vH}} = 4 k_{\mathrm{B}} T_{\mathrm{mid}}^2 \left. \frac{\mathrm{d}\theta}{\mathrm{d}T} \right|_{T=T_{\mathrm{mid}}} \ . \tag{4.4}$$

The temperature-derivative of θ can easily be computed by assuming that $\theta(T) = \langle \Delta H \rangle_T / \Delta H_{\mathrm{vH}}$ is a suitable order parameter. The derivative yields: $\mathrm{d}\theta/\mathrm{d}T|_{T=T_{\mathrm{mid}}} = C_P(T_{\mathrm{midpoint}})/\Delta H_{\mathrm{vH}}$. Inserting this expression into Equation 4.4, we readily obtain

$$\Delta H_{\mathrm{vH}} = 2 T_{\mathrm{mid}}^2 \sqrt{k_{\mathrm{B}} C_P(T_{\mathrm{mid}})} \ . \tag{4.5}$$

Dividing Equation 4.5 by the calorimetric enthalpy ΔH_{cal} and setting $T_{\mathrm{mid}} = T_{\max}$ (specific heat peak) yields the calorimetric ratio $\kappa_2 = 2T_{\max}^2 \sqrt{k_{\mathrm{B}} C_P(T_{\max})}/\Delta H_{\mathrm{cal}}$ described in Kaya and Chan [KC00b]. Both $C_P(T_{\max})$ and ΔH_{cal} can be extracted from the canonical specific heat curve: the former corresponds to the peak height and the latter to the area under the curve for which $C_P(T)$ is strongly varying (see [KC00b] for more details).

the *probability density* $p(E)$ rather than merely looking at its lowest *moments*, such as the specific heat. Such a microcanonical analysis has been applied to a wide variety of problems, e.g., spin systems [BK92, Gro01, Hül94, Des97, Jan98, HP02, PB05], nuclear fragmentation [Gro93, KR87], colloids [FMMSV10], gravitating systems [KKK09, PT06], off-lattice homo- and heteropolymer models [TPB09, CLLL07], and protein folding [HS94, SR03, JBJ06, JBJ08, HRL08, KKS09, BBD10]. Two remarks are worthwhile:

- If the transition is characterized by a substantial barrier, standard canonical sampling suffers from the usual getting-stuck-problem: during a simulation the system might not sufficiently many times cross the barrier to equilibrate the two coexisting ensembles. This, of course, is true and needs to be avoided irrespective of whether one aims at a canonical or microcanonical analysis. Many ways around this problem have been proposed, e.g., multicanonical [BN91] or Wang-Landau [WL01] sampling. Here we use the Weighted Histogram Analysis Method (WHAM) (see Appendix A and [FS89, KRB$^+$92, BS09]), a minimum variance estimator of $\Omega(E)$ that combines overlapping energy histograms sampled during different canonical simulations, coupled to a parallel tempering scheme [SW86].

- Accurately sampling the whole distribution $p(E)$ over some range of interest requires better statistics than merely sampling its lowest moments: there's a price for higher quality data. But then, a microcanonical analysis taps into this quality, while a canonical analysis of the much longer simulation run would not significantly improve the observables (see Figure 4.1). Recall that the canonical partition function $Z(T) = \int \mathrm{d}E\ \Omega(E)\mathrm{e}^{-E/k_\mathrm{B}T}$ is the *Laplace transform* of the density of states $\Omega(E)$, an operation well-known to be (*i*) strongly smoothing and thus (*ii*) hard to invert (an insightful illustration of this issue is presented in Ref. [Cha00]).

From a thermodynamic point of view, a two-state transition is characterized by two coexisting regions [CBD95]. This corresponds to the finite-size analog of a first-order phase transition. While it does not qualify as a genuine phase transition in the common terminology (because the free energy is analytical for finite systems), its finite-size equivalent can be unambiguously characterized by monitoring the entropy $S(E) = k_\mathrm{B} \ln \Omega(E)$. Such a microcanonical analysis, where the energy E is a control parameter, has shown to be more informative around finite-size first-order transitions compared to its canonical counterpart [Gro01, Hül94] (the link between microcanonical and canonical descriptions is presented briefly in Technical point 4.2). In the phase-coexistence region, the entropy will exhibit a convex intruder due to the depletion of intermediate states. This can best be observed by defining the quantity $\Delta S(E) = \mathcal{H}(E) - S(E)$, where $\mathcal{H}(E)$ corresponds to the (double-)tangent to $S(E)$ in the transition region [JBJ06, JBJ08, Des97]. In a finite system, the existence of a barrier in $\Delta S(E)$ will imply a non-zero microcanonical latent heat ΔQ, defined by the interval over which $S(E)$ departs from its convex hull, and in turn leads to a "backbending" effect (akin to a van-der-Waals loop) in the inverse microcanonical

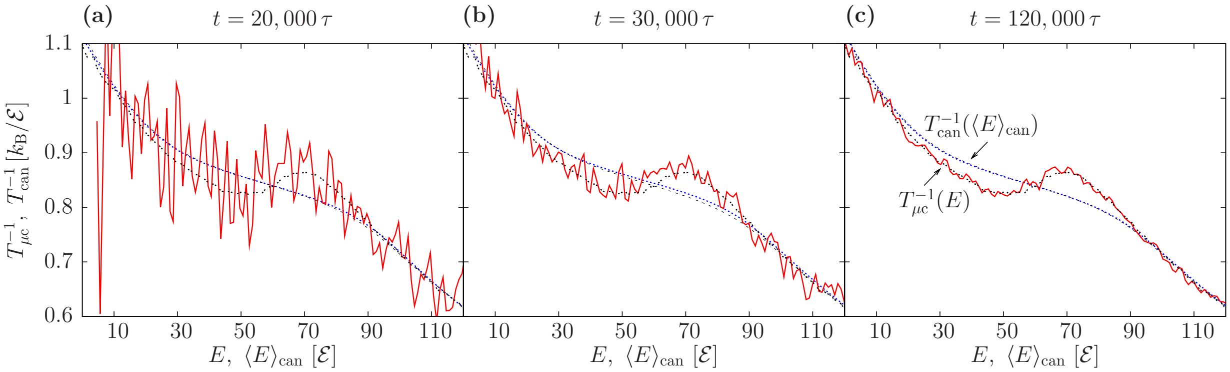


Figure 4.1: Convergence of microcanonical and canonical descriptions. Inverse temperature as a function of energy from canonical ($T_{\mathrm{can}}^{-1}(\langle E_{\mathrm{can}} \rangle)$, blue) and microcanonical ($T_{\mu c}^{-1} = \partial S/\partial E$, red) descriptions, where $\langle E_{\mathrm{can}} \rangle$ is the canonical average energy. Averages calculated over simulation times $t = 20\,000\,\tau$ (left), $t = 30\,000\,\tau$ (middle), and $t = 120\,000\,\tau$ (right). In each case, the black curves correspond to canonical and microcanonical averages calculated over a simulation time $t = 10^6\,\tau$, also reproduced in Figure 4.3 (a). While canonical averages converge rapidly (the canonical blue and black curves are indistinguishable at $t = 20\,000\,\tau$) to a monotonic—and rather uninteresting—curve, the slower convergence of the corresponding microcanonical curves finally provides a back-bending effect—a clear indicator of first-order transition (see main text).

temperature $T_{\mu c}^{-1}(E) = \partial S/\partial E$ (e.g., [Gro01, JBJ06]; experimental evidence of the back-bending effect was reported for a cluster of 147 sodium atoms [SKH+00]). A non-zero ΔQ demarcates a transition "region," whereas a downhill folder (continuous transition) will only exhibit a transition "point," where the concavity of $S(E)$ is minimal.

Figure 4.2 illustrates the relationship between different quantities in the microcanonical and canonical ensembles for a system of size N. Arrow heads denote how quantities can be determined from others. The first row links the density of states $\Omega(E)$ to the partition function $Z(T)$ via a Laplace transform. As mentioned above, the inverse operation (i.e., from $Z(T)$ to $\Omega(E)$) is difficult because of the strongly smoothing properties of the Laplace transform. Taking the logarithm yields the associated potentials: entropy $S_N(E)$ and free energy $F_N(T)$ (second row). Last, the existence of a thermodynamic limit ($N \to \infty$) implies that the intensive quantities $s_N(e) = S(E/N)/N$ and $f_N(T) = F(T)/N$ converge toward limiting functions $s_\infty(e)$ and $f_\infty(T)$, which are then linked via a Legendre transform.

In this chapter, we focus on the link between (*i*) the nature of the transition (i.e., two-state vs. downhill), (*ii*) secondary structure, and (*iii*) tertiary structure formation for several helical peptides using a high-resolution, implicit-solvent coarse-grained model. The results will be interpreted in terms of different frameworks of folding mechanisms, such as the molten globule model and simple polymer collapse models [DS95, Bal89].

Technical Point 4.2 Link between microcanonical and canonical descriptions

We will show here that canonical and microcanonical descriptions of a system are equivalent only in the thermodynamic limit.

Consider a finite-size system made of N particles. The canonical partition function $Z_N(\beta)$, where $\beta = 1/k_{\mathrm{B}}T$, is linked to its microcanonical counterpart $\Omega_N(E)$ (i.e., the density of states) by a Laplace transform

$$Z_N(\beta) = \int_E \mathrm{d}E \; \Omega_N(E)\mathrm{e}^{-\beta E}. \tag{4.6}$$

The microcanonical entropy $S_N(E) = k_{\mathrm{B}} \ln[\Omega_N(E)]$ and canonical free energy $F_N(\beta) = -\beta^{-1} \ln[Z_N(\beta)]$ can be linked together using Equation 4.6

$$\mathrm{e}^{-\beta N f_N(\beta)} = \int_E \mathrm{d}E \; \mathrm{e}^{-\beta[E - T S_N(E)]} \tag{4.7a}$$

$$= \int_e \mathrm{d}e \; \mathrm{e}^{-\beta N[e - T s_N(e)]}, \tag{4.7b}$$

where $e = E/N$, $f_N = F_N/N$, and $s_N = S_N/N$.

Equation 4.7b is equivalent to the integral

$$I_N = \int \mathrm{d}x \; \mathrm{e}^{N g(x)}, \tag{4.8}$$

for which

$$\lim_{N\to\infty} \frac{\ln I_N}{N} = \max_x g(x), \tag{4.9}$$

assuming $g(x)$ is continuous. In other words, the logarithm of the integral in Equation 4.8 can be approximated by the maximum value of the exponentiated function $g(x)$. Equation 4.9 is referred to as a “Laplace evaluation” or “method of steepest descent” [Has99].

Assuming the thermodynamic limit exists, f_N and s_N will converge to limiting functions f and s, respectively. The evaluation presented in the last paragraph yields

$$f(\beta) = \min_e \{e - T s(e)\}, \tag{4.10}$$

such that Equation 4.6 leads to a connection between the potentials f and s via a Legendre transform (see Figure 4.2).

Because proteins do not scale up to any thermodynamic limit, the connection presented in Equation 4.10 does not hold. Therefore inequivalences between microcanonical and canonical descriptions of these systems are not mere finite-size artifacts that will converge in the thermodynamic limit but true thermodynamic features.

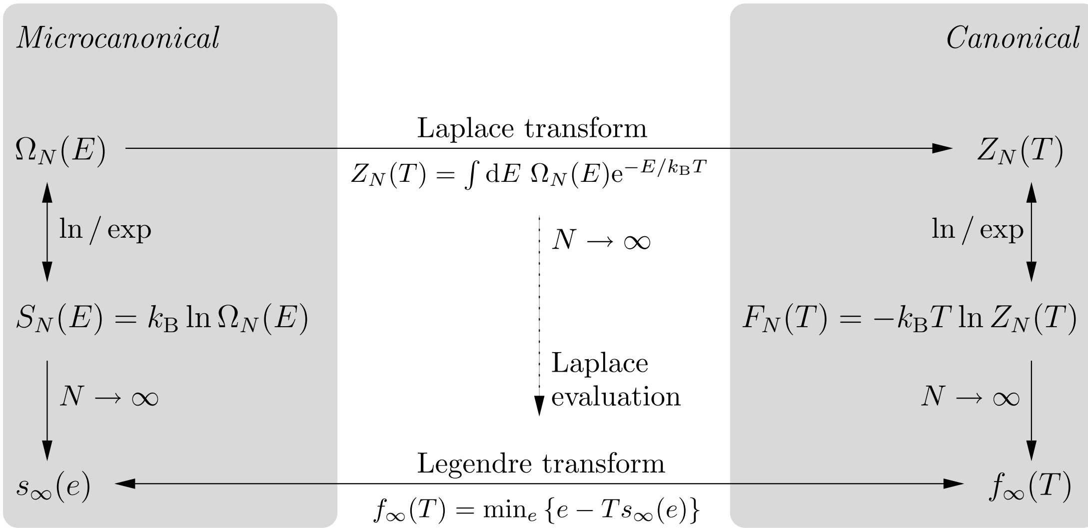


Figure 4.2: Thermodynamic quantities in the microcanonical (left) and canonical (right) ensembles. Arrow heads denote how quantities can be determined from others. For instance, the density of states determines the partition function but not the other way around (numerically, at least). For more details, see main text.

4.3 Simulation and analysis methods

In our simulations, the entropy $S(E) = k_\mathrm{B} \ln \Omega(E)$ is obtained from calculating the density of states $\Omega(E)$ by means of the Weighted Histogram Analysis Method (WHAM) (see, e.g., [KRB+92] and Appendix A). WHAM is a minimum variance estimator of the density of states which combines overlapping energy histograms sampled during different canonical simulations. These histograms will delimit the energy interval over which $\Omega(E)$ can be reconstructed. For each peptide studied (Table 4.1) a total of 36 temperatures were simulated, most of them were set close to the transition temperature of the system in order to improve the accuracy of the density of states around the transition region/point we focused on. Each simulation was run for a total time of up to $10^7\,\tau$ (depending on the peptide), where the potential energy was measured regularly (every $\approx 100\,\tau$). In order to enhance sampling, the abovementioned canonical simulations were coupled by a parallel tempering scheme where temperatures are swapped according to a Metropolis criterion [SW86]. Error bars on $\Delta S(E)$ were obtained by bootstrapping the raw energy histograms, thereby recreating distributions for each replica and calculating the corresponding density of states. This process was iterated ~ 50 times. The canonical energy $\langle E_\mathrm{can} \rangle$ as a function of temperature was also calculated from WHAM for all interpolating temperatures that were simulated.

All order parameters presented here (e.g., radius of gyration, helicity) were sampled canonically at the same frequency as E. They can be analyzed in the microcanonical ensemble by first using WHAM to calculate a two-dimensional density of states $\Omega(E, \mathcal{O})$ (where $\mathcal{O}$ denotes the order parameter of interest). The dependence of the order parameter on E is then determined by averaging $\mathcal{O}$ over the density of states at fixed E: $\langle \mathcal{O} \rangle_E \propto \sum_{\mathcal{O}} \Omega(E, \mathcal{O})\mathcal{O}$, applying suitable normalization. We have found for our simulations that simply averaging all values of $\mathcal{O}$ within a small energy interval, irrespective of the simulated temperature at which the data was sampled, gave virtually identical results compared to the proper average over $\Omega(E, \mathcal{O})$. This indicates that most values of $\mathcal{O}$ inside a given energy interval were sampled in a very narrow temperature interval, such that the Boltzmann factor associated with each data point trivially drops out of the average. The error of the mean was systematically calculated for each bin.

4.4 Secondary structure formation

We first examine the structural and energetic properties of the sequence $(\mathrm{AAQAA})_n$ with various chain lengths $n = 3$, 7, 10, 15. The $n = 3$ variant is known as a stable α-helix folder and has been studied both experimentally (e.g., [SYSB91, SDS94, ZJM+05]) and computationally (e.g., [FAC00, ZJM+05, CDL+09]). The $n = 7$ peptide has also been shown to fold into a helix [ZJM+05]. We find that all four peptides form a stable long helix in the lowest energy sector (see below), but are not aware of any structural study for $n = 10$, 15. Since we will soon show that the latter two fold differently from the shorter ones, an experimental confirmation of their ground state structure would be very useful.

For $(\mathrm{AAQAA})_3$ Figure 4.3 (a) shows a barrier in $\Delta S(E)$ as well as a backbending effect in the inverse microcanonical temperature $T_{\mu c}^{-1}(E)$, indicative of a first-order like transition. The figure also shows the inequivalence between microcanonical and canonical descriptions for finite-size systems by calculating the relationship between energy and temperature in both ensembles. The two vertical lines mark the transition region and the corresponding microcanonical latent heat ΔQ. In the region between $E = 40 - 80\,\mathcal{E}$ mostly-helical and mostly-coil conformations coexist, as can be seen from the sharp transitions in the helicity $\theta(E)$ (as determined by the STRIDE algorithm [FA95]) and the number of helices in the chain, $H(E)$. All these results point to a clear two-state folder.

Increasing the chain length from $n = 3$ to $n = 15$ (Figure 4.3 (b), (c), (d)) changes the nature of the transition significantly. While $n = 7$ still shows a (lower) barrier in $\Delta S(E)$ and a non-zero microcanonical latent heat ΔQ, $n = 10$ and $n = 15$ are downhill folders (no barrier in $\Delta S(E)$ and monotonic $T_{\mu c}^{-1}(E)$ curves). The transition region is replaced by a transition point for which the concavity of $S(E)$ is minimal and $\Delta Q = 0$. This process is associated with important structural changes around the transition region/point as seen in the number of helices $H(E)$: while the curve is monotonic for $n = 3$, it shows a *peak* with $H(E) > 1$ for bigger n. This suggests the existence of multiple helix nucleation sites upon folding (see representative conformations at the transition point for $n = 10$, 15 in Figure 4.3).

In order to further elucidate the structural features of these chains around the transition

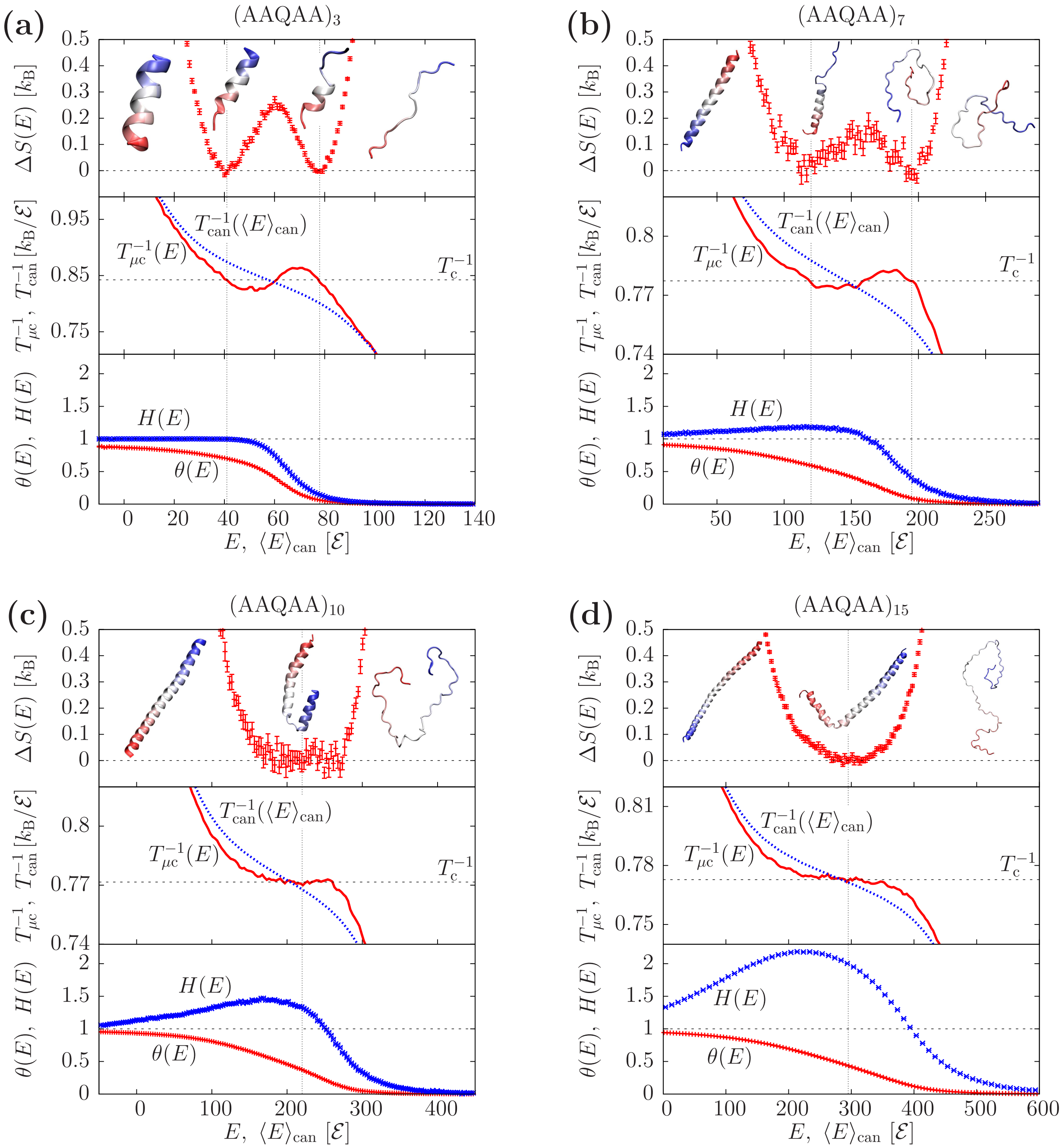


Figure 4.3: (AAQAA)$_n$ for different chain lengths: (a) $n = 3$, (b) $n = 7$, (c) $n = 10$, (d) $n = 15$. From top to bottom for each inset: $\Delta S(E)$, error bars reflect the variance of the data points ($1\,\sigma$ interval); inverse temperatures from canonical ($T_{\text{can}}^{-1}(\langle E_{\text{can}} \rangle)$, blue) and microcanonical ($T_{\mu c}^{-1} = \partial S / \partial E$, red) descriptions, where $\langle E_{\text{can}} \rangle$ is the canonical average energy; helicity $\theta(E)$ (red) and number of helices $H(E)$ (blue), both with the error of the mean. Vertical lines mark either the transition region ($n = 3$, 7) or the transition point ($n = 10$, 15). Representative conformations at different energies are shown.

Peptide	Sequence
helix $n = 3$	(AAQAA)$_3$
helix $n = 7$	(AAQAA)$_7$
helix $n = 10$	(AAQAA)$_{10}$
helix $n = 15$	(AAQAA)$_{15}$
bundle α3D	MGSWA EFKQR LAAIK TRLQA LGGSE AELAA FEKEI AAFES ELQAY KGKGN... PEVEA LRKEA AAIRD ELQAY RHN

Table 4.1: Amino acid sequences of the peptides studied in this chapter. The three helical regions of the native state (from NMR structure, PDB 2A3D) of the helix bundle α3D [WCB+99] are underlined (as predicted by STRIDE [FA95]).

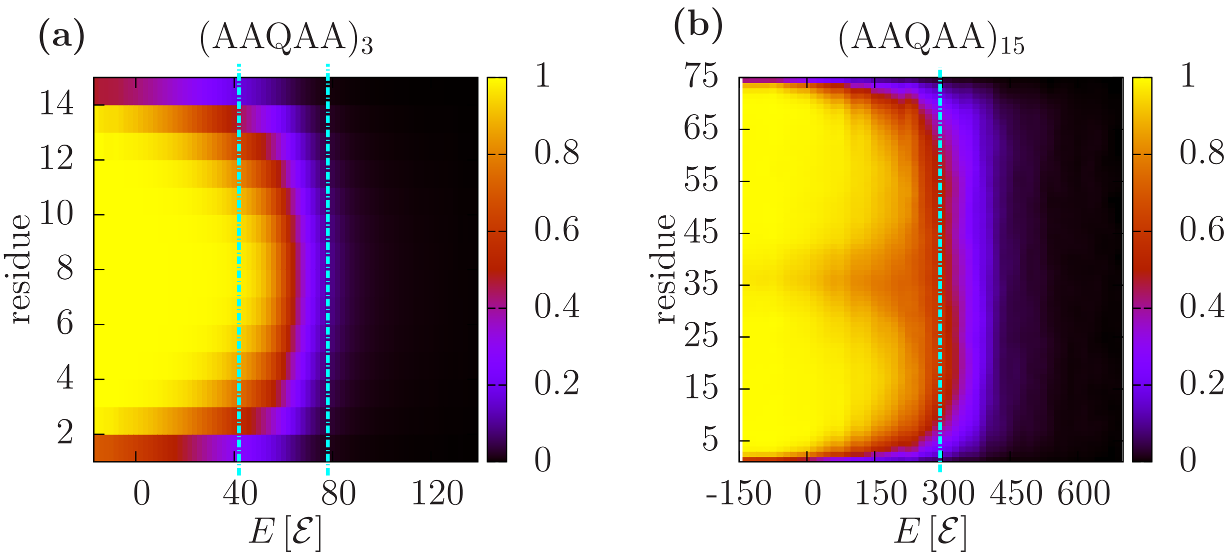


Figure 4.4: Amount of secondary structure as a function of energy and residue for (a) (AAQAA)$_3$ and (b) (AAQAA)$_{15}$. Vertical lines mark the transition region (a) and point (b), respectively.

region/point the fraction of secondary structure (i.e., helicity) was analyzed in dependence of both energy and residue index for helices $n = 3, 15$. While for $n = 3$ helix nucleation appears mostly around the center of the peptide and propagates symmetrically to the termini (Figure 4.4 (a)), $n = 15$ shows two distinct peaks at an energy E slightly below the transition point (Figure 4.4 (b)). The results suggest the formation of two individual helices placed symmetrically from the midpoint of the chain—around residue 35—which only join into one long helix significantly below the transition point. As will be discussed in section 4.6, these two helices divide the system into two distinct melting domains which fold non-cooperatively (i.e., folding one helix does not help folding the other) [Pri82, Pri89].

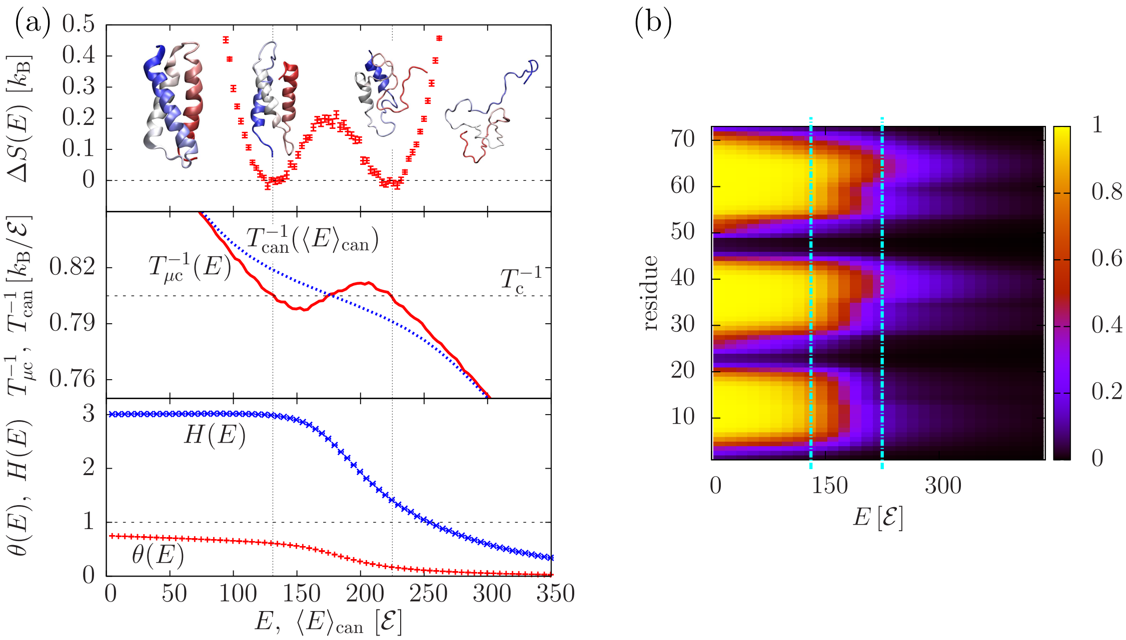


Figure 4.5: Three-helix bundle α3D. (a) From top to bottom: $\Delta S(E)$, error bars reflect the variance of the data points ($1\,\sigma$ interval); inverse temperatures from canonical ($T_{\mathrm{can}}^{-1}(\langle E_{\mathrm{can}}\rangle$, blue) and microcanonical ($T_{\mu c}^{-1} = \partial S/\partial E$, red) descriptions, where $\langle E_{\mathrm{can}}\rangle$ is the canonical average energy; helicity $\theta(E)$ (red) and number of helices $H(E)$ (blue), both with the error of the mean. Vertical lines delimit the transition region. Representative conformations at different energies are shown. (b) Amount of secondary structure as a function of energy and residue. Vertical lines mark the transition region.

To probe the behavior of simultaneous folding motifs within a chain, we performed a microcanonical analysis of the 73 residue *de novo* three-helix bundle α3D [WCB+99] (amino acid sequence given in Table 4.1). The CG model used here has been shown to fold α3D with the correct native structure, up to a root-mean-square-deviation of 4 Å from the NMR structure (Figure 2.6). While of similar length compared to $(\mathrm{AAQAA})_{15}$, it shows a *discontinuous* transition (Figure 4.5 (a)), and thus a nonzero microcanonical latent heat during folding. Inside the associated transition region the helicity increases sharply from 20% to about 65%, and the average number of helices also increases sharply—but monotonically—from 1.5 to 3. Unlike for the simple $n = 7, 10, 15$ helices, the transition region never samples more helix nucleation sites than the number of helices at lower energies. As can be seen from the representative conformations shown in Figure 4.5 (a), the ensemble of folded states ($E \approx 130\,\mathcal{E}$) consists of three partially formed helices in largely native chain topology; the coexisting unfolded ensemble ($E \approx 225\,\mathcal{E}$) consists of a compact structure containing transient helices. All these findings identify α3D as a two-state folder.

To better monitor the formation of individual helices, we measured the fraction of helicity as a function of energy and residue, see Figure 4.5 (b). Unlike $(\mathrm{AAQAA})_n$ (Figure 4.4), α3D shows strong features due to its more interesting primary sequence. The turn regions (dark color) delimiting the three helices (light color) are clearly visible at low energies and correspond well to the STRIDE prediction of the NMR structure, as shown in Table 4.1. Moreover, it is clear from this figure that secondary structure formation happens simultaneously (i.e., at the same energy) for all three helices, and most of the folding happens within the coexistence region (marked by the two vertical lines). The residues which form the native turn regions do not show any statistically significant signal of helix formation at any energy. Secondary structure has almost entirely formed close to the folded ensemble in the transition region (left-most vertical line)—in line with the representative conformations shown in Figure 4.5 (a).

4.5 Tertiary structure formation

A secondary structure analysis alone can only provide information on the local aspects of folding. Several studies have highlighted the role of an interplay between local and non-local interactions in protein folding cooperativity (see, e.g., [KC00a, GD09, BHT$^+$09, BBD10]). Here we first analyze the size and shape of the overall molecule by monitoring, respectively, the radius of gyration $R_\mathrm{g} = \sqrt{\lambda_x^2 + \lambda_y^2 + \lambda_z^2}$ and the normalized acylindricity $c = (\lambda_x^2 + \lambda_y^2)/2\lambda_z^2$ as a function of E, expressed in terms of the three eigenvalues of the gyration tensor[3] $\lambda_x^2 < \lambda_y^2 < \lambda_z^2$. The results for the single helices $n = 3$ and $n = 15$ and the three-helix bundle α3D are shown in Figure 4.6. $(\mathrm{AAQAA})_3$ shows sharp features in both order parameters within the transition region, indicating an overall structural compaction (in shape and size) of the chain as energy is lowered. Observe that c approaches 0.13 at high energy, which is close to the random walk or self-avoiding walk values, both close to $c \approx 0.15$ [Sol71, Sci96]. The longer helix $n = 15$ shows a non-monotonic behavior in both $R_\mathrm{g}(E)$ and $c(E)$: while the radius of gyration exhibits a minimum around $E = 400\,\mathcal{E}$, the normalized acylindricity displays a maximum. This indicates a structure that is most compact and spherical $100\,\mathcal{E}$ *above* the transition point. This dip in $R_\mathrm{g}(E)$ corresponds to a chain collapse into "maximally compact non-native states" [DS95] due to a non-specific compaction of the chain gradually restricted by steric clashes, at which point secondary structure becomes favorable. Upon lowering the energy, the radius of gyration increases

[3]The gyration tensor [TS85] is defined as:

$$G_{mn} = \frac{1}{N}\sum_{i=1}^{N}\left(r_m^{(i)} - \overline{r_m}\right)\left(r_n^{(i)} - \overline{r_n}\right),$$

where N is the number of particles, $r_m^{(i)}$ is the m^{th} coordinate of the position vector $\boldsymbol{r}^{(i)}$ of the i^{th} particle, and $\overline{r_m} = \sum_{i=1}^{N} r_m^{(i)}/N$ (here, we can write $\overline{r_m}$ as a center of geometry, rather than center of mass, because all masses in the present model are equal; see subsection 2.1.3 on page 17).

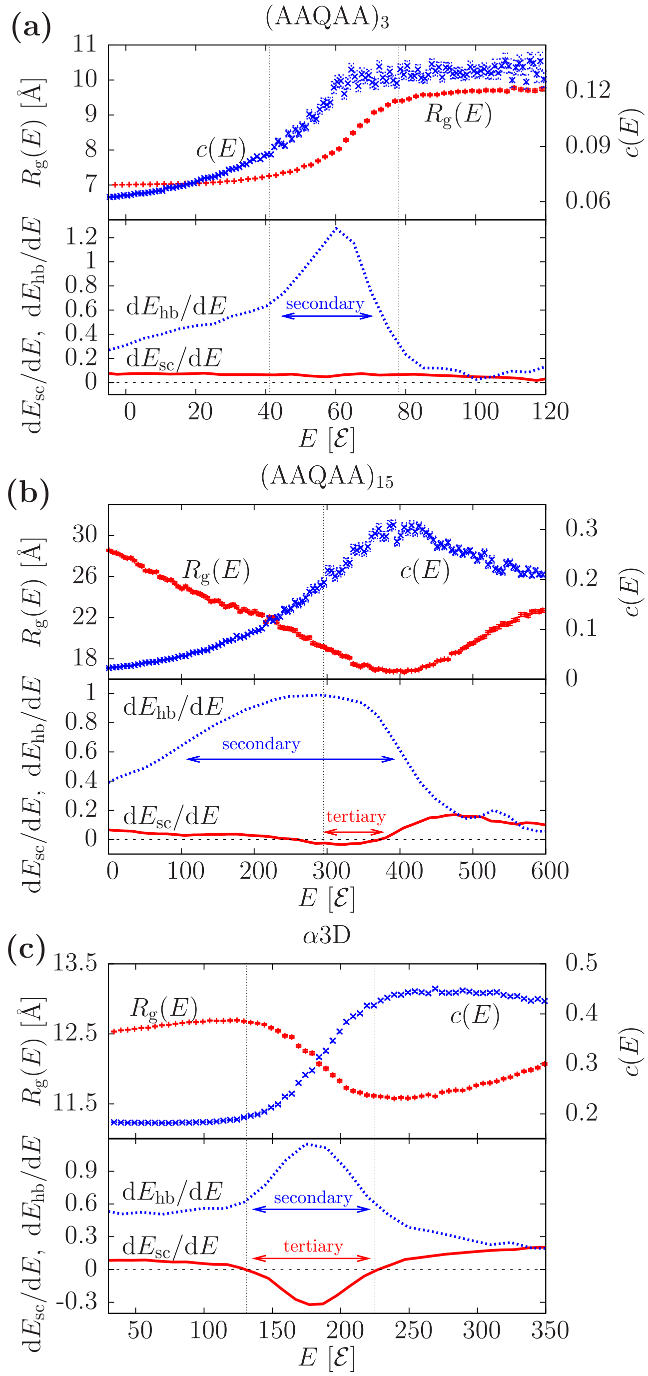


Figure 4.6: (top) Radius of gyration $R_g(E)$ (red) and normalized acylindricity parameter $c(E)$ (blue), both with the error of the mean, as well as (bottom) rates of hydrogen-bond and side-chain energies dE_{hb}/dE and dE_{sc}/dE, respectively, for (a) $(AAQAA)_3$, (b) $(AAQAA)_{15}$, and (c) α3D. Vertical lines mark either the transition region ($n = 3$, α3D) or the transition point ($n = 15$).

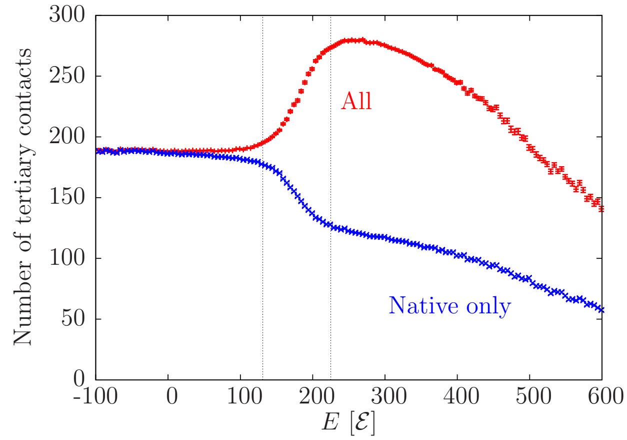


Figure 4.7: Number of tertiary contacts for α3D as a function of energy. The "All" curve (red) averages over all non-local pairs whereas the "Native only" curve (blue) only counts native pairs (see text for details) . Vertical lines mark the transition region.

and the acylindricity decreases, because the peptide elongates while folding from a compact globule into an α-helix. Results for the three helix bundle are similar: $R_{\mathrm{g}}(E)$ and $c(E)$ also show a minimum and a maximum, respectively, slightly above the transition region. This indicates a similar type of chain collapse mechanism. However, non-monotonic features appear also at the other end of the transition region ($E \approx 130\,\mathcal{E}$) where the radius of gyration shows a *maximum* and the acylindricity plateaus. The evolution of the two order parameters below the transition region is rather limited, suggesting that only minor conformational changes take place (i.e., the shape of the molecule stays steady while its size decreases slightly). In contrast, at high energy both $(\mathrm{AAQAA})_{15}$ and α3D are still far away from a random walk limit, as evidenced by the acylindricity being far away from 0.15.

Chain collapse in longer chains (such as $(\mathrm{AAQAA})_{15}$ and α3D) can readily be observed by monitoring tertiary contacts as a function of energy. Figure 4.7 shows the total number of non-local contacts (red curve) as well as the number of native contacts alone (blue curve). Tertiary contacts are defined here as pairs of residues that are more than five amino acids apart (this prevents chain connectivity artifacts) and within a 10 Å distance (these numbers are somewhat arbitrary, but their value does not affect the qualitative behavior of Figure 4.7). *Native* contacts correspond here to the set of abovementioned tertiary contacts sampled with a frequency higher than 1% from a set of 10 000 low-energy conformations ($E \leq 50\,\mathcal{E}$). While the two curves are virtually identical below the transition region (i.e., all contacts are native) and of similar trend above it, they behave very differently *inside* that interval. Although the number of native contacts monotonically increases as the energy is lowered (transition from globule to native-like structure), the total number of contacts

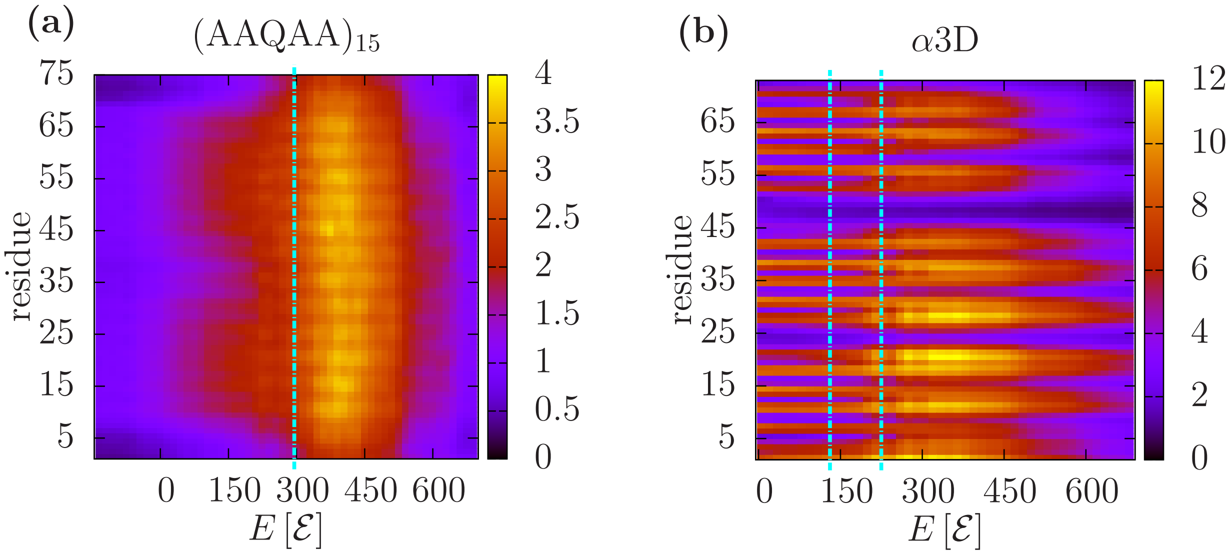


Figure 4.8: Number of tertiary contacts as a function of energy and residue for (a) $(AAQAA)_{15}$ and (b) α3D. The two plots have different depth ranges. Vertical lines mark the transition point (a) and region (b).

shows a peak above the transition region and sharply decreases inside it. To approach the native state, the peptide needs to break more contacts of non-native type than it gains contacts which are native.

The non-monotonicity of this curve, as well as the R_g data, invite a comparison with the thermodynamics of water: upon cooling liquid water expands below 4° C. Weak but isotropic van der Waals interactions are given up for strong but directional hydrogen bonds. This energy/entropy balance seems to occur in a very similar manner here, and essentially for the same reason. Weak van der Waals side-chain interactions (i.e., tertiary contacts) are replaced by hydrogen-bond interactions (i.e., secondary structure) at lower energies. This further confirms the concept of a chain collapse into maximally compact non-native states: upon lowering the energy (above the transition region) the system has accumulated a large number of non-native contacts due to a simple hydrophobicity-driven compaction mechanism. This idea was proposed early on as the "hydrophobic collapse model" or "molten globule model" [DS95, Bal89]. A similar effect was observed by Hills and Brooks using a Gō model, where out-of-register contacts had to unfold in order to reach the native state [HJBI08].

While a transient chain collapse upon cooling is present in both $(AAQAA)_{15}$ and α3D ($R_g(E)$ is non-monotonic, see Figure 4.6 (b) and (c)), its effect on tertiary structure formation will greatly depend on the amino acid sequence. Figure 4.8 shows the number of tertiary contacts of the two peptides as a function of energy and residue. The single helix $n = 15$ shows a uniformly small number of tertiary contacts in the low energy region (due to the linearity of the helix) and peaks *above* the transition point (which corresponds to the energy where $R_\mathrm{g}(E)$ is smallest). The tertiary contact distribution in the maximally compact non-native states is homogeneous along the chain (i.e., all residues have the same number of contacts). On the other hand, the number of tertiary contacts along the three-helix bundle (Figure 4.8 (b)) is highly structured, forming stripes as a function of residue that extend below the transition region. This follows directly from the amphipathic nature of the subhelices that constitute α3D: residues that form the native hydrophobic core of the bundle have a higher number of contacts. The presence of these stripes in the energetic region of collapsed structures ($E \approx 300\,\mathcal{E}$) is due to a strong selection between hydrophobic and polar amino acids during the hydrophobic collapse, burying hydrophobic groups inside the globule. The low number of tertiary contacts in the turn regions indicates that they remain on the surface of the maximally compact globule during chain collapse.

4.6 Interplay between secondary and tertiary structure

Two-state cooperativity has been characterized as a common signature of small proteins for which the transition of the cooperative domain corresponds to the whole molecule (i.e., the protein undergoes a transition as a whole) [Pri79]. While this framework applies well to the small helix $(\mathrm{AAQAA})_3$, it is difficult to predict its thermodynamic signature from other grounds: a description of the conventional helix-coil transition is not appropriate due to the small size of the system and the correspondingly important finite-size effects.

The thermodynamic signature of proteins can better be described for longer chains. Several arguments can be brought forward to explain the transition we observe for the longer helices $(\mathrm{AAQAA})_n$ for $n = 10$, 15:

- Most theoretical models of the helix-coil transition (e.g., [ZB59, LR61]) are based on the one-dimensional Ising model, which—being one-dimensional—shows no genuine phase transition but only a finite peak in the specific heat. The entropic gain of breaking a hydrogen-bond (i.e., forming two unaligned spins) outweighs the associated energetic cost for a sufficiently long chain.

- The structure of the maximally compact state right above the transition (Figure 4.8) indicates that there is no statistically significant competition between amino acids (i.e., all residues have the same number of tertiary contacts) and is therefore associated with a homopolymer-type of collapse, which is indeed barrierless [DS95, TBRP94].

- The denaturation of large proteins composed of several "melting" domains is not a two-state transition [Pri82, Pri89]. The presence of two helix nucleation sites around the transition point (Figure 4.4) indicates the existence of two such melting domains that fold *non*-cooperatively: folding one helix is not correlated with the formation of

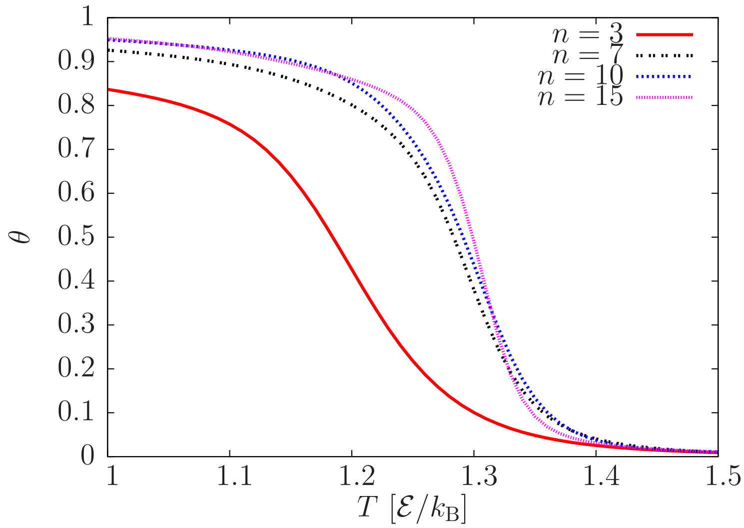


Figure 4.9: Canonical helicity $\theta(T)$ for $(\mathrm{AAQAA})_n$, $n \in \{3, 7, 10, 15\}$. The transition becomes sharper for longer chains.

the other. We have checked that there are no statistically significant helix-helix interactions between the two domains by calculating contact maps. These were averaged over the ensemble of conformations for which $50 \leq E \leq 150\,\mathcal{E}$ (data not shown).

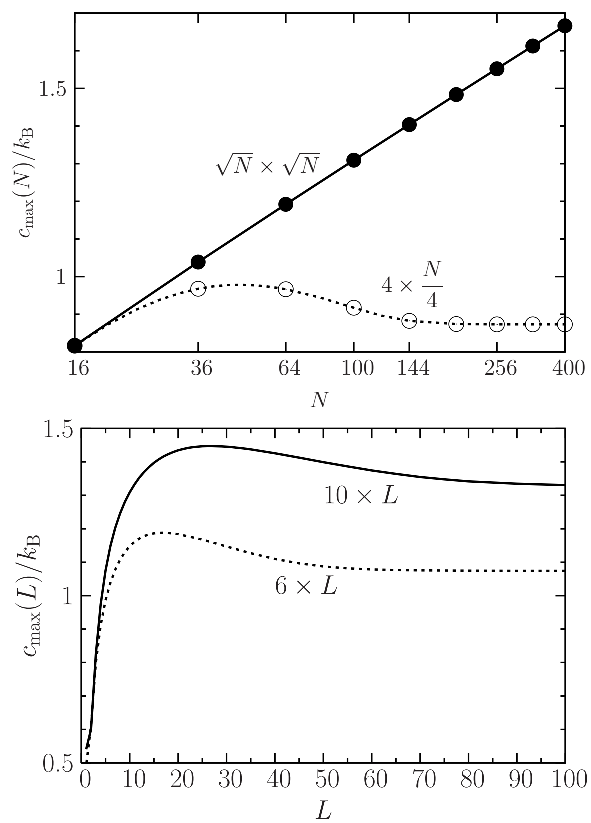
Common expectations is that bigger systems show sharper transition signals, and it might thus appear surprising that the transition of the $(\mathrm{AAQAA})_n$ sequence weakens for increasing n. However, one needs to bear two things in mind. First, size alone is not sufficient, dimensionality counts as well. In Technical Point 4.3 we show examples of quasi-one-dimensional systems for which transitions become weaker for bigger systems, because in the process of growing they become "more one-dimensional." When size is associated with cooperativity, one tends to think of globular (three-dimensional) systems, for which the size-cooperativity connection is true, but this is not the most general case. And second, the sharpness might depend on what observable one studies. The helicity θ as a function of temperature indeed varies more sharply for larger n (see, e.g., [ZDI59, ZB59, LR61]; Figure 4.9), making the response function $(\partial\theta/\partial T)_n$ peak more strongly for bigger n. While this steepening would suggest a stronger two-state nature, this goes against every other observable which suggests a downhill folder—including the calorimetric ratio (see below); observing response functions alone can thus be misleading.

The two-state signature of the helix bundle α3D can be understood from two different perspectives:

- While there are clearly three distinct folding motifs (i.e., three helices), the selective hydrophobicity (i.e., amphipathic sequence) between residues provides cooperativ-

Technical Point 4.3 Impact of dimensionality on cooperativity

Consider a two-dimensional Ising model with periodic boundary conditions in zero external field and $N = L_x \times L_y$ spins. Using the exact solution for the partition function by Onsager [Ons44], as later generalized by Kaufman [Kau49], one can determine the exact density of states [Bea96, WWWd]. We will illustrate the extent of cooperativity in such Ising systems as a function of system size and aspect ratio by calculating the specific heat peak, $c_{\max}$.



The figure on the right (top) shows $c_{\max}$ for a square system of size $\sqrt{N} \times \sqrt{N}$ as a function of N (solid line). One finds that $c_{\max}$ grows monotonically (in fact, logarithmically) with N and will ultimately diverge in the thermodynamic limit $N \to \infty$. Alternatively, a *rectangular* patch of N spins that is grown in only one direction shows a drastically different behavior: the dotted line in the figure (top) shows $c_{\max}$ for a system of size $4 \times N/4$. One can see that the curve first peaks before it *decreases*.

The bottom part of the figure provides a similar example: a rectangular patch of height 6 (dotted line) and 10 (solid line) and width L shows that $c_{\max}$ decreases after an initial peak. By considering $c_{\max}$ as a proxy for cooperativity, these results show that a system becoming increasingly one-dimensional can in fact *lose* cooperativity.

ity:[4] folding one helix helps the formation of the others.

- The barrier associated with a two-state transition is interpreted in the hydrophobic collapse model as the result of the cost of breaking hydrophobic contacts from a maximally compact state into the folded ensemble [DS95]. Experimental studies of this protein showed a fast folding rate of $1 - 5\,\mu$s and single-exponential kinetics [ZAM$^+$03], compatible with a two-state transition.

Cooperative secondary and tertiary structure formation had previously been proposed as a mechanism for two-state folding on the basis of lattice simulations [KC00a] and theoretical models [GD09]. As presented here, our results highlight also the interplay between secondary structure formation (see Figure 4.5 (b)) and the *loss* of non-native tertiary contacts (see Figure 4.7)—both occurring exactly within the coexistence region—as a possible mechanism for folding cooperativity.

This interplay between secondary and tertiary structure formation is also clearly illustrated from the energetic rates of hydrogen-bond ($\mathrm{d}E_{\mathrm{hb}}/\mathrm{d}E$) and side-chain ($\mathrm{d}E_{\mathrm{sc}}/\mathrm{d}E$) formation—assumed to be suitable proxies of secondary and tertiary contacts, respectively. Figure 4.6 shows that secondary structure formation is most pronounced within the coexistence region for $(\mathrm{AAQAA})_3$ and α3D while it occurs over a broad energy interval for the long helix $(\mathrm{AAQAA})_{15}$. The dip below zero of $\mathrm{d}E_{\mathrm{sc}}/\mathrm{d}E$ for the two longer chains is reminiscent of the sharp change in the total number of tertiary contacts shown in Figure 4.7 for α3D. It is a direct consequence of the reorganization of the maximally compact non-native states into the folded structure. The presence of such feature in both $(\mathrm{AAQAA})_{15}$ and α3D indicates that independent of its nature, the folding transition is driven by the loss of non-native tertiary contacts (i.e., the region where $\mathrm{d}E_{\mathrm{sc}}/\mathrm{d}E < 0$), which is reminiscent of the heteropolymer collapse model [DS95]. Secondary structure formation, on the other hand, shows very different signals: secondary structure formation in a downhill-folding peptide occurs over a much broader interval than for the loss of non-native tertiary contacts, whereas these two quantities are contained within the same narrow interval for a two-state peptide (for more details, see [BBD10]).

Compaction of the unfolded state upon temperature increase has been observed experimentally by Nettels *et al.* using single-molecule FRET [NMSK$^+$09]. While in our simulations the decrease in the radius of gyration (Figure 4.6) can be explained by a combination of the hydrophobic effect and the loss of helical structure, Nettels *et al.* showed similar behavior for an intrinsically disordered hydrophilic protein, where other mechanisms are likely to play a role.

The present work avoided any reference to free energy barriers so far. While the nature of the finite-size transition can unambiguously be characterized from the presence of a convex intruder in the entropy $S(E)$ [Gro01], the mere existence of a free energy barrier is not a strong criterion because, first, the definition of a free energy barrier is not unique in a finite-

[4] Here, we refer to cooperativity in the broader sense of interactions making transitions more sharply defined.

size system [Jan98, BK92] and, second, the height of the barrier depends on the reaction coordinate used. Chan [Cha00] therefore argued that the calorimetric criterion, which relates the van 't Hoff and calorimetric energies, is often more restrictive on protein models than the existence of such a free energy barrier. Still, the density of states calculations performed here correlate well with calorimetric ratios for $(\mathrm{AAQAA})_n$, $n = \{3, 7, 10, 15\}$, and α3D: $\delta = 0.78$, 0.76, 0.51, 0.52, and 0.78, respectively. These were determined by analyzing the canonical specific heat curve $C_{\mathrm{V}}(T)$ as in Kaya and Chan [KC00b] (κ_2 without baseline subtraction; see Technical Point 4.1). The value of 0.78 for the helix bundle agrees with an earlier theoretical calculation of the similar bundle α3C from Ghosh and Dill [GD09], who found $\delta = 0.72$.

5 Amyloid-β Oligomerization

Simulations of 32 Amyloid-β_{1-42} peptides stabilize large oligomers only limited in size by the simulation itself. This contradicts a recent coarse-grained study of similar resolution. Yet, the structure of a pentamer agrees remarkably well.

The field of protein aggregation—how proteins associate to form large-scale structures—has become an intense area of research due to its link to many degenerative diseases (e.g., Parkinson's [Dav08], Alzheimer's [WHK+97]).

Protein Amyloid-β_{1-42} (Aβ_{1-42}) has been repeatedly identified as the main constituent of amyloid plaques in the brains of Alzheimer's disease patients (e.g., [SLM+08]). Its sequence is shown in Table 5.1. It is formed after cleavage of the amyloid precursor protein (APP), a protein tied to the plasma membrane. The gene for APP is located on chromosome 21. Down syndrome patients—who carry three copies of chromosome 21—have been shown more susceptible to Alzheimer's [Man88]. Cleavage of APP yields peptides of different lengths due to the variety of secretases involved [TT07]. Aβ_{1-42} is the most fibrillogenic sequence out of the various isoforms created, e.g., Aβ_{1-40}. Moreover, various familial mutations have been identified and can trigger early onsets of the disease [SCLH99].

The *amyloid hypothesis* [HS02] states that the amyloids are responsible for the pathology through a cascade of events, ranging from protein aggregation to fibrillar structures.[1] Still, the causality invoked here is subject to much debate: it is not clear whether the deposits found in patients are the cause or the result of the disease. The sequential oligomerization of amyloids leads to different structures: monomers, oligomers, protofibrils, fibrils. Several recent studies suggest that the cytotoxic species of Aβ_{1-42} may be the oligomeric forms rather than the mature fibrils (e.g., [KHT+03]). The characterization of the effects of amyloid aggregation on neuronal cells is difficult to achieve experimentally because of the complexity of the system, as well as the intrinsic kinetics associated with any aggregation process. There is no X-ray or NMR structure of these oligomers because of their disorder and transient nature (i.e., they are intermediate forms of larger fibrillar structures). On the other hand, Petkova *et al.* presented a putative structural model for the isoform Aβ_{1-40} using solid-state NMR [PIB+02].

From a computational standpoint, the atomistic simulation of several Aβ_{1-42} proteins is untractable due to the large equilibration time involved. Recent studies of Aβ_{1-42} include either a single protein (e.g., [XSL+05, SYM+07]),[2] or the aggregation mechanisms of parts

[1] Concomitantly, other pieces of evidence point to the role of intra-cellular deposits of tau proteins.
[2] Even the complexity of a *single* protein is too large to perform exhaustive all-atom folding simulations.

$\overline{\texttt{D}}\texttt{A}\overline{\texttt{E}}\texttt{F}\underline{\texttt{R}}$ $\underline{\texttt{H}}\overline{\texttt{D}}\texttt{SGY}$ $\overline{\texttt{E}}\texttt{V}\underline{\texttt{H}}\underline{\texttt{H}}\texttt{Q}$ $\underline{\texttt{K}}\texttt{LVFF}$ $\texttt{A}\overline{\texttt{E}}\texttt{DVG}$ $\texttt{SN}\underline{\texttt{K}}\texttt{GA}$ $\texttt{IIGLM}$ $\texttt{VGGVV}$ $\texttt{IA}$

Table 5.1: Amino acid sequence of Aβ_{1-42}. The coding of amino acids obeys the following scheme: positively charged (overline); negatively charged (underline); hydrophobic (dark green); hydrophilic (black). Aβ_{1-40} exhibits the same sequence except for two fewer amino acids at the end.

of the sequence (e.g., [MN02, BBW$^+$06]). Here, in an attempt to characterize the large-scale oligomerization properties of Aβ_{1-42}, we present coarse-grained simulations of 32 full-length Aβ_{1-42} proteins.

5.1 Large-scale aggregation

The large-scale aggregation properties of Aβ_{1-42} were studied by performing a simulation of 32 peptides in a cubic box of side length 250 Å. To probe the temperature dependence—as well as facilitate sampling—48 replicas at different temperatures were coupled according to a parallel tempering scheme [SW86]. The 48 temperatures were uniformly distributed on a logarithmic scale between $T = 1.0\,\mathcal{E}/k_\mathrm{B}$ and $T = 2.2\,\mathcal{E}/k_\mathrm{B}$.[3] Each replica was run for $300\,000\,\tau$. Statistics were collected from 100 snapshots taken over the last $100\,000\,\tau$.

Figure 5.1 shows representative conformations at the lowest and highest temperatures simulated, i.e., $T = 1.0\,\mathcal{E}/k_\mathrm{B}$ and $T = 2.2\,\mathcal{E}/k_\mathrm{B}$, respectively. While high-temperature conformations mostly sample low-number aggregates (e.g., monomers, dimers), the low-temperature simulations show large aggregates, only limited in size by the number of peptides simulated. Figure 5.1 (a) displays the oligomerization of 30 monomers into a highly disordered structure as well as a dimer.

To characterize quantitatively oligomer-size distributions, inter-peptide hydrogen-bond patterns were analyzed from simulation snapshots: two monomers that formed inter-peptide hydrogen bonds[4] were marked as being part of the same oligomer. Oligomers were sequentially grown until no other peptide hydrogen-bonded to the aggregate. Finally, the probability of formation of an n-mer was calculated from the number of such oligomers

[3] The number of required replicas scales heavily with the complexity of the system: the energy of the system scales like system size $E \sim N$; the range of energies sampled at a given temperature is given by the root-mean-square energy fluctuations which scales like the square root of system size $\sqrt{N}$; therefore the relative size of the fluctuations compared to the energy *decreases* as $1/\sqrt{N}$ [NB99]. This explains why atomistic replica-exchange simulations of proteins use small temperature intervals, spanning only $2-5\,\mathrm{K}$.

[4] Hydrogen bonds were identified using the STRIDE algorithm [FA95]. The original implementation of the algorithm can be found at `http://webclu.bio.wzw.tum.de/stride/`. Note that the code does not handle periodic boundary conditions (i.e., hydrogen bonds between chains at two edges of the periodic box would not be considered). The analysis performed here was done using a modified version of the STRIDE package that incorporated this feature, written by the author of the present thesis. The implementation is straightforward and thus not detailed here.

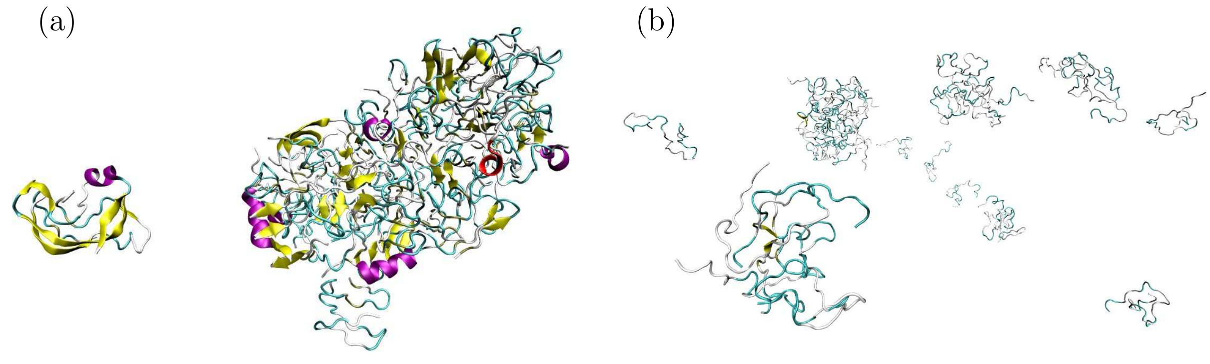


Figure 5.1: Conformations of a system of 32 Aβ_{1-42} sampled at (a) $T = 1.0\,\mathcal{E}/k_\mathrm{B}$ and (b) $T = 2.2\,\mathcal{E}/k_\mathrm{B}$.

sampled during the simulation normalized by the total number of oligomers within the ensemble of snapshots.

The distribution of oligomer sizes at various temperatures is presented in Figure 5.2. Figure 5.2 (a) shows a clear bimodal distribution: one peak around 30-mers and another one around dimers. Clearly the distribution is dominated by finite-size effects as the simulation only contains 32 peptides. Figure 5.2 (b), (c), and (d) show a progressive shift of the distribution towards lower aggregates when increasing temperature. Remarkably, none of these distributions show a non-zero peak.[5] Higher temperature distributions decay quickly as a function of oligomer size.[6]

In a recent study from Urbanc *et al.*, the same simulation setup used with a different coarse-grained model showed a rather narrow bimodal distribution that peaked around trimers and pentamers, with no significant population of large-number aggregates [UCY+04]. Corresponding simulations of Aβ_{1-40} peptides revealed a unimodal distribution that peaked at dimers. A previous experimental *in vitro* study of the oligomer size distribution of Aβ_{1-40} and Aβ_{1-42} also showed unimodal and bimodal distributions, respectively [BKL+03]. However, their data suggests that the average Aβ_{1-42} oligomer is significantly larger than for Aβ_{1-40}, as given by dynamic light scattering measurements as a function of hydrodynamic radius R_hyd. According to their results, the oligomer size distribution of Aβ_{1-42} peaks at a value of $R_{\mathrm{hyd},\,42}$ that is roughly five to ten times larger than corresponding measurements on Aβ_{1-40}. Assuming that the size of an oligomer (i.e., number of monomers) scales like R^3_hyd, one can estimate how much larger the Aβ_{1-42} oligomer must

[5] The small, intermediate peaks in Figure 5.2 (b) and (c) are not to be trusted.

[6] The data is not accurate enough to clearly identify the type of distribution (e.g., exponential).

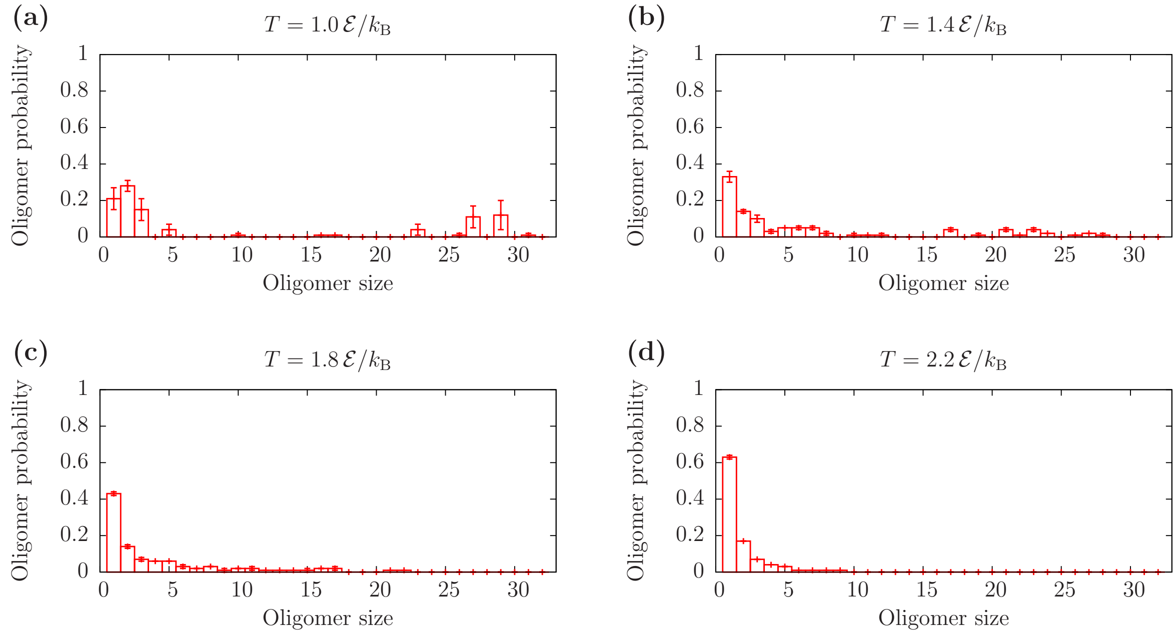


Figure 5.2: Oligomer distribution functions of a system of 32 Aβ_{1-42} at temperatures (a) $T = 1.0\,\mathcal{E}/k_\mathrm{B}$, (b) $T = 1.4\,\mathcal{E}/k_\mathrm{B}$, (c) $T = 1.8\,\mathcal{E}/k_\mathrm{B}$, and (d) $T = 2.2\,\mathcal{E}/k_\mathrm{B}$.

be. Denoting n_{40} and n_{42} the oligomer sizes of Aβ_{1-40} and Aβ_{1-42}, respectively, we find

$$\frac{R^3_{\mathrm{hyd},\,42}}{R^3_{\mathrm{hyd},\,40}} = \frac{n_{42}}{n_{40}}. \tag{5.1}$$

From their results, we estimate a lower boundary for the ratio of hydrodynamic radii $R_{\mathrm{hyd},\,42}/R_{\mathrm{hyd},\,40} \approx 7/2$. This would imply that the Aβ_{1-42} oligomers are ≈ 40 times larger (in size) than Aβ_{1-40} oligomers. This presents a serious discrepancy with Urbanc's simulation results which predict dimers and pentamers for Aβ_{1-40} and Aβ_{1-42}, respectively. The abovementioned experimental results clearly alude to much larger oligomers. Therefore, a simulation of 32 peptides would most likely be too small to probe the oligomer size distribution of Aβ_{1-42}—in line with the strong finite-size effects encountered in the present low-temperature simulations (Figure 5.2 (a)).

While the computational study of Urbanc *et al.* was performed at a single temperature, none of the distributions sampled from the present parallel tempering simulation matches their data, even qualitatively. This discrepancy may arise from two sources: insufficient sampling in any of the simulations and the accuracy of each force field:

- The complexity of the system studied here is associated with long correlation times (i.e., slow dynamics). In order to better sample phase space, we used a parallel tempering scheme over a large temperature interval. 100 snapshots were used to collect statistics.[7] On the other hand, the study of Urbanc *et al.* consisted of standard canonical simulations. Statistics were collected from eight simulations, each of which was analyzed at three different times.

- While both models use the same mapping (i.e., four beads per amino acid, including one for the side-chain) and contain explicit hydrogen bonds, amino acid specificity is more finely resolved in the model presented in this thesis (chapter 2), whereas only four types of amino acids are considered in Urbanc *et al.* Also, their model was parametrized to reproduce the conformational changes of aggregating α-helices into β-sheet structures [DBB+03].

It is unclear whether a larger simulation (i.e., more peptides) would ever yield a non-zero peak in the oligomer distribution that is not due to finite-size effects. The oligomerization process shown here describes a hydrophobicity-driven mechanism—the large aggregates stabilized at low temperatures break down upon heating. Experimental measurements indicate very different oligomer-size distributions for $A\beta_{1-40}$ and $A\beta_{1-42}$ and suggest that $A\beta_{1-42}$ oligomers may be much larger than 32 peptides [BKL+03].

5.2 Structure of the pentamer

To further elucidate the discrepancy between the model presented in chapter 2 and the one used in Urbanc *et al.* [UCY+04], a structural comparison of $A\beta_{1-42}$ pentamers was performed. These oligomers show high conformational variability—they do not stabilize a unique fold. Thus, any attempt to characterize the structure of such entities will be the result of averages over very different structures. Here, we characterized the average distance of each residue to the center of mass of $A\beta_{1-42}$ pentamers. Figure 5.3 shows a conformation of such a pentamer, sampled at $T = 1.0\,\mathcal{E}/k_\mathrm{B}$. While compact, the system lacks any structural order and is subject to strong conformational variability.[8]

Following a similar simulation setup as in section 5.1, a simulation box consisting of five $A\beta_{1-42}$ was studied. All simulation snapshots consisting of a pentamer were considered for analysis, performed at $T = 1.0\,\mathcal{E}/k_\mathrm{B}$. Figure 5.4 (left) displays the distance from center of mass of each residue, averaged over all five peptides as well as sampled conformations. Figure 5.4 (right) shows its equivalent from the study of Urbanc *et al.* (red curve). The agreement—up to a vertical shift—is remarkable. This shift indicates that the pentamer simulated in this work is more compact. It may be due to a temperature difference between the two simulations or a discrepancy between the two force-field parametrizations. A similar analysis at higher temperatures flattened the overall distribution rather than merely shifting it upwards. The positions of the dips (e.g., residues 15–20, 30–35) correlate highly with amino-acid hydrophobicity. Overall, the results can be explained, to a large extent,

[7] The number of extracted snapshots was kept low due to large auto-correlation times.

[8] Again, Figure 5.3 is only one sampled conformation out of many possible.

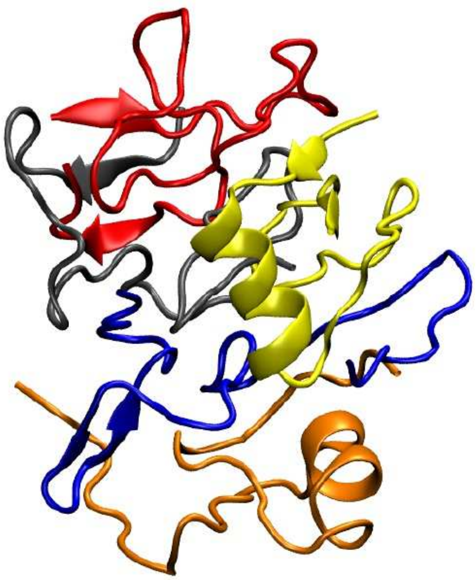


Figure 5.3: Sampled (but not necessarily representative) conformation of an Aβ_{1-42} pentamer at $T = 1.0\,\mathcal{E}/k_{\mathrm{B}}$.

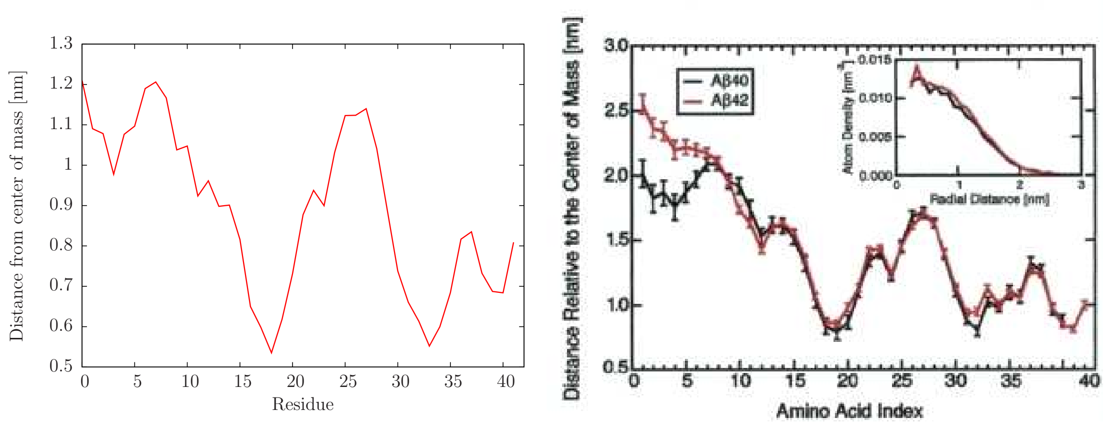


Figure 5.4: Average distance to the center of mass of the pentamer for each residue: (left) this work, conformations sampled at $T = 1.0\,\mathcal{E}/k_{\mathrm{B}}$; (right) red curve—simulation study from Urbanc *et al.* [UCY^{+}04]. Copyright (2004) National Academy of Sciences, USA.

by considering the hydrophobicity scale of amino acids alone. It might therefore not be surprising that the two models behave similarly.

6 Protein-Lipid Interactions

Many proteins function in, or close to, the cell membrane. In this chapter, we cross-parametrize the peptide model presented in chapter 2 with a high-resolution, implicit-solvent coarse-grained model for lipids. Coarse-grained potentials are tuned to reproduce atomistic potential of mean force curves for the insertion of single amino acid side chains into a DOPC bilayer. The validity of the protein-lipid model is probed by various simulations of synthetic transmembrane proteins.

Cell membranes form the boundaries between the inside and outside of the cell and, in eukaryotic cells, form the structural basis of many important cellular organelles (such as nucleus, endoplasmic reticulum, Golgi apparatus, and mitochondria). They also control the selective permeation of ions and organic molecules via channels and pores [AJL+02]. They mainly consist of (*i*) a lipid bilayer and (*ii*) embedded proteins:[1]

- The lipid bilayer stabilizes membrane shape and structure; it is made of two layers of lipid molecules. Lipids form a broad group of naturally occurring amphiphilic[2] molecules (Figure 6.1 (a)). Three classes of lipids are present in the membranes of eukaryotic cells: phospholipids (the most abundant species), glycolipids, and cholesterol. They differ both in terms of structure and function [BTS10]. The hydrophobic effect (e.g., [Tan80]; see also section 1.1 on page 1) drives the self-assembly of these molecules into various supramolecular structures (e.g., vesicle, bilayer sheet) depending on concentration and chemical environment [Isr92]. In water, a lipid bilayer will expose its polar head groups to the solvent and bury its apolar hydrocarbon tails (i.e., fatty acids). While thin ($\approx 5\,\mathrm{nm}$), this partitioning prevents the diffusion of ions and polar molecules through the bilayer.[3]

- Integral (i.e., permanently attached to the membrane) and peripheral membrane proteins (Figure 6.1 (b)) amount to half of the membrane weight [FP02]; they provide many biological functions to the cell membrane, e.g., transporters, channels, receptors, enzymes [AJL+02]. The structure of only a small number of transmembrane proteins has been solved so far (using X-ray or NMR techniques), due mainly to

[1] Cell membranes also contain carbohydrates in the form of glycoproteins and glycolipids [BTS10].

[2] Chemical compound possessing spatially separated hydrophilic and hydrophobic moieties.

[3] In terms of electrostatics, the bilayer creates a large change in dielectric constant: while $\epsilon = 80$ in water, $\epsilon \approx 3$ in a hydrophobic environment. This generates a large free energy barrier for an ion to penetrate the membrane.

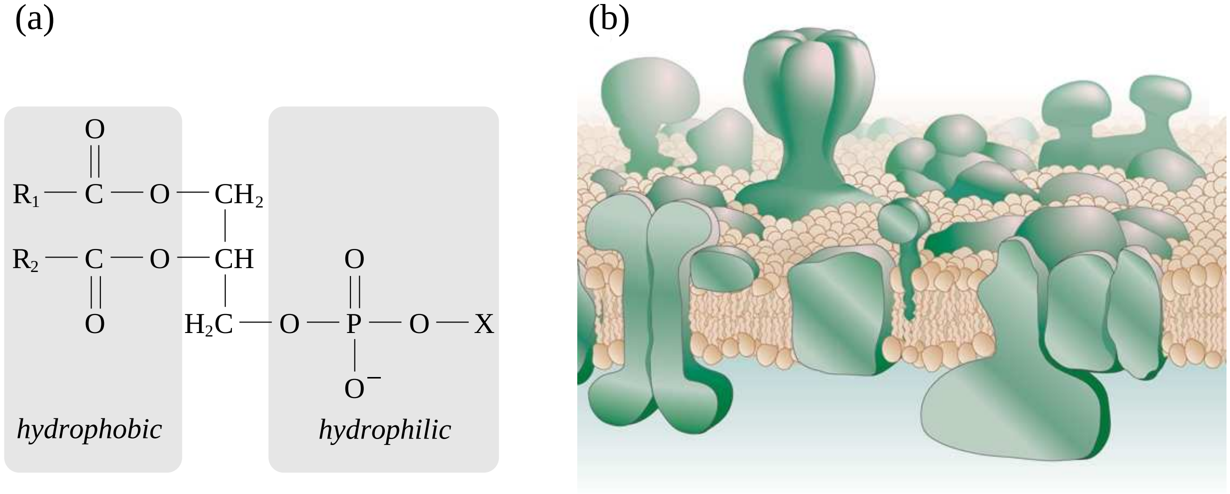


Figure 6.1: (a) Chemical representation of a phospholipid molecule. Fatty acid chains (denoted R_1 and R_2) form the hydrophobic region of the molecule. The hydrophilic region is composed of a phosphate group and a head group (denoted X; e.g., serine, choline, glycerol). An atomistic representation of a phosphatidylcholine lipid is displayed in Figure 6.2. (b) Cartoon representation of a cell membrane composed of lipid molecules (beige) crowded by many integral and peripheral proteins (green inclusions). Adapted by permission from Macmillan Publishers Ltd: *Nature* **438**, 578–580 (1 December 2005), copyright 2005 [Eng05].

(i) poor solubility in water[4] and (ii) crystallization difficulties caused by disordered associations [FP02].

Overall, the membrane behaves as a *two-dimensional fluid* of oriented lipids and proteins, where lateral diffusion is fast compared to the transition of a molecule from one leaflet to the other (i.e., "flip-flop") [SN72]. The fluidity of the membrane is largely controlled by fatty acid composition and cholesterol content. For a more detailed introduction to proteins and lipids in membranes, we refer the reader to standard biochemistry and molecular biology textbooks (e.g., [AJL$^+$02, GG08, BTS10]).

Many of the abovementioned biophysical processes involving the interaction of proteins with lipid membranes operate at time- and length-scales that are currently unattainable by atomistic computer simulations (limited to small systems in the $10\,\mathrm{ns} - 1\,\mu\mathrm{s}$ range). To cope with this difficulty, several lipid-protein coarse-grained models of various degrees of resolution have been developed and studied (e.g., [VSS05, BS06, ID08, MKP$^+$08, WBS09]).

[4] Integral proteins form a large hydrophobic surface that interacts favorably with the hydrocarbon tails of the bilayer lipids.

One of them, the MARTINI force field [MRY+07, MKP+08], has been applied to a wide variety of lipid-protein systems, such as pore formation by antimicrobial peptides [RSGM10], helix rearrangements of the ATP synthase subunit C [SRM09], association behavior of glycophorin A [SM10], and lateral organization of transmembrane helices in heterogeneous model membranes [SdJH+11]. It maps on average three to four heavy atoms into one coarse-grained bead, and parametrizes each bead according to thermodynamic data (in particular, oil/water partitioning coefficients). By dividing molecules into sets of chemical building blocks, it provides a generic force field that (ideally) does not require any reparametrization each time a new system is studied. Other properties of the MARTINI force field include:

- Explicit representation of the aqueous solvent by grouping four water molecules into one Lennard-Jones bead.
- Explicit short-range electrostatics, beyond which a uniform dielectric constant $\epsilon = 15$ is applied (i.e., "reaction field"). This value is somewhat arbitrary as it neither represents the aqueous environment ($\epsilon = 80$) nor the hydrophobic regions ($\epsilon \approx 3$). An extension of the force field involving Drude-like (i.e., polarizable) water beads was recently proposed [YSSM10].
- While protein side-chain energetics and packing are represented with high accuracy, the model *constrains* secondary structure via the dihedral potentials linking the C_α beads of the backbone.[5]

In the following, we present an alternative protein-lipid coarse-grained model which offers different features, including (*i*) implicit water—allowing for significant speedup—in such a way that important solvent effects are treated implicitly, (*ii*) no explicit electrostatics,[6] and (*iii*) *unconstrained* secondary structure formation of the protein model (chapter 2).

The implicit-solvent coarse-grained lipid model used in this work, developed by Wang and Deserno [WD10b, WD10a], provides a similar resolution as the MARTINI force field (i.e., three to four heavy atoms per bead), which fits well with the peptide model's resolution. Figure 6.2 (b) represents a coarse-grained 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC) lipid, as described in [WD10b]. It is made of 16 beads and 8 bead types: **CH** Choline; **PH** Phosphate; **GL** Glycerol; **E1** and **E2** Ester groups; **AS** Saturated

[5] There is little hope that a one-bead-per-backbone model will ever encode enough information to accurately reproduce local conformations (in particular, Ramachandran plots; see chapter 2).

[6] While essential, electrostatics is both computationally expensive and difficult to coarse-grain. We point out that (*i*) simply putting charges on the beads and working out $\frac{1}{r}$ interactions is not the right thing to do in the presence of dielectric discontinuities, (*ii*) the coarse-grained lipid model used here targets neutral lipids—the contribution from partial charges being hopefully absorbed in the coarse-grained potentials, and (*iii*) the interaction between charged amino acids is, *to some extent*, reproduced in the Miyazawa-Jernigan matrix [MJ96] used for the side-chain–side-chain interactions. That being said, we warn the reader against the use of this model for phenomena that heavily rely on electrostatics (e.g., electroporation [Tie04]).

alkyl group $-(CH_2–CH_2–CH_2)-$; **AD** Unsaturated alkyl group $-(CH_2{=}CH_2)-$; **AE** Hydrocarbon endgroup $-(CH_2–CH_3)$. It was systematically parametrized to reproduce the radial distribution functions of groups of atoms of POPC from atomistic simulations using iterative Boltzmann inversion [Sop96]. Because of the lack of solvent in the coarse-grained system, an additional effective cohesion was included to compensate for the lack of a confining pressure [WD10b]. On top of its ability to self-assemble a random dispersion of lipids into a bilayer, Wang and Deserno showed that their model can almost quantitatively reproduce many of the properties of a POPC bilayer, such as bending and stretching modulus, mass density profile, and orientational P_2 order parameter of intramolecular bonds. They showed that the construction of *other* types of lipids (e.g., DOPC,[7] DPPC[8])—starting from the same set of coarse-grained beads as POPC—provides reliable transferability in terms of structure and area per lipid [WD10a].[9]

6.1 Force-field cross-parametrization

In this section, we derive potentials of interaction for the cross-parametrization of the two force fields: the protein (see chapter 2 and [BD09]) and the lipid [WD10b, WD10a] models. The parametrization aims at reproducing the potential of mean force (PMF) curves for the insertion of individual amino acid side chains into a DOPC bilayer.[10] These PMFs provide additional spatial resolution compared to experimental hydrophobicity scales (e.g., [FP83, WW96]). Because of the planar structure of lipid bilayers, free energies as a function of the z coordinate (i.e., perpendicular to the bilayer plane), $F(z)$, incorporate much of the thermodynamic information required to describe the energetics of insertion of molecules in a membrane.[11] Moreover, the spatial ordering of lipid groups (e.g., alkyl chains, phosphate) in a bilayer provides a means to understanding the impact of each group onto $F(z)$.

The reference data is based on atomistic simulations performed by MacCallum *et al.* [MBT08], where amino acid side chains (i.e., starting at the β-carbon) were inserted in a 64-molecule DOPC bilayer using the OPLS all-atom force field [JMTR96, KFTRJ01]. Their study provided curves for the free energy of insertion of side chains for all standard amino acids except Gly, His, and Pro. While Gly and Pro were not calculated because of the chemistry of their side chain (the side chain of Gly holds no heavy atom; the side chain

[7] $1,2$-dioleoyl-*sn*-glycero-3-phosphocholine.

[8] $1,2$-dipalmitoyl-*sn*-glycero-3-phosphocholine.

[9] The difference between DPPC, POPC and DOPC lies in their lipid tails alone: (i) they comprise 0, 1, and 2 one-fold unsaturated lipid tails, respectively, and (ii) each lipid tail contains different numbers of carbon atoms: $16+16$ (DPPC), $16+18$ (POPC), and $18+18$ (DOPC).

[10] This parametrization focuses on reproducing the *energetics* of peptide-lipid interactions, as opposed to, say, the structure.

[11] This projection excludes the orientational dependence of the side chain. On the other hand, the parametrization of a one-bead-per-side-chain model using radial interactions would not allow for any angular dependence.

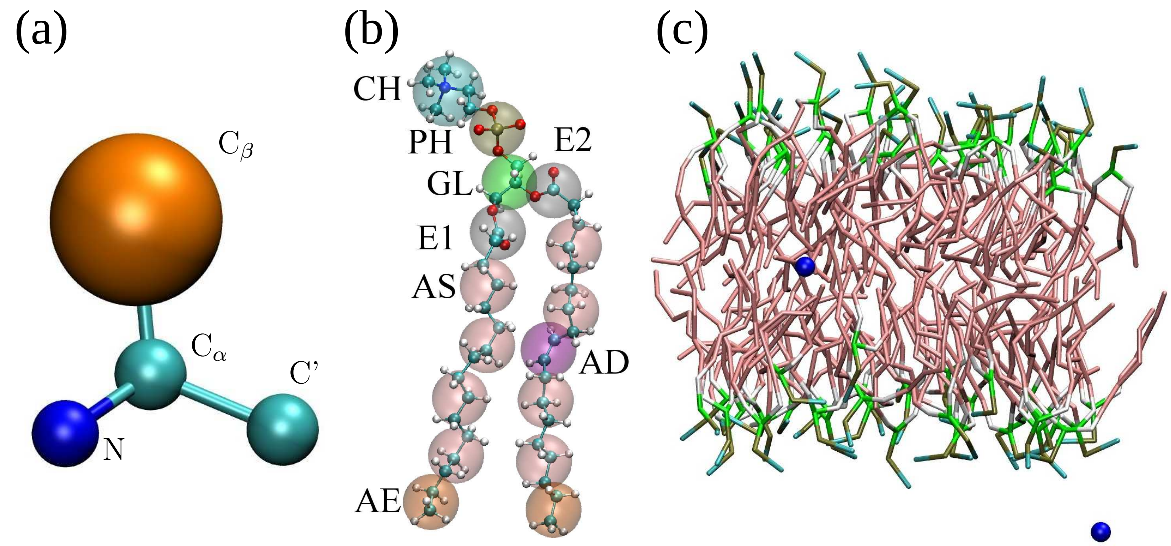


Figure 6.2: (a) Coarse-grained amino acid (colors: N dark blue, C_α and C' cyan, C_β orange). Lipids interact with backbone bead C_α and side-chain bead C_β. (b) Coarse-grained POPC lipid, as presented in [WD10b]. Reproduced in part with permission from Wang, Z.-J. and Deserno, M. *J. Phys. Chem. B* **114** (2010), no. 34, 11207–11220. Copyright 2010 American Chemical Society. (c) Side view of a 72-molecule DOPC bilayer as well as two amino acid side chains (in blue) constrained at the center and outside of the membrane.

of Pro is connected to the backbone at *both* the central carbon and the amide group, see Table 1.1 on page 2), His causes issues due to its multiple protonation states (besides, its pK_a is close to neutral pH, see Table 1.2). Ionizable residues Arg, Lys, Asp, and Glu were calculated for both the charged and neutral forms. The results are reproduced in Figures 6.3, 6.4, and 6.5 (blue, dashed).

The cross-parametrization thus consists of optimizing potentials of interaction between lipids and amino acids to reproduce the PMFs derived from MacCallum *et al.* [MBT08]. Because of the limited amount of information that can be extracted from these free energy profiles, not all cross-interactions can be determined. Specifically, the PMFs derived from atomistic simulations only provide information on the interaction between lipids and side chains, rather than the entire amino acid. The atomistic PMFs will be used as target function for the free energy of insertion of the coarse-grained side-chain bead C_β in a DOPC bilayer. Purely repulsive interactions between lipid and peptide C_α, N, and C' beads will model the excluded volume effect between lipids and the protein backbone. See Figure 6.2 (a) for details.

6.1.1 Simulation and analysis methods

Each simulation consisted of a 72-molecule DOPC bilayer as well as *two* amino acids. In addition to being computationally advantageous, simulating two amino acids allowed the insertion of a residue in each bilayer leaflet, thus avoiding differences in leaflet area due to the presence of an inclusion. See Figure 6.2 (c) for a simulation snapshot.

Simulations were run at constant temperature ($k_{\mathrm{B}}T = \mathcal{E}$) and lateral bilayer tension ($\Sigma = 0$) for a total simulation time $t = 200\,000\,\tau$ and time step $\delta t = 0.1\,\tau$.[12] Temperature control was achieved using a Langevin thermostat with friction constant $\Gamma = 0.2\,\tau^{-1}$, whereas a modified Andersen barostat allowed for box resizing in the lateral x and y directions (box friction $\Gamma_{\mathrm{box}} = 4 \times 10^{-5}\,\tau^{-1}$ and a box mass $Q = 5 \times 10^{-4}\,\mathcal{M}$) [KD99]. All simulations were performed using ESPResSo [LAMH06].

The distance between the amino acid side chain and the bilayer midplane was calculated by measuring the difference between (i) the z coordinate of the side chain bead and (ii) the z coordinate of the center of mass of the bilayer. 10 000 data points were recorded every $20\,\tau$. Umbrella sampling (see subsection A.1.5) was used to constrain the z positions of the two inclusions, applying a harmonic restraint of spring constant $k = 2\,\mathcal{E}/\text{Å}^2$. Pairs of two inclusions were systematically placed 30 Å apart (corresponding roughly to the height of a monolayer) to avoid artifacts.[13] 32 such umbrellas, each separated by 1 Å, allowed to obtain biased distributions in an interval $0 < z < 32$ Å, where $z = 0$ corresponds to the bilayer midplane. We unbiased the sampled distributions using the Weighted Histogram Analysis Method (WHAM) (see Appendix A). In particular, the individual distribution functions were combined using the optimized algorithm presented in subsection A.1.4, rather than the conventional iterative scheme (subsection A.1.3). Convergence was reached ≈ 30 times faster using the optimized technique. Error bars were calculated from bootstrapping the biased distributions (see subsection A.1.6).

6.1.2 Interaction potentials and parametrization

Potentials of interaction were set between all lipid and amino acid bead types using only Lennard-Jones (LJ) and Weeks-Chandler-Andersen (WCA) potentials as functional forms

$$U_{\mathrm{LJ}}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right], \tag{6.1}$$

$$U_{\mathrm{WCA}}(r) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} + \frac{1}{4}\right], & \text{if} \quad r < 2^{1/6}\sigma \\ 0, & \text{otherwise.} \end{cases} \tag{6.2}$$

[12] As in chapter 2, $\mathcal{E} = k_{\mathrm{B}}T_{\mathrm{room}}$ is the intrinsic unit of energy, τ of time, and $\mathcal{M}$ of mass.

[13] Two kinds of artifacts were addressed: (i) inserting two—rather than one—particles in a bilayer can prevent pressure differences between the two leaflets, and (ii) constraining two particles at different *heights* might avoid significant structural effects on the bilayer.

$U_{\mathrm{LJ}}(r)$ was consistently cut and shifted to 0 at a distance $r = 15\,\text{Å}$. During parametrization, we systematically assigned an interaction strength $\epsilon = 1.0\,\mathcal{E}$ for WCA potentials since the function is so insensitive to changes in this parameter.

Lipid bead types AS and AD were not distinguished in terms of their interactions with peptide beads. They were both denoted as AS in the following. Likewise, E1 and E2 were both associated to the same bead type ES in terms of nonbonded interactions with amino acids.

Sets of interaction potentials were iteratively refined until the resulting PMF matched the target curve. A single iteration consisted of (*i*) running 16 simulations (2 umbrellas per simulation yields a total of 32 umbrellas) with a predefined set of interaction potentials (CPU time per simulation ≈ 22 hours) and (*ii*) calculating the resulting PMF from WHAM. While devising an algorithm capable of automatically optimizing the parameters ϵ and σ of the potentials would seem fit, it has shown difficult to effectively establish proper update rules. A random search through parameter space would be intractable because of the time required for a single iteration to complete. Indeed, the parametrization between a side chain and 6 lipid bead types involves 12 free parameters (i.e., ϵ and σ for each potential) per amino acid.[14] Instead, parameter optimization was performed by hand. While the convergence of the first PMF required a large number of iterations,[15] a number of guidelines were empirically observed and subsequently applied to improve the iterative process:

- We used either LJ or WCA potentials to model attractive and repulsive interactions, respectively (i.e., only one such potential was chosen for a given lipid-protein interaction). A larger value of σ increases the excluded volume effect, thus locally shifting the PMF up. Similarly, a stronger interaction strength ϵ will provide more cohesion and affect the free energy profile downwards.

- The effect of choline beads (CH) have a comparatively weak effect on the PMF. Other lipid beads (see below) have a stronger impact on the free energy profile simply because they are further away from bulk water—set as the reference point of the PMF (i.e., $F(z \to \infty) = 0$)—and thus shift the curve over the corresponding interval.

- The interaction between side chain beads and phosphate groups (PH beads) were tuned to reproduce the shape of the PMF in the lipid head region. Many profiles display unfavorable interactions in that region (i.e., $F(z) > 0$). A WCA potential with varying radius σ allowed to reproduce this feature best.

- Glycerol groups form the core of the interfacial region ($z \approx 15 - 20\,\text{Å}$). Many PMFs display strong energetic variations in this small interval. For instance, the profiles of hydrophobic amino acids show sharp drops between the lipid head and lipid tail

[14] This number is a lower bound as it does not take into account the choice in functional form of the potential (i.e., $U_{\mathrm{LJ}}(r)$ or $U_{\mathrm{WCA}}(r)$).

[15] The author of the present thesis believes that his very first white hair grew while parametrizing alanine. C. Yolcu, personal communication.

regions. This was reproduced at the coarse-grained level by setting strong ϵ parameters for the LJ interaction. The σ parameter allowed a fine-tuning of the range over which the drop extended.

- Strong interactions between side chains and ester groups (modeled here by ES beads) have shown to flatten out the PMF curve. In many cases, sharp features between the lipid tails and interfacial region ($z \approx 10 - 15$ Å; e.g., Cys) have best been reproduced by setting weak interactions (or, alternatively, no interactions) between side chains and ES beads.
- Alkyl AS/AD and AE beads affect a remarkably large z interval of the free energy profile (up to $z \approx 18$ Å), requiring a careful adjustment of the parameters. While AE beads naturally have a lower impact on the PMF, their parametrization allowed to fine-tune the close vicinity of the bilayer midplane.

Although the peptide-lipid cross-parametrization is optimized here using a DOPC bilayer, we will assume these potentials to be transferrable across lipid types (e.g., POPC, DPPC).

6.1.3 Optimal parameters and PMFs

Side chains

Table 6.1 shows the sets of interaction potentials and parameters that yielded the closest free energies $F(z)$ to the atomistic targets. For each amino-acid–lipid-group pair, the table displays (*i*) the functional form of the interaction, (*ii*) the strength of the interaction ϵ, and (*iii*) the range of the interaction σ. Empty fields mean no interaction. Figures 6.3, 6.4, and 6.5 show both the coarse-grained (in red) and atomistic (in blue) PMFs.

While it was often possible to reach a good agreement between pairs of curves (i.e., coarse-grained and atomistic), certain features have shown difficult to reproduce:

- High free energies near the membrane center for charged residues (i.e., Arg^+, Asp^-, Glu^-, Lys^+). The atomistic PMFs in the region $z < 5$ Å display a steep, linear behavior. All attempts to reproduce this feature in the coarse-grained simulations yielded a plateau that extended from the center of the bilayer for a few Ångströms. The coarse-grained PMFs were thus shifted downwards to best reproduce the rest of the curve. It is interesting to note that such a behavior was observed previously using the MARTINI force field [MKP+08]. In this case, however, the PMFs were an *independent test* of the parametrization. This artifact is likely the result of coarse-graining, thus lowering the possibility for sharp features in the PMFs.
- Strong features in certain amphipathic and aromatic residues (Cys, Met, Trp) show large variations in rather confined regions of the bilayer. In certain cases, it has shown difficult to appropriately tune the interaction potentials to match the atomistic PMFs.

	CH			PH			GL			ES			AS			AE		
		ϵ	σ		ϵ	σ		ϵ	σ		ϵ	σ		ϵ	σ		ϵ	σ
Ala	LJ	1.0	5.96	WCA	1.0	9.54	LJ	4.5	5.26	WCA	1.0	4.61	LJ	0.85	4.21	LJ	0.85	4.21
Arg^0	LJ	1.0	6.66	WCA	1.0	9.32	LJ	7.5	5.13	WCA	1.0	1.94	WCA	1.0	4.17	WCA	1.0	11.03
Arg^+	LJ	1.2	6.66	WCA	1.0	6.66	LJ	6.5	4.17	LJ	2.0	4.52	WCA	1.0	4.17	WCA	1.0	13.71
Asn	LJ	1.1	6.21	WCA	1.0	7.45	LJ	7.5	3.86	WCA	1.0	3.01	WCA	1.0	3.86	WCA	1.0	9.92
Asp^0	LJ	1.2	6.18	WCA	1.0	7.42	LJ	6.5	3.84	WCA	1.0	1.79	WCA	1.0	2.74	WCA	1.0	4.93
Asp^-	LJ	1.2	6.18	WCA	1.0	9.27	LJ	4.0	5.48	WCA	1.0	2.39				WCA	1.0	13.7
Cys	LJ	1.0	6.16	WCA	1.0	8.32	LJ	7.0	3.82				WCA	1.0	1.09	WCA	1.0	1.37
Gln	LJ	1.2	6.45	WCA	1.0	7.74	LJ	7.5	4.03	WCA	1.0	2.50	WCA	1.0	4.60	WCA	1.0	8.05
Glu^0	LJ	1.2	6.41	WCA	1.0	7.69	LJ	7.5	4.00	WCA	1.0	2.48	WCA	1.0	4.00	WCA	1.0	9.14
Glu^-	LJ	1.2	6.41	WCA	1.0	9.94	WCA	1.0	5.71				WCA	1.0	5.71	WCA	1.0	10.28
Gly	LJ	1.0	5.63	WCA	1.0	8.73	LJ	4.5	4.44	WCA	1.0	4.34	LJ	0.7	3.94	LJ	0.7	3.94
His	LJ	1.0	6.23	WCA	1.0	9.28	LJ	7.5	4.98	WCA	1.0	0.60	WCA	1.0	4.15	WCA	1.0	7.74
Ile	LJ	1.2	6.61	WCA	1.0	9.92	LJ	5.5	5.91	WCA	1.0	6.73	LJ	1.45	3.55	LJ	1.5	3.55
Leu	LJ	1.2	6.61	WCA	1.0	10.25	LJ	5.5	5.91	WCA	1.0	6.73	LJ	1.2	3.55	LJ	1.2	3.55
Lys^0	LJ	1.0	6.63	WCA	1.0	9.02	LJ	7.5	4.39				WCA	1.0	1.48	WCA	1.0	1.19
Lys^+	LJ	1.2	6.63	WCA	1.0	6.63	LJ	6.5	4.15	WCA	1.0	2.57	WCA	1.0	4.74	WCA	1.0	10.67
Met	LJ	1.0	6.59	WCA	1.0	8.96	LJ	7.5	4.36				WCA	1.0	1.47	WCA	1.0	1.18
Phe	LJ	1.2	6.77	WCA	1.0	9.82	LJ	7.5	5.46	WCA	1.0	6.37	LJ	1.1	3.64	LJ	1.2	3.64
Pro	LJ	1.0	6.20	WCA	1.0	8.99	LJ	7.5	4.95	WCA	1.0	0.60	WCA	1.0	4.68	WCA	1.0	8.25
Ser	LJ	1.0	5.97	WCA	1.0	8.24	LJ	5.5	4.22	WCA	1.0	1.73	WCA	1.0	3.16	WCA	1.0	10.01
Thr	LJ	1.0	6.23	WCA	1.0	9.28	LJ	7.5	4.98	WCA	1.0	0.60	WCA	1.0	4.15	WCA	1.0	7.74
Trp	LJ	1.2	6.99	WCA	1.0	7.69	LJ	7.0	4.15				WCA	1.0	1.89	WCA	1.0	3.15
Tyr	LJ	1.2	6.79	WCA	1.0	8.15	LJ	6.5	4.45	WCA	1.0	1.98	WCA	1.0	3.65	WCA	1.0	4.87
Val	LJ	1.2	6.42	WCA	1.0	10.08	LJ	7.0	5.15	WCA	1.0	1.87	LJ	1.0	3.43	LJ	1.0	3.43

Table 6.1: Optimized interaction potentials for all amino acids and lipid bead types: functional form (LJ or WCA) and their parameters ϵ [$\mathcal{E}$] and σ [Å]. Empty fields mean no interaction.

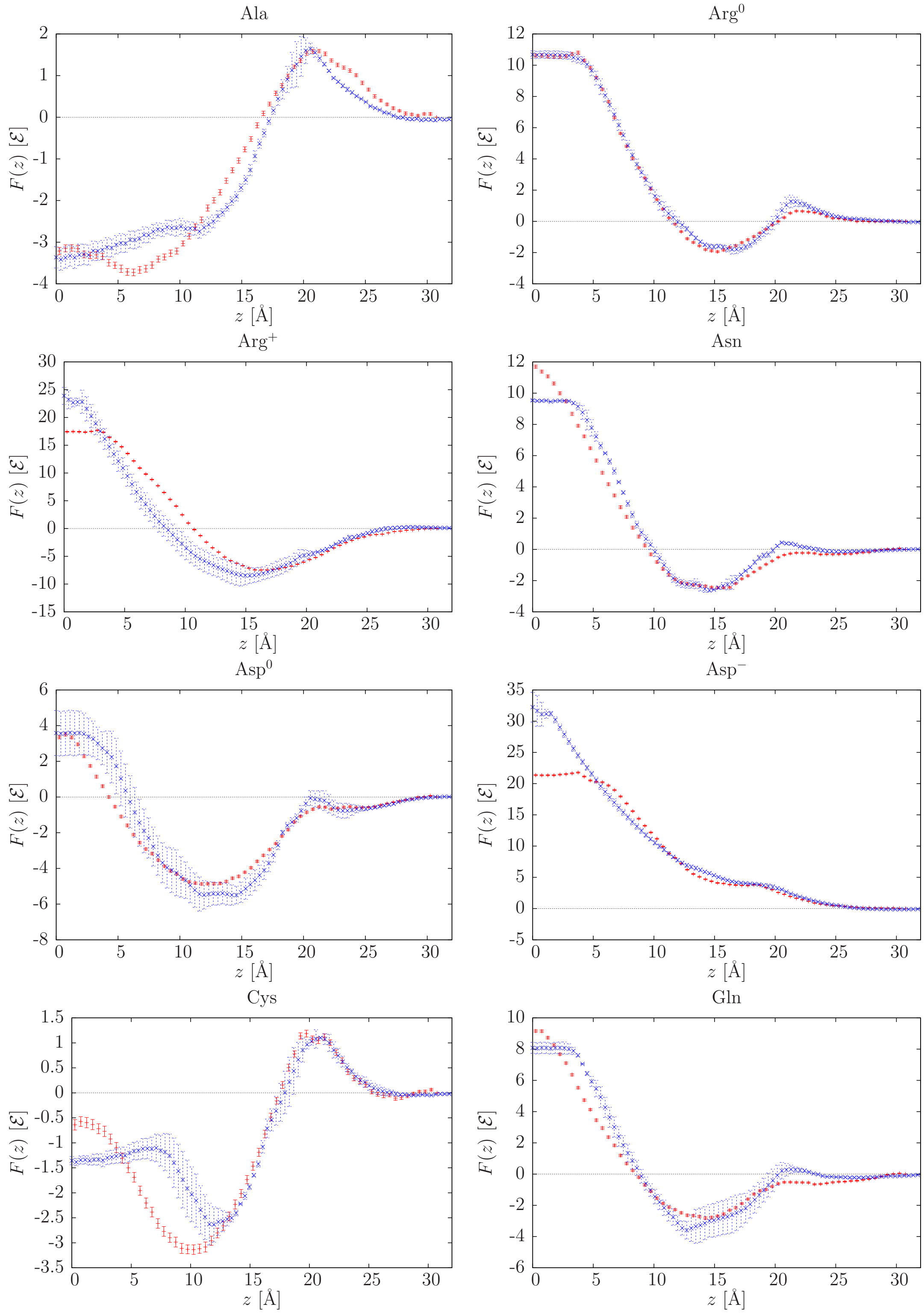


Figure 6.3: Free energy profiles for the insertion of a single amino acid in a DOPC bilayer: coarse-grained (red, solid) and atomistic (blue, dashed. From MacCallum *et al.* [MBT08].). Part 1/3.

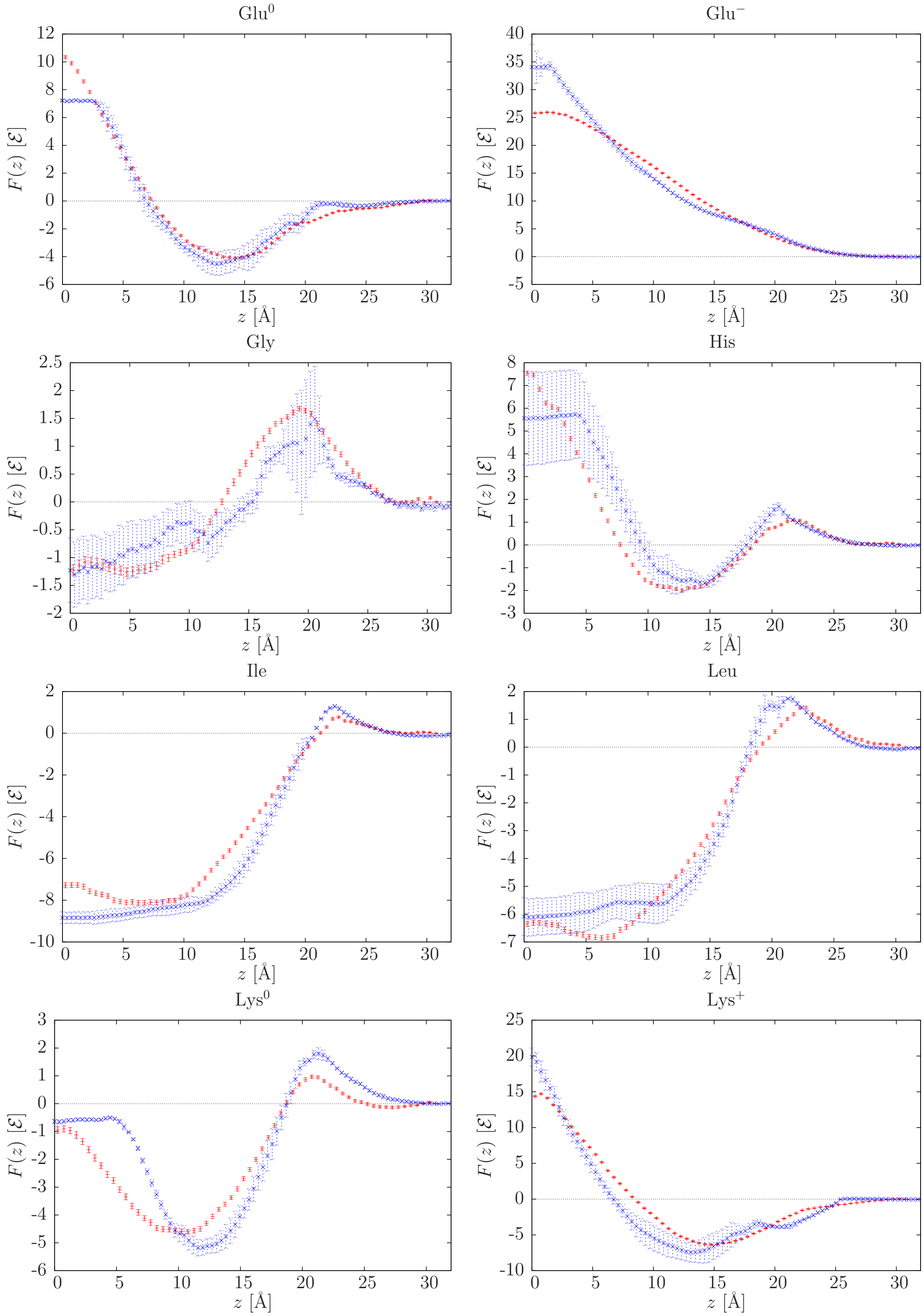


Figure 6.4: Free energy profiles for the insertion of a single amino acid in a DOPC bilayer: coarse-grained (red, solid) and atomistic (blue, dashed. From MacCallum *et al.* [MBT08].). Part 2/3. Reference data for Gly and His were extrapolated as described in the text.

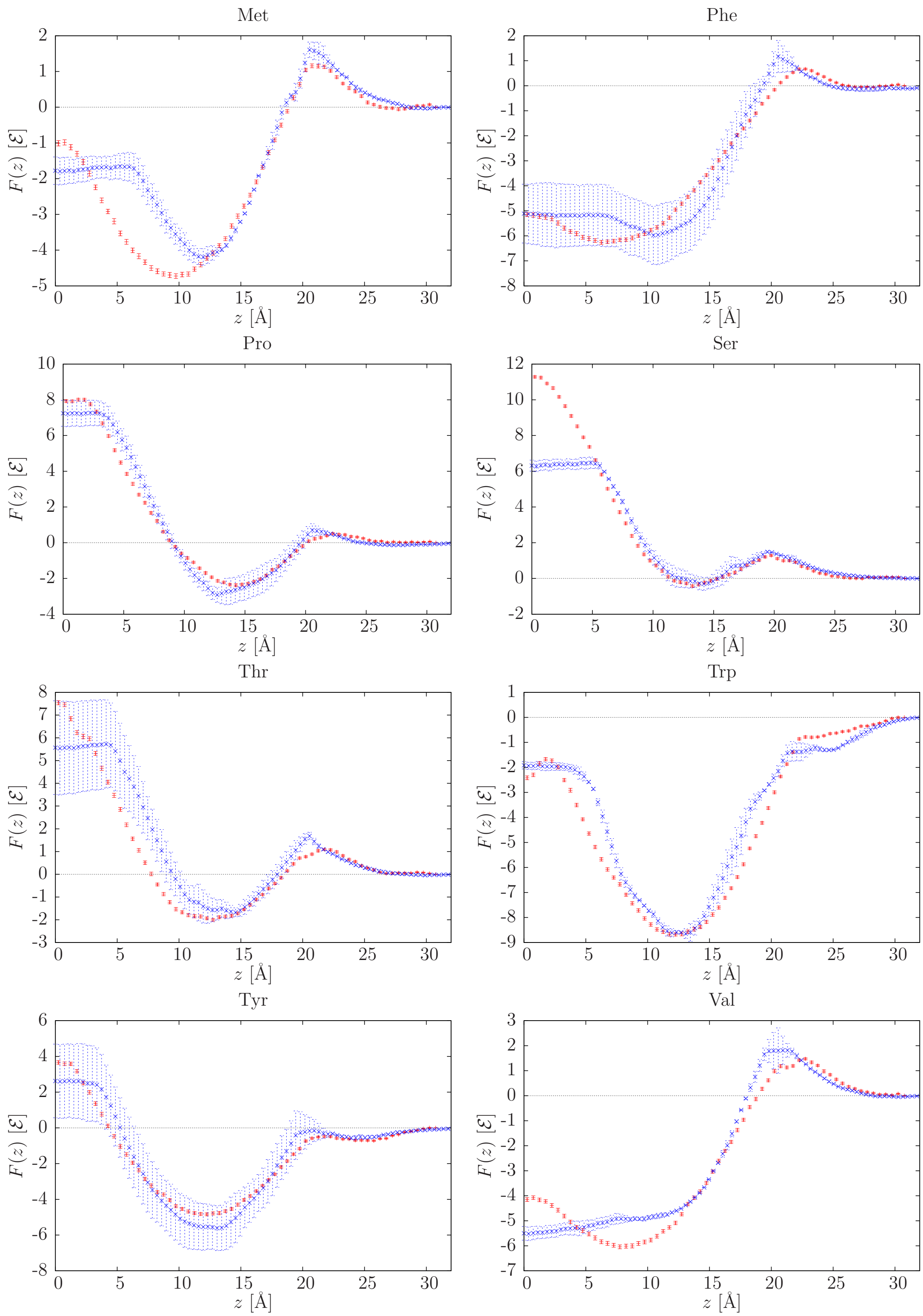


Figure 6.5: Free energy profiles for the insertion of a single amino acid in a DOPC bilayer: coarse-grained (red, solid) and atomistic (blue, dashed. From MacCallum *et al.* [MBT08]). Part 3/3. Reference data for Pro were extrapolated as described in the text.

- The sharp drops present in the PMFs of hydrophobic residues (e.g., Ala, Ile, Leu, Val) can be somewhat difficult to reproduce at the coarse-grained level. The fewer beads (and types of beads) involved in the reduced representation again seem to prevent sharp features in PMFs. This is somewhat compensated by providing strong attractions between these residues and glycerol beads (GL).

Backbone

To reproduce the excluded volume effect between lipids and the protein backbone, purely repulsive interactions (WCA) were set between all lipid bead types and peptide backbone particles N, C_α, and C' with parameters $\epsilon = \epsilon_{\mathrm{bb}} = 0.02\,\mathcal{E}$ (Table 2.3 on page 25) and $\sigma = \sigma_{\mathrm{peptide}} + \sigma_{\mathrm{lipid}}$, where $\sigma_{\mathrm{peptide}}$ corresponds to the van der Waals radius of peptide backbone particles N, C_α, or C' (see Table 2.3), and $\sigma_{\mathrm{lipid}} = 3.0\,\text{Å}$ represents an average lipid bead radius.[16] While somewhat arbitrary, this parametrization is sufficient to reproduce simple steric effects.

6.1.4 Parametrization of Glycine, Histidine, and Proline

Due to the lack of atomistic data for residues Gly, His, and Pro, their PMF curves were extrapolated from other amino acids similar in chemical structure or hydrophobicity. Details of their construction are provided in the next paragraphs. The resulting reference curves are shown in Figures 6.3, 6.4, and 6.5, as well as the corresponding optimized coarse-grained curves. Coarse-grained parameters used to reproduce the target data are shown in Table 6.1.

Glycine

The side chain of glycine contains no heavy atom: it only consists of a single H atom. The simplest side chain computed atomistically by MacCallum *et al.* is alanine, which consists of a single methyl group (see Table 1.1). The contribution of this group to the free energy profile is estimated by considering valine, which consists of two methyl groups and a central carbon. Neglecting the impact of the central carbon, we estimate the contribution of one methyl group by subtracting the PMF of alanine from the PMF of valine. Then, we subtract this quantity from the PMF of alanine in order to obtain an estimate for the PMF of glycine

$$\begin{aligned} F_{\mathrm{Gly}}(z) &= F_{\mathrm{Ala}}(z) - (F_{\mathrm{Val}}(z) - F_{\mathrm{Ala}}(z)) \\ &= 2F_{\mathrm{Ala}}(z) - F_{\mathrm{Val}}(z). \end{aligned} \tag{6.3}$$

A simple propagation of uncertainty (see Technical Point 6.1 for a derivation; we neglect cross-correlations between Ala and Val) leads to an expression for the z-dependent standard

[16] The radius of a lipid bead can be estimated from the radial distribution function describing its interaction with itself (e.g., CH–CH).

deviation of the PMF of glycine, $\sigma_{\text{Gly}}(z)$, as a function of the standard deviations of $F(z)$ for alanine $\sigma_{\text{Ala}}(z)$ and valine $\sigma_{\text{Val}}(z)$:

$$\sigma_{\text{Gly}}(z) = \sqrt{4\sigma^2_{\text{Ala}}(z) + \sigma^2_{\text{Val}}(z)}. \tag{6.4}$$

The resulting target curve (with error bars) is shown in Figure 6.4. Interaction parameters for the coarse-grained curve are reported in Table 6.1; the resulting coarse-grained PMF is shown in Figure 6.4.

One may wonder how accurate these results are, considering the PMF of glycine does not extend beyond $\pm 1\,\mathcal{E}$: is the abovementioned estimate (i.e., using the PMFs of alanine and valine) accurate within $1\,\mathcal{E}$? Alternatively, one could argue against *any* interaction between a glycine side chain and the different lipid beads, since a glycine side chain is only composed of one hydrogen atom (implying weak interactions and a small van der Waals radius).[17] Either way, the resulting PMF ought to show few features and is unlikely to play a major role in the presence of other amino acids.

Proline

The PMF of proline is estimated from its hydrophobicity, rather than chemical structure as above. For consistency with the peptide-peptide side-chain interactions, we follow the hydrophobicity scale derived in chapter 2. This normalized scale reduces the 20×20 Miyazawa-Jernigan matrix [MJ96] into a set of 20 interaction parameters ϵ_i that approximately recreate all interactions using the Lorentz-Berthelot mixing rule (see Table 2.2). The normalized hydrophobicity of proline ($\epsilon'_i = 0.14$) is located between glutamine ($\epsilon'_i = 0.13$) and threonine ($\epsilon'_i = 0.16$). Figures 6.3 and 6.5 show that glutamine and threonine have very similar PMF curves. We interpolate these curves to estimate the PMF of proline by weighting them according to the normalized hydrophobicities

$$F_{\text{Pro}}(z) = \frac{2}{3}F_{\text{Gln}}(z) + \frac{1}{3}F_{\text{Thr}}(z). \tag{6.10}$$

Error bars are calculated, as before, using propagation of uncertainty from the error bars of F_{Gln} and F_{Thr}:

$$\sigma_{\text{Pro}}(z) = \sqrt{\frac{4}{9}\sigma^2_{\text{Gln}}(z) + \frac{1}{9}\sigma^2_{\text{Thr}}(z)}. \tag{6.11}$$

The resulting target curve (with error bars) is shown in Figure 6.5. Interaction parameters for the coarse-grained curve are reported in Table 6.1; the resulting coarse-grained PMF is shown in Figure 6.5.

Histidine

Like proline, the PMF of histidine is estimated from its hydrophobicity. We compare His with other hydrophilic amino acids of similar normalized hydrophobicity parameters ϵ'_i

[17] Recall that the glycine side chain is *not* modeled in the original protein model (chapter 2).

Technical Point 6.1 Propagation of uncertainty

The following derivation estimates the uncertainty of a function based on the errors of its variables [Mey92]. Consider the distribution function $Y = f(X)$ where $f(\cdot)$ is some known function and the distribution of the random variable X is known. By Taylor expanding Y about the mean $X = \mu_X$, one gets

$$Y \approx f(\mu_X) + \left.\frac{\partial f}{\partial X}\right|_{X=\mu_X} (X - \mu_X). \tag{6.5}$$

The first and second moments of Y can then be determined

$$\begin{aligned} \mu_Y &= \mathrm{E}\left[Y\right] \approx \mathrm{E}\left[f(\mu_X) + \frac{\partial f}{\partial X}(X - \mu_X)\right] = f(\mu_X), & (6.6) \\ \sigma_Y^2 &= \mathrm{E}\left[(Y - \mu_Y)^2\right] \approx \mathrm{E}\left[\left\{\frac{\partial f}{\partial X}(X - \mu_X)\right\}^2\right] \\ &= \left(\frac{\partial f}{\partial X}\right)^2 \mathrm{E}\left[(X - \mu_X)^2\right] = \left(\frac{\partial f}{\partial X}\right)^2 \sigma_X^2. & (6.7) \end{aligned}$$

Now consider the corresponding multidimensional problem of a distribution function Y with n different random variables X_i: $Y = f(X_1, X_2, \ldots, X_n)$. A similar Taylor expansion will yield

$$Y \approx f(\mu_1, \mu_2, \ldots, \mu_n) + \sum_{i=1}^{n} \left[\frac{\partial f}{\partial X_i}(\mu_1, \mu_2, \ldots, \mu_n))\right] (X_i - \mu_i). \tag{6.8}$$

As before, the first moment will only consist of the constant $f(\mu_1, \mu_2, \ldots, \mu_n)$. The second moment yields

$$\begin{aligned} \sigma_Y^2 &= \mathrm{E}\left[(Y - \mu_Y)^2\right] \approx \mathrm{E}\left[\sum_i \frac{\partial f}{\partial X_i}\left(X_i - \mu_i\right) \sum_j \frac{\partial f}{\partial X_j}\left(X_j - \mu_j\right)\right] \\ &= \mathrm{E}\left[\sum_i \left(\frac{\partial f}{\partial X_i}\right)^2 (X_i - \mu_i)^2 + \sum_i \sum_{j \neq i} \frac{\partial f}{\partial X_i}\frac{\partial f}{\partial X_j}\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right] \\ &= \sum_i \left(\frac{\partial f}{\partial X_i}\right)^2 \mathrm{E}\left[(X_i - \mu_i)^2\right] + \sum_{i \neq j} \frac{\partial f}{\partial X_i}\frac{\partial f}{\partial X_j}\mathrm{E}\left[\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right] \\ &= \sum_i \left(\frac{\partial f}{\partial X_i}\right)^2 \sigma_i^2 + \sum_{i \neq j} \frac{\partial f}{\partial X_i}\frac{\partial f}{\partial X_j}\sigma_{ij}. \end{aligned} \tag{6.9}$$

σ_{ij} corresponds to the covariance of the random variables X_i and X_j. This term vanishes when the variables are uncorrelated.

Assuming the distribution function is of the form $g = \alpha a + \beta b$, the second moment will yield $\sigma_g^2 = \alpha^2\sigma_a^2 + \beta^2\sigma_b^2 + 2\alpha\beta\sigma_{ab}$.

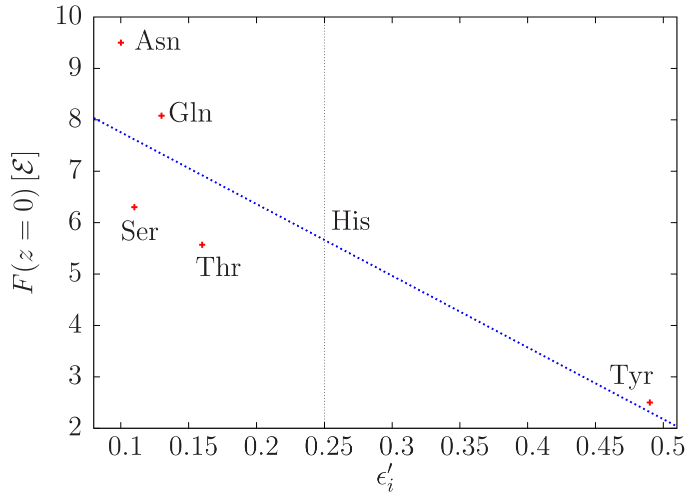


Figure 6.6: PMF height at origin $F(z = 0)$ as a function of the normalized hydrophobicity parameter ϵ'_i (defined in chapter 2) for several hydrophilic amino acids. PMF heights determined from the atomistic simulations of MacCallum *et al.* [MBT08]. Vertical line marks the hydrophobicity of histidine. Dashed line is a linear fit to the different residues. The intersection between the fit and the vertical line shows the estimate for histidine.

(shown in Table 2.2): Asn, Ser, Gln, Thr, and Tyr. These curves all show the same overall shape: a repulsive ($F > 0$) range around the lipid tails, followed by a dip in the interfacial region, and finally a peak around the head groups. Differences between these curves mostly arise from vertical shifts—especially at the origin (i.e., $F(z = 0)$). In Figure 6.6 we show the relationship between $F(z = 0)$ and the normalized hydrophobicity parameter ϵ'_i for the abovementioned hydrophilic side chains. The vertical line marks the value of ϵ'_i for His. A linear fit to the five other hydrophilic residues is represented by the dashed line (this assumes a linear relationship between hydrophobicity and free energy of insertion at the bilayer midplane, $F(z = 0)$). The intersection between the two lines is our estimate of the PMF height for histidine. Note that $F_{\text{His}}(z = 0) \approx F_{\text{Thr}}(z = 0)$. This is in good agreement with the PMFs of His and Thr derived from the Martini force field [MKP+08]. From lack of further data, we use the same free-energy profiles for His and Thr, i.e., $F_{\text{His}}(z) = F_{\text{Thr}}(z)$.

6.1.5 Structure and energetics between residues and the bilayer

At this point, we briefly discuss two specific properties of lipid-residue interaction, namely (i) the presence of water defects due to charged residues and (ii) the protonation of ionizable residues as a function of bilayer depth.

(a) (b)

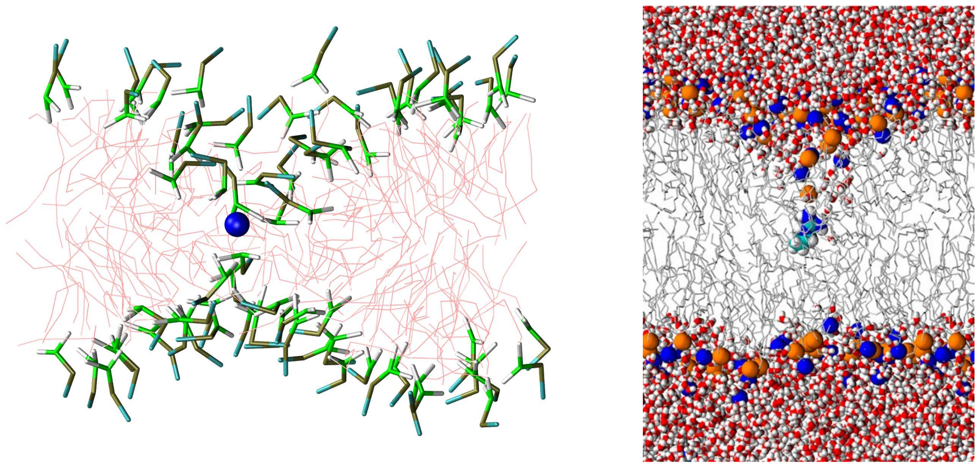


Figure 6.7: (a) Coarse-grained configuration of Arg^+ (blue sphere) in a DOPC bilayer (thin lines: hydrocarbon tails; licorice: head groups). The insertion of Arg^+ creates a strong, local deformation of the bilayer. (b) Atomistic configuration of Arg^+ in a DOPC bilayer. Reprinted from Biophysical Journal, **94**, J. L. MacCallum, W. F. D. Bennett, and D. P. Tieleman, *Distribution of amino acids in a lipid bilayer from computer simulations*, 3393–3404, Copyright (2008), with permission from Biophysical Society [MBT08].

Water defects in a solvent-free model

The partitioning of polar and charged residues into the hydrocarbon region of the bilayer have been associated with large water defects (e.g., [MRMT04, FTvHW05, MBT08]; see Figure 6.7 (b)). MacCallum *et al.* observed the stabilization of narrow pores that allow water molecules to interact with a polar/charged residue. Their simulations suggest that such a water channel persists even when a charged Arg residue is located in the bilayer midplane [MBT08].

The coarse-grained simulations show that Arg^+ strongly deform the bilayer locally (Figure 6.7). The layering of the lipids is strongly perturbed because the residue interacts more favorably with the lipid head groups than the hydrocarbon tails (see Figure 6.3). Overall, we observe a localized thinning of the bilayer.[18]

Ionizable residues

As mentioned before, ionizable residues Arg, Asp, Glu, and Lys can be found in either charged or neutral form of their acid-base pairs. Classical simulations do not allow for the explicit modeling of proton exchange between chemical species. Only one of the two

[18]It is quite remarkable that such an effect be present *without* explicit electrostatics.

conjugates is simulated, chosen by their relative population (i.e., the pK_a) in a given environment. pK_a values of these residues in bulk water are reported in Table 1.2 on page 3. MacCallum *et al.* have calculated the z-dependence of these pK_a values (Figure 6.8). These results were calculated from the PMFs presented in their paper (shown here in Figures 6.3, 6.4, and 6.5). Using a thermodynamic cycle, they calculated the free energy of protonation as a function of depth in the membrane from the PMFs and the pK_a of each side chain in bulk water (see Technical Point 6.2). Figure 6.8 shows that

- Asp and Glu remain charged until they reach a bilayer depth $z \approx 20\,\text{Å}$.
- Lys becomes neutral very close to the membrane center ($z < 4\,\text{Å}$).
- the pK_a of Arg remains above 7.0 over the entire bilayer except in the close vicinity of the membrane center. At this point, its value (≈ 7.0) suggests that the residue might remain charged.[19]

Overall, these provide guidelines as to which protonation state should be selected provided the depth of an amino acid in the bilayer. Due to the lack of reference data, His will systematically be modeled as neutral in the coarse-grained simulations.

6.2 Simulations of transmembrane helices

In the following, we perform simulations of several transmembrane proteins and compare them with either experimental or atomistic data. To do so, we first need to model the N- and C-termini that are present at the ends of a protein. While this issue *should* have been addressed in the presentation of the coarse-grained protein model (chapter 2), their effects on the conformations sampled is rather limited—the termini are thus usually neglected in coarse-grained peptide models parametrized in an aqueous environment. However, these termini may play an important role for transmembrane proteins for which most amino acids are hydrophobic, but the termini are either polar or charged. This scenario ensures that the protein is *integral* to the membrane, i.e., it spans the bilayer thickness (rather than "dive" inside the membrane if all amino acids were hydrophobic). Common examples of N- and C-termini include the acetyl and n-methyl amide groups, respectively (Figure 6.9 (e) and (f)). Similarly, Figure 6.9 (d) represents the chemical structure of a C-terminal phenylalaninol found in the helix-forming alamethicin peptide [TSB99].

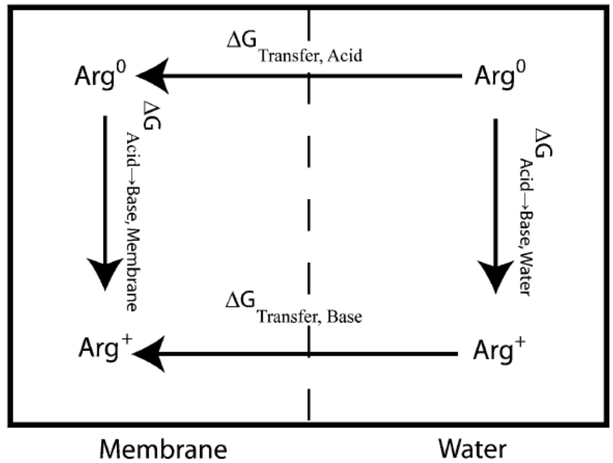
Because there exist different kinds of N- and C-termini, and because we are only interested in reproducing their "hydrophilicity," we model these groups in a very simple way: both N- and C-termini are represented by an entire amino acid. In particular:

- They are parametrized as Gly residues (i.e., no heavy atom on the side chain; see chapter 2) in terms of peptide-peptide interactions. This provides flexibility in the backbone, while preventing side chain–side chain interactions.

[19] As mentioned previously, the presence of charged residues in the membrane center can be explained by stabilizing water defects (e.g., [MRMT04, FTvHW05, MBT08]).

Technical Point 6.2 Thermodynamic cycle for pK_a calculation in the bilayer

The free energy of protonation *inside* the bilayer $\Delta G_{\text{Acid}\to\text{Base, Membrane}}$ is calculated using the thermodynamic cycle shown here. The upper and lower branches correspond to the free energy difference between a residue (either charged or neutral) in water and in the membrane, as given by the PMFs in Figures 6.3, 6.4, and 6.5. The last branch—on the right—describes the free energy of protonation in bulk water $\Delta G_{\text{Acid}\to\text{Base, Water}}$. Its value can be obtained from the pK_a in water (see Table 1.2). The following derivation provides the relationship between pK_a and free energy difference $\Delta G_{\text{Acid}\to\text{Base}}$.



Reprinted from Biophysical Journal, **94**, J. L. MacCallum, W. F. D. Bennett, and D. P. Tieleman, *Distribution of amino acids in a lipid bilayer from computer simulations*, 3393–3404, Copyright (2008), with permission from Biophysical Society [MBT08].

By definition, the pK_a is a logarithmic measure of the acid dissociation constant K_a: p$K_a = -\log_{10} K_a$, where

$$K_a = \frac{[\mathrm{A}^-][\mathrm{H}^+]}{[\mathrm{AH}]}. \tag{6.12}$$

pH is associated with the quantity $-\log_{10}[\mathrm{H}^+]$ (strictly speaking, pH is defined by the *activity* of hydrogen ions) while the free energy $\Delta G_{\text{Acid}\to\text{Base}}$ is given by

$$\Delta G_{\text{Acid}\to\text{Base}} = -k_{\mathrm{B}}T \ln\left(\frac{[\mathrm{A}^-]}{[\mathrm{AH}]}\right). \tag{6.13}$$

Finally, we obtain

$$\mathrm{p}K_a = \frac{1}{k_{\mathrm{B}}T\ln 10}\Delta G_{\text{Acid}\to\text{Base}} + \mathrm{pH}, \tag{6.14}$$

which is equivalent to the Henderson-Hasselbalch equation [GG08].

Equation 6.14 can easily be inverted to express $\Delta G_{\text{Acid}\to\text{Base}}$ as a function of pK_a. Given pK_a values for each side chain in water (Table 1.2), one can calculate $\Delta G_{\text{Acid}\to\text{Base, Water}}$. The thermodynamic cycle shown in the figure illustrates that the free energy of protonation *inside* the bilayer is given by

$$\Delta G_{\text{Acid}\to\text{Base, Membrane}} = -\Delta G_{\text{Transfer, Acid}} + \Delta G_{\text{Acid}\to\text{Base, Water}} + \Delta G_{\text{Transfer, Base}}. \tag{6.15}$$

Using Equation 6.14 once more provides the pK_a of all ionizable residues at *any* depth z in the bilayer. The results are shown in Figure 6.8.

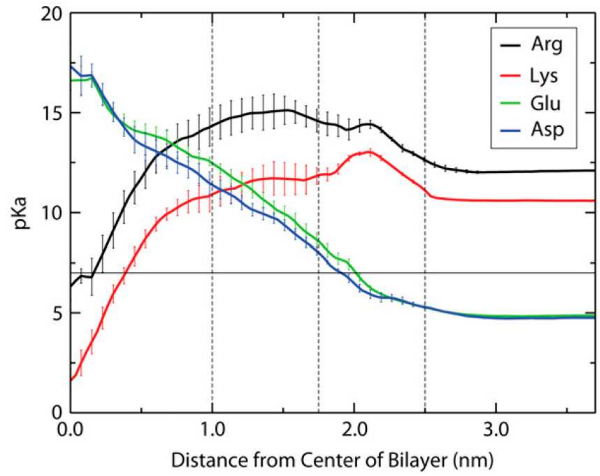


Figure 6.8: pK_a of ionizable residues as a function of bilayer depth z. pK_a values at large z (bulk water) correspond to the values reported in Table 1.2. Reprinted from Biophysical Journal, **94**, J. L. MacCallum, W. F. D. Bennett, and D. P. Tieleman, *Distribution of amino acids in a lipid bilayer from computer simulations*, 3393–3404, Copyright (2008), with permission from Biophysical Society [MBT08].

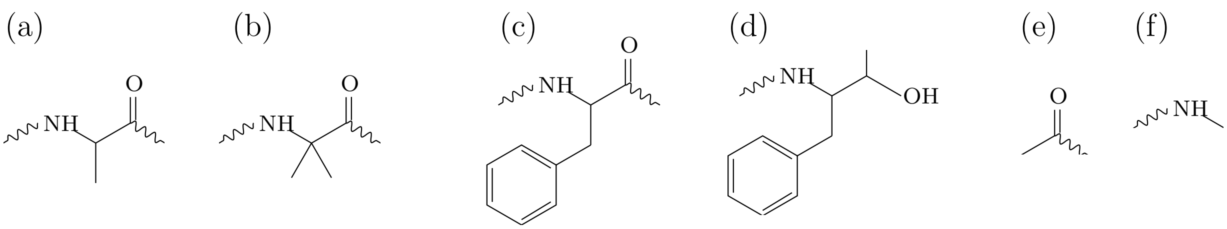


Figure 6.9: Chemical structures of (a) alanine (Ala), (b) α-aminoisobutyric acid (Aib) [IBKS01, TSB99], (c) phenylalanine (Phe), and (d) C-terminal phenylalaninol (Phol) [TSB99] residues, and (e) acetyl and (f) n-methyl amide groups. Squiggly lines represent the peptide bonds connecting neighboring amino acids.

- All protein-lipid interactions (i.e., between the terminus' side chain and all lipid bead types) are parametrized as repulsive. WCA potentials are used with strength $\epsilon = 1.0\,\mathcal{E}$ and radius $\sigma = 4.0\,\text{Å}$. This imprints enough hydrophilicity in these termini, as can be seen in its PMF (Figure 6.10). While this choice of parametrization may seem arbitrary, we note that (*i*) the lack of reference PMF data prevents a proper optimization of the potentials and (*ii*) commonly used acetyl and n-methyl amide groups are *charged* at neutral pH [FP02]. While these groups might neutralize upon insertion into the bilayer, the atomistic PMFs of charged amino acid side chains from MacCallum *et al.* [MBT08] all display very large free energy barriers ($\approx 20-25\,k_\mathrm{B}T$). On the other hand, the PMF of our coarse-grained N- and C-termini shows a much smaller barrier ($\approx 7\,\mathcal{E}$), which suggests a reasonable parametrization considering that it is repulsive enough to keep transmembrane helices integral to the membrane (see below).

- The presence of extra amino acids at the ends of the chain has an impact on the number of groups that can form hydrogen bonds: recall from chapter 2 the multi-body nature of the hydrogen-bond potential (Equation 2.8 on page 24)—the very first amide group and very last carbonyl groups of a protein chain cannot form hydrogen bonds due to a lack of neighboring backbone beads. The presence of amino-acid-like N- and C-termini allows the formation of hydrogen bonds for all groups along the chain except for the amide and carbonyl groups of the N- and C-termini, respectively.[20]

While it would certainly be possible to refine this model depending on the chemical structure of a given terminus, this parametrization has proven successful in keeping hydrophobic transmembrane helices stably oriented. We have found that fine-tuning the parameters of the WCA potentials (i.e., strength of the repulsion) did not affect the *equilibrium* results presented below in any qualitative way (data not shown).[21] Of course, the proper tuning of these termini would be essential to accurately reproduce the free energy of insertion of transmembrane proteins.

6.2.1 Fluctuations in and out of the bilayer

A peptide that folds into an α-helix both in water and in the membrane will likely *not* have the same flexibility in the two environments: the free energy of breaking a peptide-peptide hydrogen bond is higher in the membrane because there are no available hydrogen-bond donors/acceptors in the apolar solvent. One thus expects a helix to be stiffer in the membrane. This is indeed what was observed from atomistic simulations for alamethicin (Alm), a channel-forming, fungal peptide (Tieleman *et al.* [TSB99]; sequence shown in Table 6.2): root-mean-square fluctuations[22] (RMSF) of the helix in water and a POPC bilayer (reproduced in Figure 6.11; denoted "water AA" and "membrane AA," respectively) clearly

[20] The addition of these N- and C-termini in the coarse-grained simulations has allowed to significantly reduce artificial fluctuations at the two ends of the chain (data not shown).

[21] Surely the free energy of insertion of these termini will have an impact on the *kinetics* of insertion of transmembrane proteins.

[22] The root-mean-square fluctuation $\sqrt{\overline{\Delta r_i^2}}$ measures the deviation of particle i with respect to its average position, averaged over time:

$$\overline{\Delta r_i^2} = \frac{1}{N_t}\sum_{k=1}^{N_t}\left(\boldsymbol{r}_i(t_k) - \overline{\boldsymbol{r}_i}\right)^2$$

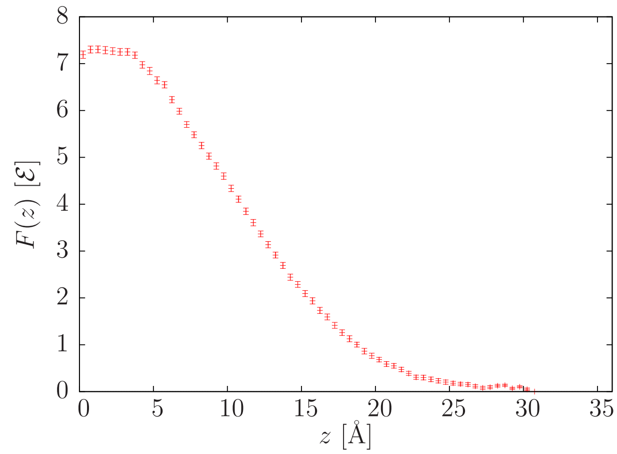


Figure 6.10: Free energy profile for the insertion of a coarse-grained N- or C-terminus side chain in a DOPC bilayer.

Name	Sequence
Alamethicin	`APAAA AQAVA GLAPV AAEQF`
WALP16	`GWWLA LALAL ALAWW A`
WALP23	`GWWLA LALAL ALALA LALAL WWA`
WALP27	`GWWLA LALAL ALALA LALAL ALALW WA`

Table 6.2: Amino acid sequences of alamethicin and several WALP peptides. Underlined amino acids are non-natural: `A` and `F` refer to α-aminoisobutyric acid and C-terminal phenylalaninol, respectively (e.g., [TSB99, IBKS01]; Figure 6.9). All sequences form transmembrane helices [TSB99, MKP$^+$08, KI10]. The N-terminus and C-terminus of each peptide is blocked by an acetyl and n-methyl amide groups, respectively (except alamethicin, for which the C-terminus is embedded in the last phenylalaninol).

make the point. One wonders, though, whether the same should hold at the coarse-grained level:

- The strength of the hydrogen-bond interaction ϵ_{hb} was parametrized to reproduce the structure of helical proteins in *water* (see section 2.4 on page 31). This strongly suggests the need for a reparametrization of ϵ_{hb} in a membrane environment (the value would likely go *up* to reproduce the abovementioned change in free energy).

with $\overline{\boldsymbol{r}_i} = \sum_{k=1}^{N_t} \boldsymbol{r}_i(t_k)/N_t$, which corresponds to a time average over N_t measurements, and $\boldsymbol{r}_i(t_k)$ is the position of atom i at time t_k [Kuc96]. Here, we calculate the fluctuations of a residue by monitoring its C_α position *after* alignment (i.e., translation and rotation) of all snapshots with a reference conformation.

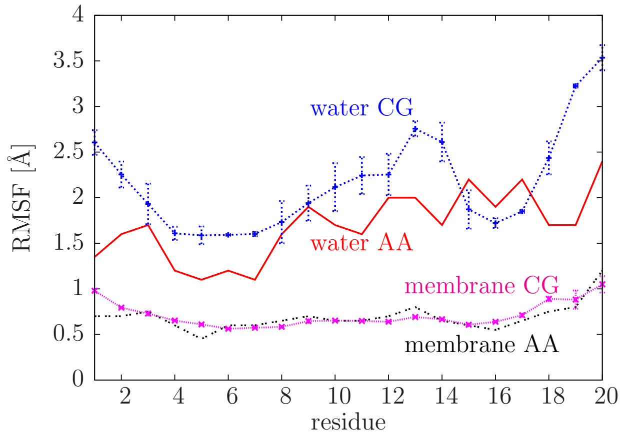


Figure 6.11: Root-mean-square fluctuations of alamethicin in water and in a POPC bilayer (denoted "membrane"). "AA" and "CG" correspond to atomistic and coarse-grained simulations, respectively. Atomistic data reproduced from Tieleman *et al.* [TSB99].

- This coarse-grained model couples an *implicit*-water solvent with an *explicit*-membrane environment. The associated change in terms of sterics—and thus fluctuations—is difficult to predict.

In the following, we repeat the simulations of Tieleman *et al.* using the present coarse-grained model.

Table 6.2 indicates that several residues of Alm are not part of the set of the twenty naturally occurring amino acids (Tables 1.1 and 1.2 on pages 2–3), namely α-aminoisobutyric acid (Aib) and C-terminal phenylalaninol (Phol):[23]

- Aib is close in structure to Ala. As illustrated in Figure 6.9, Aib contains a methyl group instead of the C_α-bound hydrogen. It has been shown to promote 3_{10}-helix formation (rather than α-helix for polyalanine) [IBKS01]. The native structure of Alm, nevertheless, stabilizes to an α-helix. For simplicity, we therefore model Aib residues as Ala (the coarse-grained simulations correctly stabilized an α-helix in both water and the membrane; see below).

- Phol incorporates a CH_2OH C-terminal group into the backbone of a Phe residue [TSB99] (Figure 6.9). We parametrize Phol in the simulations as a standard Phe residue followed by a (generic) coarse-grained C-terminal.[24]

[23] Alm is part of a family of fungal peptides that produces nonstandard amino acids [CWN06].

[24] Recall that we model only one type of terminus in the coarse-grained simulations. See above for more details on termini modeling.

Overall, we believe that the assumptions made here are reasonable considering the level of resolution of the coarse-grained model, and that such details are unlikely to play a major role in the resulting RMSF.[25]

Replica-exchange coarse-grained simulations in implicit water were run at temperatures $k_{\mathrm{B}}T/\mathcal{E} \in \{1.0, 1.05, 1.1, 1.2, 1.3, 1.4, 1.6, 1.9\}$ over a total simulation time of $10^7\,\tau$ in each replica. Initial conformations were randomly selected, and the low temperature replicas quickly stabilized an α-helix. We extracted the RMSF of the first replica ($k_{\mathrm{B}}T = 1.0\,\mathcal{E}$) by discarding the first $2 \times 10^6\,\tau$ and splitting the remaining data into two independent sets (from which we can calculate a mean and standard deviation for each residue). The results are shown in Figure 6.11 ("water CG"). We note that the corresponding atomistic and coarse-grained simulations in water agree well considering that no temperature calibration was applied.[26] It isn't clear why the coarse-grained simulations show enhanced fluctuations at the ends of the chain (i.e., residues 1–4 and 17–20), compared to the atomistic data: all backbone-hydrogen bonds are modeled in the coarse-grained system (see above). The peak around residues 12–15 illustrates the added flexibility due to both Gly_{11} and Pro_{14}—the role of proline in the conformation of the helix is explained in detail in Tieleman *et al.* [TSB99].

Coarse-grained simulations of Alm in a 72-POPC lipid membrane[27] were run at constant temperature, $k_{\mathrm{B}}T = \mathcal{E}$, and zero lateral tension, $\Sigma = 0$, for $500\,000\,\tau$ (see subsection 6.1.1 for more details on the simulation protocol).[28] A helical conformation of Alm (sampled from the previous water simulation) was inserted in an equilibrated lipid bilayer, *without* removing lipids. To relax the strong steric clashes due to the insertion of Alm, the peptide was initially restrained while the lipids were first warmed up, and then evolved freely for $100\,\tau$.[29] The peptide was subsequently unrestrained and the production run followed. Alm did not show any significant change of secondary structure over the entire simulation. Like

[25]It is worth noting that the atomistic simulation of Tieleman *et al.* is more than a decade old. Force fields evolve at a fast rate and one should be careful when relying on "old" simulation data. Here, we are only interested in the RMSF, a rather coarse-grained order parameter of the system. We assume it remains robust against small force field inaccuracies.

[26]The limited transferability of generic coarse-grained models—such as the one presented in chapter 2—often require a small temperature rescaling of the simulation in order to best reproduce experimental/atomistic data.

[27]While the use of a larger membrane would have been beneficial, a smaller system allows for better statistics. We assume here that the results presented below would not vary significantly by using a larger number of lipids.

[28]Kinetic simulations of folding and lateral diffusion for the peptide and lipid models, respectively, both yielded a speedup factor of 10^3. Since $\tau \sim 0.1\,\mathrm{ps}$ for both models [BD09, WD10b], the simulation time used here roughly corresponds to $\sim 50\,\mu\mathrm{s}$ of "real" time.

[29]While the peptide force field requires an integration time-step $\delta t = 0.01\,\tau$ (see chapter 2; [BD09]), the lipid force field can be run at $\delta t = 0.1\,\tau$ [WD10b]. This means that (*i*) protein-lipid simulations must be run at the smaller time-step, and (*ii*) simulations of restrained proteins in a membrane may be run at the faster time-step. Even though a multiple-time-step algorithm [KPD97] would be appropriate in the present context, the proper modification of the modified Andersen barostat [KD99] is (up to the author's knowledge) still an open question.

the simulation in water, the RMSF was extracted by discarding the first $100\,000\,\tau$ and splitting the remaining data into two data sets. The results, shown in Figure 6.11, agree remarkably well with the atomistic data of Tieleman *et al.* [TSB99]. We point out that no free parameter was tuned to reproduce the atomistic curve.

On the bright side, the results show that the coarse-grained model is robust enough to reproduce the difference in fluctuations between water and membrane environments, even though it couples an implicit with an explicit solvent. On the other hand, two very different processes are strongly contributing to the results: (*i*) the change in hydrogen-bond strength in water and the membrane, and (*ii*) interactions between the peptide and ordered lipid tails (i.e., nonzero orientational P_2 order parameter; this effect is present in the coarse-grained system [WD10b]) which might help aligning the helix. These contributions are difficult to disentangle on the basis of the results presented in Figure 6.11 alone. Increasing the hydrogen-bond strength $\epsilon_{\rm hb}$ in a membrane environment would likely increase the fraction of α-helices sampled, since they maximize the number of backbone hydrogen bonds [PCB51, PE90]. This is in agreement with the observation that most transmembrane proteins are helical (e.g., [Xio06]). Because the model was parametrized against helical protein structures, and because β-sheets are, in general, more difficult to stabilize (due to end effects; see chapter 2), we argue that the coarse-grained peptide force field parametrized for water is also adequate for the membrane environment.[30]

6.2.2 Tilt and hydrophobic mismatch

The tilt angle of model transmembrane proteins—an indicator of its orientation relative to the membrane normal—has recently been the subject of detailed studies in order to better understand hydrophobic mismatch between lipids and proteins [MB94]. The orientation of transmembrane WALP peptides, which consists of a hydrophobic stretch of alternating Leu and Ala bound by two pairs of Trp residues (sequences shown in Table 6.2), was estimated experimentally from ^{2}H solid-state NMR experiments [ÖRLK05, SÖR$^+$04]. The results of these authors showed an increase of the tilt angle upon membrane thinning (i.e., *positive* hydrophobic mismatch) even though the tilt angle values they measure were surprisingly small ($\approx 5\,^\circ$ for DMPC and DOPC). Atomistic and coarse-grained simulations, on the other hand, predicted much larger tilt angles ($\approx 15-30\,^\circ$ for the same lipids) (e.g., [ÖEKF07, KI10, MTF10, MKP$^+$08]). Özdirekcan *et al.* showed that the discrepancy was due to an averaging artifact of the NMR data [ÖEKF07]. Fluorescence spectroscopy measurements later confirmed the predictions from simulations [HKRM$^+$09].

Coarse-grained simulations of the WALP23 and WALP27 peptides were run in a 72-POPC lipid bilayer. Constant temperature ($k_{\rm B}T = \mathcal{E}$) and tension ($\Sigma = 0$) simulations were performed as above (subsection 6.2.1). Because of the long autocorrelation time involved in the relaxation of the abovementioned tilt angle—$50\,000\,\tau$ and $70\,000\,\tau$ for WALP23 and WALP27, respectively (data not shown)—simulations were run for $2-3\,\times$

[30]Future simulations may indicate the need for a careful reparametrization of the force field.

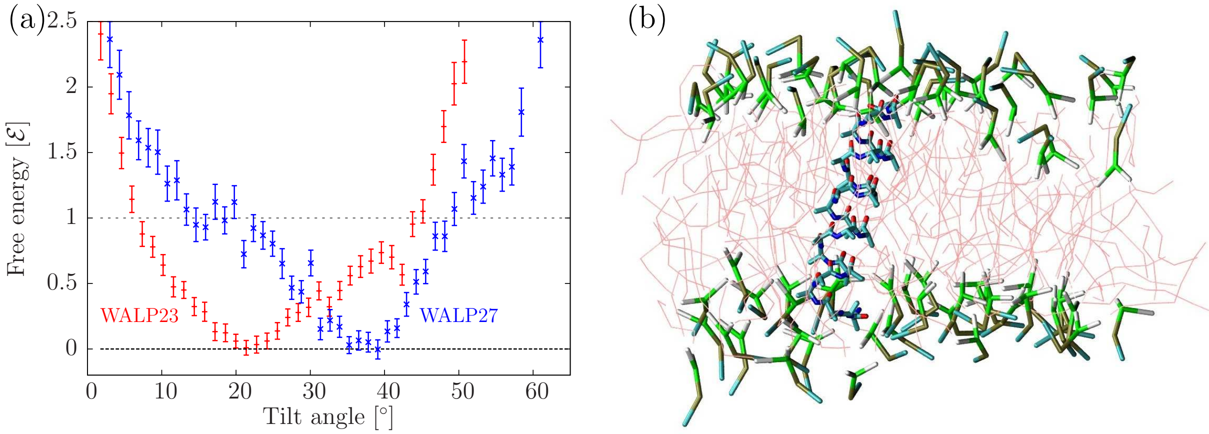


Figure 6.12: (a) Free energy as a function of tilt angle for WALP23 (red) and WALP27 (blue) in POPC. Horizontal lines denote $F = 0$ and $F = 1\,\mathcal{E}$ (i.e., thermally accessible interval). Error bars reflect the variance of the data points ($1\,\sigma$ interval). (b) Representative conformation of WALP23 (thin lines: lipid tails; licorice: lipid head groups and protein).

$10^6\,\tau$ to allow proper sampling.[31] The tilt angle was measured from the orientation of the helical principal axis (as calculated from the gyration tensor) and the unit vector along the membrane normal. Sampled distributions were inverted into free energies, as shown in Figure 6.12.[32] Error bars were estimated from bootstrap resampling (subsection A.1.6). The results clearly illustrate the impact of hydrophobic mismatch on the orientations of the two peptides: WALP27, being longer (i.e., larger positive hydrophobic mismatch), shows a more pronounced tilt to optimize hydrophobic matching [MB94]. The results are in good agreement with the atomistic, umbrella sampling simulations of Kim and Im [KI10], who measured thermally accessible tilt angles in the range $7-26\,°$ and $14-46\,°$ for WALP23 and WALP27, respectively, in POPC. While using different lipids, several independent, experimental and simulation studies point to an average tilt angle of $\approx 15-25\,°$ for WALP23 in DOPC [MTF10, HKRM+09]. Similar results were obtained in DMPC [ÖEKF07, KI10]. While small deviations are observed from one experimental method or simulation force field to the other, the results presented here are in good agreement with the published data. This shows that the model is capable of reproducing simple structural aspects of transmembrane protein orientation and, more generally, hydrophobic mismatch.

[31] Because of these long autocorrelation times, all-atom standard canonical simulations are unable to converge the distribution of tilt angles [MKP+08]. The use of more sophisticated sampling techniques, such as umbrella sampling, seems to alleviate the problem [KI10].

[32] A quadratic expansion of the WALP23 free energy profile around its minimum, $F = \frac{1}{2}c(\alpha - \alpha_0)^2$, where α is the tilt angle, α_0 the minimum value, and c the associated modulus, allows to estimate the restoring torque $N = \partial F/\partial\alpha$ exerted by the helix. We find $c \approx 30\,\mathcal{E}/\mathrm{rad}^2$ (in comparison, bond angles in the protein model exert a modulus $k_{\mathrm{angle}} = 300\,\mathcal{E}/\mathrm{rad}^2$; see Table 2.1 on page 17).

6.2.3 Helix-helix interactions

The aggregation of proteins in, or close to, the lipid bilayer may have important biological consequences for the membrane, e.g., membrane-curving proteins and vesicle budding [BV06, RIH+07], pore formation [OS99]. These phenomena depend not only on protein-lipid interactions, but also protein-protein interactions *in the membrane environment*. The self-association of WALP peptides in model membranes—studied both experimentally and computationally [SAN+05, MKP+08]—provides an appropriate benchmark to test the coarse-grained force field by studying the distance and crossing (i.e., relative) angle of WALP dimers.

We simulated WALP23 dimers in a 72-POPC lipid bilayer at constant temperature and tension. All simulation conditions were the same as above. Two independent simulations, totaling $10^6\,\tau$, were run with helical peptides initially placed in *parallel*, integral to the bilayer, and at a 13 Å distance of one another. Helix-helix distances were measured from the centers of mass of the two peptides, while the crossing angle was defined from the angle between the two helical principal axes. Figure 6.13 shows the free energies as a function of the helix-helix distance (a) and crossing angle (b) between the two peptides. These results compare well with atomistic simulations of WALP dimers in DOPC which report an average distance of 11 Å and angle $15-25\,^\circ$ [SAN+05] for two parallel dimers. Monticelli *et al.* performed similar test simulations with the MARTINI force field for the *antiparallel* configuration and found an average distance of 7 Å, compared to $8-9$ Å atomistically [MKP+08, SAN+05]. The difference in helix-helix distance between parallel and antiparallel dimers has been argued to stem from dipolar interactions between the two helices—which should indeed favor antiparallel dipoles (e.g., [Hol85, SAN+05]). Because the present force field does not model explicit electrostatics, we expect this model *not* to reproduce this feature. Indeed, dimer simulations of the antiparallel configuration show that the average distance and angle (Figure 6.13 (c) and (d)) are virtually identical to the abovementioned parallel packing scenario. Differences in the shape of the distributions between parallel and antiparallel packing may be caused by a smaller amount of statistics in the latter case (only $625\,000\,\tau$ of simulation time).

6.2.4 Insertion and folding

The ability of the model to fold simple peptides (chapter 2) provides the means to study interfacial folding and membrane insertion. This approach has been used before on the WALP peptide using atomistic simulations with implicit [IBI05] and explicit [NWG05] membrane environments. In both cases, a parallel tempering scheme was applied to improve sampling. In the *explicit* membrane case, restraining potentials were applied to the lipid head groups in order to keep the bilayer stable. The results showed that folding is *not* required for bilayer insertion (unlike what others have proposed, e.g., [JW89]). Instead, the unstructured peptide first inserts into the bilayer and then folds into an α-helix. Because of obvious computational limitations, the system was limited to a short peptide (WALP16;

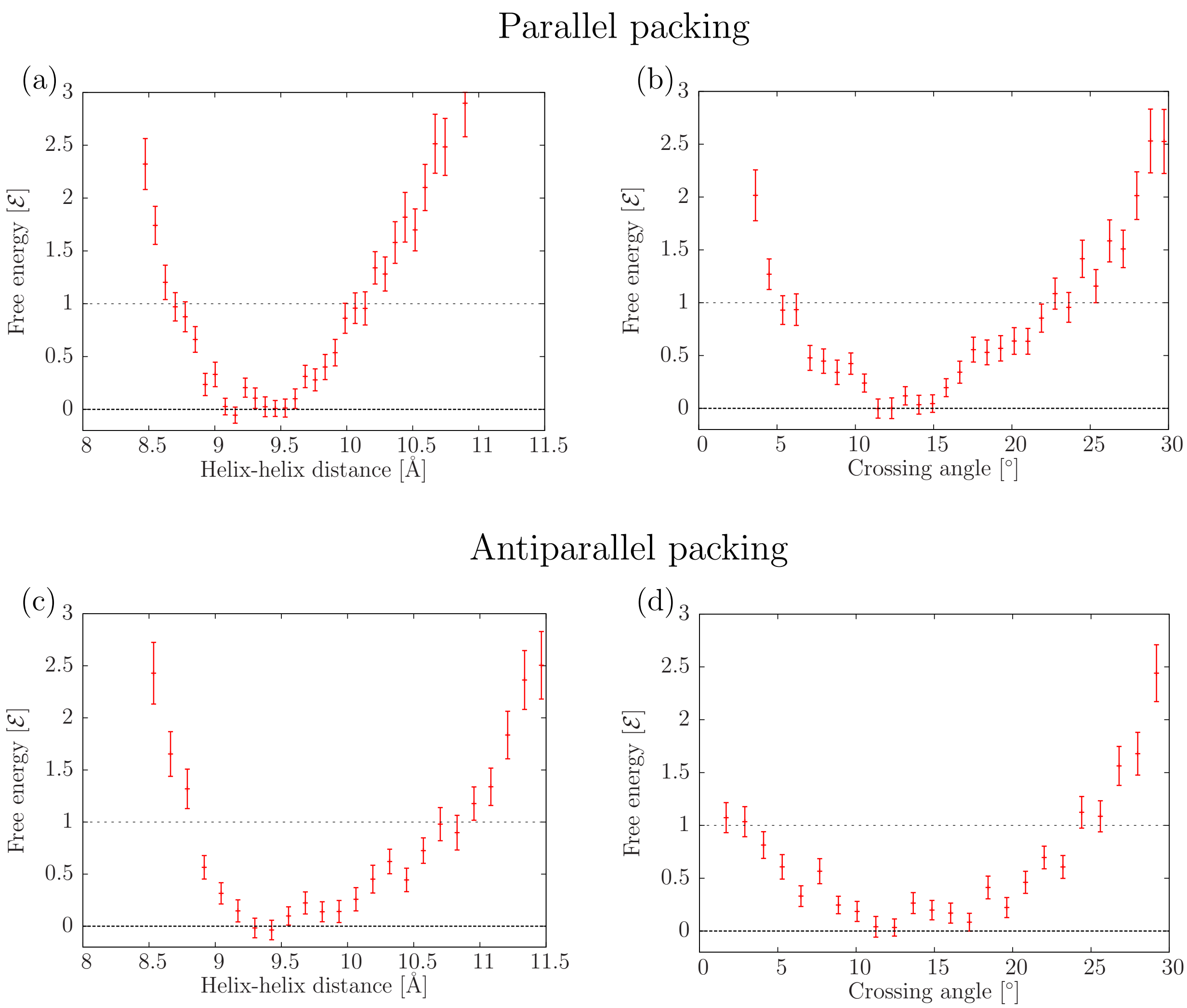


Figure 6.13: Free energies between two WALP23 peptides in a POPC bilayer as a function of the helix-helix distance, (a) and (c), and crossing angle, (b) and (d), for parallel and antiparallel packing, respectively. Horizontal lines denote $F = 0$ and $F = 1\,\mathcal{E}$ (i.e., thermally accessible interval). Error bars reflect the variance of the data points ($1\,\sigma$ interval). Helix-helix distance distributions average over all crossing angles, and vice-versa. The amount of sampling obtained here is not enough to produce two-dimensional free energy surfaces as a function of both helix-helix distance and crossing angle.

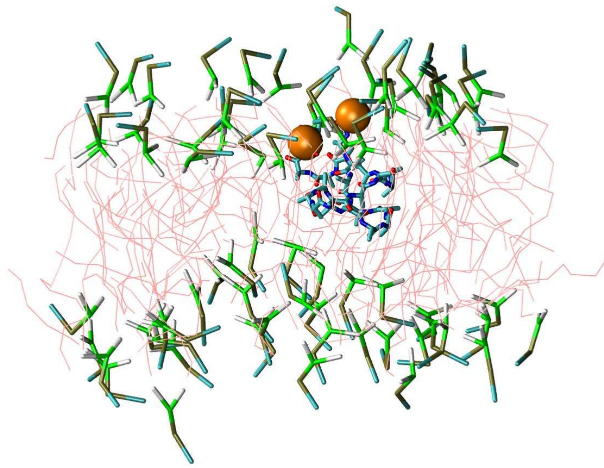


Figure 6.14: Insertion and folding of one WALP peptides on a 72-POPC lipid bilayer. Simulation snapshots recorded at $t = 1.5 \times 10^6\,\tau$. Thin lines: lipid tails; licorice: lipid head groups and protein; orange beads: N- and C-termini.

Table 6.2), a small bilayer (36 DPPC lipids), and 3.5 ns of simulation time. The implicit membrane simulation, on the other hand, studied various peptides for longer simulation times (besides, reaching equilibrium was facilitated by the absence of molecular friction in the membrane environment).

Here, we study the insertion of unfolded WALP23 peptides in a 72-lipid POPC bilayer. Two independent canonical simulations (same conditions as above) were first run with a single WALP peptide. Each of them was initially set in an unfolded conformation, placed in the aqueous environment (i.e., above—but close to—the lipid bilayer). While the peptide quickly binds to the bilayer, it does not insert easily. Figure 6.14 shows the conformation of the system after $t = 1.5 \times 10^6\,\tau$: most of the peptide inserts into the hydrophobic region of the bilayer[33] while the N- and C-termini remain in the lipid head-group region, due to their hydrophilic nature (Figure 6.10); the other simulation evolved similarly. In comparison with the average folding time of $(\mathrm{AAQAA})_3$ in an aqueous environment ($\lesssim 10\,000\,\tau$ at $T = \mathcal{E}/k_\mathrm{B}$; Table 3.1 on page 53), folding in a membrane environment is substantially slowed down. Neither simulation shows any sign of peptide folding *at the membrane interface*. The folding/insertion process is frozen by the free energy barrier of carrying one (N- or C-)terminus across the bilayer.[34]

A second set of two simulations were run with four WALP peptides on the same lipid bi-

[33] Note that Leu and Ala, which make for most of the WALP peptide (Table 6.2), are hydrophobic residues.

[34] Obviously, the kinetics of insertion of the N- and C-termini will strongly depend on the associated PMF (Figure 6.10). A more accurate parametrization of these groups may lower the PMF and the average insertion time.

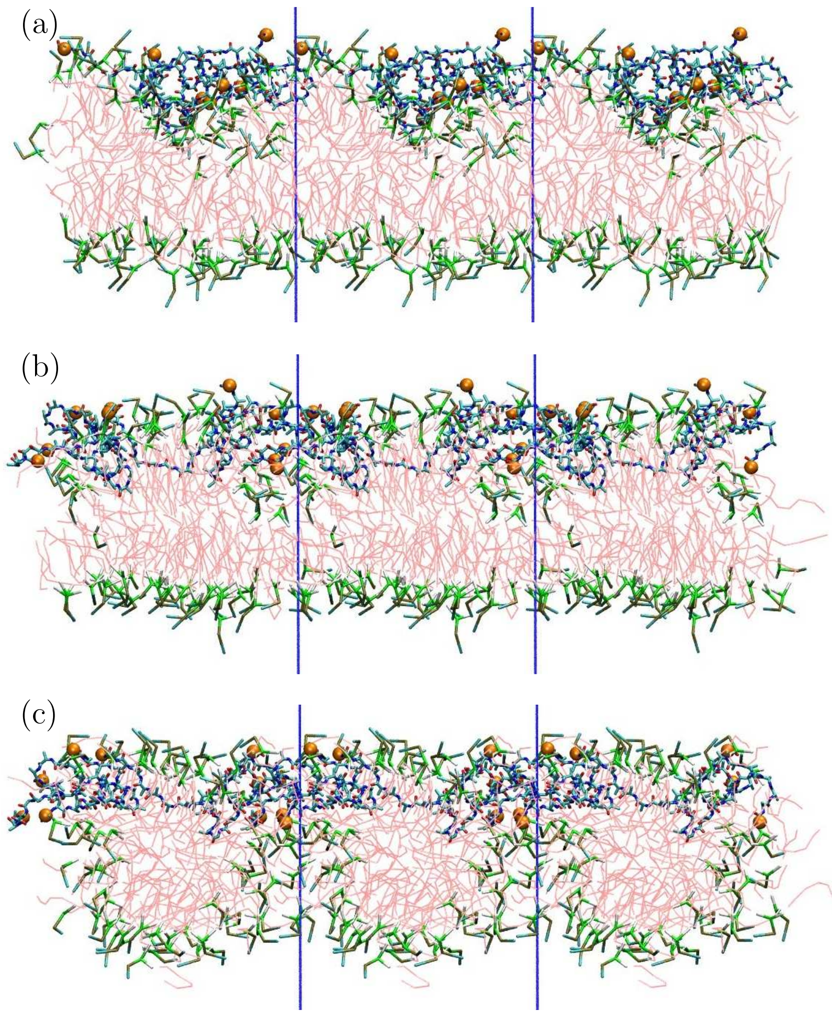


Figure 6.15: Insertion and folding of four WALP peptides on a 72-POPC lipid bilayer. Simulation snapshots recorded at $t = 50\,000\,\tau$ (a), $t = 450\,000\,\tau$ (b), and $t = 600\,000\,\tau$ (c), respectively. Note the increasing deformation of the lower leaflet of the bilayer. Blue, vertical lines mark the periodic boundaries of the simulation box (3 identical unit cells are shown). Thin lines: lipid tails; licorice: lipid head groups and protein; orange beads: N- and C-termini.

layer. As shown in the conformation of one of the simulations at $t = 450\,000\,\tau$ (Figure 6.15), the higher peptide density provides two key features:

- Intermolecular peptide-peptide interactions tend to favor extended conformations inside the bilayer. The larger number of neighboring peptide backbone chains seems to be stabilizing more hydrogen bonds, compared to the one-peptide simulation. The peptides have an increased tendency to align *along* the membrane interface. Nevertheless, the simulations show no sign of helix formation while peptides are adsorbed on the lipid bilayer.

- The peptides strongly affect the stability of the bilayer. We observe strong displacements of termini-neighboring lipid head groups towards the hydrocarbon region of the bilayer (see Figure 6.15). These head groups *could* provide a hydrophilic shell around a peptide N- or C-terminus across the bilayer, thereby reducing the associated free energy of penetration (Figure 6.10). Whether properly inserted peptides would subsequently fold in the membrane remains an open question.

While running longer simulations would ultimately answer these questions,[35] the significant correlation times involved seem prohibitive. Notice that none of the insertion simulations presented here showed any sign of helix formation in the membrane, which suggests inadequate sampling. Helix formation alone might be more easily observed by initially setting a random coil conformation with its hydrophilic end groups located on each side of the bilayer. In terms of insertion, an ingenious protocol was recently proposed in which peptides are initially placed in a random dispersion of lipids, and stabilize either in or out of the self-assembling bilayer [EMS07]. Averaged over many simulations, this method allows to calculate the ratio of surface-bound vs. transmembrane proteins [MKP$^+$08]. The kinetics of insertion might be best probed by the use of sophisticated techniques such as transition path sampling [BCDG02] or forward-flux sampling [AWtW05]. In terms of thermodynamic properties, parallel tempering has shown valuable for protein-lipid systems (e.g., [NWG05]), even though the use of restraining potentials for the bilayer is likely to prevent potentially informative membrane deformations (see Figure 6.15). Finally, an elegant solution might be the use of Hamiltonian replica exchange molecular dynamics (HREMD) [BD00]. While similar in spirit to parallel tempering, HREMD only decouples the important degrees of freedom (rather than all of them—through the temperature). One could imagine a scheme in which the protein-lipid interactions are successively lowered: the first replica would represent the original Hamiltonian, whereas proteins and lipids would not interact with one another in the last one. This scheme remains to be implemented in the present context.

[35] But then, *how long* is long enough?

Conclusions

The formation of structure in protein molecules was investigated using coarse-grained simulations. The level of resolution of the model—which allows to capture local conformations—is able to fold simple peptides without any primary sequence dependent bias, while gaining much computational efficiency compared to atomistic models. The model was first applied to various biophysical problems: kinetics of folding, protein folding cooperativity, and amyloid aggregation. In the last chapter, the peptide model was cross-parametrized with a high-resolution, solvent-free coarse-grained lipid model in an attempt to study the interactions of proteins with the lipid membrane.

One recurring question associated with the development and use of a coarse-grained model concerns its range of applicability and accuracy. The set of assumptions that are put in the model (or, in general, any theory) must be compatible with the problem at hand. For instance, it would be inconceivable to use the present peptide model to study hydrodynamic properties of proteins, side-chain hydrogen bonding, or electrostatics. In terms of accuracy, the success of a coarse-grained model will rely on the judicious choice and parametrization of its degrees of freedom. It can then be tested against simple scenarios for which the result is known, before applying the model to new problems. It is worth noting that the proper sampling of a potentially inaccurate coarse-grained simulation still provides a *falsifiable*—and thus scientific—result. On the other hand, the corresponding atomistic simulation may only allow for very poor statistics (for large systems), overall providing uncontrolled errors.

Clearly, the goal of biophysics is to shed light on biological phenomena using physics-related tools (e.g., statistical mechanics, thermodynamics). The author hopes that the present thesis complies with this objective, at least partially. Certainly, the *physics* involved is interesting in its own right. For instance, detailed mechanisms of protein folding cooperativity were clearly identified using a microcanonical analysis (chapter 4), a technique which originated from the thermodynamic analysis of finite-size transitions. These are very exciting times for biophysics: recent technological advancements in terms of experimental resolution and computational capabilities are rapidly closing the gap (i.e., experimental techniques can resolve finer details while simulations reach longer time- and length-scales) to provide an unprecendented amount of insight and understanding.

A Histogram Reweighting Techniques

A.1 Formalism

The output data of computer simulations (e.g., Monte Carlo, molecular dynamics) consists of trajectories of the system studied. These may then be analyzed in terms of time or ensemble averages of various observables. Both types of averages will coincide assuming (*i*) sufficient sampling and (*ii*) ergodicity: observing a process for a long time is equivalent to sampling many independent realizations of the same process.[1] The calculation of statistical quantities is often characterized by the determination of moments X^n of the (unknown) underlying distribution function p_X. For instance, the *Binder cumulant* allows to locate the critical point of a statistical system from ratios of moments. In the case of the Ising model with zero external field, the Binder cumulant is obtained from the second and fourth moments of the magnetization [Bin81]. Rather than characterizing a distribution through its moments, more recent developments aims to determine the distribution itself, namely by sampling suitable *histograms*. The present chapter summarizes important concepts underlying so-called *histogram reweighting techniques* and provides implementation details of several algorithms.

A.1.1 Estimators and distribution functions

We consider the simulation of a system at constant temperature T. While all microstates have equal weight(s) in the microcanonical ensemble, their weight in the canonical ensemble is proportional to the Boltzmann factor $\exp(-\beta E)$, where $\beta = 1/k_\mathrm{B}T$. The calculation of the expectation value of any observable Q at inverse temperature β can formally be expressed as

$$\langle Q\rangle(\beta) = \frac{\sum_\mu Q_\mu \mathrm{e}^{-\beta E_\mu}}{\sum_\mu \mathrm{e}^{-\beta E_\mu}}, \tag{A.2}$$

[1] For a random variable X with probability density p_X, the ergodic theorem [Pet83] implies for the n^{th} moment of X:

$$\mathrm{E}[X^n] = \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathrm{d}t\; x^n(t) = \int_{-\infty}^{\infty} \mathrm{d}x\; x^n p_X(x), \tag{A.1}$$

where the first and second integrals describe a time and ensemble average, respectively. Molecular dynamics simulations numerically integrate equations of motion, thus providing a time average. Monte Carlo simulations rely on a Markov process to draw configurations from a predefined probability density—they also generate a time average even though the trajectories may contain unphysical moves [NB99].

where the sums run over all microstates μ of the system, and $\langle \cdot \rangle$ corresponds to a canonical average.[2] The expectation value is expressed as the average over all Q_μ weighted by the Boltzmann factor. The denominator in Equation A.2 is called the (canonical) partition function of the system. Explicitly calculating $\langle Q \rangle$ by summing over all microstates is only tractable for very small systems.

In larger systems, the calculation of the average must be restricted over a subset of states $\{\mu_1, \ldots, \mu_M\}$. This introduces a *new* distribution function p_μ which specifies the probability with which a state μ is picked during sampling. The quantity

$$Q^{(M)}(\beta) = \frac{\sum_{i=1}^{M} Q_{\mu_i} p_{\mu_i}^{-1} \mathrm{e}^{-\beta E_{\mu_i}}}{\sum_{j=1}^{M} p_{\mu_j}^{-1} \mathrm{e}^{-\beta E_{\mu_j}}} \tag{A.3}$$

is an *estimator* of $\langle Q \rangle$. Its properties include (i) absence of bias, such that the error between an ensemble average over $Q^{(M)}$ and the parameter being estimated, $\langle Q \rangle$, is zero, (i.e., $\langle Q^{(M)} \rangle = \langle Q \rangle$), and ($ii$) consistency, where the estimator converges in probability to $\langle Q \rangle$ (i.e., $\lim_{M\to\infty} Q^{(M)} = \langle Q \rangle$) [Sha03, NB99]. Unlike Equation A.2, the sum in Equation A.3 does not run over all microstates but only over *sampled* microstates.

It is important to distinguish between the Boltzmann factor—inherent to any canonical sampling—and p_μ, which is an artificial property used to estimate Equation A.2. The distribution p_μ can therefore be freely chosen.[3] It is often set equal to the Boltzmann factor such that Equation A.3 reduces to

$$Q^{(M)}(\beta) = \frac{1}{M} \sum_{i=1}^{M} Q_{\mu_i, \beta}, \tag{A.4}$$

where $Q_{\mu_i,\beta}$ corresponds to observables sampled canonically at inverse temperature β.

However, many methods rely on a cleverer choice of this distribution function to enhance statistical sampling, such as several equilibrium and non-equilibrium free energy calculation techniques (e.g., Umbrella sampling [TV77], Multicanonical [BN91] and Wang-Landau [WL01], Adaptive Biasing Force [DP01], and Metadynamics [MLP04]). Alternatively, setting p_μ to the Boltzmann factor may be used to estimate properties of the system at a *different* state point than the one used in the simulation. This forms the basis of the single and multiple histogram methods presented next.

A.1.2 The single histogram method

The single histogram method, introduced by Ferrenberg and Swendsen [FS88], exploits the abovementioned duality between the distribution function that describes the thermodynamic state point (e.g., Boltzmann factor in the canonical ensemble) and the sampling

[2] Here we mostly follow the notation of Newman and Barkema [NB99].

[3] Practically, imposing a probability distribution function p_μ in a simulation is not necessarily trivial. While Monte Carlo simulations provide a direct control of the acceptance criterion (e.g., $p_\mu = \exp(-\beta E)$ corresponds to the Boltzmann factor as proposed in the original Metropolis algorithm), molecular dynamics requires the expression of p_μ into biasing *forces*.

scheme. The key concept in the method is that Equation A.3—which provides the exact value of $\langle Q \rangle$ at inverse temperature β in the limit $M \to \infty$ [NB99]—contains a large amount of information about the expectation value of Q at a *neighboring* inverse temperature β'.[4] Using Equation A.3 we formulate an expression for the estimator of $\langle Q \rangle$ *sampled* at inverse temperature β (i.e., $p_\mu \propto \exp(-\beta E)$) but *analyzed* at a neighboring inverse temperature β' as

$$Q^{(M)}(\beta') = \frac{\sum_{i=1}^{M} Q_{\mu_i,\beta}\, \mathrm{e}^{-(\beta'-\beta)E_{\mu_i}}}{\sum_{j=1}^{M} \mathrm{e}^{-(\beta'-\beta)E_{\mu_j}}}. \tag{A.5}$$

While Equation A.5 represents the most fundamental equation of the single histogram method, it is best illustrated by replacing the instantaneous measurements recorded during the simulation (i.e., $Q_{\mu_i,\beta}$) by histograms. We present the special case where the observable is the energy (or one of its derivatives). The equation can be written (for $\lim_{M\to\infty}$)

$$\langle E \rangle(\beta') = \frac{\sum_E E\, H(E;\beta) e^{-(\beta'-\beta)E}}{\sum_E H(E;\beta) e^{-(\beta'-\beta)E}}, \tag{A.6}$$

where $H(E;\beta)$ is the histogram of the energies of the states sampled at inverse temperature β. Since evidently $H(E;\beta) \propto \Omega(E)\exp(-\beta E)$, where $\Omega(E)$ is the density of states, we see that Equation A.6 can also be written as:

$$\langle E \rangle(\beta') = \frac{\sum_E E\, \Omega(E) e^{-\beta' E}}{\sum_E \Omega(E) e^{-\beta' E}}. \tag{A.7}$$

It becomes clear that Equation A.6 divides the Boltzmann factor $\exp(-\beta E)$ from $H(E;\beta)$ in order to estimate the density of states $\Omega(E)$ alone, which is then multiplied by $\exp(-\beta' E)$ to reweight the overall distribution at inverse temperature β'.

A few remarks on the method must be made at this point:

- it is possible to estimate *how far* in temperature difference the single histogram method will yield reasonable results. Obviously, only regions which were significantly sampled ($H(E;\cdot) \gg 1$) can be reweighted (see [NB99] for more details),

- the single histogram method not only provides an analytic expression for the average energy $\langle E \rangle$ at a neighboring temperature, *any* derivative can be calculated analytically (i.e., without requiring numerical differentiation) because Equation A.6 is a sum of ratios of exponentials, and derivatives of exponentials merely bring additional multiplicative factors (i.e., $\exp(u)' = u' \exp(u)$). This proves useful in a variety of scenarios, e.g., calculating the temperature at which the specific heat curve peaks.

[4] The method is readily applicable to any other intensive parameter such as pressure or chemical potential. Here we use temperature as a didactic example.

A.1.3 The multiple histogram method (also known as WHAM)

The single histogram method is based on the estimation of the density of states $\Omega(E)$ from canonical energy histograms $H(E)$, where $H(E) \propto \Omega(E)\exp(-\beta E)$. Reweighting the histograms to neighboring temperatures is naturally limited by the width of the histogram itself. In order to extend the method to wider energy intervals, a naïve approach would consist of performing several simulations with overlapping histograms $H_i(E)$ such that $H_i(E) \gg 1$ for any E within an interval of interest. This method is suboptimal, in the sense that it does not take advantage of overlapping data: the density of states at an energy E significantly sampled by more than one simulation will be evaluated using only one histogram. One can intuitively recognize that complementary information is contained in overlapping histograms. Weighted averages of single histogram extrapolations have shown to be error-prone [NB99]. Rather than patching together individual estimates of the density of states, the error can be greatly minimized by setting up a framework in which all histograms contribute to the estimation of the *same* density of states (i.e., $\Omega(E)$ depends only on the system, and not on the temperature at which it is studied) such that the error of the estimated density of states is minimized.[5] This framework consists of a minimum variance estimator for the density of states, coined multiple histogram method in the physics literature [FS89] and Weighted Histogram Analysis Method (WHAM) in the "bio" community [KRB+92]. We skip the derivation of the method which can be found elsewhere, e.g., Newman and Barkema [NB99], Ferrenberg and Swendsen [FS89], Kumar *et al.* [KRB+92], Souaille and Roux [SR01], Bartels and Karplus [BK98].

For a set of R simulations at different inverse temperatures β_i, each of which sampled an energy histogram $H_i(E;\beta_i)$ with N_i data points, the estimator for the density of states yields[6]

$$\Omega(E) = \frac{\sum_i H_i(E;\beta_i)}{\sum_j N_j \mathrm{e}^{-\beta_j E - f_j}}. \tag{A.8}$$

where $f_j = -\beta_j F_j = \ln Z_j$ is the (unknown) *scaled* free energy of simulation j (F_j is the free energy and Z_j its partition function). These quantities can be determined iteratively by simple definition of the partition function

$$e^{f_k} = Z_k = \sum_E \Omega(E) e^{-\beta_k E} = \sum_E \frac{\sum_i H_i(E;\beta_i)}{\sum_j N_j e^{(\beta_k-\beta_j)E - f_j}}. \tag{A.9}$$

Having thus determined the f_j, the partition function[7] Z can now be evaluated at *any*

[5] The best estimate of $\Omega(E)$ is obtained by weighing each individual contribution (i.e., simulation) according to the number of samples in the corresponding histogram at that energy [NB99].

[6] Here we will neglect terms describing the correlation between data points $g_i = 1 + 2\tau_i$, where τ is the auto-correlation time. Because every histogram H_i gets multiplied by an associated term g_i, it is easy to see that all of them cancel when they are equal (this is, in general, not true).

[7] This quantity does not correspond to the *true* partition function of the system because one never exhaustively samples all of phase space. It rather indicates its value *relative* to the other simulations. Such an offset factor in the partition function is, of course, irrelevant when calculating thermodynamic observables.

interpolating temperature

$$Z(\beta) = \sum_E \frac{\sum_i H_i(E;\beta_i)}{\sum_j N_j e^{(\beta-\beta_j)E-f_j}}, \tag{A.10}$$

and provide continuous approximations to canonical averages for any sampled observables. Moreover, the scaled free energies $f_j = \ln Z_j$ give access to the density of states, as shown in Equation A.8. From the density of states, any thermodynamic quantity can now be calculated using a *microcanonical* description (i.e., the energy E is a control parameter).[8]

Consider now the canonical interpolation of an observable Q (i.e., expectation of Q at various temperatures). The evaluation of $\langle Q\rangle(\beta)$ will rely on the calculation of the probability distribution $p(Q) \propto \sum_E \Omega(E,Q)\exp(-\beta E)$,[9] where $\Omega(E,Q)$ is a two-dimensional density of states. Following Equation A.8, $\Omega(E,Q)$ may be evaluated from a series of two-dimensional histograms $H_i(E,Q;\beta_i)$ at different temperatures

$$\Omega(E,Q) = \frac{\sum_i H_i(E,Q;\beta_i)}{\sum_j N_j \mathrm{e}^{-\beta_j E-f_j}}. \tag{A.11}$$

The evaluation of $\langle Q\rangle(\beta)$ would then require sums over both E and Q:

$$\begin{aligned} \langle Q\rangle(\beta) &= \frac{1}{Z(\beta)} \sum_Q \sum_E Q\,\Omega(E,Q)\mathrm{e}^{-\beta E} & \text{(A.12)} \\ &= \frac{1}{Z(\beta)} \sum_Q \sum_E Q \frac{\sum_i H_i(E,Q;\beta_i)}{\sum_j N_j \mathrm{e}^{-(\beta-\beta_j)E-f_j}}. & \text{(A.13)} \end{aligned}$$

It is straightforward to rewrite Equation A.13 in terms of the instantaneous (sampled) states s during the i^{th} simulation, rather than the histograms $H_i(E,Q;\beta_i)$, such that

$$\langle Q\rangle(\beta) = \frac{1}{Z(\beta)} \sum_{i,s} \frac{Q_{i,s}}{\sum_j N_j e^{(\beta-\beta_j)E_{i,s}-f_j}}. \tag{A.14}$$

Avoiding the explicit use of histograms (and their inherent binning artifacts) is especially useful when dealing with continuous spectra (e.g., E and Q). As a practical example, the calculation of the average energy $\langle E\rangle$ at inverse temperature β is given by

$$\langle E(\beta)\rangle = \frac{1}{Z(\beta)} \sum_{i,s} \frac{E_{i,s}}{\sum_j N_j e^{(\beta-\beta_j)E_{i,s}-f_j}}. \tag{A.15}$$

Likewise, the canonical specific heat is easily obtained using moments of the energy distribution, $C_V(\beta) = k_{\mathrm{B}}\beta^2(\langle E^2\rangle - \langle E\rangle^2)$:

$$C_V(\beta) = \frac{k_{\mathrm{B}}\beta^2}{Z(\beta)} \left[\sum_{i,s} \frac{E_{i,s}^2}{\sum_j N_j e^{(\beta-\beta_j)E_{i,s}-f_j}} - \left(\sum_{i,s} \frac{E_{i,s}}{\sum_j N_j e^{(\beta-\beta_j)E_{i,s}-f_j}} \right)^2 \right]. \tag{A.16}$$

[8] Such an approach is used in chapter 4 to study thermodynamic aspects of protein folding cooperativity.
[9] $p(Q) \propto \int \mathrm{d}E\,\Omega(E,Q)\exp(-\beta E)$ for continuous energy spectra.

Finally, the multiple histogram method gives access to free energies as a function of any parameter Q at inverse temperature β (this is often referred to as a potential of mean force, or PMF). This can easily be derived by noting that the canonical probability density as a function of Q will only select conformations for which $Q_{i,s} = Q$, such that

$$F(Q) = -\beta^{-1} \ln p(Q) = -\beta^{-1} \ln \left[\frac{1}{Z(\beta)} \sum_{i,s} \frac{\delta(Q - Q_{i,s})}{\sum_j N_j e^{(\beta - \beta_j)E_{i,s} - f_j}} \right]. \tag{A.17}$$

This equation requires a binning of the variable Q within the interval over which the free energy will be calculated.

A.1.4 Optimized convergence of the free energies

The iterative convergence of the set of free energies f_k described in Equation A.9 may become slow for large systems (though it always converges exponentially). An alternative solution focuses on the free energy *difference* between neighboring histograms [BS09]. It is straightforward to show from Equation A.9 that a system consisting of only two simulations will yield

$$1 = \sum_E \frac{H_1(E;\beta_1) + H_2(E;\beta_2)}{N_1 + N_2 \exp[-\Delta\beta\, E - \Delta f]} \tag{A.18}$$

where $\Delta\beta = \beta_2 - \beta_1$ and $\Delta f = f_2 - f_1$ is the only unknown parameter.[10] The solution to the equation does not require any iteration and can simply be obtained numerically. The generalization of Equation A.18—originally derived by Bennett [Ben76]—to q neighboring histograms (i.e., q histograms on the left *and* q histograms on the right of Δf) leads to an efficient iterative solution for the full set of histograms. Rather than taking into account all histograms at once (as in the iterative scheme), each free energy difference $\Delta f_k = f_{k+1} - f_k$ will be iteratively refined by including additional neighboring histograms until full convergence (see Figure A.1 for details). Unlike the original Bennett equation, the incorporation of more neighbors will require an iteration of the solution in order for all values Δf_k to be consistent. The generalized Bennett equation for Δf_k with q neighboring histograms can be written

$$1 = \sum_E \frac{\sum_{i=k-q}^{k+1+q} H_i(E;\beta_i)}{\Omega_k(E)} \tag{A.19}$$

[10] There is only *one* unknown—rather than two—because free energies are only determined up to an arbitrary constant.

and

$$\begin{aligned}\Omega_k(E) &= N_k + N_{k+1}\exp\left[-\Delta\beta_k\, E - \Delta f_k^{\text{new}}\right] \\ &+ \sum_{m=k-q}^{k-1} \exp\left[(\beta_k-\beta_m)E + \sum_{j=m}^{k-1}\Delta f_j^{\text{old}}\right] \\ &+ \exp\left[-\Delta f_k^{\text{new}}\right] \times \sum_{m=k+2}^{k+1+q} N_m \exp\left[(\beta_k-\beta_m)E - \sum_{j=k+1}^{m-1}\Delta f_j^{\text{old}}\right]. \end{aligned} \tag{A.20}$$

Δf_k^{old} represents the previous evaluation of Δf_k whereas Δf_k^{new} represents the current one. The pair of equations A.19 and A.20 contains only a single unknown Δf_i^{new}, making it easy to solve numerically. The last iteration of the algorithm, which includes all neighbors, is equivalent to the solution of Equation A.9.

The algorithm performs especially well for histograms that have smaller overlaps because the correlation between neighboring histograms decays quickly. Apart from chapter 4,[11] all free energy calculations in the present thesis were converged using this method, achieving speedups of up to sixty-fold compared to the iterative scheme.

A.1.5 Umbrella sampling

Umbrella sampling [TV77], as mentioned in subsection A.1.1, offers a way to improve statistical sampling in regions of low probability by replacing the standard Boltzmann weight by a more appropriate function w. The estimator for the canonical expectation value of observable Q evaluated at inverse temperature β, but *sampled* from a distribution w, becomes (see Equation A.3)

$$Q^{(M)}(\beta) = \frac{\sum_{i=1}^{M} Q_{\mu_i} w_{\mu_i}^{-1} e^{-\beta E_{\mu_i}}}{\sum_{j=1}^{M} w_{\mu_j}^{-1} e^{-\beta E_{\mu_j}}}, \tag{A.21}$$

where w_{μ_i} is the evaluation of the function w for microstate μ_i.

To further introduce the Umbrella sampling technique, recall the problem of calculating the free energy of a system as a function of an order parameter Q (i.e., PMF, see subsection A.1.3) $F(Q) = -\beta^{-1}\ln p(Q)$. Let's assume the PMF to be strongly varying as a function of Q, such that a standard canonical simulation, with probability distribution $p(Q)$, fails to sample all values of the order parameter within an interval of interest. The introduction of a biasing weight function w is equivalent to an additional term $V(Q)$ in the Hamiltonian of the system $\mathcal{H} = \mathcal{H}_0 + V(Q)$,[12] such that the standard Boltzmann weight

[11] This work, which is based on a microcanonical analysis of helical peptides, required the accurate evaluation of $\Omega(E)$. This was achieved by simulating many replicas close to the transition temperature and thus lead to strong overlap between histograms. It has proven difficult to converge strongly overlapping histograms using the method presented in subsection A.1.4 due to stability issues.

[12] We assume Q to be a function of the system's coordinates (e.g., particle positions).

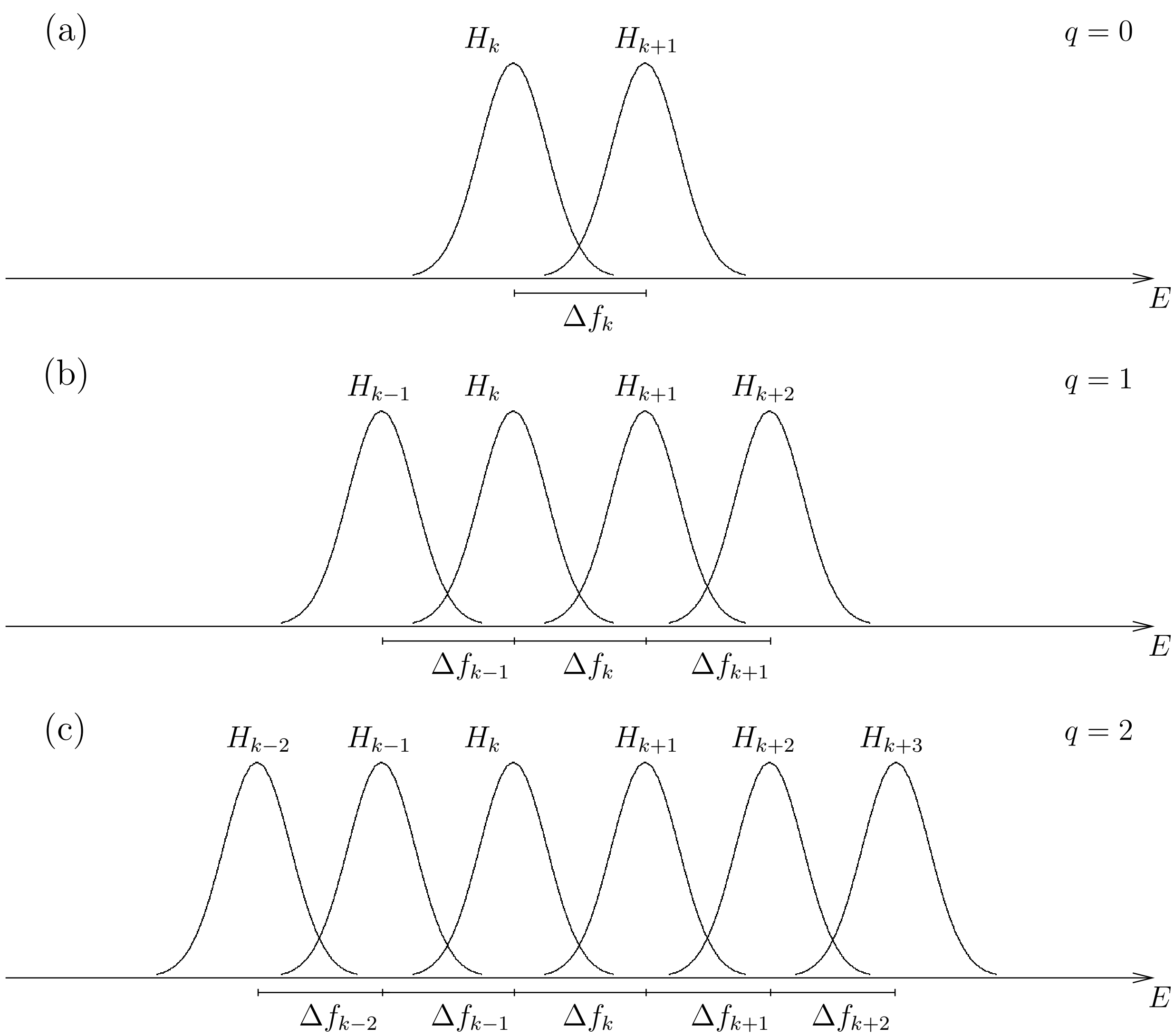


Figure A.1: Iterative refinement of the free energy difference Δf_k for (a) $q = 0$ which corresponds to the original Bennett equation (Equation A.18), (b) $q = 1$, and (c) $q = 2$, corresponding to one and two additional histograms on each side, respectively.

is replaced by $\exp(-\beta E) \times \exp(-\beta V)$. By properly choosing V, we can enhance sampling in the regions of low probability. Ideally, setting the potential $V(Q)$ such that it cancels $F(Q)$ will allow to sample all values Q with equal probability. This is easier said than done as $F(Q)$ is often the very quantity we wish to determine from this technique. Instead, the range of the order parameter is often split into small windows, each of which is sampled by the use of a harmonic restraint $V_i(Q) = \frac{1}{2}k_Q(Q - Q_i)^2$, where i runs over the number of windows.

Once the biased distributions $p_i^w(Q)$ have been sampled, it is possible to calculate the corresponding *unbiased* PMFs $F_i(Q)$ through the following expression

$$F_i(Q) = -V_i(Q) - k_{\mathrm{B}}T \ln p_i^w(Q) + C_i, \tag{A.22}$$

where C_i represent constant shifts between windows. Because we are only interested in free energy differences, the curves may be shifted such that the resulting PMF $F(Q)$ is continuous. While it is possible to shift the different curves by hand, the results tend to suffer from artifacts and do not take advantage of the data available in neighboring umbrellas. Instead, we again rely on results from the multiple histogram method, presented below.

Since its original introduction, Umbrella sampling has greatly benefitted from the development of the multiple histogram method. Kumar *et al.* showed that recovering the unbiased distribution $p(Q)$ by means of the multiple histogram method offers (i) the optimal set of free energies so as to minimize statistical errors and (ii) allows multiple overlaps of probability distributions for obtaining better estimates of $F(Q)$ [KRB+92].

The equations derived from the multiple histogram method to unbias Umbrella sampling simulations, analogous to Equation A.9, are

$$e^{f_k} = \sum_{i,s} \frac{1}{\sum_j N_j \exp[\beta_i V_{i,s} - \beta_j V_{j,s} - f_j]}, \tag{A.23}$$

where $V_{i,s}$ is the s^{th}-sampled value of the i^{th}-restraining potential V. This equation was used in chapter 6 to calculate the PMF curves of insertion of single amino acid side chains into a DOPC bilayer.

A.1.6 Error estimation and bootstrap resampling

The multiple histogram method provides an estimate for the relative error on the density of states, $\delta\Omega/\Omega$. In the case of unbiased energy histograms (subsection A.1.3) the following can be shown [FS89]

$$\frac{\delta\Omega(E)}{\Omega(E)} = \left[\sum_{i=1}^{R} H_i(E;\beta_i)\right]^{-1/2}, \tag{A.24}$$

such that the error is simply related to the amount of sampling. Similarly, the error on the density of states from Umbrella sampling simulations reads [KRB+92]

$$\frac{\delta\Omega(Q)}{\Omega(Q)} = \left[\sum_{i=1}^{R} H_i(Q;V_i)\right]^{-1/2}, \tag{A.25}$$

where $H_i(Q;V_i)$ corresponds to bin Q of the i^{th}-histogram with restraining potential V_i. In both cases the error on the density of states scales like $1/\sqrt{N}$, where N is the number of data points. Unfortunately, the multiple histogram method offers no expression for the *absolute* error on Ω.

Formally, error bars on $\delta\Omega$ can be estimated by running the same simulation a large number of times, determine Ω for each of them, and then calculate the standard deviation at small intervals of the associated variable (e.g., E, Q). This is often impractical due to limited computational resources. One alternative is to estimate the variance by using a statistical resampling method. Such methods use subsamples of the available data to calculate robust estimates of certain statistical estimators, such as the average and the variance. Famous examples include (*i*) the "bootstrap" method which estimates the robustness of a distribution by sampling with replacement from the original sample [Che08]—more precisely, the bootstrap method randomly draws N data points from a sample containing N elements, allowing points to be drawn multiple times, and (*ii*) the "jackknife" method where the precision of the data set is probed by systematically recomputing the variance leaving out one data point. The jackknife was shown to be a linear approximation method for the bootstrap [Efr79]. Both methods will yield exact estimates for the error of a measured quantity in the limit of infinite data set [NB99].

The bootstrap method was used to estimate error bars on densities of states (as well as related quantities such as the entropy $S(E) = k_{\mathrm{B}} \ln(E)$) and PMFs calculated in the present thesis.

A.2 Implementation of the multiple histogram method

This section describes several implementation aspects of the multiple histogram method and focuses specifically on the convergence of the unknown free energies f_k (Equation A.9). The calculation requires much care due to the exponential functions which appear in the denominator. Because computers can only represent numbers within a certain range, overflow errors easily arise when implementing the multiple histogram method.[13] Two complementary approaches have shown useful when implementing the algorithm:

- A constant shift applied to all the free energies will neither affect the final answer nor the rate of the convergence of the whole set. Shifting all f_k such that $\sum_k f_k = 0$ can avoid drifts towards unreasonably low or high values.

[13] These implementation issues can be avoided by using recently developed, arbitrary precision arithmetic libraries. See, e.g., `http://gmplib.org`.

- The program will result in an overflow error whenever the exponent $(\beta_k - \beta_j)E_{i,s} - f_j$ becomes large.[14] A simple technique easily solves the problem:[15] for any fixed values of k, i and s, scan the variable j in order to obtain the largest contribution $(\beta_k - \beta_j)E_{i,s} - f_j$, and keep track of this index $\zeta := j$. Now recast Equation A.9 in the following way

$$e^{f_k} = \sum_{i,s} \frac{1}{\sum_j N_j \exp\left[(\beta_k - \beta_j)E_{i,s} - f_j\right]} \frac{\exp\left[-(\beta_k - \beta_\zeta)E_{i,s} + f_\zeta\right]}{\exp\left[-(\beta_k - \beta_\zeta)E_{i,s} + f_\zeta\right]} \tag{A.26}$$

$$= \sum_{i,s} \frac{\exp\left[-(\beta_k - \beta_\zeta)E_{i,s} + f_\zeta\right]}{\sum_j N_j \exp\left[\{(\beta_k - \beta_j)E_{i,s} - f_j]\} - \{(\beta_k - \beta_\zeta)E_{i,s} - f_\zeta\}\right]} \tag{A.27}$$

  This makes sure the exponential factor in the denominator will be less than 1. The exponential in the numerator should be small since it corresponds to the inverse of the largest exponential factor (by definition of ζ). While this method is likely to create *underflows* due to contributions from small exponentials, it should not affect the final result. This technique was also applied on Equation A.23 to unbias umbrella simulations.

A.2.1 Iterative convergence of the free energies

Algorithm A.1 shows a `C` implementation of the iterative convergence of the free energies as described in Equation A.27. The `f_new` array (underlined throughout the code) contains the updated values of the quantities f_k. `argarray` stores the values of $(\beta_k - \beta_j)E_{i,s} - f_j$ for all j and `arg` represents the largest contribution (i.e., $j = \zeta$). `deltaF` calculates the summed absolute difference between all f_k values of step $n-1$ and step n. It is used on line 1 as a termination criterion of the `while` loop.

In order to speed up the calculation, the routine was parallelized using the OpenMP application programming interface.[16] The following piece of code was inserted between line 1 and 2 of Algorithm A.1.

```
#pragma omp parallel for private(j,i_HE,sumNum,sumDen,↩
                                  arg,k,argarray)
```

The arrow symbol "↩" means that the line continues (i.e., no line break). Parallelization in this case is trivial because the calculation of Δf_k is independent of $\Delta f_{k'}$ at a given iteration step.

[14] *How large* may depend on the programming language, compiler, and computer architecture. Also, note that the implementation avoids the use of histograms, and the equations sum over states rather than energy (see Equation A.14).

[15] R. H. Swendsen, personal communication.

[16] OpenMP (`www.openmp.org`) provides a remarkably simple interface to perform shared-memory parallel programming. The compiler flag required to activate it is `-fopenmp` in `gcc` and `-openmp` in `icc`.

Algorithm A.1: Convergence of the free energies by the iterative method

```
while (deltaF>TOL_ITER) {
  for (k = 0; k<N_SIMS; ++k){
    f_new[k] = 0.;
    for (i = 0; i<N_SIMS; ++i){
      for (s = 0; s<HIST_SIZES[i]; ++s){
        denominator  = 0.;
        arg          = -1e300;
        argarray     = calloc (N_SIMS, sizeof *argarray);
        /* Overflow trick: determine largest contribution */
        for (j = 0; j<N_SIMS; ++j){
          argarray[j] = (BETAS[k] - BETAS[j])
                        * HIST[i][s] - f_current[j];
          if (argarray[j]>arg)
            arg=argarray[j];
        }
        /* Calculation of f_new */
        for (j = 0; j<N_SIMS; ++j)
          denominator += HIST_SIZES[j]
                         * exp(argarray[j]-arg);
        numerator = exp(-arg);
        f_new[k] += numerator/denominator;
      }
    }
    f_new[k]      = log(f_new[k]);
    f_previous[k] = f_current[k];
    f_current[k]  = f_new[k];
  }
  deltaF = 0.;
  for (k = 0; k<N_SIMS; ++k )
    deltaF += fabs(f_new[k]-f_previous[k]);
}
```

A.2.2 Optimized convergence of the free energies

While extremely efficient, the implementation of the generalized Bennett equation (Equations A.19 and A.20) involves a number of technical caveats—many of which are described in [BS09]. Here we present a `C` implementation of the algorithm. The main routine that calculates the generalized Bennett equations is shown on Algorithm A.2. The outer-most `for` loop (line 1) iterates over the number of neighbors q incorporated in the evaluation of Δf_k. Previous and current evaluations of the k^{th}-free energy difference are stored in `f_new[k]` and `f_current[k]`, respectively. The next evaluation of Δf_k is provided by the function `halfinterval()`. Iterative refinements of the free energy differences are weighted by an update factor `UPDATE_COEFF` to avoid instabilities.

The evaluation of Δf_k is presented in Algorithm A.3. It numerically finds the solution to Equations A.19 and A.20, as calculated from the function `fermi()`[17] using the false-position method (see Technical Point A.1). Compared to other numerical root finding schemes, the false position method

- does not require the evaluation of derivatives (unlike, e.g., Newton-Raphson), and
- will always converge (unlike, e.g., the secant method).

The two initial conditions required to use the algorithm are determined by coming back to the original Bennett equation (Equation A.18). Note that there is a fundamental asymmetry in the equation when exchanging indices 1 and 2 (or, more generally, k and $k+1$):

$$1 = \sum_E \frac{H_k(E) + H_{k+1}(E)}{N_k + N_{k+1}\exp[-\Delta\beta_k\, E - \Delta f_k]}, \tag{A.28}$$

$$1 = \sum_E \frac{H_k(E) + H_{k+1}(E)}{N_k \exp[\Delta\beta_k\, E + \Delta f_k] + N_{k+1}}. \tag{A.29}$$

The two initial conditions for the false position method were set to simplified (and therefore approximate) analytical solutions of these equations, such that

$$\Delta f_k^a \approx -\ln \sum_E \frac{H_{k+1}(E)}{N_{k+1}\exp(-\Delta\beta_k\, E)}, \tag{A.30}$$

$$\Delta f_k^b \approx \ln \sum_E \frac{H_k(E)}{N_k \exp(-\Delta\beta_k\, E)}. \tag{A.31}$$

This is precisely what the function `init_fermi()` calculates,[18] as shown in Algorithm A.4. Note that while we do not prove here that Δf_k^a and Δf_k^b bracket the true solution (as required by the false position method), the method has, so far, never failed.

---

[17] The name is due to the functional similarity between Equation A.19 and the famous Fermi function in solid-state physics.

[18] Again, the function implements a variant that does not refer explicitly to energies but states in order to avoid histogram binning artifacts.

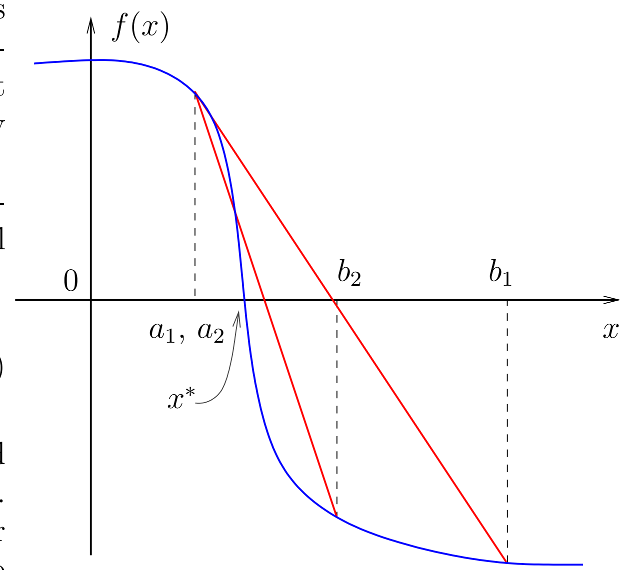
Technical Point A.1 The false position method

The false position method [BF05] is a root-finding algorithm that iteratively converges towards the solution $f(x^*) = 0$ from two initial values $x = a_1$ and $x = b_1$ such that $f(a_1)f(b_1) < 1$, and provided there is only one root in the interval $[a_1;\, b_1]$.

Iteration n of the algorithm first determines a new value d_n within the interval $[a_n;\, b_n]$

$$d_n = \frac{f(b_n)a_n - f(a_n)b_n}{f(b_n) - f(a_n)}, \tag{A.32}$$

where the line between $(a_n,\, f(a_n))$ and $(b_n,\, f(b_n))$ goes through $y = 0$ at $x = d_n$. This point is used alongside either a_{n+1} or b_{n+1} in the next iteration such that the two points are on opposite sides of the $y = 0$ line.



Last, the `fermi()` function, shown in Algorithm A.5, implements the calculation of Δf_k. The variable `den`, which calculates $\Omega_k(E)$ in Equation A.20, was underlined throughout the function. The function subtracts the result by 1 because the false position method looks for the solution of the equation $f(x^*) = 0$. An overall vertical shift of the function does not affect the result whatsoever.

Finally, the overall speedup of the algorithm can be greatly enhanced by performing parallel computation. This is simply done by adding the following piece of code between lines 6 and 7 of Algorithm A.2

```
#pragma omp parallel for reduction(+:converg_rate)
```

As before, the parallelization is trivial because the calculation of Δf_k is independent of $\Delta f_{k'}$ at a given iteration step.

Algorithm A.2: Convergence of the free energies by the optimized method

```
for (q=0;q<N_SIMS-1;++q) {
  /* q: neighbor level */
  converg_rate = 1.;
  iter_q       = 0;
  while (converg_rate>TOL_ITER){
    converg_rate=0.;
    for (k=0;k<N_SIMS-1;++k){
      if (iter==0 && q==0)
        /* Very first iteration */
        f_new[k] = halfinterval(q,k);
      else {
        f_current[k] = f_new[k];
        /* update factor to avoid unstability */
        f_new[k] = UPDATE_COEFF * halfinterval(q,k)
                   + (1-UPDATE_COEFF) * f_current[k];
      }
      converg_rate += fabs(f_new[k]-f_current[k]);
    }

    /* Test early exit condition
     * (more neighbors do not contribute)
     */
    if (q > 0 && converg_rate < TOL_ITER && iter_q == 0)
      q = N_SIMS;
    ++iter;
    ++iter_q;
  }
}
```

Algorithm A.3: Function `halfinterval()` used in the optimized method

```
double halfinterval(int q, int k)
{
  double a, b, fa, fb, d, fd;

  /* initial conditions */
  a=init_fermi(k, 1);
  b=init_fermi(k, 0);

  /* False position method */
  do{
    fa=fermi(q, k, a);
    fb=fermi(q, k, b);
    d=(fb*a-fa*b)/(fb-fa);
    fd=fermi(q, k, d);
    if (fa*fd>0) a=d;
    else         b=d;
  } while (fabs(fd) > TOL_FERMI);
  return d;
}
```

Algorithm A.4: Function `init_fermi()` used in the optimized method

```
double init_fermi(int k, int left)
{
  int s, j; double func, den, arg;

  if (left)  j = k+1;
  else       j = k;
  func=0.;

  for (s=0;s<HIST_SIZES[j];++s){
    arg=(BETAS[k]-BETAS[k+1])*HIST[j][s];
    den=NORM_HIST[j];
    if (left)  den*=exp(arg);
    else       den*=exp(-arg);
    func+=1./den;
  }
  if (left)    return -log(func);
  else         return log(func);
}
```

Algorithm A.5: Function `fermi()` used in the optimized method

```
double fermi(int q, int k, double x)
{
  double func, dbeta, den, dbm, deltafj;
  int i, s, m, k;

  func=0.;
  for (i=k-q; i<=k+1+q; ++i){
    if (i>=0 && i<N_SIMS){
      for (s=0; s<HIST_SIZES[i]; ++s)
        dbeta = (BETAS[k]-BETAS[k+1]) * HIST[i][s];
      den = NORM_HIST[k] + NORM_HIST[k+1] * exp(dbeta-x);
      if (q>0){
        for (m=k-q; m<=k-1; ++m){
          if (m>=0){
            dbm = (BETAS[k]-BETAS[m]) * HIST[i][s];
            deltafj = 0.;
            for (j=m; j<=k-1; ++j)
              deltafj += FENERGIES[j];
            den += NORM_HIST[m] * exp(dbm+deltafj);
          }
        }
        for (m=k+2; m<=k+1+q; ++m){
          if (m<N_SIMS){
            dbm = (BETAS[k]-BETAS[m])*HIST[i][s];
            deltafj = 0.;
            for (j=k+1; j<=m-1; ++j)
              deltafj += FENERGIES[j];
            den += NORM_HIST[m] * exp(dbm-deltafj-x);
          }
        }
      }
      func += 1./den;
    }
  }
  func -= 1.;
  return func;
}
```

List of Tables

List of Figures

List of Technical Points

List of Algorithms

Bibliography

[ACCDP10] D. Alemani, F. Collu, M. Cascella, and M. Dal Peraro, *A nonradial coarse-grained potential for proteins produces naturally stable secondary structure elements*, J. Chem. Theory Comput. **6** (2010), 315–324.

[AFS06] A. Arkhipov, P. L. Freddolino, and K. Schulten, *Stability and dynamics of virus capsids described by coarse-grained modeling*, Structure **14** (2006), no. 12, 1767–1777.

[AJL$^+$02] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell*, fourth ed., Garland Science, 2002.

[Anf72] C. Anfinsen, *The formation and stabilization of protein structure*, Biochem. J. **128** (1972), no. 4, 737–749.

[ANV07] G. S. Ayton, W. G. Noid, and G. A. Voth, *Multiscale modeling of biomolecular systems: in serial and in parallel*, Curr. Opin. Struct. Biol. **17** (2007), no. 2, 192–198.

[AT93] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*, Clarendon Pr., 1993.

[AWtW05] R. J. Allen, P. B. Warren, and P. R. ten Wolde, *Sampling rare switching events in biochemical networks*, Phys. Rev. Lett. **94** (2005), 018104 1–4.

[AYS08] A. Arkhipov, Y. Yin, and K. Schulten, *Four-scale description of membrane sculpting by BAR domains*, Biophys. J. **95** (2008), no. 6, 2806–2821.

[Bal89] R. L. Baldwin, *How does protein folding get started?*, Trends Biochem. Sci. **14** (1989), no. 7, 291–294.

[BBD10] T. Bereau, M. Bachmann, and M. Deserno, *Interplay between secondary and tertiary structure formation in protein folding cooperativity*, J. Am. Chem. Soc. **132** (2010), no. 38, 13129–13131.

[BBW$^+$06] A. Baumketner, S. L. Bernstein, T. Wyttenbach, N. D. Lazo, D. B. Teplow, M. T. Bowers, and J. E. Shea, *Structure of the 21-30 fragment of amyloid β-protein*, Prot. Sci **15** (2006), no. 6, 1239–1247.

[BCDG02] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Transition path sampling: Throwing ropes over rough mountain passes, in the dark*, Annu. Rev. Phys. Chem. **53** (2002), 291–318.

[BD00] A. Bunker and B. Dünweg, *Parallel excluded volume tempering for polymer melts*, Phys. Rev. E **63** (2000), 016701.

[BD09] T. Bereau and M. Deserno, *Generic coarse-grained model for protein folding and aggregation*, J. Chem. Phys. **130** (2009), no. 23, 235106–235120.

[Bea96] P. D. Beale, *Exact distribution of energies in the two-dimensional Ising model*, Phys. Rev. Lett. **76** (1996), 78–81.

[Ben76] C. H. Bennett, *Efficient estimation of free energy differences from Monte Carlo data*, J. Comput. Phys. **22** (1976), no. 2, 245–268.

[Bet09] M. R. Betancourt, *Coarse-grained protein model with residue orientation energies derived from atomic force fields*, J. Phys. Chem. B **113** (2009), 14824–14830.

[BF05] R. Burden and J. D. Faires, *Numerical analasis*, eighth ed., Brooks Cole, 2005.

[BHT+09] A. V. Badasyan, G. N. Hayrapetyan, S. A. Tonoyan, Y. S. Mamasakhlisov, A. S. Benight, and V. F. Morozov, *Intersegment interactions and helix-coil transition within the generalized model of polypeptide chains approach*, J. Chem. Phys. **131** (2009), no. 11, 115104–115111.

[Bin81] K. Binder, *Finite size scaling analysis of Ising model block distribution functions*, Z. Phys. B - Condensed Matter **43** (1981), no. 2, 119–140.

[BK92] C. Borgs and S. Kappler, *Equal weight versus equal height: a numerical study of an asymmetric first-order transition*, Phys. Lett. A **171** (1992), no. 1-2, 37–42.

[BK98] C. Bartels and M. Karplus, *Probability distributions for complex systems: Adaptive umbrella sampling of the potential energy*, J. Phys. Chem. B **102** (1998), no. 5, 865–880.

[BKL+03] Gal Bitan, Marina D. Kirkitadze, Aleksey Lomakin, Sabrina S. Vollers, George B. Benedek, and David B. Teplow, *Amyloid β-protein (Aβ) assembly: Aβ40 and Aβ42 oligomerize through distinct pathways*, Proceedings of the National Academy of Sciences of the United States of America **100** (2003), no. 1, 330–335.

[BN91] B. A. Berg and T. Neuhaus, *Multicanonical algorithms for first order phase transitions*, Phys. Lett. B **267** (1991), no. 2, 249–253.

[BO09] M. R. Betancourt and S. J. Omovie, *Pairwise energies for polypeptide coarse-grained models derived from atomic force fields*, J. Chem. Phys. **130** (2009), no. 19, 195103–195113.

[BOSW95] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Funnels, pathways, and the energy landscape of protein folding: a synthesis*, Proteins: Struct. Func. Genet. **21** (1995), no. 3, 167–195.

[BS06] P. J. Bond and M. S. P. Sansom, *Insertion and assembly of membrane proteins via simulation*, J. Am. Chem. Soc. **128** (2006), no. 8, 2697–2704.

[BS07] G. Bellesia and J.-E. Shea, *Self-assembly of β-sheet forming peptides into chiral fibrillar aggregates*, J. Chem. Phys. **126** (2007), no. 24, 245104–245114.

[BS09] T. Bereau and R. H. Swendsen, *Optimized convergence for multiple histogram analysis*, J. Comput. Phys. **228** (2009), no. 17, 6119–6129.

[BT99] M. R. Betancourt and D. Thirumalai, *Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes*, Prot. Sci. **8** (1999), no. 2, 361–369.

[BTS10] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, seventh ed., W. H. Freeman, 2010.

[BV06] P. D. Blood and G. A. Voth, *Direct observation of Bin/amphiphysin/Rvs (BAR) domain-induced membrane curvature by means of molecular dynamics simulations*, Proc. Natl. Acad. Sci. USA **103** (2006), 15068–15072.

[BVP11] G. R. Bowman, V. A. Voelz, and V. S. Pande, *Taming the complexity of protein folding*, Curr. Opin. Struct. Biol. **21** (2011), no. 1, 4–11.

[Caf06] A. Caflisch, *Network and graph analyses of folding free energy surfaces*, Curr. Opin. Struct. Biol. **16** (2006), no. 1, 71–78.

[CBD95] H. S. Chan, S. Bromberg, and K. A. Dill, *Models of cooperativity in protein folding*, Philos. Trans. R. Soc. Lond. B Biol. Sci. **348** (1995), no. 1323, 61–70.

[CDL+09] Y. Chebaro, X. Dong, R. Laghaei, P. Derreumaux, and N. Mousseau, *Replica exchange molecular dynamics simulations of coarse-grained proteins in implicit solvent*, J. Phys. Chem. B **113** (2009), no. 1, 267–274.

[CGO02] M. S. Cheung, A. E. Garcia, and J. N. Onuchic, *Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse*, Proc. Natl. Acad. Sci. **99** (2002), no. 2, 685–690.

[CH02] S. Cho and D. W. Hoffman, *Structure of the β subunit of translation initiation factor 2 from the Archaeon Methanococcus jannaschii: A representative of the eIF2 β/eIF5 family of proteins*, Biochemistry **41** (2002), no. 18, 5730–5742.

[Cha00] H. S. Chan, *Modeling protein density of states: Additive hydrophobic effects are insufficient for calorimetric two-state cooperativity*, Proteins: Struct. Funct. Genet. **40** (2000), no. 4, 543–571.

[Che08] M. R. Chernick, *Bootstrap methods: A guide for practitioners and researchers*, second ed., Wiley-Interscience, 2008.

[Cle08] C. Clementi, *Coarse-grained models of protein folding: toy models or predictive tools?*, Curr. Opin. Struct. Biol. **18** (2008), 10–15.

[CLLL07] T. Chen, X. Lin, Y. Liu, and H. Liang, *Microcanonical analysis of association of hydrophobic segments in a heteropolymer*, Phys. Rev. E **76** (2007), no. 4, 046110–046113.

[Cre92] T. E. Creighton, *Proteins: Structures and molecular properties*, second ed., W. H. Freeman, 1992.

[CSM06] N. Y. Chen, Z. Y. Su, and C. Y. Mou, *Effective potentials for folding proteins*, Phys. Rev. Lett. **96** (2006), no. 7, 078103–078106.

[CWN06] K. T. Chu, H. X. Wang, and T. B. Ng, *Fungal peptides with antifungal activity*, Handbook of Biologically Active Peptides (A. J. Kastin, ed.), Academic Press, 2006, pp. 125–130.

[Dav08] C. A. Davie, *A review of Parkinson's disease*, Br. Med. Bull. **86** (2008), no. 1, 109–127.

[DBB+03] F. Ding, J. M. Borreguero, S. V. Buldyrey, H. E. Stanley, and N. V. Dokholyan, *Mechanism for the α-helix to β-hairpin transition*, Proteins: Struct. Func. Genet. **53** (2003), no. 2, 220–228.

[Des97] M. Deserno, *Tricriticality and the Blume-Capel model: A Monte Carlo study within the microcanonical ensemble*, Phys. Rev. E **56** (1997), no. 5, 5204–5210.

[DM07] P. Derreumaux and N. Mousseau, *Coarse-grained protein molecular dynamics simulations*, J. Chem. Phys. **126** (2007), 025101.

[DMS+06] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, *Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction*, Proc. Natl. Acad. Sci. **103** (2006), no. 26, 9885–9890.

[DOW+07] K. A. Dill, S. B. Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz, *The protein folding problem: when will it be solved?*, Curr. Opin. Struct. Biol. **17** (2007), no. 3, 342–346.

[DP01] E. Darve and A. Pohorille, *Calculating free energies using average force*, J. Chem. Phys. **115** (2001), no. 20, 9169–9183.

[DS95] K. A. Dill and D. Stigter, *Modeling protein stability as heteropolymer collapse*, Adv. Protein Chem. (C. B. Anfinsen, F. M. Richards, J. T. Edsall, and D. S. Eisenberg, eds.), vol. 46, Academic Press, 1995, pp. 59–104.

[DŠK98] C. M. Dobson, A. Šali, and M. Karplus, *Protein folding: A perspective from theory and experiment*, Angew. Chem. Int. Ed. **37** (1998), no. 7, 868–893.

[DTF+02] Q. H. Dai, C. Tommos, E. J. Fuentes, M. R. A. Blomberg, P. L. Dutton, and A. J. Wand, *Structure of a de novo designed protein model of radical enzymes*, J. Am. Chem. Soc. **124** (2002), no. 37, 10952–10953.

[Efr79] B. Efron, *Bootstrap methods: Another look at the jackknife*, Ann. Statist. **7** (1979), 1–26.

[EMS07] S. Esteban-Martín and J. Salgado, *Self-assembling of peptide/membrane complexes by atomistic molecular dynamics simulations*, Biophys. J. **92** (2007), no. 3, 903–912.

[EMT+98] W. A. Eaton, V. Munoz, P. A. Thompson, E. R. Henry, and J. Hofrichter, *Kinetics and dynamics of loops, α-helices, β-hairpins, and fast-folding proteins*, Acc. Chem. Res. **31** (1998), no. 11, 745–753.

[Eng05] D. M. Engelman, *Membranes are more mosaic than fluid*, Nature **438** (2005), 578–580.

[FA95] D. Frishman and P. Argos, *Knowledge-based protein secondary structure assignment*, Proteins: Struct. Func. Genet. **23** (1995), no. 4, 566–579.

[FAC00] P. Ferrara, J. Apostolakis, and A. Caflisch, *Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations*, J. Phys. Chem. B **104** (2000), no. 20, 5000–5010.

[Fer98] A. Fersht, *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding*, first ed., W. H. Freeman, 1998.

[Fin99] A. L. Fink, *Chaperone-mediated protein folding*, Physiological Reviews **79** (1999), no. 2, 425–449.

[FIW02] G. Favrin, A. Irbäck, and S. Wallin, *Folding of a small helical protein using hydrogen bonds and hydrophobicity forces*, Proteins: Struct. Func. Genet. **47** (2002), no. 2, 99–105.

[FMMSV10] L. A. Fernández, V. Martín-Mayor, B. Seoane, and P. Verrocchio, *Separation and fractionation of order and disorder in highly polydisperse systems*, Phys. Rev. E **82** (2010), no. 2, 021501–021507.

[FOYHG07] N. L. Fawzi, Y. Okabe, E. H. Yap, and T. Head-Gordon, *Determining the critical nucleus and mechanism of fibril elongation of the Alzheimer's A β_{1-40} peptide*, J. Mol. Biol. **365** (2007), no. 2, 535–550.

[FP83] J. L. Fauchere and V. Pliska, *Hydrophobic parameters pi of amino-acid side-chains from the partitioning of N-acetyl-amino-acid amides*, Eur. J. Med. Chem. **18** (1983), no. 4, 369–375.

[FP02] A. V. Finkelstein and O. B. Ptitsyn, *Protein physics*, Academic Press, 2002.

[FPRS09] P. L. Freddolino, S. Park, B. Roux, and K. Schulten, *Force field bias in protein folding simulations*, Biophys. J. **96** (2009), 3772–3780.

[FS88] A. M. Ferrenberg and R. H. Swendsen, *New Monte Carlo technique for studying phase transitions*, Phys. Rev. Lett. **61** (1988), no. 23, 2635–2638.

[FS89] ———, *Optimized Monte Carlo data analysis*, Phys. Rev. Lett. **63** (1989), no. 12, 1195–1198.

[FS01] D. Frenkel and B. Smit, *Understanding molecular simulation: From algorithms to applications*, second ed., Academic Press, 2001.

[FTLSW04] Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes, *Optimizing physical energy functions for protein folding*, Proteins: Struct. Func. Genet. **54** (2004), no. 1, 88–103.

[FTvHW05] J. A. Freites, D. J. Tobias, G. von Heijne, and S. H. White, *Interface connections of a transmembrane voltage sensor.*, Proc. Natl. Acad. Sci. USA **102** (2005), no. 42, 15059–15064.

[GCLK02] C. L. Guo, M. S. Cheung, H Levine, and D. A. Kessler, *Mechanisms of cooperativity underlying sequence-independent β-sheet formation*, J. Chem. Phys. **116** (2002), no. 10, 4353–4365.

[GD09] K. Ghosh and K. A. Dill, *Theory for protein folding cooperativity: Helix bundles*, J. Am. Chem. Soc. **131** (2009), no. 6, 2306–2312.

[Gel98] S. H. Gellman, *Minimal model systems for β sheet secondary structure in proteins*, Curr. Opin. Chem. Biol. **2** (1998), no. 6, 717–725.

[GG08] R. H. Garrett and C. M. Grisham, *Biochemistry*, fourth ed., Brooks Cole, 2008.

[GHC03] J. Gsponer, U. Haberthur, and A. Caflisch, *The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35*, Proc. Natl. Acad. Sci. **100** (2003), no. 9, 5154–5159.

[GKBH08] A. Go, S. Kim, J. Baum, and M. H. Hecht, *Structure and dynamics of de novo proteins from a designed superfamily of 4-helix bundles*, Prot. Sci. **17** (2008), no. 5, 821–832.

[Gō83] N. Gō, *Theoretical studies of protein folding*, Annu. Rev. Biophys. Bioeng. **12** (1983), 183–210.

[Gro93] D. H. E. Gross, *Multifragmentation, link between fission and the liquid-gas phase-transition*, Prog. Part. Nucl. Phys. **30** (1993), 155–164.

[Gro01] D. H. E. Gross, *Microcanonical thermodynamics: Phase transitions in 'small' systems*, 1st ed., World Scientific Publishing Company, January 2001.

[Ham87] R. W. Hamming, *Numerical methods for scientists and engineers*, second ed., Dover Publications, 1987.

[Has99] S. Hassani, *Mathematical physics: A modern introduction to its foundations*, first ed., Springer, 1999.

[Hay10] W. M. Haynes, *CRC handbook of chemistry and physics*, 91st ed., CRC Press, 2010.

[HB94] J. D. Hirst and C. L. Brooks, III, *Helicity, circular dichroism and molecular dynamics of proteins*, J. Mol. Biol. **243** (1994), no. 2, 173–178.

[HDS96] W. Humphrey, A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics*, J. Mol. Graphics **14** (1996), no. 1, 33–38.

[HGB03] T. Head-Gordon and S. Brown, *Minimalist models for protein folding and design*, Curr. Opin. Struct. Biol. **13** (2003), no. 2, 160–167.

[HGZ+02] C. Y. Huang, Z. Getahun, Y. J. Zhu, J. W. Klemke, W. F. DeGrado, and F. Gai, *Helix formation via conformation diffusion search*, Proc. Natl. Acad. Sci. **99** (2002), no. 5, 2788–2793.

[HJBI08] R. D. Hills Jr and C. L. Brooks III, *Subdomain competition, cooperativity, and topological frustration in the folding of CheY*, J. Mol. Biol. **382** (2008), no. 2, 485–495.

[HKRM+09] A. Holt, R. B. M. Koehorst, T. Rutters-Meijneke, M. H. Gelb, D. T. S. Rijkers, M. A. Hemminga, and J. A. Killian, *Tilt and rotation angles of a transmembrane model peptide as studied by fluorescence spectroscopy*, Biophys. J. **97** (2009), 2258–2266.

[HKvdSL08] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, *GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation*, J. Chem. Theory Comput. **4** (2008), no. 3, 435–447.

[Hol85] H. G. J. Hol, *The role of the α-helix dipole in protein function and structure*, Prog. Biophys. Mol. Biol. **45** (1985), 149–195.

[HP02] A. Hüller and M. Pleimling, *Microcanonical determination of the order parameter critical exponent*, Int. J. Mod. Phys. C **13** (2002), no. 7, 947–956.

[HRL08] J. Hernandez-Rojas and J. M. Gomez Llorente, *Microcanonical versus canonical analysis of protein folding*, Phys. Rev. Lett. **100** (2008), no. 25, 258104–258107.

[HS94] M. H. Hao and H. A. Scheraga, *Monte Carlo simulation of a first-order transition for protein folding*, J. Phys. Chem. **98** (1994), no. 18, 4940–4948.

[HS02] J. Hardy and D. J. Selkoe, *The amyloid hypothesis of alzheimer's disease: Progress and problems on the road to therapeutics*, Science **297** (2002), no. 5580, 353–356.

[HSS+07] F. Huang, S. Sato, T. D. Sharpe, L. Ying, and A. R. Fersht, *Distinguishing between cooperative and unimodal downhill protein folding*, Proc. Natl. Acad. Sci. USA **104** (2007), no. 1, 123–127.

[HTB03] B. K. Ho, A. Thomas, and R. Brasseur, *Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the α-helix*, Prot. Sci. **12** (2003), no. 11, 2508–2522.

[Hül94] A. Hüller, *First order phase transitions in the canonical and the microcanonical ensemble*, Zeit. Phys. B **93** (1994), 401–405, 10.1007/BF01312712.

[HWJW10] W. Han, C.-K. Wan, F. Jiang, and Y.-D. Wu, *Pace force field for protein simulations. 1. full parameterization of version 1 and verification*, J. Chem. Theory Comput. **6** (2010), 3373–3389.

[HWW08] W. Han, C.-K. Wan, and Y.-D. Wu, *Toward a coarse-grained protein model coupled with a coarse-grained solvent model: Solvation free energies of amino acid side chains*, J. Chem. Theory Comput. **4** (2008), 1891–1901.

[IBI05] W. Im and C. L. Brooks III, *Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations*, Proc. Natl. Acad. Sci. USA **102** (2005), no. 19, 6771–6776.

[IBKS01] R. Improta, V. Barone, K. N. Kudin, and G. E. Scuseria, *Structure and conformational behavior of biopolymers by density functional calculations employing periodic boundary conditions. I. The case of polyglycine, polyalanine, and poly-α-aminoisobutyric acid in vacuo*, J. Am. Chem. Soc. **123** (2001), 3311–3322.

[ID08] G. Illya and M. Deserno, *Coarse-grained simulation studies of peptide-induced pore formation*, Biophys. J. **95** (2008), no. 9, 4163–4173.

[Isr92] J. N. Israelachvili, *Intermolecular and surface forces*, second ed., Academic Press, 1992.

[ISW00] A. Irbäck, F. Sjunnesson, and S. Wallin, *Three-helix-bundle protein in a Ramachandran model*, Proc. Natl. Acad. Sci. **97** (2000), no. 25, 13614–13618.

[Jac98] S. E. Jackson, *How do small single-domain proteins fold?*, Fold. Des. **3** (1998), no. 4, R81–R91.

[Jan98] W. Janke, *Canonical versus microcanonical analysis of first-order phase transitions*, Nucl. Phys. B Proc. Suppl. **63** (1998), no. 1-3, 631–633, Proceedings of the XVth International Symposium on Lattice Field Theory.

[JBJ06] C. Junghans, M. Bachmann, and W. Janke, *Microcanonical analyses of peptide aggregation processes*, Phys. Rev. Lett. **97** (2006), no. 21, 218103–218106.

[JBJ08] ———, *Thermodynamics of peptide aggregation processes: An analysis from perspectives of three statistical ensembles*, J. Chem. Phys. **128** (2008), no. 8, 085103–085111.

[JMTR96] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids*, J. Am. Chem. Soc. **118** (1996), no. 45, 11225–11236.

[JW89] R. E. Jacobs and S. H. White, *The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices*, Biochemistry **28** (1989), 3421–3437.

[KAB96] I.L. Karle, S.K. Awasthi, and P. Balaram, *A designed β-hairpin peptide in crystals*, Proc. Natl. Acad. Sci. **93** (1996), no. 16, 8189–8193.

[Kat65] J. J. Katz, *Chemical and biological studies with deuterium*, pp. 1–110, 39th Annual Priestly Lecture, Pennsylvania State University, University Park, PA, 1965.

[Kau49] B. Kaufman, *Crystal statistics. 2. Partition function evaluated by spinor analysis*, Phys. Rev. **76** (1949), 1232–1243.

[KC00a] H. Kaya and H. S. Chan, *Energetic components of cooperative protein folding*, Phys. Rev. Lett. **85** (2000), no. 22, 4823–4826.

[KC00b] ______, *Polymer principles of protein calorimetric two-state cooperativity*, Proteins: Struct. Func. Genet. **40** (2000), no. 4, 637–661.

[KC03] ______, *Origins of chevron rollovers in non-two-state protein folding kinetics*, Phys. Rev. Lett. **90** (2003), 258104–258107.

[KD99] A. Kolb and B. Dünweg, *Optimized constant pressure stochastic dynamics*, J. Chem. Phys. **111** (1999), no. 10, 4453–4459.

[KFTRJ01] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides*, J. Phys. Chem. B **105** (2001), no. 28, 6474–6487.

[KHT+03] R. Kayed, E. Head, J. L. Thompson, T. M. McIntire, S. C. Milton, C. W. Cotman, and C. G. Glabe, *Common Structure of Soluble Amyloid Oligomers Implies Common Mechanism of Pathogenesis*, Science **300** (2003), no. 5618, 486–489.

[KI10] T. Kim and W. Im, *Revisiting hydrophobic mismatch with free energy simulation studies of transmembrane helix tilt and rotation*, Biophys. J. **99** (2010), 175–183.

[KKK09] N. Komatsu, S. Kimura, and T. Kiwata, *Negative specific heat in self-gravitating n-body systems enclosed in a spherical container with reflecting walls*, Phys. Rev. E **80** (2009), no. 4, 041107–041115.

[KKS09] J. Kim, T. Keyes, and J. E. Straub, *Relationship between protein folding thermodynamics and the energy landscape*, Phys. Rev. E **79** (2009), 030902.

[KPD97] A. Kopf, W. Paul, and B. Dünweg, *Multiple time step integrators and momentum conservation*, Comput. Phys. Commun. **101** (1997), 1–8.

[KR87] S. E. Koonin and J. Randrup, *Microcanonical simulation of nuclear disassembly*, Nucl. Phys. A **474** (1987), no. 1, 173–192.

[KRB+92] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *The weighted histogram analysis method for free-energy calculations on biomolecules: I. The method*, J. Comput. Chem. **13** (1992), no. 8, 1011–1021.

[KRB+95] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P.A. Kollman, *Multidimensional free-energy calculations using the weighted histogram analysis method*, J. Comput. Chem. **16** (1995), no. 11, 1339–1350.

[Kuc96] K. Kuczera, *Dynamics and thermodynamics of globins*, Advanced Series in Physical Chemistry: Recent Developments in Theoretical Studies of Proteins (R. Elber, ed.), vol. 7, World Scientific, 1996, pp. 1–64.

[Lak99] J. R. Lakowicz, *Principles of fluorescence spectroscopy*, second ed., Springer, 1999.

[LAMH06] H.-J. Limbach, A. Arnold, B. A. Mann, and C. Holm, *ESPResSo - an extensible simulation package for research on soft matter systems*, Comput. Phys. Comm. **174** (2006), no. 9, 704–727.

[LBZ+00] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Mol. cell biol.*, W.H. Freeman & Company, New York, 2000.

[LD89] K. F. Lau and K. A. Dill, *A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins*, Macromolecules **22** (1989), no. 10, 3986–3997.

[Lev69] C. Levinthal, *How to Fold Graciously*, Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois (J. T. P. Debrunnder and E. Munck, eds.), University of Illinois Press, 1969, pp. 22–24.

[LL06] P. T. Lansbury and H. A. Lashuel, *A century-old debate on protein aggregation and neurodegeneration enters the clinic*, Nature **443** (2006), no. 7113, 774–779.

[LR61] S. Lifson and A Roig, *On the theory of helix-coil transition in polypeptides*, J. Chem. Phys **34** (1961), 1963–1874.

[Man88] D. M. A. Mann, *The pathological association between down syndrome and alzheimer disease*, Mech. Ageing Dev. **43** (1988), no. 2, 99–136.

[Mar07] Y. Marechal, *The hydrogen bond and the water molecule: The physics and chemistry of water, aqueous and bio-media*, first ed., Elsevier Science, 2007.

[MB94] O. G. Mouritsen and M. Bloom, *Mattress model of lipid-protein interactions in membranes*, Biophys. J. **46** (1994), 141–153.

[MBF+00] H. Meyer, O. Biermann, R. Faller, D. Reith, and F. Müller-Plathe, *Coarse graining of nonbonded inter-particle potentials using automatic simplex optimization to fit structural properties*, J. Chem. Phys. **113** (2000), no. 15, 6264–6275.

[MBT08] J. L. MacCallum, W. F. D. Bennett, and D. P. Tieleman, *Distribution of amino acids in a lipid bilayer from computer simulations*, Biophys. J. **94** (2008), no. 9, 3393–3404.

[MC06] S. Matysiak and C. Clementi, *Minimalist protein model as a diagnostic tool for misfolding and aggregation*, J. Mol. Biol. **363** (2006), no. 1, 297–308.

[MC07] S. Muff and A. Caflisch, *Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein*, Proteins: Struct. Funct. Bioinf. **70** (2007), 1185–1195.

[Mey92] S. L. Meyer, *Data analysis for scientists and engineers*, Peer Management Consultants, Ltd., 1992.

[MG07] Y. Mu and Y. Q. Gao, *Effects of hydrophobic and dipole-dipole interactions on the conformational transitions of a model polypeptide*, J. Chem. Phys. **127** (2007), no. 10, 105102–105111.

[MGK77] J. A. McCammon, B. R. Gelin, and M. Karplus, *Dynamics of folded proteins*, Nature **267** (1977), 585–590.

[MJ96] S. Miyazawa and R. L. Jernigan, *Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading*, J. Mol. Biol. **256** (1996), no. 3, 623–644.

[MJ99] ———, *An empirical energy potential with a reference state for protein fold and sequence recognition*, Proteins: Struct. Func. Genet. **36** (1999), no. 3, 357–369.

[MKP+08] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, *The MARTINI coarse-grained force field: Extension to proteins*, J. Chem. Theory Comput. **4** (2008), no. 5, 819–834.

[MLP04] C. Micheletti, A. Laio, and M. Parrinello, *Reconstructing the density of states by history-dependent Metadynamics*, Phys. Rev. Lett. **92** (2004), no. 17, 170601–170604.

[MN02] B. Ma and R. Nussinov, *Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers ($A\beta_{16-22}$, $A\beta_{16-35}$, and $A\beta_{10-35}$): Sequence effects*, Proc. Natl. Acad. Sci. **99** (2002), no. 22, pp. 14126–14131 (English).

[MRMT04] L. Monticelli, K. M. Robertson, J. L. MacCallum, and D. P. Tieleman, *Computer simulation of the KvAP voltage-gated potassium channel: Steered molecular dynamics of the voltage sensor*, FEBS Lett. **564** (2004), no. 3, 325–332.

[MRY+07] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, *The MARTINI force field: Coarse grained model for biomolecular simulations*, J. Phys. Chem. B **111** (2007), no. 27, 7812–7824, PMID: 17569554.

[MS96] L. A. Mirny and E. I. Shakhnovich, *How to derive a protein folding potential? A new approach to an old problem*, J. Mol. Biol. **264** (1996), no. 5, 1164–1179.

[MS97] D. A. McQuarrie and J. D. Simon, *Physical chemistry: A molecular approach*, first ed., University Science Books, 1997.

[MSA+89] A. Mondragon, S. Subbiah, S. C. Almo, M. Drottar, and S. C. Harrison, *Structure of the amino-terminal domain of phage-434 repressor at 2.0 Å resolution*, J. Mol. Biol. **205** (1989), no. 1, 189–200.

[MTF10] L. Monticelli, D. P. Tieleman, and P. F. J. Fuchs, *Interpretation of ^{2}H-NMR experiments on the orientation of the transmembrane helix WALP23 by computer simulations*, Biophys. J. **99** (2010), 1455–1464.

[NB99] M. E. J Newman and G. T. Barkema, *Monte Carlo methods in statistical physics*, first ed., Oxford University Press, 1999.

[NMSK+09] D. Nettels, S. Müller-Späth, F. Küster, H. Hofmann, D. Haenni, S. Rüegger, L. Reymond, A. Hoffmann, J. Kubelka, B. Heinz, K. Gast, R. B. Best, and B. Schuler, *Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins*, Proc. Natl. Acad. Sci. **106** (2009), no. 49, 20740–20745.

[NWG05] H. Nymeyer, T. B. Woolf, and A. E. Garcia, *Folding is not required for bilayer insertion: Replica exchange simulations of an α-helical peptide with an explicit lipid bilayer*, Proteins: Struct. Funct. Bioinf. **59** (2005), 783–790.

[ÖEKF07] S. Özdirekcan, C. Etchebest, J. A. Killian, and P. F. J. Fuchs, *On the orientation of a designed transmembrane peptide: Toward the right tilt angle?*, J. Am. Chem. Soc. **129** (2007), 15174–15181.

[Ons44] L. Onsager, *Crystal statistics. 1. A two-dimensional model with an order-disorder transition*, Phys. Rev. **65** (1944), 117–149.

[ÖRLK05] S. Özdirekcan, D. T. S. Rijkers, R. M. J. Liskamp, and J. A. Killian, *Influence of flanking residues on tilt and rotation angles of transmembrane peptides in lipid bilayers. a solid-state ^2H NMR study*, Biochemistry **44** (2005), no. 3, 1004–1012.

[OS99] Z. Oren and Y. Shai, *Mode of action of linear amphipathic α-helical antimicrobial peptides*, Peptide Sci. **47** (1999), 451–463.

[OW04] J. N. Onuchic and P. G. Wolynes, *Theory of protein folding*, Curr. Opin. Struct. Biol. **14** (2004), no. 1, 70–75.

[Pac90] C. N. Pace, *Conformational Stability of Globular Proteins*, Trends Biochem. Sci. **15** (1990), no. 1, 14–17.

[PB05] M. Pleimling and H. Behringer, *Microcanonical analysis of small systems*, Phase Transitions **78** (2005), no. 9-11, 787–797.

[PCB51] L. Pauling, R. B. Corey, and H. R. Branson, *The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain*, Proc. Natl. Acad. Sci. USA **37** (1951), 205–211.

[PdlPL+02] M. T. Pastor, M. L. de la Paz, E. Lacroix, L. Serrano, and E. Perez-Paya, *Combinatorial approaches: A new tool to search for highly structured β-hairpin peptides*, Proc. Natl. Acad. Sci. **99** (2002), no. 2, 614–619.

[PDU+04] S. Peng, F. Ding, B. Urbanc, S. V. Buldyrev, L. Cruz, H. E. Stanley, and N. V. Dokholyan, *Discrete molecular dynamics simulations of peptide aggregation*, Phys. Rev. E **69** (2004), no. 4, Part 1, 041908–041914.

[PE90] J. L. Popot and D. M. Engelman, *Membrane protein folding and oligomerization: the two-stage model*, Biochemistry **29** (1990), no. 17, 4031–4037.

[Pet83] K. Petersen, *Ergodic theory*, Cambridge University Press, 1983.

[PHR+05] J. Parsons, J. B. Holmes, J. M. Rojas, J. Tsai, and C. E. M. Strauss, *Practical conversion from torsion space to cartesian space for in silico protein synthesis*, J. Comput. Chem. **26** (2005), no. 10, 1063–1068.

[PIB+02] A. T. Petkova, Y. Ishii, J. J. Balbach, O. N. Antzutkin, R. D. Leapman, F. Delaglio, and R. Tycko, *A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR*, Proc. Natl. Acad. Sci. **99** (2002), no. 26, 16742–16747.

[Pri79] P. L. Privalov, *Stability of proteins: Small globular proteins*, Adv. Protein Chem. (C. B. Anfinsen, J. T. Edsall, and F. M. Richards, eds.), vol. 33, Academic Press, 1979, pp. 167–241.

[Pri82] ———, *Stability of proteins: Proteins which do not present a single cooperative system*, Adv. Protein Chem. (C. B. Anfinsen, J. T. Edsall, and F. M. Richards, eds.), vol. 35, Academic Press, 1982, pp. 1–104.

[Pri89] ———, *Thermodynamic problems of protein structure*, Annu. Rev. Biophys. Biophys. Chem. **18** (1989), no. 1, 47–69.

[PS04] S. Park and K. Schulten, *Calculating potentials of mean force from steered molecular dynamics simulations*, J. Chem. Phys. **120** (2004), no. 13, 5946–5961.

[PSMG96] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala, *Forces contributing to the conformational stability of proteins*, FASEB J. **10** (1996), no. 1, 75–83.

[PT06] H. A. Posch and W. Thirring, *Thermodynamic instability of a confined gas*, Phys. Rev. E **74** (2006), no. 5, 051103–051108.

[Pti95] O. B. Ptitsyn, *Molten globule and protein folding*, Adv. Protein Chem. (C. B. Anfinsen, J. T. Edsall, and F. M. Richards, eds.), vol. 47, Academic Press, 1995, pp. 83–229.

[RIH+07] B. J. Reynwar, G. Illya, V. A. Harmandaris, M. M. Müller, K. Kremer, and M. Deserno, *Aggregation and vesiculation of membrane proteins by curvature-mediated interactions*, Nature **447** (2007), 461–464.

[RJL+09] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, *Versatile object-oriented toolkit for coarse-graining applications*, J. Chem. Theory Comput. **5** (2009), no. 12, 3211–3223.

[RRS63] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, *Stereochemistry of Polypeptide Chain Configurations*, J. Mol. Biol. **7** (1963), no. 1, 95–99.

[RSGM10] A. J. Rzepiela, D. Sengupta, N. Goga, and S. J. Marrink, *Membrane poration by antimicrobial peptides combining atomistic and coarse-grained descriptions*, Faraday Disc. **144** (2010), 431–443.

[SAN+05] E. Sparr, W. L. Ash, P. V. Nazarov, D. T. Rijkers, M. A. Hemminga, D. P. Tieleman, and J. A. Killian, *Self-association of transmembrane alpha-helices in model membranes: Importance of helix orientation and role of hydrophobic mismatch*, J. Biol. Chem. **280** (2005), 39324–39331.

[Sci96] S. J. Sciutto, *The shape of self-avoiding walks*, J. Phys. A: Math. Gen. **29** (1996), 5455–5473.

[SCLH99] H. Steiner, A. Capell, U. Leimer, and C. Haass, *Genes and mechanisms involved in beta-amyloid generation and alzheimer's disease*, Eur. Arch. Psychiatry Clin. Neurosci. **249** (1999), 266–270.

[SdJH+11] L. V. Schäfer, D. H. de Jong, A. Holt, A. J. Rzepiela, A. H. de Vries, B. Poolman, J. A. Killian, and S. J. Marrink, *Lipid packing drives the segregation of transmembrane helices into disordered lipid domains in model membranes*, Proc. Natl. Acad. Sci. USA **108** (2011), no. 4, 1343–1348.

[SDS94] W. Shalongo, L. Dugad, and E. Stellwagen, *Distribution of helicity within the model peptide acetyl(AAQAA)$_3$amide*, J. Am. Chem. Soc. **116** (1994), no. 18, 8288–8293.

[SE08] B. Schuler and W. A. Eaton, *Protein folding studied by single-molecule FRET*, Curr. Opin. Struct. Biol. **18** (2008), no. 1, 16–26, Folding and Binding / Protein-nucleic acid interactions.

[SEB10] M. Schor, B. Ensing, and P. G. Bolhuis, *A simple coarse-grained model for self-assembling silk-like protein fibers*, Faraday Discuss. **144** (2010), 127–141.

[SH01] A. V. Smith and C. K. Hall, *α-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model*, Proteins: Struct. Func. Genet. **44** (2001), no. 3, 344–360.

[Sha03] J. Shao, *Mathematical statistics*, second ed., Springer, 2003.

[SHG00] J. M. Sorenson and T. Head-Gordon, *Matching simulation and experiment: A new simplified model for simulating protein folding*, J. Comput. Biol. **7** (2000), no. 3-4, 469–481.

[SHHB09] D. Sadava, D. M. Hillis, H. C. Heller, and M. Berenbaum, *Life: the science of biology*, ninth ed., W. H. Freeman, 2009.

[SJKG97] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, *Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?*, Prot. Sci. **6** (1997), no. 3, 676–688.

[SKH+00] M. Schmidt, R. Kusche, T. Hippler, J. Donges, W. Kronmüller, B. von Issendorff, and H. Haberland, *Negative heat capacity for a cluster of 147 sodium atoms*, Phys. Rev. Lett. **86** (2000), no. 7, 1191–1194.

[SLM+08] G. M. Shankar, S. Li, T. H. Mehta, A. Garcia-Munoz, N. E. Shepardson, I. Smith, F. M. Brett, M. A. Farrell, M. J. Rowan, C. A. Lemere, C. M. Regan, D. M. Walsh, B. L. Sabatini, and D. J. Selkoe, *Amyloid-β protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory*, Nat. Med. **14** (2008), 837–842.

[SM10] D. Sengupta and S. J. Marrink, *Lipid-mediated interactions tune the association of glycophorin A helix and its disruptive mutants in membranes*, Phys. Chem. Chem. Phys. **12** (2010), 12987–12996.

[SMLL+10] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, *Atomic-level characterization of the structural dynamics of proteins*, Science **330** (2010), no. 6002, 341–346.

[SN72] S. J. Singer and G. L. Nicolson, *The fluid mosaic model of the structure of cell membranes*, Science **175** (1972), no. 4023, 720–731.

[SO94] N. D. Socci and J. N. Onuchic, *Folding kinetics of proteinlike heteropolymers*, J. Chem. Phys. **101** (1994), no. 2, 1519–1528.

[Sol71] K. Solc, *Shape of a random-flight chain*, J. Chem. Phys. **55** (1971), no. 1, 335–344.

[Sop96] A. K. Soper, *Empirical potential monte carlo simulation of fluid structure*, Chem. Phys. **202** (1996), no. 2-3, 295–306.

[SÖR+04] E. Strandberg, S. Özdirekcan, D. T. S. Rijkers, P. C. A. van der Wel, R. E. Koeppe II, R. M. J. Liskamp, and J. A. Killian, *Tilt angles of transmembrane model peptides in oriented and non-oriented lipid bilayers as determined by ^{2}H solid-state NMR*, Biophys. J. **86** (2004), no. 6, 3709–3721.

[SR01] M. Souaille and B. Roux, *Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations*, Comp. Phys. Comm. **135** (2001), 40–57.

[SR03] A. Sikorski and P. Romiszowski, *Thermodynamical properties of simple models of protein-like heteropolymers*, Biopolymers **69** (2003), no. 3, 391–398.

[SRM09] D. Sengupta, A. Rampioni, and S. J. Marrink, *Simulations of the c-subunit of ATP-synthase reveal helix rearrangements*, Mol. Membr. Biol. **26** (2009), no. 8, 422–434.

[SW86] R. H. Swendsen and J. S. Wang, *Replica Monte-Carlo simulation of spin-glasses*, Phys. Rev. Lett. **57** (1986), no. 21, 2607–2609.

[SYM+07] N. G. Sgourakis, Y. Yan, S. A. McCallum, C. Wang, and A. E. Garcia, *The alzheimer's peptides Aβ40 and 42 adopt distinct conformations in water: A combined MD/NMR study*, J. Mol. Biol. **368** (2007), no. 5, 1448–1457.

[SYSB91] J.M. Scholtz, E.J. York, J.M. Stewart, and R.L. Baldwin, *A neutral, water-soluble, α-helical peptide: the effect of ionic-strength on the helix-coil equilibrium*, J. Am. Chem. Soc. **113** (1991), no. 13, 5102–5104.

[Tan80] C. Tanford, *The hydrophobic effect: Formation of micelles and biological membranes*, second ed., John Wiley & Sons Inc, 1980.

[TBRP94] E. I. Tiktopulo, V. E. Bychkova, J. Ricka, and O. B. Ptitsyn, *Cooperativity of the coil-globule transition in a homopolymer: Microcalorimetric study of poly(N-isopropylacrylamide)*, Macromolecules **27** (1994), no. 10, 2879–2882.

[Tie04] D. P. Tieleman, *The molecular basis of electroporation*, BMC Biochemistry **5** (2004), no. 10, 1–12.

[TLSW99] S. Takada, Z. Luthey-Schulten, and P. G. Wolynes, *Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer*, J. Chem. Phys. **110** (1999), no. 23, 11616–11629.

[Toz05] V. Tozzini, *Coarse-grained models for proteins*, Curr. Opin. Struct. Biol. **15** (2005), no. 2, 144–150.

[TPB09] M. P. Taylor, W. Paul, and K. Binder, *All-or-none proteinlike folding transition of a flexible homopolymer chain*, Phys. Rev. E **79** (2009), no. 5, 050801–050804.

[TS85] D. N. Theodorou and U. W. Suter, *Shape of unperturbed linear polymers: polypropylene*, Macromolecules **18** (1985), no. 6, 1206–1214.

[TSB99] D. P. Tieleman, M. S. P. Sansom, and H. J. C. Berendsen, *Alamethicin helices in a bilayer and in solution: Molecular dynamics simulations*, Biophys. J. **76** (1999), no. 1, 40–49.

[TSV+08] L. Thøgersen, B. Schiøtt, T. Vosegaard, N. C. Nielsen, and E. Tajkhorshid, *Peptide aggregation and pore formation in a lipid bilayer: A combined coarse-grained and all atom molecular dynamics study*, Biophys. J. **95** (2008), no. 9, 4337–4347.

[TT07] M. Tabaton and E. Tamagno, *The molecular link between β- and γ-secretase activity on the amyloid β precursor protein*, Cell Mol. Life Sci. **64** (2007), 2211–2218.

[TV77] G. M. Torrie and J. P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling*, J. Comput. Phys. **23** (1977), no. 2, 187–199.

[TZV08] I. F. Thorpe, J. Zhou, and G. A. Voth, *Peptide folding using multiscale coarse-grained models*, J. Phys. Chem. B **112** (2008), no. 41, 13079–13090.

[UCY+04] B. Urbanc, L. Cruz, S. Yun, S. V. Buldyrev, G. Bitan, D. B. Teplow, and H. E. Stanley, *In silico study of amyloid β-protein folding and oligomerization*, Proc. Natl. Acad. Sci. **101** (2004), no. 50, 17345–17350.

[VBBP10] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, *Molecular simulation of ab initio protein folding for a millisecond folder NTL9$_{1-39}$*, J. Am. Chem. Soc. **132** (2010), no. 5, 1526–1528.

[Vot08] G. A. Voth (ed.), *Coarse-graining of condensed phase and biomolecular systems*, Taylor & Francis, Inc, Sep 2008.

[VSS05] M. Venturoli, B. Smit, and M. M. Sperotto, *Simulation studies of protein-induced bilayer deformations, and lipid-induced protein tilting, on a mesoscopic model for lipid bilayers with embedded proteins*, Biophys. J. **88** (2005), no. 3, 1778–1798.

[WBS09] B. West, F. L. H. Brown, and F. Schmid, *Membrane-protein interactions in a generic coarse-grained model for lipid bilayers*, Biophys. J. **96** (2009), no. 1, 101–115.

[WCB+99] S. T. R. Walsh, H. Cheng, J. W. Bryson, H. Roder, and W. F. DeGrado, *Solution structure and dynamics of a de novo designed three-helix bundle protein*, Proc. Natl. Acad. Sci. **96** (1999), no. 10, 5486–5491.

[WD10a] Z.-J. Wang and M. Deserno, *Systematic implicit solvent coarse-graining of bilayer membranes: lipid and phase transferability of the force field*, New J. Phys. **12** (2010), 095004.

[WD10b] ______, *A systematically coarse-grained solvent-free model for quantitative phospholipid bilayer simulations*, J. Phys. Chem. B **114** (2010), no. 34, 11207–11220.

[WHK+97] D. M. Walsh, D. M. Hartley, Y. Kusumoto, Y. Fezoui, M. M. Condron, A. Lomakin, G. B. Benedek, D. J. Selkoe, and D. B. Teplow, *Amyloid β-protein fibrillogenesis*, J. Biol. Chem. **272** (1997), no. 35, 22364–22372.

[WKF+03] Y. N. Wei, S. Kim, D. Fela, J. Baum, and M. H. Hecht, *Solution structure of a de novo protein from a designed combinatorial library*, Proc. Natl. Acad. Sci. **100** (2003), no. 23, 13270–13273.

[WL00] Z. H. Wang and H. C. Lee, *Origin of the native driving force for protein folding*, Phys. Rev. Lett. **84** (2000), no. 3, 574–577.

[WL01] F. Wang and D. P. Landau, *Efficient, multiple-range random walk algorithm to calculate the density of states*, Phys. Rev. Lett. **86** (2001), no. 10, 2050–2053.

[WW96] W. C. Wimley and S. H. White, *Experimentally determined hydrophobicity scale for proteins at membrane interfaces*, Nature Struct. Biol. **3** (1996), 842–848.

[WW10] D. Wei and F. Wang, *Mimicking coarse-grained simulations without coarse-graining: Enhanced sampling by damping short-range interactions*, J. Chem. Phys. **133** (2010), 084101–084106.

[WWWa] `http://www.rcsb.org/`.

[WWWb] `http://www.bipm.org/en/si/si_brochure/`.

[WWWc] `http://www.espressomd.org/`.

[WWWd] `http://spot.colorado.edu/~beale/IsingExactMathematica.html`.

[WZG+04] T. Wang, Y. J. Zhu, Z. Getahun, D. G. Du, C. Y. Huang, W. F. DeGrado, and F. Gai, *Length dependent helix-coil transition kinetics of nine alanine-based peptides*, J. Phys. Chem. B **108** (2004), no. 39, 15301–15310.

[Xio06] J. Xiong, *Essential bioinformatics*, first ed., Cambridge University Press, 2006.

[XPG06] Y. Xu, P. Purkayastha, and F. Gai, *Nanosecond folding dynamics of a three-stranded β-sheet*, J. Am. Chem. Soc. **128** (2006), no. 49, 15836–15842.

[XSL+05] Y. Xu, J. Shen, X. Luo, W. Zhu, K. Chen, J. Ma, and H. Jiang, *Conformational transition of amyloid β-peptide*, Proc. Natl. Acad. Sci. **102** (2005), no. 15, 5403–5407.

[YFHG08] E. H. Yap, N. L. Fawzi, and T. Head-Gordon, *A coarse-grained alpha-carbon protein model with anisotropic hydrogen-bonding*, Proteins: Struct. Func. Bioinf. **70** (2008), no. 3, 626–638.

[YH95] A. S. Yang and B. Honig, *Free-energy determinants of secondary structure formation: II. Antiparallel β-sheets*, J. Mol. Biol. **252** (1995), no. 3, 366–376.

[YSSM10] S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta, and S. J. Marrink, *Polarizable water model for the coarse-grained MARTINI force field*, PLoS Comp. Bio. **6** (2010), no. 6, e1000810.

[ZAM+03] Y. Zhu, D. O. V. Alonso, K. Maki, C. Y. Huang, S. J. Lahr, V. Daggett, H. Roder, W. F. DeGrado, and F. Gai, *Ultrafast folding of α3D: A de novo designed three-helix bundle protein*, Proc. Natl. Acad. Sci. **100** (2003), no. 26, 15486–15491.

[ZB59] B. H. Zimm and J. K. Bragg, *Theory of the phase transition between helix and random coil in polypeptide chains*, J. Chem. Phys. **31** (1959), no. 2, 526–535.

[ZDI59] B. H. Zimm, P. Doty, and K. Iso, *Determination of the parameters for helix formation in poly-γ-benzyl-L-glutamate*, Proc. Natl. Acad. Sci. USA **45** (1959), 1601–1607.

[ZHK99] Y. Q. Zhou, C. K. Hall, and M. Karplus, *The calorimetric criterion for a two-state process revisited*, Prot. Sci. **8** (1999), no. 5, 1064–1074.

[ZJM+05] B. Zagrovic, G. Jayachandran, I. S. Millett, S. Doniach, and V. S. Pande, *How large is an α-helix? studies of the radii of gyration of helical peptides by small-angle X-ray scattering and molecular dynamics*, J. Mol. Biol. **353** (2005), 232–241.

[ZTIV07] J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth, *Coarse-grained peptide modeling using a systematic multiscale approach*, Biophys. J. **92** (2007), no. 12, 4289–4303.