

Univerzita Karlova v Praze  
Přírodovědecká fakulta

Modelování chemických vlastností nano- a biostruktur



Mgr. Jiří Vymětal

Výpočetní studie krátkých peptidů a miniproteinů a vliv prostředí  
na jejich konformaci

Computational study of short peptides and miniproteins in different  
environments

Disertační práce

Školitel: RNDr. Jiří Vondrášek, CSc.

Praha 2014

**Prohlášení:**

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 25.6.2014

Jiří Vymětal

## Abstrakt

Peptidy, kromě své biologické funkce, představují také důležité modely nesbalených, denaturovaných nebo nestrukturovaných proteinů. Poborně důležitými modely pro experimentální i teoretické studium sbalování proteinů jsou miniproteiny, jako např. Trp-cage. Chování peptidů i proteinů lze studovat v počítačových simulacích pomocí metod molekulární dynamiky, které umožňují sledovat děje v atomistickém rozlišení. Tyto metody však čelí dvěma zásadním problémům – přesnosti používaných energetických funkcí a nedostatečnému vzorkování konformačních stavů. V této disertaci jsem se zabýval oběma okruhy problémů.

Vliv rozdílných, běžně používaných energetických funkcí („force fields“) byl testován na modelu aminokyselinových dipeptidů. Žádná sada parametrů však nedokázala konzistentně reprodukovat konformační preference jednotlivých aminokyselin. Výsledky simulací byly mezi sebou srovnány a byly hledány příčiny jejich vzájemných odlišností.

Abychom odhalili, jakým způsobem různé podmínky ovlivňují konformační stavy peptidů, zkoumali jsme vlastnosti aminokyselin v AAXAA peptidech. Simulace odhalily zásadní rozdíl ve vlivu tepelné a chemické denaturace (močovinou) na charakter a zastoupení konformací peptidů, stejně jako konformačních preferencí jednotlivých aminokyselin.

K problematice vzorkování konformačního prostoru jsem přispěl zavedením kolektivních souřadnic pro metadynamiku odvozených z gyračního tenzoru a tenzoru setrvačnosti. Efektivita těchto kolektivních souřadnic popisujících velikost a tvar molekul byla testována v simulacích alaninových polypeptidů a Trp-cage miniproteinu. V těchto simulacích bylo úspěšně dosaženo reprodukovatelného nalezení nativní konformace miniproteinu a podstatného zlepšení ve vzorkování konformačního prostoru flexibilních polyalaninových peptidů.

Zcela nový miniprotein byl vytvořen obrácením sekvence Trp-cage. Tento umělý konstrukt však narozdíl od Trp-cage nevytváří stabilní třídídimenzionální strukturu v běžných pufrách, ale strukturuje se až po přidání 2,2,2-trifluoethanolu (TFE). Stabilita a další vlastnosti molekuly retro Trp-cage byly studovány v MD simulacích, ale nepodařilo se nalézt strukturu indukující efekt TFE. Proto se stalo TFE předmětem našeho dalšího zájmu.

Nové parametry pro TFE, založené na předchozím modelu, byly optimalizovány, aby lépe a kvalitněji popsali vlastnosti nejen samotného TFE, ale i jeho vodných roztoků. Tento nový model realističtěji zachycuje chování směsných roztoků v rámci Kirkwood-Buffovy teorie.

## Abstract

Apart from biological functions, peptides are of uttermost importance as models for unfolded, denatured or disordered state of the proteins. Similarly, miniproteins such as Trp-cage have proven their role as simple models of both experimental and theoretical studies of protein folding. Molecular dynamics and computer simulations can provide an unique insight on processes at atomic level. However, simulations of peptides and miniproteins face two cardinal problems—inaccuracy of force fields and inadequate conformation sampling. Both principal issues were tackled in this theses.

Firstly, the differences in several force field for peptides and proteins were questioned. We demonstrated the inability of the used force fields to predict consistently intrinsic conformational preferences of individual amino acids in the form of dipeptides and the source of the discrepancies was traced.

In order to shed light on the nature of conformational ensembles under various denaturing conditions, we studied host–guest AAXAA peptides. The simulations revealed that thermal and chemical denaturation by urea produces qualitatively different ensembles and shift propensities of individual amino acids to particular conformers.

The problem of insufficient conformation sampling was dealt by introducing gyration- and inertia-tensor based collective coordinates to metadynamics. We validated this newly implemented size- and shape- descriptors in simulations of alanine peptides and Trp-cage miniprotein. Such facilitated dynamics led to reproducible folding of miniprotein and extensive conformational sampling of flexible polyalanines.

A novel miniprotein were designed by idea of retro transformation of protein sequence. The resulting retro Trp-cage molecule does not fold in water but the structure emerges upon addition of a cosolvent—2,2,2-trifluoroethanol (TFE) into buffer. However, this behavior was not observed in simulations and therefore the force field model of TFE were questioned.

We further developed a novel model of TFE based on generalized amber force field by exhaustive optimization of force field parameters. The resulting model reproduces excellently the liquid state properties of pure TFE and behaves realistically in TFE/water mixtures as we investigated by means of Kirkwood–Buff theory of solutions.

## Acknowledgement

I thank to Dr. Jiří Vondrášek for his guidance and full support of my research activities. I am very grateful to all members of group Bioinformatics at Institute of Organic Chemistry and Biochemistry for friendly and inspiring atmosphere. I acknowledge kindly the fruitful collaboration with the group of prof. V. Daggett at University of Washington and the group of prof. V. Sklenář at National Center for Biomolecular Research.

# Contents

<b>1</b>	<b>Introduction and Methods</b>	<b>9</b>
1.1	Changing paradigms in protein science . . . . .	9
1.1.1	Sequence to Structure to Function . . . . .	9
1.1.1.1	Globular Proteins and Their Folding . . . . .	9
1.1.1.2	Intrinsically disordered proteins . . . . .	10
1.1.2	Structures, Dynamics, Ensembles . . . . .	12
1.1.2.1	Structure . . . . .	12
1.1.2.2	Dynamics . . . . .	12
1.1.2.3	Ensembles . . . . .	13
1.2	Force fields . . . . .	15
1.2.1	Amber force fields . . . . .	16
1.2.2	CHARMM . . . . .	19
1.2.3	OPLS . . . . .	21
1.2.4	GROMOS . . . . .	22
1.2.5	ENCAD . . . . .	23
1.2.6	Performance of force fields . . . . .	23
1.2.7	Remarks to force field development. . . . .	24
1.3	Metadynamics . . . . .	26
1.3.1	Overview of the method . . . . .	26
1.3.2	Algorithms . . . . .	27
1.3.3	Collective coordinates . . . . .	29
1.3.4	Selection of parameters . . . . .	30
1.3.5	Problems of metadynamics and advanced computational schemes . . . . .	31
<b>2</b>	<b>Aims of the thesis</b>	<b>33</b>
<b>3</b>	<b>Intrinsic conformational preferences of amino acids</b>	<b>34</b>
3.1	Motivation . . . . .	34
3.2	Intrinsic propensities of dipeptides in different force fields . . . . .	35
3.3	The difference between thermal and chemical denaturation of AAXAA host-guest peptides and their conformational preferences . . . . .	40
<b>4</b>	<b>Metadynamics in gyration-tensor-based collective coordinates</b>	<b>43</b>
4.1	Motivation . . . . .	43
4.2	Theory . . . . .	43
4.3	Application . . . . .	46
<b>5</b>	<b>Computational study of retro Trp-cage miniprotein</b>	<b>50</b>
5.1	Motivation . . . . .	50
5.2	Experimental characterization of retro Trp-cage . . . . .	51
5.3	Modeling of retro Trp-cage . . . . .	51
5.4	Conclusion . . . . .	53

## *Contents*

<b>6</b>	<b>Optimization of force field parameters for 2,2,2-trifluoroethanol</b>	<b>57</b>
6.1	Motivation . . . . .	57
6.2	Parametrization . . . . .	58
6.3	Results . . . . .	59
<b>7</b>	<b>Concluding remarks</b>	<b>63</b>
	<b>References</b>	<b>67</b>
	<b>Appendices</b>	<b>79</b>

## Preface

This thesis was composed for the purpose of obtaining PhD degree in chemistry. It concludes my work at Institute of Organic Chemistry and Biochemistry AS CR under supervision of Dr. J. Vondrášek in years 2009–2014. I tackled several topics and participated in several projects in course of the doctoral study programme Modelling of Chemical Properties of Nano- and Biostructures at Faculty of Science, Charles University in Prague. However, these projects were always linked conceptually to the goals of my thesis and my research interests—behavior and properties of peptides and proteins scrutinized by computational methods.

This thesis is divided in 6 chapters. The first, introductory chapter familiarizes reader with basic paradigms in protein science, force fields used in computer simulations of proteins and peptides and metadynamics as a method for free energy calculations and accelerated conformational sampling.

The second chapter states the aims of this theses.

The third chapter provides an overview of my contribution to modeling intrinsic conformational preferences of short peptides. The behavior of several force fields is demonstrated on smallest possible model peptides with implication for modeling of unfolded or intrinsically disordered proteins. The shift of propensities upon chemical denaturation investigated in collaboration with prof. Daggett helps to reconcile the random coil model in NMR spectroscopy.

The fourth chapter introduces collective coordinates for metadynamics based on tensor of gyration. It also provides an application on conformational mapping of alanine peptides and folding of a miniprotein.

The fifth chapter is dedicated to design and characterization of *de novo* retro Trp-cage miniprotein. This artificial construct folds to transiently stable structure only in the presence of helix promoting agents (2,2,2-trifluoroethanol, TFE). The performance of structure prediction methods on this novel miniprotein is investigated. Similarly, the ability of force field simulations to maintain the transient fold were questioned.

The last chapter provides results of our effort to improve TFE model for amber force fields. The goal was to provide force field parameters which model reliably the properties of TFE/water mixtures—the necessary condition for intended simulation of proteins and peptides in mixture solvents.



# 1 Introduction and Methods

## 1.1 Changing paradigms in protein science

It is unnecessary to stress here the role that proteins play in maintaining a life as well as importance of their research. Although proteins have been intensively studied since their discovery in eighteenth century, novel breakthroughs are continuously changing paradigms in protein science. This section will briefly illustrate how the perspective on key topics evolved to the current opinions.

### 1.1.1 Sequence to Structure to Function

Every textbook of biochemistry repeats a dogma that function (biological activity) of a protein is facilitated by its three-dimensional structure which is coded by its amino acid sequence. The way how the sequences find their native conformation is the fundamental protein folding problem. However, this dogma were seriously disrupted by the uncovering of intrinsically disordered proteins and their large prevalence in eukaryotic proteomes.<sup>1</sup> Intrinsically disordered proteins contains regions which lacks uniquely folded structure but are critical for function like association or ligand binding.

#### 1.1.1.1 Globular Proteins and Their Folding

Despite the fact that protein folding has been intensively studied for more than 50 years,<sup>2</sup> only an agreement on general principles has been reached by the research community.<sup>3</sup> Controversy still remains on cardinal issues like the importance of individual physical mechanisms taking part in the folding process, existence of an unifying approach reconciling different folding strategies and experimental observations,<sup>4</sup> presence and role of folding intermediates,<sup>5</sup> residual structure in the unfolded ensemble<sup>6</sup> and a role of chaperons.<sup>7</sup> Additionally, it should be noted that most studies of protein folding were conducted on representative soluble globular proteins and much less is known e.g. about folding of membrane proteins.<sup>8</sup> Moreover, the conditions *in vivo* differ from those *in vitro* and the relevance of *in vitro* experiments for processes in live cells could be questioned.<sup>9–11</sup>

Historically, yet the first folding experiments pioneered by Anfinsen demonstrated that protein function and structure is coded in the sequence.<sup>12</sup> It was recognized very soon that a specific mechanism of folding must exist, otherwise a single protein would never fold simply in biologically relevant time by a random sampling of possible conformations (Levinthal's paradox).<sup>13</sup> Since then, protein folding problem raised three fundamental questions which has been tackled by generations of scientist:<sup>2</sup>

- Folding code—How the amino acids code the unique three-dimensional structure? Which physical forces or interactions are responsible for selection and stabilization of the fold?

- Kinetics—What is the mechanism that allows proteins to fold in amenable time?
- Predictability—Can the fold be predicted *a priori* by knowledge of the sequence?

The early proposed model of protein folding were hierarchic framework model.<sup>14</sup> It supposes formation of secondary structure prior to the tertiary. Under the framework model the folding is initiated by local establishment of secondary structure elements which are brought into contact by diffusion-collision mechanism and further stabilized by mutual interactions in the tertiary structure. However, this model postulates existence of intermediates in course of folding. Therefore it is inconsistent with behavior of two-state folders—proteins which show simple kinetics of folding without detectable intermediates.<sup>15</sup>

Highly cooperative two-state folding could be interpreted by means of nucleation-condensation mechanism.<sup>16</sup> It anticipates formation of nucleus involving even the long range interactions. The final tertiary structure is accommodated immediately by fast collapse around the folding nucleus.

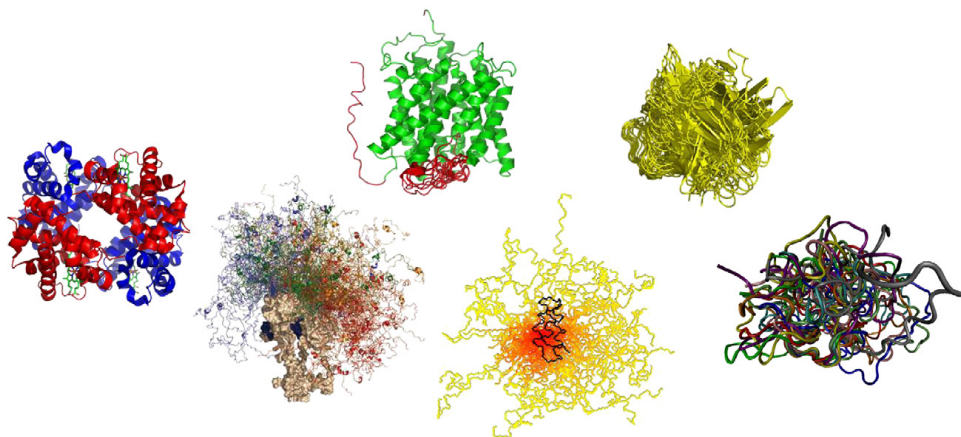
Another approach how to deal with protein folding is based on the thermodynamic and energetic view of the conformational space. Funnel-like shaped energy landscape for protein folding was suggested.<sup>17</sup> The native ordered conformation possesses the lowest free energy content and represents the global minimum on the energy landscape. The native structure could not be the only low energy region on the landscape and the other minima can be considered as folding intermediates or traps. The complexity and ruggedness of the free energy funnel determine the folding mechanisms. The folding on complex landscapes introduces inevitably folding intermediates and also alternative folding paths. In contrast, smooth funnels explains the cases of the two state folding or downhill folders without any apparent barriers in folding pathways.<sup>18</sup>

The recent advances in experimental single-molecule methods can probe energy landscapes and provide valuable information on protein folding.<sup>19</sup> Nevertheless, the most detailed view of folding events can be delivered by computer simulations.<sup>20</sup> The continuous progress in technology and performance of computers (predicted by Moore law) allow to run molecular dynamics on timescales up to one millisecond in full atom resolution.<sup>21</sup> However, the quality of computer simulations depends critically on the quality of a force field. The commonly used force fields has been shown to catch the process of folding and describe its reversibility for model proteins in extensive simulations.<sup>22</sup> However, results may differ in provided mechanistic and thermodynamic predictions.<sup>23</sup>

### 1.1.1.2 Intrinsically disordered proteins

Disorder has been recognized as a vital property of many biologically active proteins. Proteins lacking well formed tertiary structures was found in all kingdoms of life (eukaryotes, eubacteria and archea). However, proteins of eukaryotes are estimated to contain more disordered regions (10-45%) than the prokaryotes. The high content of disorder correlates highly with regulatory and signaling functions, especially for proteins involved in transcription and translation. On the other hand, disorder is rarely exhibited by enzymes.<sup>1,24-26</sup>

Disorder as well as a structure is coded in primary protein sequence and can be predicted.<sup>27</sup> The amino-acid composition of unstructured proteins/regions has typical features—depletion in order-promoting (Ile, Leu, Val, Tyr, Trp, Phe, Cys and Asn) and enrichment



**Figure 1.1:** Current view on intrinsically disordered proteins.

Disorder can affect the whole protein or only its parts. Various regions can be disordered at different extent. The picture were adopted from Ref. 28.

---

in disorder-promoting (Ala, Arg, Gly, Gln, Ser, Glu, Lys and Pro) amino acids. Such sequences possess statistically lower hydrophobicity and higher charge density due to the abundance of charged residues. It is generally accepted that disordered proteins/regions do not manifest funnel-like energy landscape as structured proteins but populate a range of isoenergetic conformers.<sup>28</sup>

The nature and the extent of disorder differ broadly among proteins. Disorder may be manifested in the range of small fluctuating regions to completely unstructured protein molecules. As number of characterized disordered proteins increased, it was recognized that proteins may exhibit whole spectrum of states from fully ordered to the fully disordered (see Fig. 1.1).

The structural disorder can be further classified by functionality into 6 categories.<sup>29</sup> i) Entropic chains utilize the conformational flexibility and work as entropic springs, linkers or spacers. The other 5 categories participate in molecular recognition and binding, both transient and permanent. ii) Display sites for post-translational modification employ the transient but specific recognition by modifying enzymes. iii) Similarly chaperones recognize target protein or RNA molecules. The disordered proteins binding permanently can be referred as effectors, assemblers and scavengers. iv) Effectors modify the activity of partner enzymes and usually act as inhibitors. v) Assemblers are found as structural parts of macromolecular complexes such as ribosom, cytoskeleton or chromatin. vi) Scavengers bind and store small molecular ligands, e.g. casseins in milk protect calcium cations by this mechanism.

Some disordered regions are able to structurize upon a binding to the ligand or interaction partner.<sup>30</sup> Disordered proteins can often interact with several partners and thus provide promiscuous binding.<sup>31,32</sup> This unique ability is attributed to the structural heterogeneity in the unbound state. However, some proteins may still keep significant disorder upon binding or may adapt structurally to different functions.<sup>33</sup>

Because of lack of tertiary structure, disordered proteins/regions are inaccessible for usual structure determining methods as X-ray diffraction and NMR spectroscopy. However, ad-

vances in NMR techniques allow to overcome major difficulties in interpretation of spectra and provide structural details at atomic resolution.<sup>34</sup> The chemical shifts and scalar coupling report about local structural features;<sup>35</sup> Residual Dipolar Couplings,<sup>36</sup> Nuclear Overhauser Effect and Paramagnetic Relaxation Enhancements may reveal long-range contacts.<sup>37</sup> In order to get illustrative representation of the disorder the observed local and long-range contacts are used for generation of structural ensemble model compatible with experimental data.<sup>38,39</sup>

Simulations of protein disorder by classic molecular dynamics are very computationally demanding because the proper sampling of the ensemble needs at least time scales of microseconds.<sup>40</sup> However, biases in force fields towards different secondary structure elements can produce simulations incompatible with experimental data.<sup>41</sup> The possible solution of the problem could be overcome by using of experimental constrains during the simulation. This drives the sampling to relevant regions of conformational space and thus suppresses inaccuracies of energetic functions.<sup>42</sup>

## **1.1.2 Structures, Dynamics, Ensembles**

### **1.1.2.1 Structure**

Since the first three dimensional structures of myoglobin and hemoglobin were deciphered by Kendrew and Perutz in 50's,<sup>43</sup> protein X-ray diffraction crystallography has evolved into the the most precise source of structural information about proteins. Synchrotron radiation and cryogenics allowed to reach even subatomic resolution that provide electron density with recognizable valence electron shells.<sup>44</sup> Similarly, the progress in Nuclear Magnetic Resonance (NMR) techniques led to the first de novo solved 3D structure in 1984.<sup>45</sup> NMR is recently a method of first choice because it does not need crystallized proteins and it is carried out in solution under physiological conditions. Moreover, NMR can be applied on disordered or denatured proteins.<sup>34</sup> It provides information about dynamics on broad range of times scales<sup>46</sup> and NMR spectra can be even measured in living cells.<sup>47</sup>

Collection and deposition of solved protein structures in databases like Protein Data Bank (PDB)<sup>48</sup> facilitated the survey and classification of protein folds.<sup>49,50</sup> The existence of the conserved structural features of proteins supports the structure-centered view on the protein function—one structure exerts one function. This observation seems to be valid for large portion of proteins, mainly enzymes. For these proteins their functions can be predicted even by knowledge of structural homologue.<sup>51</sup>

### **1.1.2.2 Dynamics**

Although a static structures give an impression of proteins as rigid molecules, proteins are in a constant motion. They are inherently flexible and they employ broad repertoire of movements significant for their functions.<sup>52,53</sup> The protein dynamics is a consequence of the complex energy surface with many close or far isoenergetic basins allowing conformational transitions.<sup>54</sup> As a result, different conformational transition on time scales spanning 13 orders can be observed.<sup>55</sup> The fastest motions involve bond vibration and angle bending on femto- and picosecond time scales, followed by the fast rotations of side chains realized in nanoseconds. The longer times are necessary for collective motion of

the backbone and flanking of the loops. Folding or ligand binding or unbinding stay on the opposite site of the time scales, because they may take up to several hours. Dynamics plays important role in regulation of protein functions by allostery—mechanism which propagates information about binding of effector molecule to distant regions on the molecule.<sup>56</sup>

The comprehensive picture of protein dynamics can be naturally obtained by molecular simulations. In order to get a dynamic complement to the structure of known folds, large scale simulations of protein motions were conducted by Dynameomics project<sup>57</sup> and MoDEL database.<sup>58</sup> The analysis of such simulations revealed e.g. how different amino acids influence the rigidity of secondary structure elements or an existence of surprisingly rigid loops.<sup>59</sup>

### 1.1.2.3 Ensembles

If the molecule exhibits large conformational flexibility or significant heterogeneity in the sample the high resolution techniques for structure determination are barely applicable. In such situation the molecules cannot be obviously represented just by a single average structure. The representative set of plausible structures—ensemble has to be generated in order to provide insight on the real behavior of the molecule. The valid ensemble must reproduce the available experimental data as a whole rather than as an individual structure.<sup>60</sup> These constraints can be obtained by NMR parameters and small angle X-ray scattering (SAXS) that provide information about size distribution of the molecules in solution.<sup>61</sup>

The typical cases that demand ensemble approach are unfolded and denatured states of proteins. Nevertheless, this concept may be applied also on folded proteins like ubiquitin that exhibits surprisingly large conformation heterogeneity in solution.<sup>62</sup> The ensemble framework is currently the most general way how to describe and merge different conception of protein disorder, dynamics and allosteric effects.<sup>63,64</sup>

### Unfolded and denatured states

The folded proteins exist in continuous equilibrium with its unfolded state that is usually strongly biased toward the folded one under physiological conditions. However, the equilibrium can be shifted by change of thermodynamic conditions as for example by temperature and pressure. The dependence of the folded protein population on temperature provides denaturation and stability curves which are essential tools for estimation of thermodynamic parameters associated with protein folding.<sup>65</sup> Typical proteins manifest a thermal stability maximum implying that the folded protein can be destabilized either by high or low temperature (*cold denaturation*).

The equilibrium between folded and unfolded state can be also influenced by chemical composition of the environment. The chemical compounds in solution (*cosolvents*) that destabilize the folded state are called *denaturants* and those with the opposite effect are *protecting osmolytes*.<sup>66</sup>

Urea and guanidinium hydrochloride belong to the most commonly used denaturants. Effect of both compounds was intensively studied experimentally and theoretically. The

early studies by Tanford suggested that proteins in high molar solution of guanidinium expand their hydrodynamic radii to the values expected for *random coils*.<sup>67</sup> These result have been often misinterpreted and generalized for any kind of denaturation. In the fact, later study proved that the behavior of random coils can be achieved only exceptionally and the denatured molecules stay much more compact with significant residual structure.<sup>68</sup> The secondary structure elements can be still present under the mild denaturation conditions but the global tertiary structure loses its compactness—this particular state is referred as *molten globule*.<sup>69</sup>

The effect of protecting osmolytes is the opposite. They help in stabilizing of the native state and functionality of the protein under non-physiological conditions. They can be naturally found in cells and organisms as a response on stress or adaptation on presence of denaturation agents. The most common protective osmolytes include trimethylamine N-oxide (TMAO), betaine, sarcosine, taurine and glycerol.<sup>70</sup> Apart from these natural osmolytes some other compounds are routinely utilized for their stabilizing effect on the protein and peptide structure e.g. trifluoroethanol (TFE) and hexafluoro-2-propanol (HFIP).<sup>71</sup>

The effects of the cosolvents on stability of proteins can be analyzed thermodynamically. The change of the free energy of folding upon addition of cosolvent can be incorporated into thermodynamic cycle and then separated using Tanford transfer model.<sup>72</sup> This allows to evaluate the individual contribution of amino acid side chains and the backbone. High contribution of the backbone and its hydrogen bonding suggest that it plays a prominent role in the folding process and protein stability.<sup>73</sup>

The competition between internal protein–protein, protein–solvent and solvent–solvent interactions determine preferences for compact (native) or extended (unfolded) structure. Solvation in *good solvents* results in preference of protein-solvent interaction and promotes unfolding, whereas the *poor solvents* favour formation of compact structures.<sup>74</sup> Hence, destabilizing denaturants seem to increase the solvent “goodness” whereas protecting osmolytes show the opposite effect. The interactions of cosolvents with proteins can be further quantified by preferential interaction coefficients  $\Gamma$  that express their affinities.<sup>75</sup>

The mechanisms of protein thermal unfolding and chemical denaturation were extensively studied by molecular dynamics. The extensive simulations of the unfolding process were employed in Dynameomics project for characterization of folding intermediates and folding pathways based on the hypothesis of microscopic reversibility.<sup>76</sup> Molecular dynamics also revealed that urea and guanidinium chloride actively disrupt the tertiary structure by binding to amino acid side chains and backbones.<sup>77</sup> The explanation of atomistic mechanism of protective osmolytes like TMAO or TFE is not completely clear because their effect on protein behavior in simulations depends on the selected force field parameters.<sup>78</sup>

To conclude, the thorough characterization of ensembles of unfolded and denatured states is necessary for our complete understanding of folding process since they represent the reference state for thermodynamic models. Importantly, progress in modeling of these states will influence significantly practical applications like engineering of protein stability.<sup>79</sup>

## 1.2 Force fields

The potential energy functions (force fields) used in this work for description of proteins, peptides and solvents employ the most simplistic mathematical forms. They include a harmonic potential for simulation of bond stretching and angle bending and a variant of trigonometric series for representation of torsional terms. The non-bonded interactions are uniformly modeled by Lennard-Jones *6-12* potential and electrostatics using partial atomic charges interacting by Coulomb's law. The complete energy function follows typically:

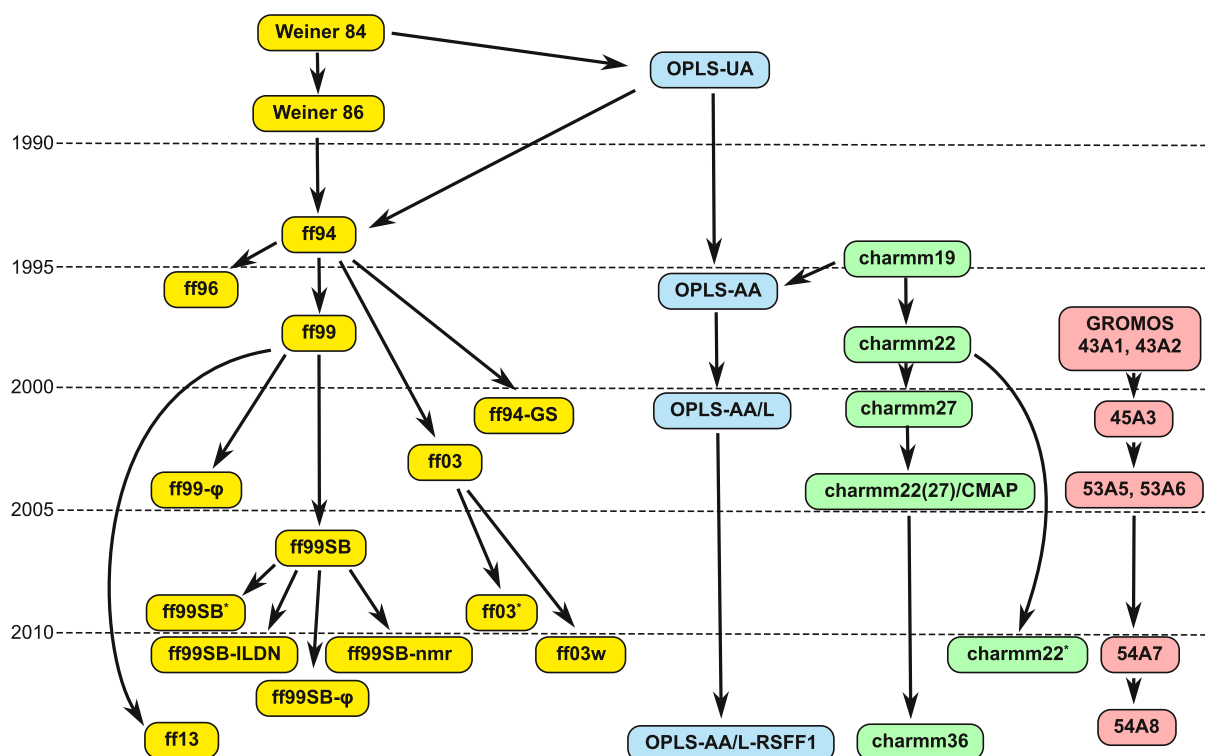
$$E = \sum_{\text{bonds}} \frac{K_B}{2} (r - r_0)^2 + \sum_{\text{angles}} \frac{K_A}{2} (\theta - \theta_0)^2 + \sum_{\text{torsions}} \sum_n \frac{V_n}{2} (1 + \cos(n\phi - \delta_n)) \\ + \sum_{i < j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{\epsilon r_{ij}} \right),$$

where  $r_0$  and  $\theta_0$  stands for equilibrium bond lengths and angles, respectively and  $K_B$  and  $K_A$  for corresponding force constants. Torsion potential is usually expanded in cosine terms with different multiplicities  $n$ , phase offsets  $\delta_n$  and magnitudes  $V_n$ . The atoms separated by more than two covalent bonds are allowed to influence each other via nonbonded interactions prescribed for each pair by partial atomic charges  $q$  or Lennard-Jones parameters  $A_{ij}$  and  $B_{ij}$ . The common practice is to treat non-bonded interactions between atoms separated by 3 bonds (1-4 interaction) specially. They are often scaled down because of their strength.

These simple forms were already applied in the first molecular mechanics and dynamics studies pioneered on proteins<sup>80,81</sup> and still are massively used for simulations of biomolecules although more complicated and more physically sounded terms have been developed, e.g. polarizable multipoles for electrostatics.<sup>82</sup> The most obvious shortcoming of this class of force fields stems from the absence of explicit electrostatic polarization terms.<sup>83</sup> On the other hand, these forms still approximate the potential energy surface sufficiently well and in a robust manner what can be judged by the number of successful applications, e.g. protein folding simulations. Additionally, the simplicity of the functional forms enables their rapid calculation on modern computers and it allows to perform molecular dynamics simulations of time scales yet relevant for biological processes.<sup>84</sup>

Although different force fields share more or less the same energy function forms, they may differ substantially in parameters for individual types of interactions and the way how they were obtained. This reflects the distinct parametrization strategies and properties addressed during parameter development. Although the force fields was reviewed in the past,<sup>85,86</sup> I regard as useful to summarize briefly the history and put into context the recent development. The rest of this section will shed light on the differences between the major families of the force fields for biomolecules, their variants and the origin of parameters.

In the beginning the development of particular force fields were connected tightly to the development of computer codes for simulations. However, the current simulations engines allow using of various parameters deposited to their libraries and hence the ties between force field parameters and the simulation codes were weakened.



**Figure 1.2:** Force field genealogy.

The main force field families (AMBER, OPLS, CHARMM and GROMOS) comprise several versions of published parameters. This chart expresses the protein-centric view on relationships between individual variants and the year of the publication.

The roots of the current force fields for biomolecular simulations—AMBER, CHARMM, OPLS, GROMOS and ENCAD can be traced mostly to the 80’s, when researchers gained enough experience with force fields for organic molecules.<sup>87,88</sup> Nevertheless, the first highly influential works on force fields for proteins and peptides were pioneered in groups of S. Lifson and H. Scheraga which resulted in Consistent Force Field (CFF)<sup>89</sup> and Empirical Conformational Energy Program for Peptides (ECEPP).<sup>90,91</sup>

The comprehensive picture of force field genealogy for major force field families is presented in Fig 1.2.

### 1.2.1 Amber force fields

AMBER force fields were designed for AMBER (Assistant Model Building with Energy Refinement) modeling package,<sup>92</sup> which is still actively developed<sup>93</sup> and belongs to the most popular simulation software.

**Weiner’s force field(84)**<sup>94</sup> This is the first complete set of parameters for simulation of proteins and nucleic acids released for AMBER modeling program. It replaced and improved the parameters for proteins previously used in Kollman’s group.<sup>95</sup>

In order to spare computation resources the aliphatic carbon groups were treated as united atoms but polar hydrogens were represented explicitly. The parameters for



bonds and angles originated from x-ray structures, microwave spectra and *ab initio* calculations of model compounds. Torsional parameters were fitted to reproduce data on conformational equilibria of small model molecules. The partial charges were consistently derived by electrostatic potential fit procedure (ESP) on charge density calculated by HF STO-3G method.<sup>96</sup> To mimic missing explicit solvation the electrostatic energies were evaluated by distance-dependent dielectric constant. Lennard-Jones parameters of atoms were initially adopted from the CFF<sup>89</sup> or Transferable Interaction Potentials (TIP)<sup>97</sup> for Monte Carlo simulations and further adjusted to reproduce molecular geometries. Additionally, scaling of Lennard-Jones non-bonded 1–4 interactions was introduced in order to decrease overestimated torsional barriers. Hydrogen bonds were treated separately by *10-12* potential.

The parameters of protein backbone were developed using N-acetyl-N'methylglycinamid (glycin dipeptide), N-acetyl-N'methylalaninamid (alanine dipeptide) and N-methylacetamide (NMA) as the model compounds. The partial charges for backbone atoms were fitted on HF/6-31G electrostatic potential of NMA. However, it was necessary to scale electrostatic 1–4 interaction by factor 0.5 to reproduce geometries and energy of all local minima of alanine dipeptide.

**Weiner's force field(86)**<sup>98</sup> The united atoms in the previous Weiner's force field(84) were soon found lowering the quality of the force field and they were replaced by explicit atomistic representation. Novel parameters for explicit hydrogens and carbons were included as well as updated partial charges.

**ff94**<sup>99</sup> Next generations of amber force field were parametrized reflecting the feasibility of simulation with explicit solvent molecules. The philosophy of the former versions was kept but the ESP procedure was replaced by RESP<sup>100</sup> at HF/6-31G\* level and an updated 1–4 scaling electrostatic factor of  $\frac{1}{1.2}$  was introduced. Distance dependent dielectric constant and special treatment of hydrogen bonds were found unnecessary and removed in order to make the force field compatible with TIP3P water model.<sup>101</sup>

The RESP at HF/6-31G\* was shown to overestimates dipole moments but this effect fortuitously substitutes the polarization of molecules in condense phase implicitly. In addition to new partial charges new Lennard-Jones parameters were adopted in order to reproduce liquid state properties. The parameters for carbon and hydrogen atoms were adjusted to reproduce density and vaporization enthalpy of simple alkanes and benzene. Parameters for other atom types were taken from OPLS force field.<sup>102</sup> Several atom types were devoted for hydrogens having different van der Waals radii. They were assigned according to the number of geminal electronegative atoms.

The bond and angle parameters were adopted from the previous force field and adjusted for fit of important experimental frequencies. Some torsion parameters were refitted on basis of MP2/6-31\* calculations, but most of them were acquired by the same procedure as discussed above.

Unlike in the previous versions of amber force fields the behavior of protein backbone was adjusted by explicit torsional terms. The potential energy of model compounds (alanine and glycine dipeptides) were fitted to reproduce *ab initio* conformation energies of totally 7 minima calculated on MP2/TZP level.

**ff96 (C96)**<sup>103</sup> This variant based on ff94 introduced new parametrization of backbone torsions. It was recognized that the ff94 was strongly biased to form helical structures. The more thorough fit was performed with aim to reproduce better the difference between  $\alpha$ -helix and  $\beta$ -sheet conformations modeled by alanine tetrapeptides (11 conformers in total).

**ff99**<sup>104</sup> This force field extends the ff94 by additional torsion terms that were fitted on experimental data or high level *ab initio* data set of 82 training organic molecules.

The torsion parameters of protein backbone were also revised in order to reproduce as close as possible the relative conformational energies of alanine dipeptide (7 conformers) and alanine tetrapeptide (11 conformers). The other parameters remained untouched with exception of 2 novel atom types which do not concern the protein or peptide molecules significantly.

**ff03**<sup>105</sup> The force field adopted almost all parameters from ff94 but introduced new methodology for calculation of partial charges. Instead of RESP at HF/6-31G\* level, the RESP procedure was performed at higher level (B3LYP/cc-pVTZ) and the condense phase polarization effects in the fit were treated by implicit continuum solvent. The chosen continuum method mimicked organic solvent with relative permittivity  $\epsilon=4$ .

Afterwards, the torsional parameters of peptide backbone were fitted to reproduce  $\varphi/\psi$  *ab initio* scan of alanine and glycine dipeptide at MP2/cc-pVTZ level with the same continuum solvent as were used for RESP.

**GAFF**<sup>106</sup> Generalized Amber Force Field (GAFF) was intended to cover pharmaceutically interesting substances in computer simulations with amber force fields. It provided additional atom types to describe various organic compounds, parameters for interaction potentials and heuristic models for their estimation.

The Lennard-Jones parameters were transferred from ff99 force field. Standard RESP procedure at HF/6-31G\* level and alternative cheaper AM1-BCC method was recommended for calculation of partial charges. Bond and angle parameters were compiled from different sources—x-ray or neutron diffraction data and *ab initio* calculations at MP2 level. Because gaff tried to be complete force field, i.e. to have parameters for all possible combination of atom types, the missing parameters were estimated by means of heuristics. The torsion parameters were determined by *ab initio* energy profiles of more than 200 compounds at MP4/6-311G(d,p) level.

**ff94-GS**<sup>18</sup> This variant of ff94 removed completely potential for  $\varphi$  and  $\psi$  backbone torsions. It was demonstrated that such *ad hoc* modified force field better described thermodynamics of formation  $\alpha$ -helices in model Fs peptide.

**ff99- $\varphi$** <sup>107</sup> Variant of ff99 force field with backbone  $\varphi$  torsion adopted from ff94. It was shown that this change qualitatively and quantitatively better describes thermodynamics and kinetics of  $\alpha$ -helix folding.

**ff99SB**<sup>108</sup> The different amber force field were critically compared by Hornak *et al.* They revealed inconsistency in *ad hoc* attempts to improve  $\alpha$ -helix or  $\beta$ -strand forming propensities by turning off or transferring  $\varphi$  or  $\psi$  torsional parameters.

Additionally the refitting of backbone torsion parameters in consisted fashion was achieved for ff99. The resulting ff99SB variant contains torsion parameters optimized to match 28 glycine tetrapeptide and 51 alanine tetrapeptide conformers at LMP2/cc-pVTZ level. The novel parameters showed better agreement with experimental NMR data on probe proteins and peptides and improved  $\alpha$ -helix/ $\beta$ -strand balance.

**ff99SB\***, **ff03\***<sup>109</sup> The ff99SB and ff03 force fields were modified by small correction term to  $\psi$  backbone torsion. The resulting variants better described fraction of helical residues in a model helix-forming peptide and they also improved in reproduction of its NMR spectra.

**ff03w**<sup>110</sup> The shortcomings of simple TIP3P water model were recognized and more advanced TIP4P/2005<sup>111</sup> model were used for simulation of helix-coil transition. Upon minor correction to backbone  $\psi$  torsion parameter ff03 manifested more cooperative transition with agreement in experimental data. The corrected ff03w accompanied by TIP4P/2005 seemed to better describe unfolded or disorder states of proteins.

**ff99SB-ILDN**<sup>112</sup> Isoleucine, leucine, aspartate and asparagine were identified as a amino acid whose distribution of rotamers differ significantly between Protein Data Bank and simulation with ff99SB. Torsional parameters for these residues were refitted to match energy profiles at MP2/aug-cc-pVTZ level. The novel parameter improved the agreement with structural database and also reproduced better NMR vicinal J-couplings and RDC of small proteins.

**ff99SB-nmr**<sup>113</sup> The optimized  $\varphi$  and  $\psi$  backbone torsion parameters of ff99SB force field were further refined to reproduce more precisely NMR chemical shifts and crystal structures of proteins.

**ff99SB- $\varphi$** <sup>114</sup> The effort to replace TIP3P water by more realistic model requested minor changes in backbone potential. The modified ff99SB force field accompanied by TIP4P/Ew<sup>115</sup> water model improved description of NMR parameters of small peptides and ubiquitin protein.

**ff13**<sup>116</sup> The upcoming force field introduces novel procedure for calculation of partial atomic charges—IPolQ method which iteratively computes solvent density around parametrized solute and use it for partial charge fitting at MP2/cc-pV(T+d)Z level. TIP4P/Ew was chosen as default water model since it performs better for water properties than TIP3P. In order to rectify hydration free energies of model compounds Lennard-Jones parameters of several common polar atom types were further optimized.

### 1.2.2 CHARMM

CHARMM is a branch of force field developed tightly with CHARMM package of programs (Chemistry at Harvard using Molecular Mechanics).<sup>117,118</sup> The numbering of individual version of force fields copies version of the software release.

**charmm19**<sup>119,120</sup> This parameters updated the united atom force field initially used in the original code<sup>117</sup> based on the former studies.<sup>121</sup> However, charmm19 still employed

united atoms for aliphatic carbons and sulfurs with nonpolar hydrogens but polar hydrogen atoms capable of hydrogen bonding were modeled explicitly.

Similarly as the early amber force fields it was originally intended for simulations and refinement without solvent which was substituted by distance dependent dielectric constant for electrostatic interactions. No specialized hydrogen bond terms are utilized since the partial charges were designed to agree with *ab initio* interaction energies. The TIP3P water model was chosen as a standard for calibration of the interactions at HF/6-31G level. Because of the fortuitous balance between protein-protein, protein-water and water-water interactions charmm19 worked well in simulation with explicit water molecules.

The Lennard-Jones parameters were adjusted to fit crystal packing and liquid densities of model compounds. The bond and angle parameters were adjusted for reproduction of molecular geometries and vibrational spectra of nucleic acid bases and amino acids analogs.

**charmm22**<sup>122</sup> The need for balanced interaction between biomolecules and water resulted in all atom charmm22 force field. The parameters for bonds, angles and torsion were refitted iteratively to match the experimental geometries, vibrational frequencies as well as *ab initio* calculated vibrations, torsion profiles and conformational energies. The Urey-Bradley bending terms were added in certain cases to improve quality of simulated spectra.

The improved protocol for assignment of partial charges was developed. They were solely fitted on *ab initio* interaction energies between the particular molecular fragment and water molecule as a probe in different positions and orientation. The calculations used HF/6-31G\* method and scaling factor of 1.16 that empirically provided suitable values due to the error compensation. The Lennard-Jones parameters were adjusted in liquid phase simulations of aliphatic and polar compounds to reproduce experimental densities and heats of vaporization. Additionally, crystal simulations were utilized for parametrization on experimental lattice constants and sublimation enthalpy. In contrast to the previous versions of force field the Lorentz-Berthelot combination rules were adopted for Lennard-Jones parameters.

The parameters for protein backbone were based on optimization of NMA interaction energies and liquid state properties. The particular attention was devoted to reproduction of geometric, spectral and energetic parameters of NMA and alanine dipeptide which were mainly taken from experimental data. The resulting protein parameters were comprehensively tested in a gas phase and crystal simulations of model peptides and proteins, namely crambin, BPTI and myoglobin.

**charmm27**<sup>123</sup> This release of charmm force field focused on improvement of nucleic acids parameters. The protein part was not modified and it is identical with charmm22. However, sometimes is charmm22/CMAP referred as charmm27.

**charmm22/CMAP**<sup>124</sup> The backbone torsional parameters were revised to match *ab initio* energy surface of alanine, glycine and proline dipeptide at LMP2/cc-pVQZ level. Because the sufficient fit could not be reached without cross-terms a new functional form was introduced as the two dimensional correction grid for  $\varphi$  and  $\psi$  torsion (CMAP). The *ab initio* fit was further manually adjusted to remove systematic bias observed in control simulations of crystalline proteins.

**charmm36**<sup>125</sup> The most recent release of charmm force field improved internal parameters associated with peptide bonds and amino acid side chains. The non-bonded parameters were not updated and remain the same as in charmm22. The CMAPs of proline and glycine were refitted on *ab initio*  $\varphi/\psi$  scans at RIMP2/CBS level. The CMAP assigned to other amino acids was optimized to match NMR chemical shifts and coupling constants of alanine pentapeptide and helical Ac-(AAQAA)<sub>3</sub>-NH<sub>2</sub> peptide.

This effort was aiming to remove helical bias present in the charmm22/CMAP force field and to prepare more suitable force field for simulation of unfolded or disordered proteins. Furthermore, the  $\chi_1$  and  $\chi_2$  torsions of amino acid side chains were optimized to reproduce RIMP2/cc-pVTZ energy scans. The improved parameters were shown to reproduce experimental NMR parameters significantly better for several tested proteins (ubiquitin, protein G, cold shock protein A, apocalmodulin, intestinal fatty acid binding protein and lysozyme).<sup>126</sup>

**charmm22\***<sup>127</sup> Because charmm22/CMAP was found to favour helices in folding simulations a more balanced modification were suggested. The CMAP correction was replaced by new backbone parameters, based on LMP2 energy scan of alanine dipeptide and NMR data of polyalanine peptides. Additionally, the partial charges on Asp, Glu and Arg were modified to improve description of salt bridges. The other changes involved reparametrization of Asp side chain torsions similarly for ff99SB-ILDN force field.

**charmm general force field**<sup>128</sup> The detailed procedure how to prepare parameters for drug-like compounds compatible with charmm force fields was developed and explained in details. The parameters for large number of scaffolds and fragments were already prepared by authors and can be assigned automatically.<sup>129</sup>

### 1.2.3 OPLS

Optimized Potentials for Liquid Simulations (OPLS)<sup>130</sup> started as an united atom force field for Monte Carlo simulation of liquids—the successor of previously developed transferable interaction potentials (TIP).<sup>97</sup> An emphasis was placed on reproduction of liquid state properties—density and heat of vaporization. OPLS was later extended to treat biomolecules such as proteins and nucleic acids.

**OPLS-UA**<sup>102</sup> First published version of parameters for proteins included the previously elaborated parameters for model organic compounds such as alkanes,<sup>130</sup> amides<sup>131</sup> and amino acid side chain analogs.

The corresponding charges and Lennard-Jones parameters were systematically optimized to reproduce density and heat of vaporization. Only polar hydrogen atoms were treated explicitly, the others were involved in united atom types. TIP4P potential<sup>101</sup> was chosen as the water model.

The OPLS-UA force field adopted parameters for bonds, angles and torsions from amber,<sup>94</sup> since the OPLS were previously developed on rigid molecules. Therefore, the resulting force field is sometimes referred as AMBER/OPLS. Due to the different non-bonded parameters the 1–4 scaling factors from amber had to be reoptimized.

This was achieved by matching conformational energies and geometries of butane, ether, alanine and glycine dipeptides.

**OPLS-AA**<sup>132</sup> It was soon recognized that all atom representation allows more flexibility in parametrization. The partial charges and Lennard-Jones parameters were therefore refitted in Monte Carlo simulations of 34 model liquids. Obtained parameters were considered to be highly transferable on level of functional groups. This principle remained one of the corner stones of OPLS philosophy. Additionally, new torsion parameters were developed to match *ab initio* gas-phase conformational energies at HF/6-31G\* level.

Special attention was paid to protein backbone resulting in specific torsional parameters fitted on aminoaldehydes and alanine dipeptide. All atom representation also required new scaling factors and value of 0.5 was chosen for electrostatic and Lennard-Jones terms. The bond and angle parameters were not changed and remained the same as in the Weiner *et al.*<sup>94</sup> force field with the exception of those for aliphatic hydrocarbons adopted from charmm22.

**OPLS-AA/L**<sup>133</sup> An improved variant of OPLS-AA force field utilized *ab initio* quantum calculation at LMP2/cc-pVTZ(-f) level to obtain new parameters for amino acids and peptides. The backbone torsions were refitted on alanine dipeptide and tested on alanine tetrapeptide. In addition, the torsion parameters of amino acid side chains in dipeptide models were refined to match *ab initio* conformational energies. The non-bonded parameters for sulfur in cysteine and methionine were revised and updated considering *ab initio* calculations.

**OPLS-AA/L-RSFF1**<sup>134</sup> The backbone and side chain torsions for each amino acid were modified to match  $\varphi/\psi$  and rotamer propensities derived from coil library. Each amino acid was optimized independently and possesses individual parameters. In several cases Lennard-Jones interaction parameters were modified to describe correctly the coupling between side chain rotamers and backbone conformers. The resulting force field was shown to reproduce reasonably NMR J-couplings of amino acid dipeptides and a ability to fold both  $\alpha$ -helical and  $\beta$ -strand proteins.

#### 1.2.4 GROMOS

GROMOS (GRoningen MOlecular Simulation) is an united atom force field developed together with the simulation software of the same name.<sup>135</sup> The force field is continuously improved. The initial parameter set in GROMOS87 was based on the work of Dunfield *et al.*<sup>136</sup> Parameters were initially optimized on crystal lattice constants and lattice energies and further tested in simulations. GROMOS force fields are designed for use with SPC water model.<sup>137</sup>

The rather unusual approach of GROMOS force fields can be demonstrated on a treatment of Lennard-Jones parameters. In general, each atom has defined radius for different types of interaction with polar and non-polar partners as well as for hydrogen bonding. The same holds for Lennard-Jones parameters of vicinal atoms (1–4) that often substitute torsional terms. Dihedral potentials are employed only in necessary cases.

The individual parameter sets are named according to the number of atom types presented in the force field.

**43A1, 43A2 and 45A3**<sup>138,139</sup> These force fields contain revised Lennard-Jones parameters. They were obtained by matching state properties and thermodynamics of model compounds at 298K. In particular, the parameters of hydrocarbons were adjusted to reproduce free energy of hydration.

**53A5 and 53A6**<sup>140</sup> These parameter sets resulted from extensive optimization of polar atom types. It was achieved by fitting of liquid properties for 28 small molecules containing important functional groups followed by matching solvation free energies of 14 amino acid analogs in cyclohexane (53A5) or water(53A6).

**54A7**<sup>141</sup> This variant introduced new torsional angle terms for protein backbone that improves stability of proteins and peptides in simulations. Hydrogen bonding between peptide bond moieties was optimized simultaneously. The other changes involved new parameters for ions and one extra atom type for better treatment of phospholipides.

**54A8**<sup>142</sup> The non-bonded parameters for charged amino acid were revised and calibrated on experimental thermodynamic data of hydration.

### 1.2.5 ENCAD

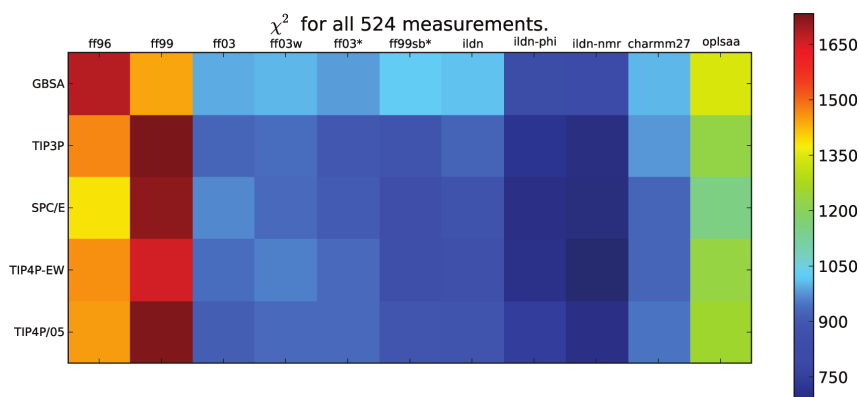
Encad force field<sup>143</sup> was developed for use in the computer simulation program ENCAD (Energy Calculation and Dynamic).<sup>144</sup> This force field is also massively used in Dyanameomics project.<sup>57</sup>

ENCAD continued in legacy of original CFF<sup>89</sup> and employed simplistic design and generic parameters. Emphasis was placed on consistent treatment of solutes and solvent, energy conservation and truncation scheme for calculation of non-bonded interactions. The bond and angle parameters were obtained from survey of crystallographic data and non-bonded parameters were optimized to reproduce lattice constant and sublimation enthalpies. Water molecules in simulations with ENCAD should be modeled by flexible F3C potential.<sup>145</sup>

### 1.2.6 Performance of force fields

An important question raised naturally considering the number of different force fields and their variant is: Which of them is the most reliable for simulation of proteins and peptides? This results in massive comparison and test studies. The first validations are usually provided by authors of the force field in question to show stability of folded globular proteins. As a result all-atom force fields together with explicit representation of solvent provide similar picture of stable globular proteins under physiological conditions.<sup>146</sup>

The more stringent tests of force fields represent simulations of peptides. They can be verified against experimental observables such as NMR shifts and coupling constants or thermodynamics of secondary-structure formation. It was demonstrated soon that different force fields predict very distinct conformations to be populated.<sup>108,147–150</sup> Amber ff94, ff99, ff94-GS, ff99- $\phi$  and ff03 have been shown to be biased toward helical conformations.<sup>108</sup> ff99SB, charmm22/CMAP, OPLSA-AA/L, GROMOS 43a1 and 53a6 produce distributions underestimating the content of polyproline-like conformers and are biased



**Figure 1.3:** Comparison of force fields against NMR experimental data. This plot expresses the overall  $\chi^2$  for 524 NMR measurements. The lower the  $\chi^2$  value, the better agreement for given combination of force field and water model was achieved. This figure was adopted from Ref.151

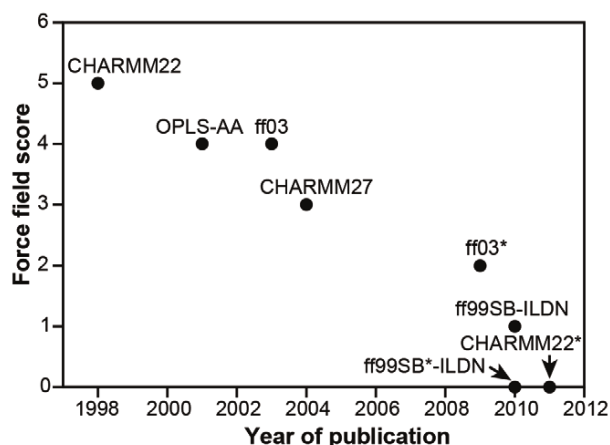
toward helical or extended structures.<sup>41</sup> The effort to match correctly helical content and thermodynamics of helix formation led to specific corrections and “star” versions of the established force fields, i.e. ff99SB\*, ff03\*, charmm22\* and ff03w. The performance of contemporary force fields was recently benchmarked against experimental NMR measurements on dipeptides, tripeptides, tetra-alanine and ubiquitin.<sup>151</sup> The best agreement were achieved by ff99sb- $\phi$  and ff99sb-nmr variants updated by ff99sb-ILDN parameters for side chains. See Fig. 1.3 for comprehensive summary of their analysis.

The atomistic description of protein folding process and unfolded states could be the ultimate benchmark of force field quality.<sup>127</sup> It is known that force fields preferentially fold either  $\alpha$ -helical or  $\beta$ -strand structures and precise balancing must be undertaken to improve this behavior.<sup>152–154</sup> Optimistic outlooks were presented recently by Lindorff-Larsen *et al* ranking the current force fields according their ability to fold small proteins and peptides and reproduce the structure and fluctuation of ubiquitin.<sup>155</sup> Their results suggested that force fields are continuously getting better (see Fig. 1.4) in description of structural and dynamic features of proteins as well as protein folding.

### 1.2.7 Remarks to force field development.

The presented overview of the empirical simple point-charge non-polarizable force fields unmasked the trends in their developments in the last decades. As the extensive simulations of proteins began to be feasible due to the enormous progress in computational hardware the need for better calibrated force field intensified. Firstly, the force field were improved by more advanced *ab initio* quantum calculations of backbone conformers or  $\phi/\psi$  profiles. This way led to many improvements, however, only the expensive high level *ab initio* methods can produce precise reference values capable of matching experimental data.<sup>156</sup> The current force fields still do not reach the precision needed for qualitative and quantitative reproduction of experimental thermodynamic or structural data on protein folding in solution. To address this problem directly the experimental data such as NMR shifts or coupling constants started to be utilized in parametrization process.





**Figure 1.4:** Force fields getting better.

The quality of reproduced data on protein folding, structure and fluctuations of proteins were ranked. The lower rank means better agreement with experiments. The figure was adopted from Ref. 155.

The shortcomings in parametrization of amino acid side chain propensities were recently recognized and tackled in several works.<sup>112,125</sup> However, also the assumption that single backbone parameters can describe propensities of all amino acids (neglecting Gly and Pro) started to be questioned.<sup>134</sup>

*The parametrization of a force field is an extremely laborious activity. Therefore their development is gradual and the newer version often still contains parameters two or three decades old. There is a clear imbalance between force field development and the fast progress in the field of computational methods and algorithms. Nevertheless, the recent activities indicate an efforts to revise aged Lennard-Jones parameters,<sup>157</sup> replace outdated water models<sup>114</sup> or to develop new techniques for assignment of partial charges.<sup>116</sup> These changes seem to be necessary since the protein folding simulations and characterization of disordered or unfolded ensembles became the new frontiers for force fields.*

## 1.3 Metadynamics

### 1.3.1 Overview of the method

Metadynamics is a method for calculation of free energy profiles and accelerating of rare events in molecular dynamics simulations. Since it was proposed by Laio and Parrinello,<sup>158</sup> metadynamics has been successfully applied on various problems and fields of interest. The versatile usage of metadynamics has been already extensively reviewed elsewhere.<sup>159–163</sup> However, it is beneficial to illustrate at least the fields where metadynamics has been successfully applied:

- Exploring of reaction mechanisms and pathways<sup>164–166</sup>
- Structural transition of biomolecules<sup>167</sup>
- Conformational preferences of flexible molecules such as peptides<sup>149,150,168</sup>
- Docking ligands to proteins<sup>169,170</sup>
- Protein–protein interactions<sup>171 172</sup>
- Protein folding<sup>173–176</sup>
- Packing molecules in crystals<sup>177,178</sup>
- Study of phase transition in solids or liquids<sup>179–181</sup>
- Study of adsorption of molecules on surfaces<sup>182</sup>
- Generation conformational ensembles of biomolecules constrained by experimental data<sup>42</sup>
- Design molecular machines<sup>183</sup>

Metadynamics possesses some characteristics common to other methods for calculation of free energy but also some unique features. The heart of the algorithm is dimensional reduction. It assumes that the processes or system of interest can be described by small number of parameters which are called collective coordinates (variables) or order parameters. Nevertheless, these descriptors must be specified and provided on the beginning of the simulation but no *a priori* knowledge about the underlying free energy profile is needed. Afterwards, the selected descriptors are efficiently biased in course of simulation what significantly improves sampling. The bias potential acting on collective coordinates is built continuously in an adaptive fashion in order to explore unvisited regions. The updates of the bias potential destabilize immediately the current configuration of the system and force the transitions between one state to another one.

The fundamental idea of metadynamics is that the bias potential itself can be used as an unbiased estimator of the free energy which is a function of the collective coordinates. Such concept was firstly recognized heuristically because it did not result obviously from any established relations of thermodynamics or statistical mechanics. However, after successful practical demonstrations on various systems<sup>158</sup> the rigorous derivations and proofs for Langevin dynamics were provided.<sup>184</sup> Recently, the equivalence between metadynamics and established Wang-Landau method<sup>185</sup> for enhanced sampling in Monte Carlo simulations was demonstrated.<sup>186</sup>

One of the main advantage of metadynamics is the straightforward parallelization of the algorithm. The same free energy landscape can be simultaneously reconstructed by several metadynamics simulations sharing the same bias potential. This method—multiple walkers metadynamics<sup>187</sup>—is approaching the desirable linear scaling limit.

### 1.3.2 Algorithms

The most general way how to introduce the metadynamics and its variant is to follow the extended Lagrangian approach.<sup>164</sup> The Lagrangian of the simulated system can be extended by set of extra degrees of freedom (auxiliary variables)  $\mathbf{s}$  with fictitious mass  $M$  and kinetic energy  $\frac{1}{2}M\dot{\mathbf{s}}^2$ . The coupling between auxiliary variables and the actual collective coordinate of the system  $\mathbf{S}(\mathbf{x}, t)$  is established by means of harmonic potential  $\frac{1}{2}k(\mathbf{S}(\mathbf{x}) - \mathbf{s})^2$  with stiffness  $k$ . The extended Lagrangian of the system can be written as:

$$L^{ext}(\mathbf{x}, \dot{\mathbf{x}}, \mathbf{s}, \dot{\mathbf{s}}) = T(\dot{\mathbf{x}}) - V(\mathbf{x}) + \frac{1}{2}M\dot{\mathbf{s}}^2 - \frac{1}{2}k(\mathbf{S}(\mathbf{x}) - \mathbf{s})^2, \quad (1.1)$$

where  $T(\dot{\mathbf{x}})$  states for the kinetic and  $V(\mathbf{x})$  for the potential energy of an unextended uncoupled system. Afterward, the free energy as a function of the auxiliary variables follows:

$$A^{ext}(\mathbf{s}) = -\frac{1}{\beta} \ln \int e^{-\beta(V(\mathbf{x}) + \frac{1}{2}k(\mathbf{S}(\mathbf{x}) - \mathbf{s})^2)} d\mathbf{x}. \quad (1.2)$$

Finally it was proven that the free energy of an enhanced system  $A^{ext}(\mathbf{s})$  approaches the free energy of uncoupled system  $A(\mathbf{S}(\mathbf{x}))$  for large values of  $k$ :

$$A(\mathbf{S}(\mathbf{x})) = \lim_{k \rightarrow \infty} A^{ext}(\mathbf{s}). \quad (1.3)$$

Simultaneously, metadynamics extends the Lagrangian by additional history dependent bias potential modified in course of simulation. In this scheme the bias potential exerts influence solely upon the auxiliary variables. Coupling of auxiliary variables to collective coordinates mediates indirectly the effect of bias potential to the rest of the system. The bias potential is usually built from gaussian functions with the same dimensionality ( $d$ ) as the space of the auxiliary variables. Finally the bias potential is given by:

$$V^{bias}(t, \mathbf{s}) = \sum_{\tau < t} W \prod_{i=1}^{i \leq d} e^{-\frac{|s_i(t) - s_i(\tau)|^2}{2\sigma_i^2}}, \quad (1.4)$$

where  $W$  controls the “height” and  $\sigma$  the “width” of the bell-shaped gaussian functions.

Metadynamics is based on validity of the assumption

$$A(\mathbf{S}(\mathbf{x})) \approx - \lim_{t \rightarrow \infty} V^{bias}(\mathbf{s}, t) + \text{const.} \quad (1.5)$$

The obtained bias potential  $V^{bias}(\mathbf{s})$  plays a role of an estimator of unbiased free energy  $A(\mathbf{S}(\mathbf{x}))$ . This key idea makes metadynamics different from other methods for free energy calculations. Importantly, the relation 1.5 suggests that the free energy as an equilibrium quality can be evaluated from non-equilibrium dynamics under the continuously updated bias potential.<sup>184</sup>

The placement of an updating gaussian is determined by the actual coordinates of the auxiliary variables and it is recorded. If the system gets trapped in a deep minimum the destabilizing effect of the stored gaussians acts as an energetic penalty and the system is facilitated to escape from the current minimum by crossing the lowest barriers. Each perturbation caused by a change of the bias potential pulls out the system from an equilibrium. Consequently, the system tends to relax and adapt to new bias potential. If the changes are gradual but continuous the configuration space can be explored very efficiently. After visiting all accessible regions the collected bias potential should completely compensate the underlying free energy landscape. The dynamics of auxiliary variables become diffusive and ideally independent on the original potential. It means that the biased probability distribution gets flat after deposition of sufficient amount of bias potential. However, because of finite size of a gaussian each additional deposition disturbs the exact compensation and introduces an error. Fortunately, metadynamics tends to cancel this kind of error by preferential deposition of other gaussians on the distant positions. As a result, the bias potential oscillates around the correct solution but does not converge in time.

The solution to this fundamental convergence issue is provided by well-tempered metadynamics.<sup>188</sup> The novel algorithm rescales the weight of the newly added gaussians by a term dependent on the magnitude of the actual bias potential acting on the current position:

$$W(\mathbf{s}) = W e^{-\frac{V^{bias}(\mathbf{s},t)}{k_B \Delta T}} \quad (1.6)$$

The direct consequence of the scaling is that the bias potential converges smoothly as the simulation progresses and hence the sampling converges to defined statistical ensemble. However, under such conditions the bias potential does not fully mirror the underlying free energy profile, but is related as:

$$V^{bias}(\mathbf{s}, t \rightarrow \infty) = -\frac{T + \Delta T}{\Delta T} A(\mathbf{S}(\mathbf{x})) + \text{const}, \quad (1.7)$$

where  $T$  stands for temperature of the system and  $\Delta T$  for a parameter that regulate the extent of sampling. Classical molecular dynamics is obtained for  $\Delta T \rightarrow 0$  and standard metadynamics is recovered for  $\Delta T \rightarrow \infty$ .

If the dynamics of the system and the auxiliary variables span different time scales, the separation of both components can be achieved and the approximation of a mean field can be formally introduced for dynamics of both parts. This procedure leads to two approaches—the discrete and direct metadynamics.<sup>159</sup>

The direct metadynamics is the far most popular version of metadynamics. Allowing formally the fictitious mass of auxiliary variables to be zero, their dynamics follow exactly the actual collective coordinates of the system of interest. Similarly, the forces originated

from bias potential act straightforwardly on the real particles associated with collective coordinates. No parameters for auxiliary variables or the coupling are required. However, efficient implementation of such approach requires modification of the MD code, because the instantaneous evaluation of bias potential and associated forces is needed in each step of molecular dynamic.

### 1.3.3 Collective coordinates

A choice of suitable collective coordinates (CVs) is a necessary condition for correct functionality of metadynamics and its successful application on system of interest. For the best performance the following properties and requirements on CVs must be fulfilled:<sup>159</sup>

- CVs must be continuous and differentiable function of atomic coordinates
- CVs should distinguish all important states, conformers, structures, phases, reactants, products or intermediates of the reactions or transitions.
- CVs should describe important slow modes, which limit the rate of studied events
- Number of used CVs should not exceed the limit given by unfavorable scaling of metadynamics for each additional dimension in space of auxiliary variables.

The basic type of collective coordinates are based on geometrical parameters. The internal coordinates such as distances, angles, dihedral angles between atoms or groups of atoms can be straightforwardly used if they are able to distinguish different states or conformers:

$$r_{ab} = |\vec{r}_a - \vec{r}_b| = \sqrt{(r_{ax} - r_{bx})^2 + (r_{ay} - r_{by})^2 + (r_{az} - r_{bz})^2} \quad (1.8)$$

$$\theta_{abc} = \cos^{-1} \left( \frac{\vec{r}_{ab} \cdot \vec{r}_{bc}}{|\vec{r}_{ab}| |\vec{r}_{bc}|} \right) \quad (1.9)$$

$$\phi_{abcd} = \tan^{-1} \left( \frac{|\vec{r}_{bc}| \vec{r}_{ab} \cdot (\vec{r}_{bc} \times \vec{r}_{cd})}{(\vec{r}_{ab} \times \vec{r}_{bc}) \cdot (\vec{r}_{bc} \times \vec{r}_{cd})} \right) \quad (1.10)$$

Furthermore, the mutual orientation of molecules can be described using Euler angles, e.g. for ligand and receptor in course of binding.<sup>169</sup>

Apart from the internal coordinates the coordination numbers contain information about spatial distribution and contact between particles:

$$C = \sum_{a,b} \frac{1 - (r_{ab}/r_0)^n}{1 - (r_{ab}/r_0)^m}, \quad (1.11)$$

where parameter  $r_0$  determines the reference length for the contact and exponents  $n$  and  $m$  control diminishing at larger distances. For well chosen  $n$  and  $m$  the value of the fraction in eq. 1.11 goes to one or zero rapidly for distance  $r_{ab}$  lesser or greater than the reference, respectively. It may be employed to count chemical bonds between atoms,<sup>165</sup> hydrogen bonds or hydrophobic contacts.<sup>171</sup>

Analysis of normal or essential modes can identify independent collective motions in biomolecules. The slowest ones are usually associated with important conformation or functional changes. Metadynamics in essential coordinates allows to map such modes and accelerate collective motions comprising many atoms.<sup>189</sup>

Particular molecules, e.g. proteins or peptides offer natural and convenient descriptors for their conformational states. It might be a content of the secondary structure elements, sequential correlations between Ramachandran angles or radius of gyration what can be used for construction of free energy landscape of folding.<sup>173,174</sup>

The complex process cannot be often described by simple combination of geometrical criteria but it is still well characterized structurally as a series of consecutive events. The class of path CVs<sup>190</sup> provides a solution for these cases. Path CVs are capable of tracing the free energy minimum pathways from the state A to B, using the 2 variables defined as:

$$s(\mathbf{x}) = \frac{1}{P-1} \frac{\sum_{l=1}^P (l-1) e^{-\lambda \|\mathbf{S}(\mathbf{x}) - \mathbf{S}(\mathbf{x}(l))\|^2}}{\sum_{l=1}^P e^{-\lambda \|\mathbf{S}(\mathbf{x}) - \mathbf{S}(\mathbf{x}(l))\|^2}} \quad (1.12)$$

$$z(\mathbf{x}) = -\frac{1}{\lambda} \ln \sum_{l=1}^P e^{-\lambda \|\mathbf{S}(\mathbf{x}) - \mathbf{S}(\mathbf{x}(l))\|^2} \quad (1.13)$$

Here, the reference frames  $\mathbf{x}(l)$  are still required to delimit the path and guide the transitions. The progress on the path is monitored by variable  $s(\mathbf{x})$  and similarly, the distance from the laid out pathway describes variable  $z(\mathbf{x})$ .

An interesting method emerges if the potential energy of the system is used as a collective coordinate. The resulting well-tempered ensemble<sup>191</sup> provides extended sampling in comparison to the canonical ensemble from classical MD and does not distort substantially ensemble averages of other properties. Another usage involves study of phase transitions that can be induced in collective fashion.<sup>181</sup>

### 1.3.4 Selection of parameters

The quality and reliability of reconstructed free energy profiles are dependent on parameters that drive building of the bias potential. The shape of the hills corresponding to the height of the gaussian  $W$  and its width  $\sigma$  influence directly the spatial and energetic resolution of the resulting profile. The finite height of the gaussian determines the error in the energy domain. It can be significantly reduced by employing well-tempered metadynamics<sup>188</sup> that guarantees convergence of the resulting profiles in the chosen range of energies due to the auto-adaptive scaling of  $W$ .

Similarly, the  $\sigma$  parameter regulates the spatial resolution of the free energy profiles for collective coordinates. The large  $\sigma$  produces broad gaussians which fill the regions on a landscape quickly but smear out the fine details. On the other hand, using narrow gaussians leads to rough and spiky profiles and low efficiency of filling free energy basins. The recently proposed algorithm<sup>192</sup> can deal with these issues and adaptively controls the resolution. The dimensions and placements of the hills are chosen according to the

dynamics of collective coordinates in the actual region. If the dynamics is diffusive, the broad gaussian is deposited there because of flat character of underlying free energy landscape and *vice versa*.

The third parameter—the deposition interval  $\tau$ , controls the deposition rate and hence the convergence and error of resulting free energy profiles. The deposition of gaussian cannot be too fast because the system must relax sufficiently after each perturbation. If this condition is not met an artificial behavior can manifest completely distorted dynamics. The errors in the reconstructed landscapes are reduced by longer time gaps which also unfortunately result in longer simulations. The proper relaxation facilitates the correct sampling of microstates necessary for reliable calculation of the free energy.

The mean error ( $\bar{\epsilon}$ ) in the free energy estimates for Langevin metadynamics was approximated as:<sup>184,193</sup>

$$\bar{\epsilon} \propto \sqrt{\frac{W\sigma^d}{D\tau}}, \quad (1.14)$$

where  $D$  states for an intrinsic diffusion coefficient of the system in the space of collective coordinates and  $d$  for the dimensionality of CVs.

This formula provides clear suggestion for reducing of the mean error. However, decrease of  $W$  and  $\sigma$  implies higher computational costs as well as increase of deposition interval  $\tau$ . Not surprisingly, the suitable parameters must be chosen as a compromise between accuracy, stability and computational efficiency.

### 1.3.5 Problems of metadynamics and advanced computational schemes

Usually some experience and knowledge of the system of interest is necessary for making efficient computational setup. It involves a correct choice of the collective coordinates and corresponding parameters for construction of the bias potential. The correct decision could not be obvious from the beginning if the properties of the system remains *a priory* unknown. Therefore, a method of trial and errors often precedes the production runs of the metadynamics.

The inappropriate choice of collective variables causes the most serious problems, particularly if the CVs corresponding to the slow modes are omitted. Metadynamics is capable of acceleration of rare events and extended sampling only for the biased set of CVs. If they cannot capture an important slow process, a hysteresis in the simulations appears. As a direct consequence, free energy profiles show poor convergence that depends on the starting conditions. These limitations can be overcome by combining metadynamics with other approaches that help to sample the unbiased “transverse coordinates”.<sup>174</sup>

The algorithm of metadynamics performs excellently in combination with replica exchange framework.<sup>194</sup> If each replica simulates the same system in different temperatures the

Parallel Tempered Metadynamics (PTMetaD)<sup>184</sup> is obtained. The regular attempts to switch systems between replicas are undertaken and accepted with probability given by:

$$P(i \rightarrow j) = \min \left\{ 1, \exp \left[ \left( \frac{1}{T_j} - \frac{1}{T_i} \right) (V(\mathbf{x}_j) - V(\mathbf{x}_i)) + \frac{1}{T_i} (V_i^{bias}(\mathbf{s}(\mathbf{x}_i)) - V_i^{bias}(\mathbf{s}(\mathbf{x}_j))) + \frac{1}{T_j} (V_j^{bias}(\mathbf{s}(\mathbf{x}_j)) - V_j^{bias}(\mathbf{s}(\mathbf{x}_i))) \right] \right\} \quad (1.15)$$

This rule grants the correct statistical sampling in each replica. Moreover, the crossing of all barriers is facilitated due to the higher temperature in some replicas. The exchange of structures between replicas then prevents the low-temperature replicas from getting stuck in one minimum and significantly improves sampling.

The bias exchange metadynamics<sup>174</sup> uses replicas which are simulated at the same temperature but biased in different collective coordinates. Therefore, different barriers may be crossed in each replica according to the utilized bias. Similarly to parallel tempered metadynamics the exchange of systems between replicas facilitates the sampling in each of them. Acceptance probability of exchange is given by detailed balance rule as:

$$P(i \rightarrow j) = \min \left\{ 1, \exp \left[ \frac{1}{T} (V_i^{bias}(\mathbf{s}(\mathbf{x}_i)) - V_i^{bias}(\mathbf{s}(\mathbf{x}_j)) + V_j^{bias}(\mathbf{s}(\mathbf{x}_j)) - V_j^{bias}(\mathbf{s}(\mathbf{x}_i))) \right] \right\}. \quad (1.16)$$

The another benefit of the method follows from the set of different free energy profiles obtained from each replica.

From practical point of view, the parallel tempered metadynamics suffers from the same unfavorable scaling as standalone parallel tempering method—the bigger the system, the smaller difference in temperature is allowed for acceptable exchange ratio between them. As a result prohibitively large number of replicas is needed for large systems. On the other hand bias replica exchange does not manifest this behavior but various collective coordinates must be employed for proper and balanced sampling.

Since metadynamics biases the probability distribution of selected collective coordinates, the distribution of other microscopic and macroscopic properties is distorted in metadynamics simulation. Recovering their unbiased distribution is non-trivial task in contrast to those described by collective coordinates. However, the reweighting scheme was proposed<sup>195</sup> and allows to reconstruct the Boltzman distribution. This scheme provides the opportunity to calculate equilibrium distribution of any quality without respect to the collective coordinates used in simulations.<sup>196</sup>



## 2 Aims of the thesis

The computational studies of peptides and miniproteins are challenging tasks for two principal reasons: i) uncertain or unsatisfactory precision of the force fields and ii) high demands on the conformational sampling in the computer simulations. In the present thesis I attempted to deal with the following challenges:

1. Assessment of force fields for simulation of peptides.
2. Investigation of conformational preferences of amino acids in water environment and in solutions with cosolvents.
3. Design and testing of collective coordinates for metadynamics of peptides.
4. Characterization and reproduction of experimental properties of designed miniprotein.
5. Thorough parametrization of a cosolvent for correct water-mixtures properties and reliable usage for simulations of proteins and peptides.

Each topic ●1–5 was elaborated as an individual study published or submitted to peer-reviewed scientific journal. All related published articles or manuscripts are attached to this thesis as Appendices A–F in the following order:

**Appendix A:** Vymětal, J.; and Vondrášek, J. Critical Assessment of Current Force Fields. Short Peptide Test Case. *Journal of Chemical Theory and Computation* **2013**, *9*, 441–451.

**Appendix B:** Vymětal, J.; and Vondrášek, J. The DF-LCCSD(T0) correction of the  $\varphi/\psi$  force field dihedral parameters significantly influences the free energy profile of alanine dipeptide. *Chemical Physics Letters* **2011**, *503*(4-6), 301-304.

**Appendix C:** Towse, C.-L.; Vymětal, J.; Vondrášek, J.; and Daggett, V. Potential for underestimating residual structure in denatured states and intrinsically disordered proteins (*submitted*)

**Appendix D:** Vymětal, J.; and Vondrášek, J. Gyration- and inertia-tensor-based collective coordinates for metadynamics. Application on the conformational behavior of polyalanine peptides and Trp-cage folding. *The journal of physical chemistry. A* **2011**, *115*, 11455–65.

**Appendix E:** Vymětal, J.; Reddy, B.S.; Černý, J.; Chaloupková, R.; Žídek, L.; Sklenář, V.; and Vondrášek, J. Retro Operation on the Trp-cage Miniprotein Sequence Produces an Unstructured Molecule Capable of Folding Similar to the Original Only upon 2,2,2-trifluoroethanol Addition. (*submitted*)

**Appendix F:** Vymětal, J.; and Vondrášek, J. Parametrization of 2,2,2-trifluoroethanol based on Generalized Amber Force Field provides realistic agreement between experimental and calculated properties of pure liquid as well as water mixed solutions. (*submitted*)

The following Chapter 3 summarizes the work on the topics ●1 and ●2. The subsequent chapters originate from the other studies in Appendices and are devoted one by one to the topics ●3, ●4, ●5.

## 3 Intrinsic conformational preferences of amino acids

### 3.1 Motivation

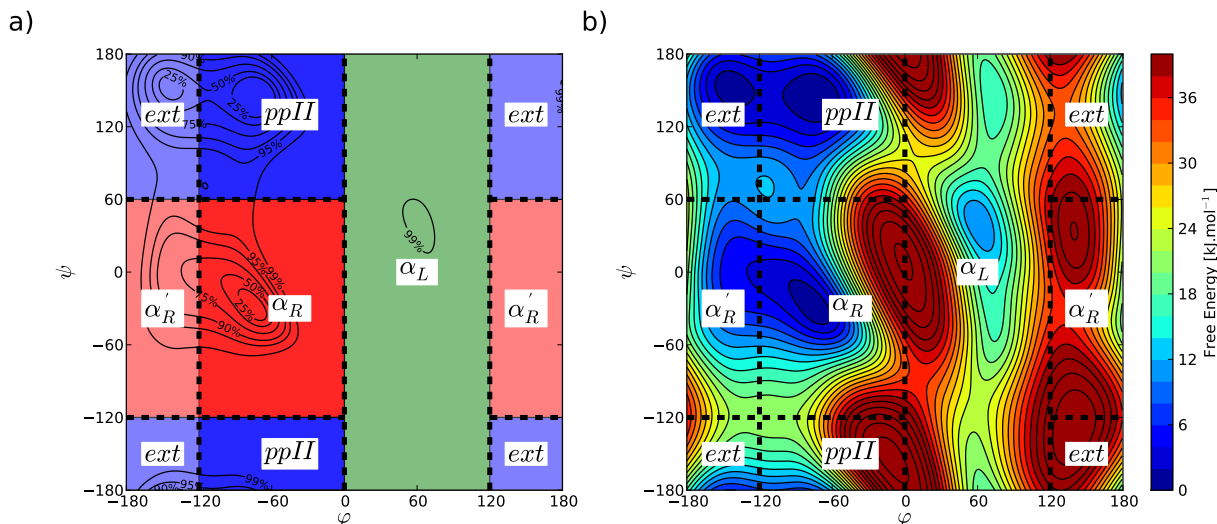
Short peptides are used traditionally as experimental and theoretical models of the unfolded, denatured or intrinsically disordered states of proteins.<sup>197,198</sup> Various studies<sup>199–203</sup> proved that the sampling of conformational space by peptide backbone is neither random nor uniform as could be inferred from random coil model<sup>204</sup> employed in the last decades.

However, the conformational states preferred by short flexible peptides cannot be directly and unambiguously determined by any experimental method. The experiments on alanine based peptides initiated intense debate about the character of favored conformations until the prevalence of the lefthanded polyproline-helix-II-like (ppII) conformer was broadly accepted.<sup>205–210</sup> ppII conformer is supposed to contribute significantly to structural diversity of amino acid dipeptides and short host peptides.<sup>202,211–213</sup> Furthermore, the content of ppII seems to be induced by some denaturants—for example by urea.<sup>214,215</sup> Therefore, ppII is supposed to participate prevalently in unfolded and denatured states of proteins<sup>215–218</sup> although this topic is still the subject of a controversy.<sup>207,219</sup>

Since the first three-dimensional structures of proteins were elucidated, different propensities of amino acids to form  $\alpha$ -helices and  $\beta$ -strands were noticed.<sup>220</sup> Distinct populations of basins on Ramachandran plot were further observed for amino acids in irregular structure elements in proteins compiled in coil libraries.<sup>221</sup> Different trends in conformational preferences can be observed by spectroscopic methods in various host–guest peptides such as GGXGG,<sup>211</sup> GXG,<sup>202</sup> AXA<sup>199</sup> or GPPXPPGY.<sup>213</sup> However, amino acids themselves in the form of dipeptides manifest distinguishable intrinsic backbone preferences.<sup>212,222</sup>

The individual intrinsic conformational preferences of amino acids are assumed to play a role in protein folding and residual structure of unfolded proteins.<sup>198</sup> The propensities of the most amino acids obtained from peptide models correlate with those derived from carefully constructed coil libraries<sup>223</sup> or  $\beta$ -strand statistics.<sup>211</sup> Nevertheless, the conformational preferences can be modulated by effects of sequence neighbors.<sup>203,221,223</sup> The importance of the systematic experiments mapping such interactions has been recognized and the first studies appeared recently.<sup>218,224</sup>

Molecular dynamics can provide an insight into repertoire of conformations sampled by short peptides. The systematic studies of amino acids in different sequential contexts have been already conducted.<sup>201,225,226</sup> However, these results critically depend on the quality of used force fields and cannot be validated without direct confrontation with experimental data. We already tested the performance of force fields on the model molecule of alanine dipeptide<sup>149</sup> and demonstrated that high level *ab initio* calculation as reference for parametrization can improve agreement with experiment.<sup>156</sup> In order to point out the differences between force fields we compared comprehensively backbone and side chain conformational preferences of amino acid dipeptides in four common force fields.<sup>150</sup>



**Figure 3.1:** Definition of backbone conformers.

The definition of conformers was based of free energy profiles in terms of  $\phi/\psi$  torsions. The boundaries of the individual regions (*ext*, *ppII*,  $\alpha_R$ ,  $\alpha'_R$  and  $\alpha_L$ ) are depicted in panel (a) and superposed on a typical free energy profile (b).

Additionally, our interest in propensities under different conditions resulted in a collaborative computational study that mapped the effect of denaturant on amino acids in the host peptides (*submitted*, Appendix C).

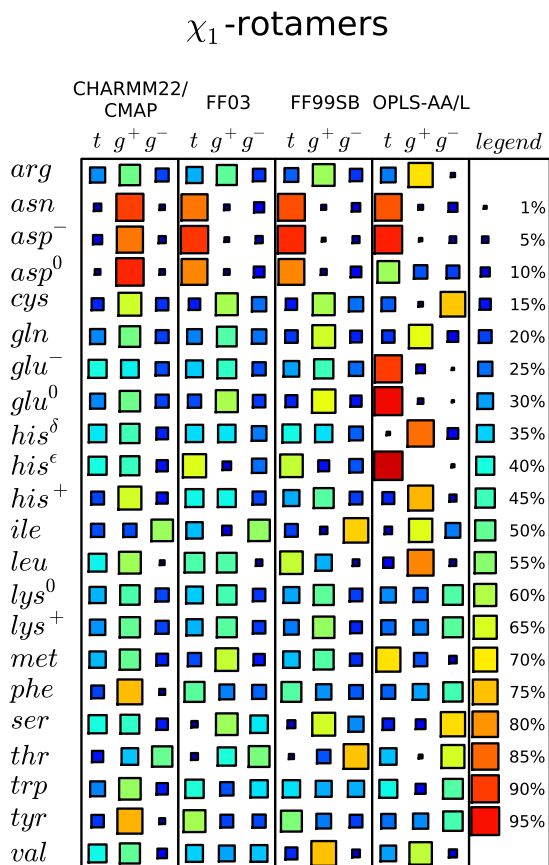
## 3.2 Intrinsic propensities of dipeptides in different force fields

Although the amino acids dipeptides seem to be trivial targets for molecular dynamics the reliable and converged sampling of all relevant conformational states (including backbone and side chain conformers) was found to be harder than it is broadly expected. Because the backbone preferences are inevitably coupled with the side chain conformations (rotamers), both of them have to be sampled properly. To achieve this goal, we used bias exchange metadynamics<sup>174</sup> that significantly improved the sampling due to the exchange of replicas. Each replica was biased in different combination of collective variables represented by backbone torsions  $\phi$ ,  $\psi$  and the two most important side chain torsion  $\chi_1$  and  $\chi_2$  (if present in the particular dipeptide).

The typical  $\phi/\psi$  profile obtained by metadynamics for all residues (except glycine and proline) is depicted in Fig. 3.1 together with a scheme how we partitioned the Ramachandran plot and defined individual backbone conformers in terms of  $\phi$  and  $\psi$  torsions. The designed 5-state model covered all regions of the Ramachandran plot as well as the reduced 3 state model that merges the *ext* and *ppII* region (E) and analogously  $\alpha'_R$  and  $\alpha_R$  (H). The coarser partitioning resulted from insufficiently delimited minima and low barriers manifested by several amino acids.

The output of bias exchange metadynamics was analyzed in terms of population of individual regions. We were interested primarily in the relative trends in propensities of non-alanine amino acids. Since the alanine dipeptide represents standard benchmark



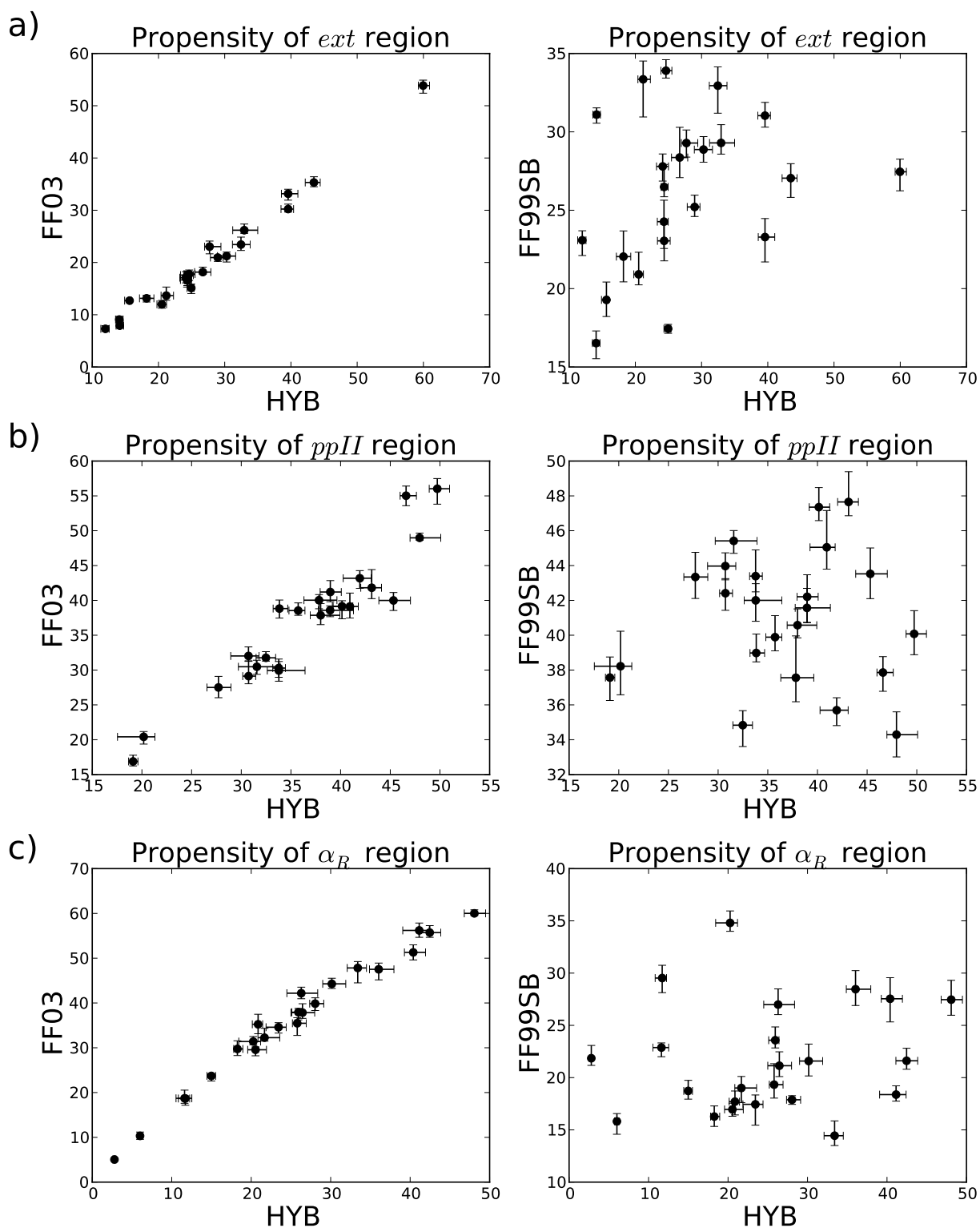


**Figure 3.3:** Propensities for  $\chi_1$  torsion.

Propensities for individual  $\chi_1$  rotamers, *trans* (*t*), *gauche*<sup>+</sup> (*g*<sup>+</sup>), and *gauche*<sup>-</sup> (*g*<sup>-</sup>), are represented in Hinton plot. The area of the square is proportional to the population of the given rotamer. The differences are also emphasized by colors.

of the partial charges became apparent if their direct influence on torsional potentials *via* 1–4 electrostatic interactions is realized. These interactions are usually scaled by factors without any physical justification. Therefore, the different charge distributions and inconsistent treatment of 1–4 electrostatics can be a plausible explanation for disagreement of predicted propensities.

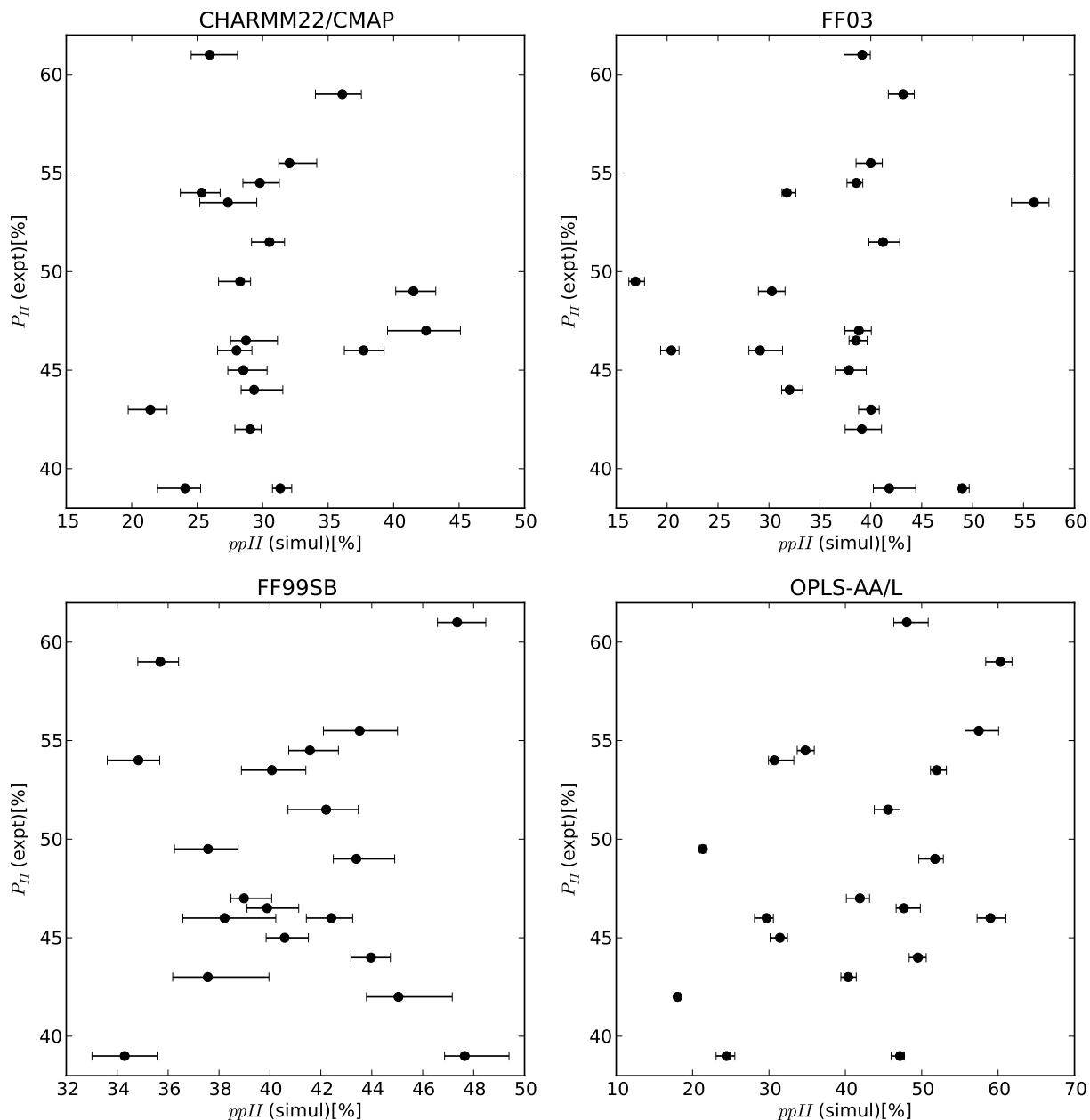
Nevertheless, the final assessment of force fields should involve comparison with experimental data. We compared the propensity scales reported by Grdadolnik *et al*<sup>212</sup> and those obtained from simulations. None of the examined force fields was able to capture the experimental trends correctly, see Fig 3.5. The inability of force fields to reproduce experimental intrinsic propensities of amino acids does not imply that these force fields must fail in description processes such as protein folding. The distinct propensities are determined by very small free energy differences that can be easily overcome by another driving forces maintaining protein stability. However, the intrinsic propensities may still contribute significantly to a character of disordered proteins. Therefore the further force field development should take them into account and focus on properties of the individual amino acids.



**Figure 3.4:** Trends in the hybrid force field (HYB), FF99SB and FF03.

The propensities of the most important conformers—*ext* (a), *ppII* (b) and  $\alpha_R$  (c) correlate obviously between FF03 and HYB, contrary to the FF99SB. The error bars express the range of values obtained from different replicas.

The correlation between experimental and force field propensities



**Figure 3.5:** The correlation between experimental ppII content and predictions of individual force fields.

The very similar pictures were obtained for examined conformers and regions on the Ramachandran plot. The error bars express the range of values obtained from different replicas.

### **3.3 The difference between thermal and chemical denaturation of AAXAA host-guest peptides and their conformational preferences**

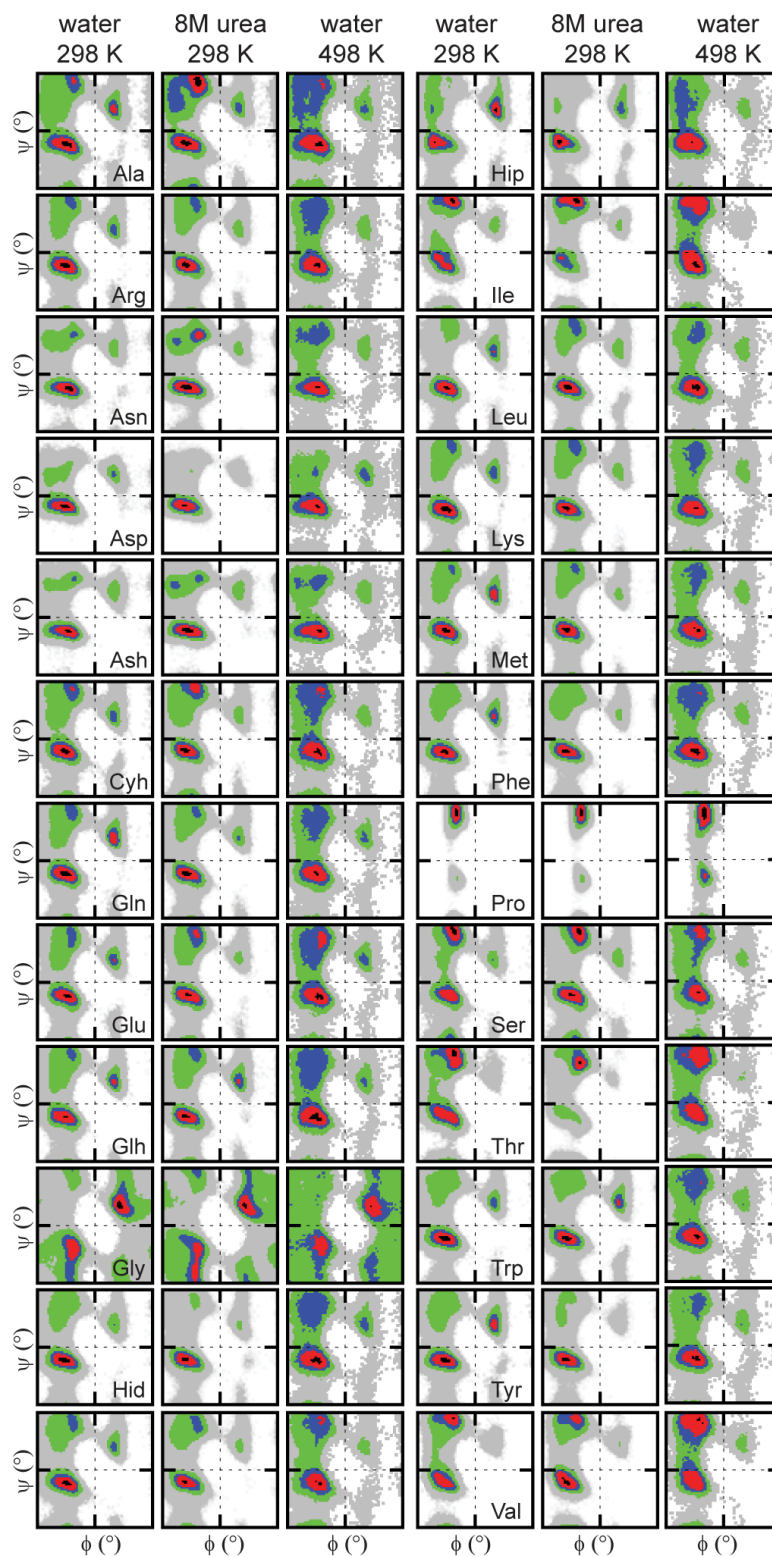
We investigated conformational preferences of amino acids on X position in AAXAA host peptide under different conditions – physiological, thermal (498K) and chemical (8M urea) denaturation. The AAXAA host peptides were chosen as a model of unfolded or denatured protein sequence. The small side chain of alanine restricts the flexibility in respect to the established GGXGG host peptides and thus may provide more realistic picture of real protein chains.

The obtained conformational preferences showed clear bias to certain regions of conformational space rather than random sampling under all conditions (see comparison in Fig. 3.6). However, the conformational bias for individual amino acids was modulated by specific conditions of denaturation. The increased temperature shifted the preferences toward more uniform sampling of all regions for all amino acids. However, the effect of urea on propensities did not result in uniform response. The population of ppII conformers increased or decreased for particular amino acids. Nevertheless, the ppII regions were never sampled exclusively by any amino acid.

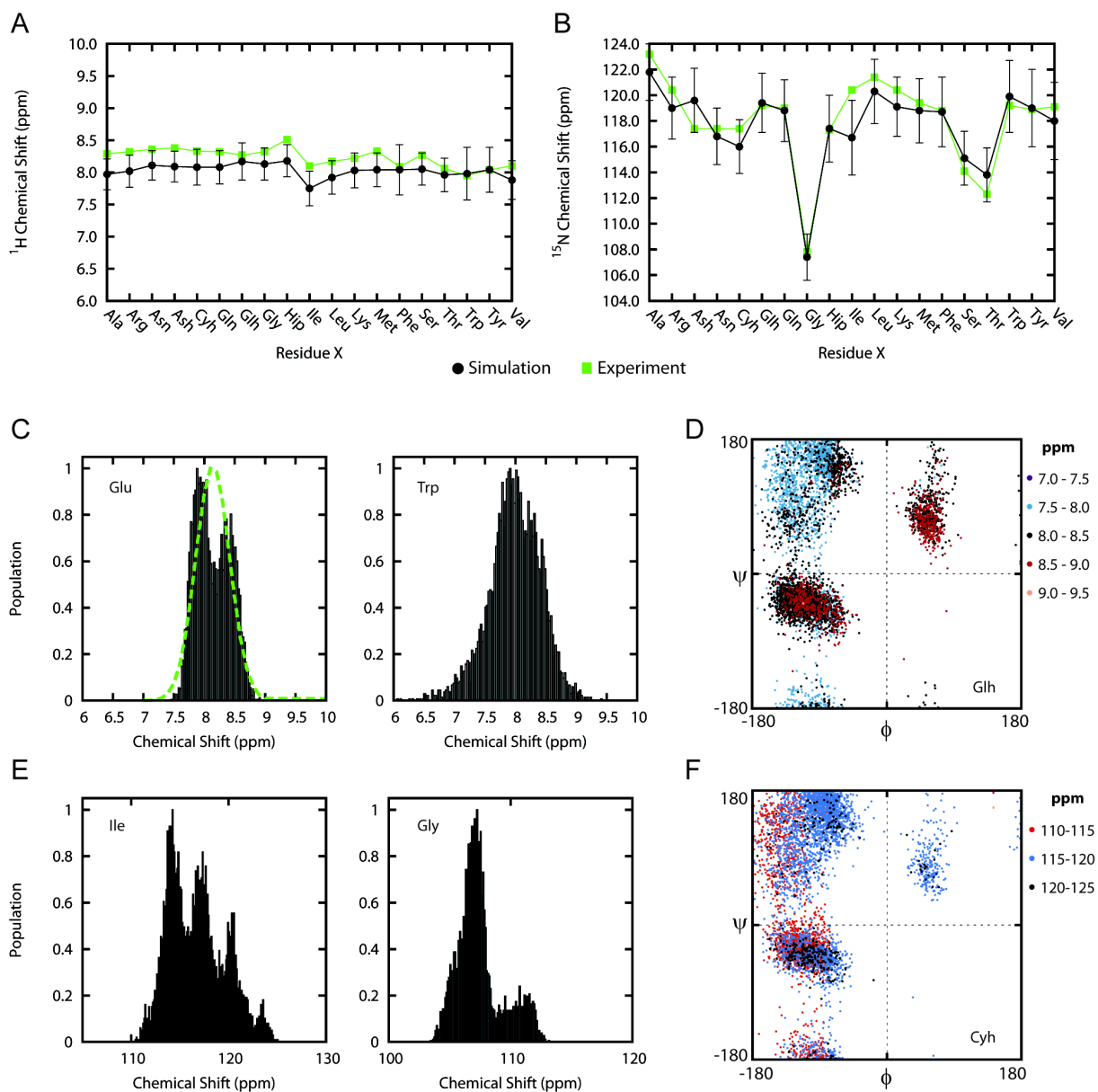
The simulated MD ensembles were validated by chemical shifts calculated by SHIFTX2 program.<sup>227</sup> The predicted shifts were in a close agreement with experimental  $^1\text{H}_\text{N}$  and  $^{15}\text{N}_\text{H}$  chemical shifts obtained from all AAXAA peptides in urea at pH 3 (see Fig. 3.7 panel A and B). Because of rapid interconversion between states, the experimental observables represents average values of the conformational dependent chemical shifts. The calculated chemical shifts from simulations exhibited bi-modal or multi-modal distributions for each amino acids (panel C and E) in apparent accordance with the presence of sampled conformational basins. However, the detailed analysis revealed that there was no correspondence of regions on Ramachandran plot to particular values of chemical shifts, as it is demonstrate in panels D and F). Therefore, it is difficult to elucidate local conformational propensities on basis of “random coil” chemical shifts. This statement was further supported by artificial increasing of ppII populations in the input data. The resulting calculated chemical shifts manifested very low sensitivity on ppII content.

The results suggest that different denaturation conditions specifically shift conformational preferences of amino acids. The content of ppII structures can be affected but still may not become dominant. The surprising insensitivity of simulated chemical shifts on population of ppII region admits the possibility that content of the residual structure in denatured and intrinsically disorder proteins might be underestimated.





**Figure 3.6:** Conformational preferences of amino acids in AAXAA host peptide. Ramachandran plots of the conformational populations of the guest residues in different environments: pure water at 298 K, 8M urea at 298 K, and pure water at 498 K. The conformational regions are colored by increasing percentage population from gray, green, blue, red to black.



**Figure 3.7:**  $^1\text{H}_\text{N}$  and  $^{15}\text{N}_\text{H}$  chemical shifts for the guest X residues.

The calculated data from simulation in 8M urea are compared with the experimental values (A, B). Error bars reflect standard deviation. Distributions of chemical shifts from simulations were examined. Both  $^1\text{H}_\text{N}$  and  $^{15}\text{N}_\text{H}$  provided bi-modal or multi-modal distribution demonstrated for exemplary amino acids in panels C and E, respectively. However, the peaks in distributions do not correspond to the unique basins on Ramachandran plot as it is shown in panels D and F.

## 4 Metadynamics in gyration-tensor-based collective coordinates

### 4.1 Motivation

Simulations of complex biomolecular phenomena such as folding of proteins and peptides can be very difficult task for classical molecular dynamics because of intrinsic barriers of these processes. Methods like metadynamics allow to accelerate rare events on the reaction coordinate very efficiently. However, the focus of the problem is then shifted toward a choice of proper reaction coordinates for the complex events. The quest for universal protein folding coordinates has succeeded only partially. It was demonstrated that many protein folding/unfolding processes can be represented on one-dimensional coordinate given by multidimensional reduction or embedding of relevant descriptors.<sup>228–231</sup> However, such characteristics always have to involve the detailed knowledge of the particular folded state and its native contacts.

Bias exchange metadynamics<sup>174</sup> offers an alternative way how to avoid complex reaction coordinates. Each parallel replica can efficiently sample one or small number of simple collective coordinates. Due to the exchanges between replicas the overall conformation space sampled in the simulation exceeds the extent of any individual collective coordinate. Moreover, the system specific collective coordinates may be inferred later in the post-production stage of simulation.<sup>232</sup> The free-energy profiles in the additional CVs can be then reconstructed from metadynamics reliably by reweighting procedure.<sup>195</sup> We focused on search of simple, robust and molecule-type independent collective variables that could extend the repertoire of the basic CVs applicable in bias exchange metadynamics.

From a general point of view a flexible molecule such as polypeptide can be naturally characterized by its size and shape. This approach has a long history in polymer physics and chemistry.<sup>233–235</sup> The effective size is commonly described by radius of gyration, however, this property is only one of many descriptors derived from gyration tensor. The others are able to express not only size but also proportions and shape of the molecule. In this study we implemented the size and the shape descriptors—components of gyration tensor, principal radii of gyration, asphericity, acylindricity and relative shape anisotropy in a computer code for metadynamics and tested their performance on model peptides and miniproteins.<sup>168</sup>

### 4.2 Theory

Gyration tensor (**S**) and tensor of inertia (**I**) describe the distribution of mass in the molecule determined by position of atoms:

$$\mathbf{S} = \frac{1}{N} \begin{pmatrix} \sum x_i^2 & \sum x_i y_i & \sum x_i z_i \\ \sum x_i y_i & \sum y_i^2 & \sum y_i z_i \\ \sum x_i z_i & \sum y_i z_i & \sum z_i^2 \end{pmatrix}, \quad (4.1)$$

$$\mathbf{I} = \begin{pmatrix} \sum m_i(y_i^2 + z_i^2) & \sum -m_i x_i y_i & \sum -m_i x_i z_i \\ \sum -m_i x_i y_i & \sum m_i(x_i^2 + z_i^2) & \sum -m_i x_i z_i \\ \sum -m_i x_i z_i & \sum -m_i y_i z_i & \sum m_i(x_i^2 + y_i^2) \end{pmatrix}. \quad (4.2)$$

All summations are performed over  $N$  atoms with individual masses  $m_i$ . The cartesian coordinates  $x_i$ ,  $y_i$  and  $z_i$  must be related to the geometrical center and center of mass of the molecule for gyration and inertia tensor, respectively.

Since both tensors are symmetric and positive semidefinite they can be diagonalized with non-negative eigenvalues. The corresponding eigenvectors determine important axes of the object—the geometric axes of an ellipsoid approximating the shape for gyration tensor or principal axes of inertia for tensor of inertia. The three eigenvalues  $S_1$ ,  $S_2$ ,  $S_3$  (sorted in descending order) are related to the length of individual elliptic semi axes and  $I_1$ ,  $I_2$ ,  $I_3$  represent principal moments of inertia around corresponding principal axes.

Moments of inertia provide information on distribution of mass in perpendicular directions to the corresponding axes. For more or less homogeneous objects it also reflects the effective size that can be better expressed by radius of gyration around particular axis:

$$r_g^{ax} = \sqrt{\frac{I^{ax}}{m}}. \quad (4.3)$$

Here  $I^{ax}$  and  $m$  stand for moment of inertia around the given axis and the total mass, respectively. This descriptor should be distinguished from the commonly used radius of gyration in molecular simulations:

$$R_g = \sqrt{\frac{\sum_i m_i r_i^2}{m}}. \quad (4.4)$$

Useful relations emerge for both tensors and their eigenvalues if they are applied on the systems composed from particles of the same mass. However, these relations are retained if the components of gyration tensor get mass-weighted, i.e. multiplied by term  $m_i * N/m$ . This assumption results in the same system of eigenvectors and links all eigenvalues as well as traces of both tensors:

$$I_1 = m(S_1 + S_2), \quad (4.5)$$

$$I_2 = m(S_1 + S_3), \quad (4.6)$$

$$I_3 = m(S_2 + S_3), \quad (4.7)$$

$$\text{Tr } \mathbf{I} = 2m \text{Tr } \mathbf{S} = 2mR_g^2. \quad (4.8)$$

The principal radii of gyration  $r_{gi}$  therefore follow and obey:

$$r_{g1} = \sqrt{\frac{I_1}{m}} = \sqrt{S_1 + S_2}, \quad (4.9)$$

$$r_{g1}^2 + r_{g2}^2 + r_{g3}^2 = \frac{R_g^2}{2}. \quad (4.10)$$

In addition to the principal radii of gyration, eigenvalues of gyration tensor allow straightforward calculation of shape descriptors proposed by Theodorou and Suter<sup>235</sup>—asphericity  $b$ , acylindricity  $c$  and relative shape anisotropy  $\kappa^2$ :

$$b = S_1 - \frac{1}{2}(S_2 + S_3), \quad (4.11)$$

$$c = S_2 - S_3, \quad (4.12)$$

$$\kappa^2 = 1 - 3 \frac{S_1 S_2 + S_1 S_3 + S_2 S_3}{(S_1 + S_2 + S_3)^2}. \quad (4.13)$$

Asphericity and acylindricity express obviously the deviation from spherical and cylindrical symmetry. Interpretation of the last descriptor  $\kappa^2$  also respects symmetry. It reaches 0 for spherical objects and approaches 1 for linear shapes. Anisotropy of planar objects converges to  $\frac{1}{4}$ .

The physical dimension of squared length for  $S_i$ ,  $b$  and  $c$  follows directly from their definitions. This can be disadvantageous in a practical application such as CVs for metadynamics. Therefore we recommend to use adapted versions:

$$S'_i = \sqrt{S_i}, \quad (4.14)$$

$$b' = \sqrt{b}, \quad (4.15)$$

$$c' = \sqrt{c}. \quad (4.16)$$

### 4.3 Application

The gyration- and inertia-tensor-based collective coordinates ( $S'_{1-3}$ ,  $r_{g1-3}$ ,  $b'$ ,  $c'$ ,  $\kappa^2$ ) were implemented and contributed to PLUMED<sup>236</sup>—the popular open source plugin for metadynamics.

Firstly, the performance of newly introduced collective coordinates was tested on polyalanine peptides of four different lengths (3, 6, 9, and 12 residues). Because we did not expect that any of the collective coordinates in question could unambiguously describe all important conformation states the bias exchange metadynamics was used. Each simulation comprised of 8 replicas either biased in different CVs or unbiased (dubbed as neutral replicas). Neutral replica approximates the canonical distribution but profits from extended sampling due to the exchanges with biased replicas. We finally combined different CVs in four resulting protocols are summarized in Tab.4.1.

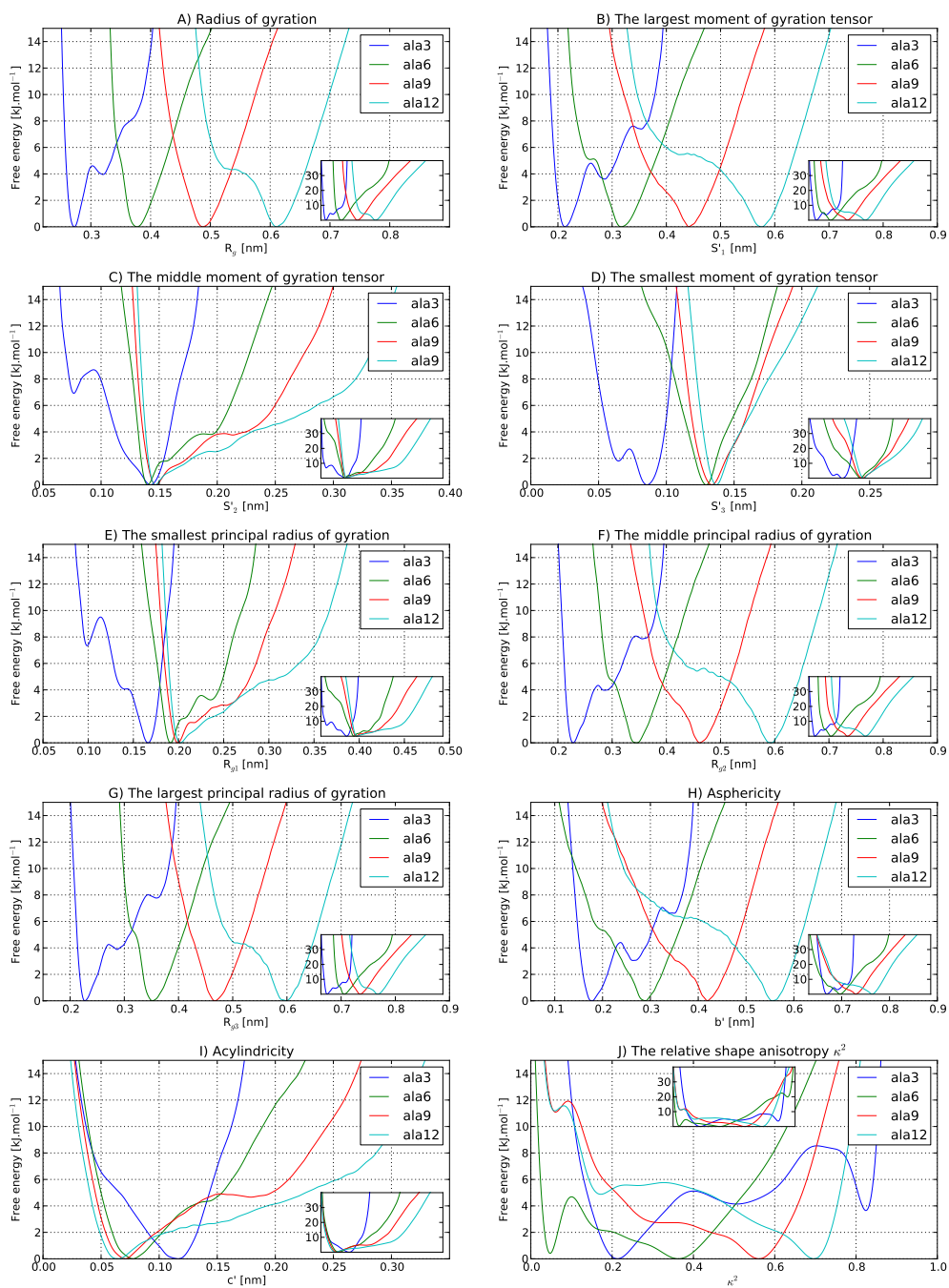
The one-dimensional free energy profiles as function of the collective coordinates in question are presented in Fig.4.1. The smooth and often featureless character of all profiles reflects clearly the flexible nature of alanine peptides. Only the shortest Ala<sub>3</sub> peptide provided profiles with 3 distinguishable minima that corresponds to different conformational families. The other peptides favored  $\alpha$ -helix as the most stable conformer that always correspond to the profound minimum on the free energy profiles. The increasing length of the helix due to elongation of sequence for Ala<sub>6-12</sub> is reflected by shift of the minima in  $R_g$ ,  $S'_1$ ,  $r_{g1}$ ,  $r_{g2}$ ,  $b'$  and  $\kappa^2$  in panels A,B,F,G, H and J in Fig.4.1. On the other hand, the cylindrical symmetry of the helix must result in constant (length-independent) values of  $S'_2$ ,  $S'_3$ ,  $r_{g3}$  and  $c'$  as it is correctly depicted in panels C, D, E and I. We did not observed any difference in performance of protocol P1 and P2. The rationale is that all employed CVs are internally linked and always represent a distinct transformation of 3 unique eigenvalues of gyration or inertia tensors. However, such transformation are still very useful for interpretation of molecular shapes.

The more stringent test was attempted to examine a folding of Trp-cage miniprotein using only gyration-tensor-based CVs. It succeeded in almost each simulation employing protocol P1–P4 for simulations 200 ns long. Once the native state appeared it had been captured by neutral replica and usually remained folded to the end of the simulation. The illustrative picture of the folded Trp-cage molecules is provided in Fig.4.2. Apart from this structures with low backbone RMSD ( $<2\text{\AA}$ ) from the experimental reference, simulations produced also relatively stable compact structures with non-native core packing and higher RMSD ( $2\text{--}4\text{\AA}$ ), see Fig. 4.3. It is questionable if they have any relevance to folding or are only artifacts of chosen CVs.

---

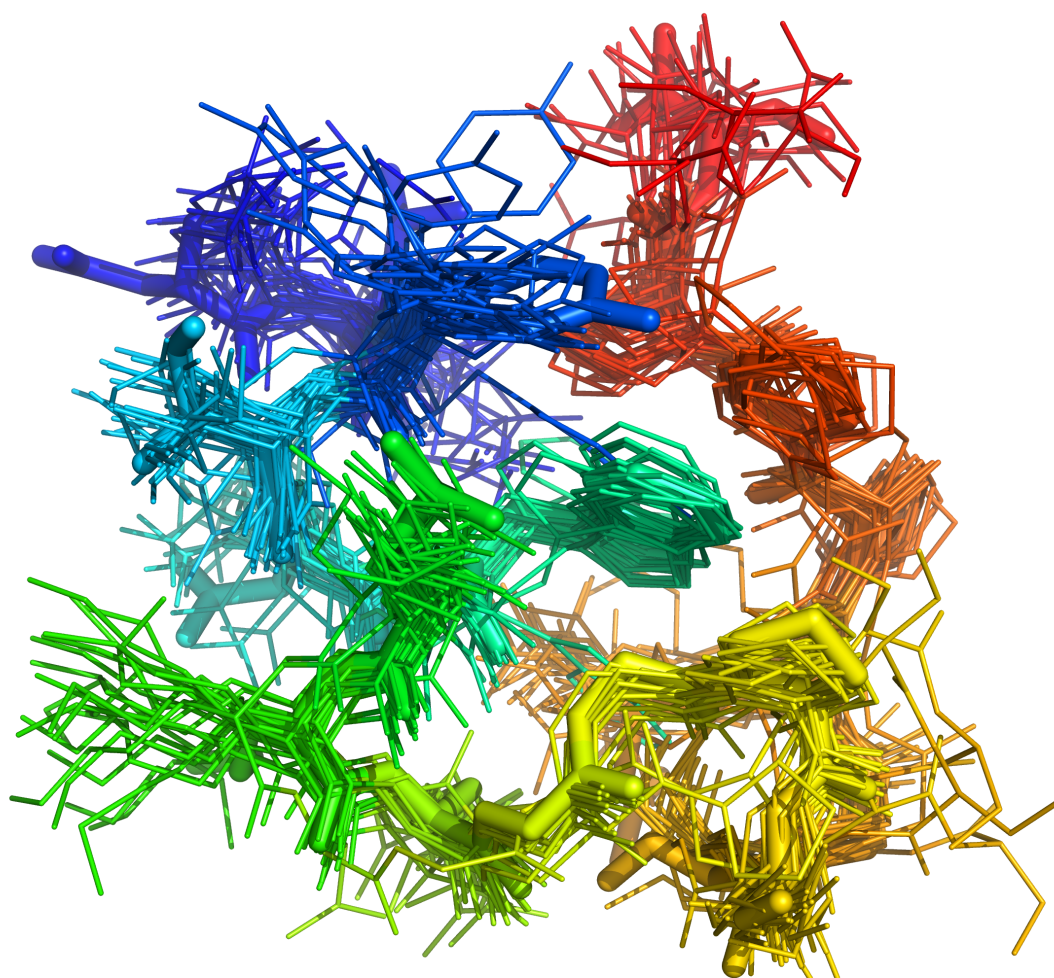
**Table 4.1:** Collective coordinates involved in bias exchange protocols.

Protocol	Collective coordinates
P1	$R_g, S'_1, S'_2, S'_3, b', c', \kappa^2$ , 1 neutral replica
P2	$R_g, S'_1, S'_2, S'_3, r_{g1}, r_{g2}, r_{g3}$ , 1 neutral replica
P3	$R_g, S'_1, S'_2, S'_3$ , 4 neutral replicas
P4	$R_g, r_{g1}, r_{g2}, r_{g3}$ , 4 neutral replicas



**Figure 4.1:** Free energy profiles of alanine peptides (Ala<sub>3-12</sub>).

The free energy profiles obtained for polyalanine peptides Ala<sub>3</sub>, Ala<sub>6</sub>, Ala<sub>9</sub> and Ala<sub>12</sub>. The individual plots show the free energy as a function of  $R_g$  (A),  $S'_1$  (B),  $S'_2$  (C),  $S'_3$  (D),  $r_{g3}$  (E),  $r_{g2}$  (F),  $r_{g1}$  (G),  $b'$  (H),  $c'$  (I) and  $\kappa^2$  (J).

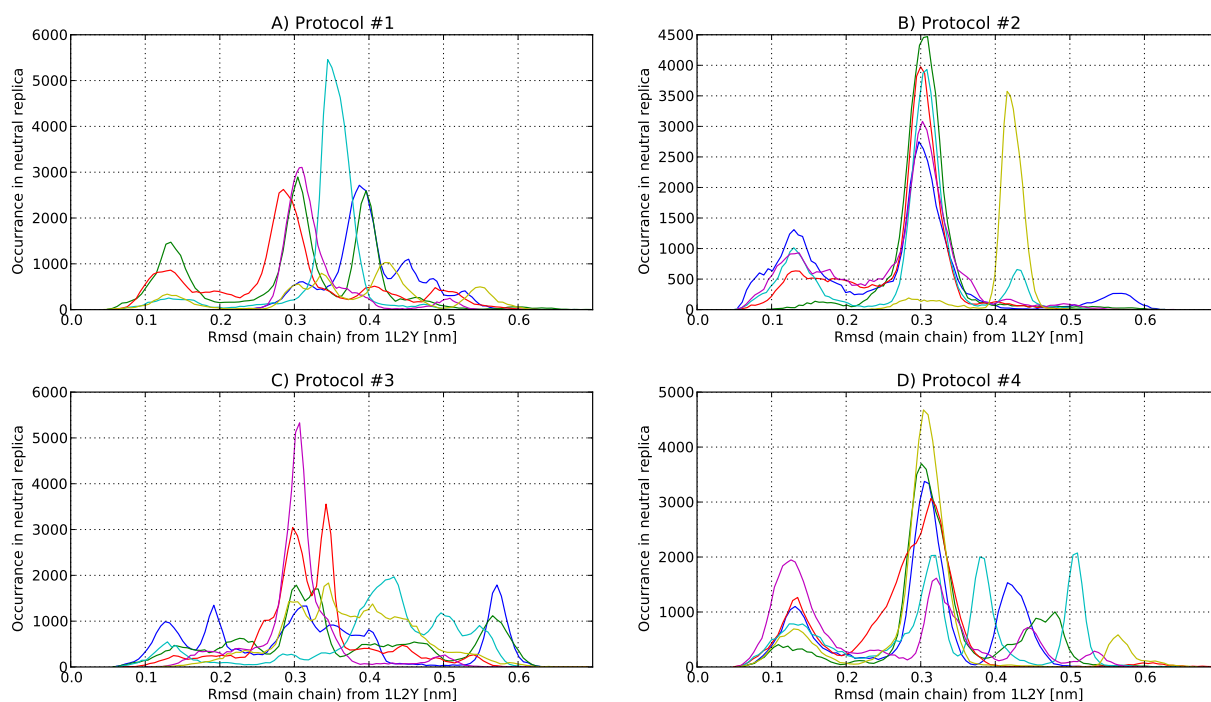


**Figure 4.2:** Superposition of Trp-cage structures.

The native-like structure obtained from simulations are compared with experimental structure of Trp-cage (thick).

---





**Figure 4.3:** Distribution of folded molecules in simulations.

RMSD of structures in neutral replicas and the native structure of Trp-cage (1L2Y) was binned and plotted for all simulation using protocol P1-P2. Native-like structures with low RMSD ( $<2\text{\AA}$ ) were present almost in all simulations.

The utilization of protocols P3 and P4 followed from testing whether more neutral replicas facilitate the correct packing of core residues. Since the gyration-tensor-based CVs take into account only the global shape of the molecule they cannot directly influence the respective orientation of side chains in the core of compact structure. However, additional neutral replicas did not increase fraction of native state molecules but increased chances of their finding.

Although the folding simulations were successful we expect even better performance if the protein-specific collective coordinates are added in the bias exchange metadynamics. The fact that the gyration-tensor-based CVs are able to facilitate folding of Trp-cage only on basis of molecular shape makes them valuable for bias exchange approach to protein folding problem.

## 5 Computational study of retro Trp-cage miniprotein

### 5.1 Motivation

The structure of folded proteins has unique features—network of intramolecular hydrogen bonds maintained by secondary structure elements and native contacts between side chains of amino acids. The planar peptide bonds can act both as hydrogen bond donors and acceptors but the order of donor/acceptor pattern is governed by direction of polypeptide chain. Interestingly, almost symmetrical image of the three-dimensional structure of a protein can be constructed from natural L-amino acids, but the following criteria must be taken into account:<sup>237</sup>

1. The direction of the backbone is reversed, i.e. the protein sequence is reversed from N-terminal to C-terminal.
2. The position of N-H and C=O groups is swapped. As a consequence the direction of hydrogen bonds is reversed.
3. Backbone torsion angles must be transformed by rules:  $\phi \rightarrow -\psi$  and  $\psi \rightarrow -\phi$ .

All these conditions can be easily fulfilled by  $\alpha$ -helical structures considering that the transformation of torsions projects them back to the same helical region on Ramachandran plot.

The effect of reverse protein sequence has been already tested experimentally. The “retro transformation” can lead to different outputs. The reverse protein sequences can either loss their ability to fold<sup>238,239</sup> or the same (or topologically similar) fold can be preserved.<sup>240–242</sup>

Trp-cage is a miniprotein derived from exendin-4 saliva protein from Gila monster lizard.<sup>243</sup> Trp-cage has become rapidly a model protein due to its tiny size (20 residues), cooperative and ultra-fast folding kinetics.<sup>244</sup> The fold of Trp-cage consists of one  $\alpha$ -helix, short loop and polyproline stretch which interacts with a side chain of tryptophan in the helix. This particular structural motif possesses surprisingly large energetic stabilization by non-covalent interactions.<sup>245</sup> Nevertheless, the stability of the Trp-cage is considerably linked also with stability of the N-terminal helix.<sup>246</sup>

The existence of small stabilizing core rises a question if it can be restored after retro transformation of the original sequence. Although the Trp-cage miniprotein ranks among the most studied proteins both experimentally and theoretically, no efforts to characterize its retro-variant have been yet reported. This challenge was addressed by our collaborative study (*submitted*, Appendix E).

## 5.2 Experimental characterization of retro Trp-cage

The amino acid sequence of retro Trp-cage was synthesized and studied by NMR and CD spectroscopy. Both methods confirmed the unstructured character of the molecule in aqueous buffer. However, changes were observed upon addition of 2,2,2-trifluoroethanol (TFE). CD spectroscopy revealed an increase of helical content and provided typical reversible heat denaturation curve with melting temperature about 32°C. The structure promoting effect of TFE allowed to conduct all NMR experiments necessary for determination of 3D structure of the miniprotein. The resulting structure was deposited in the Protein Databank under the accession code 2LUF.

The elucidated structure of retro Trp-cage resembles the structure of original Trpcage ( $C_{\alpha}$ -RMSD 3.3Å) mostly in the corresponding helical region. N-terminal helix of Trp-cage aligns with the helix in C-terminal part of its retro-variant. However, both structures differ significantly in the other parts. The typical proline–tryptophan structural motif known from Trp-cage was replaced by novel arrangement of core in retro Trp-cage. It involves the dominant tryptophan–arginine side chain stacking familiar from other protein structures. The both structures are compared in Fig. 5.1.

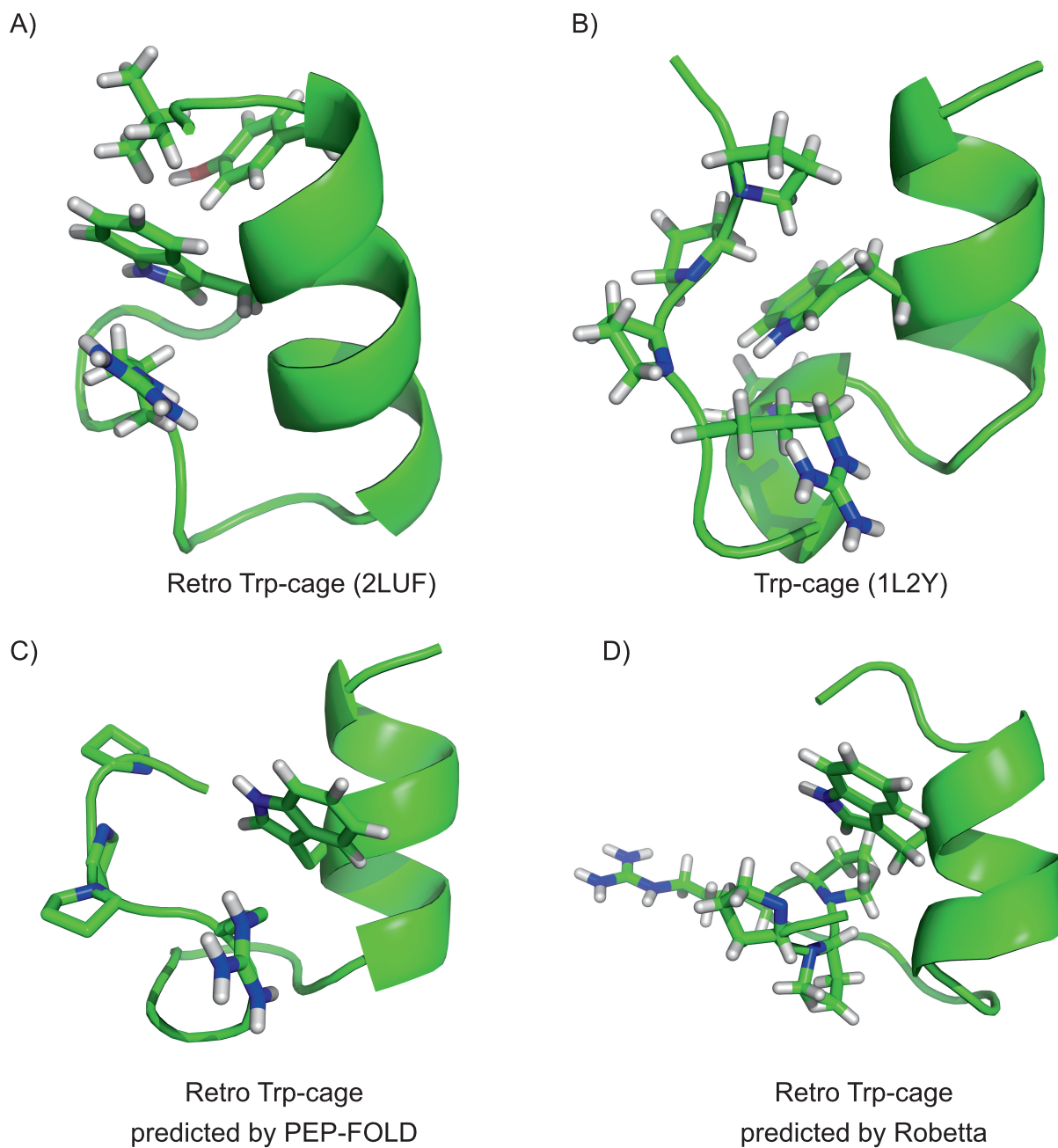
## 5.3 Modeling of retro Trp-cage

Prediction of three-dimensional protein structure from its sequence still belongs to the most challenging tasks in protein modeling.<sup>247</sup> Because the reverse sequence of the Trp-cage did not share any similarity to known proteins only *ab initio* template-free structure prediction methods could be utilized for *de novo* modeling of retro Trp-cage. Such methods as Robetta<sup>248</sup> or PEP-FOLD<sup>249</sup> employ a fragment based approach. The local structure elements are first determined using fragment libraries and then assembled into final model by Monte Carlo method or alternative techniques.

We used the online version of both structure prediction methods—Robetta and PEP-FOLD to obtain three-dimensional model before the experimental structure was solved. Both services provided sets of models accompanied by their ranking. Robetta and PEP-FOLD succeeded in prediction of C-terminal helix. The overall fold of all models resembled the structure of original Trp-cage (1L2Y) as well as its retro variant (2LUF) which is apparent from Fig. 5.1. However, no model correctly assembled the core of the miniprotein. The distorted tryptophan–arginine motif was predicted only once by PEP-FOLD. Similarly, the canonical tryptophan–proline motif known from Trp-cage molecule was not found by any of the predicting methods. In all cases, non-native arrangements of tryptophan and proline side chains were proposed.

In order to shed light on stability and character of disordered state of retro Trp-cage in water environment, we conducted a series of equilibrium MD simulations. All simulations started from the experimental structure (2LUF) but used different force fields in order to assess reproducibility of results.

All simulations rapidly deviated from the starting NMR structure.  $C_{\alpha}$ -RMSD around 2Å could be still considered as a fluctuation around the experimental structure. The higher values of  $C_{\alpha}$ -RMSD were connected to the refolding of N-terminal tail or repacking



**Figure 5.1:** The experimental structure of retro Trp-cage (A), Trp cage (B) and retro Trp-cage models (C, D).

---

of the core. However, the  $C_{\alpha}$ -RMSD values exceeding  $5\text{\AA}$  indicated the very loosened structures or unfolding of C-terminal helix.

All four force fields involved in our study (ff03<sup>105</sup>, ff99SB-ILDN<sup>112</sup>, charmm22/CMAP<sup>124</sup> and OPLS-AA/L<sup>133</sup>) provided different results in the simulations. Unfortunately, it was not possible to clearly distinguish the effects of the force fields from insufficient sampling as can be inferred from Tab. 5.1. Four independent runs for each force field manifested high diversity in maintaining structural features of retro Trp-cage and related properties. Nevertheless, all force field were able to keep, at least transiently, the fold of the miniprotein in water as proved by  $C_{\alpha}$ -RMSD. This fact indicates that force fields are able to capture the stabilizing interactions found in experimental NMR structure.

The helix favoring ff03 and charmm22/CMAP force fields kept higher helical content than ff99SB-ILDN and OPLS-AA/L (see Fig. 5.2), but all force fields provided similarly compact structures as can be deduced from radii of gyration in Tab. 5.1. Interestingly, the presence of the tryptophan–arginine structural motif did not correlate strongly with RMSD from the experimental structure (2LUF). This founding suggests independence of binding motif on presence of helical framework (see Fig.5.3).

Since retro Trp-cage forms a stable structure only upon addition of TFE in buffer we tested the effect of this cosolvent on stability of miniprotein in simulations. The standardly parametrized GAFF model was used in simulations with ff03 and ff99SB-ILDN force fields. Considering the high variability of individual runs, no significant changes were observed for ff03. However, the structures simulated by ff99SB-ILDN were dramatically affected. The corresponding  $C_{\alpha}$ -RMSD from experimental structure increased as well as radii of gyration (see Tab. 5.1). This changes indicated the rather denaturing effect of TFE on protein structure in clear disagreement with experiment. It is not obvious what caused the incorrect behavior in the solvent mixture. The probable explanations could be incompatibility of force field parameters for protein, water model and TFE cosolvent, which we decided to attempt as our next goal.


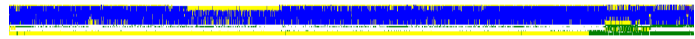
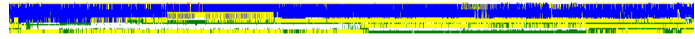

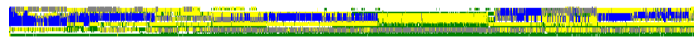
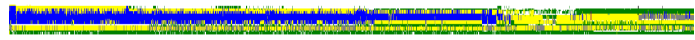
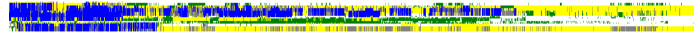
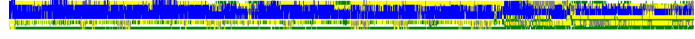




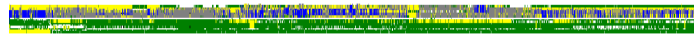
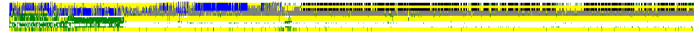



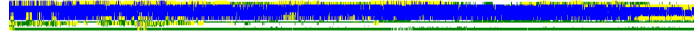

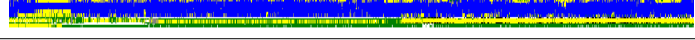
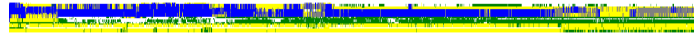
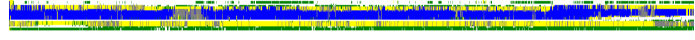


## 5.4 Conclusion

This study of retro Trp-cage confirmed the importance of fine structural details on stability of proteins. The same composition and the respective order of amino acids do not guarantee the same ability to fold if the direction of protein backbone is reversed. On the other hand, a latent structure can be still preserved in the sequence and manifested under structure promoting conditions such as in TFE mixture solvents.

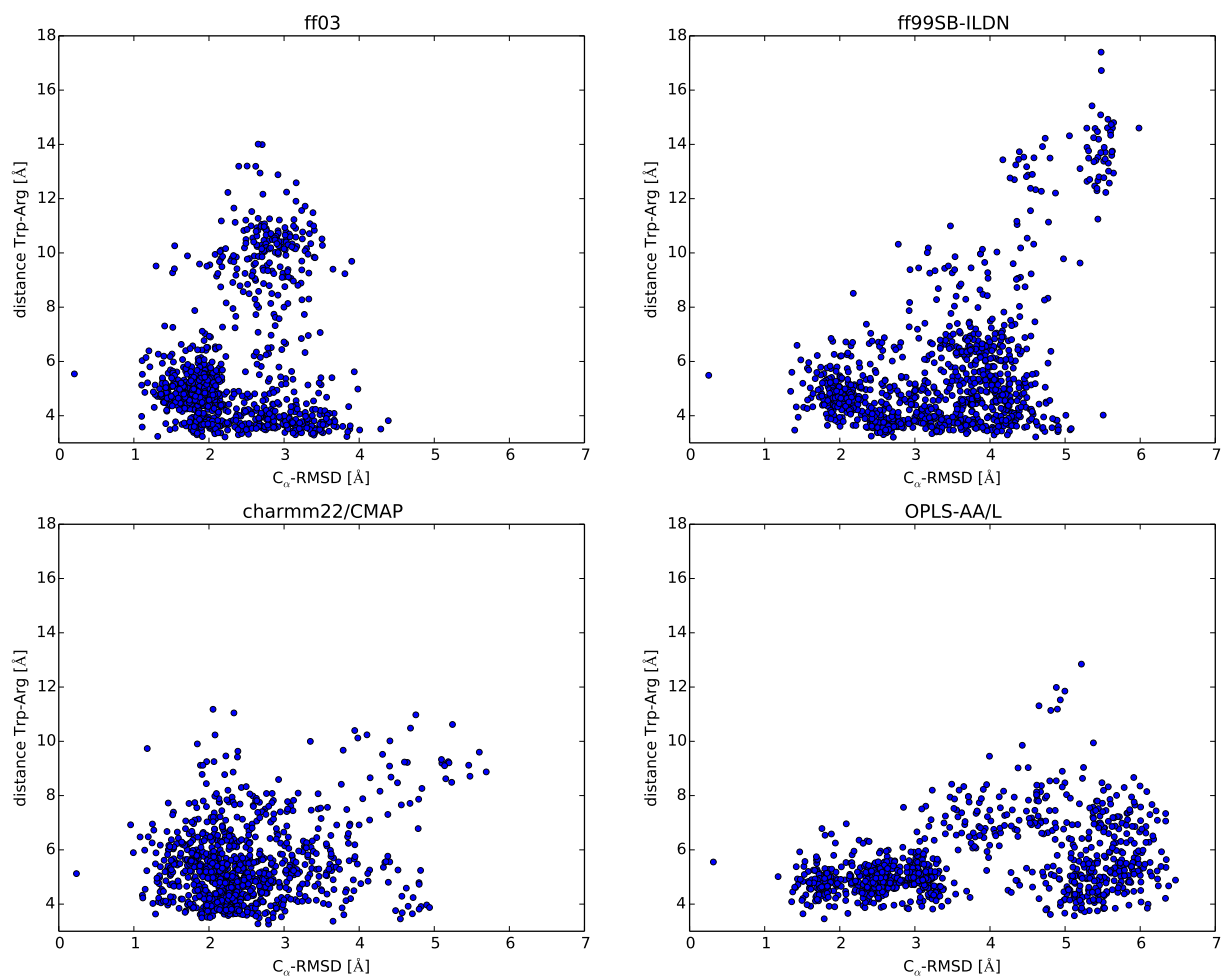
The delicate structure/disorder transitions are notoriously difficult task for force fields. We believe that constructs like retro Trp-cage can help in the force field development as a suitable targets for calibrating and balancing protein–water–cosolvent interactions.

**Table 5.1:** Result of retro Trp-cage simulations.  $C_{\alpha}$ -RMSD from reference structure 2LUF, the radius of gyration, helical content in the helical part of retro Trp-cage and occurrence of the Trp-Arg contact during MD simulations in water and 30% TFE. Each 200-ns MD simulation is divided into four 50-ns frames.

Solvent	Force field	Simulation	RMSD [Å]		Radius of gyration [Å]				Helical content [res]				Trp-Arg contact [%]					
Water	ff03	1	2.2	1.9	1.9	2.9	7.4	7.2	7.3	7.5	9.3	10.1	9.9	8.9	84	81	90	99
		2	1.8	1.8	1.8	1.6	7.0	7.0	7.0	7.0	9.4	8.0	9.5	9.2	99	99	99	95
		3	2.1	2.5	3.1	2.9	7.2	7.3	7.3	7.4	8.1	6.1	6.9	7.2	37	11	0	7
		4	2.2	2.7	2.8	2.8	7.6	8.0	7.7	8.0	9.4	9.3	7.9	9.6	100	99	96	95
	ff99SB-ILDN	1	3.1	4.0	3.8	3.4	7.8	7.6	7.9	7.8	6.2	3.4	1.6	4.3	97	72	30	36
		2	2.8	2.6	3.5	4.2	7.4	7.3	7.5	8.1	6.9	6.0	4.7	1.1	100	94	96	71
		3	3.7	4.2	4.4	5.2	7.5	7.7	8.1	7.4	6.7	4.7	3.3	0.2	63	71	68	0
		4	2.0	2.0	1.9	3.2	7.4	7.5	7.3	8.0	8.5	7.6	7.1	5.7	94	96	85	82
	charmm22/CMAP	1	3.8	2.8	2.3	2.2	8.5	7.9	7.6	7.6	7.5	6.1	6.3	6.3	27	94	100	100
		2	2.2	2.3	2.1	2.3	7.5	7.5	7.6	7.6	8.4	9.2	8.6	9.4	81	97	91	92
		3	3.1	2.9	2.9	2.7	7.8	7.6	7.6	7.5	9.6	9.5	9.4	9.1	63	64	53	50
		4	1.8	1.9	2.2	2.3	7.4	7.4	7.6	7.8	8.8	8.3	8.1	8.5	61	92	51	74
OPLS-AA/L	1	4.2	5.3	5.2	5.8	7.6	8.0	8.1	7.5	4.1	4.7	5.1	3.4	96	81	71	82	
	2	1.8	2.3	3.1	3.1	7.2	7.2	7.2	7.2	4.2	4.7	2.8	2.8	99	96	97	95	
	3	2.3	2.5	2.5	2.5	7.4	7.5	7.4	7.3	3.5	2.9	2.6	3.0	92	99	97	99	
	4	3.4	5.9	4.7	4.1	7.5	7.5	7.5	7.8	4.0	4.3	3.7	3.7	62	40	1	2	
30%TFE	ff03	1	3.5	3.1	3.0	2.2	7.9	7.6	7.8	7.7	8.3	8.5	7.9	9.5	94	99	99	58
		2	2.4	2.6	2.7	3.0	7.5	7.5	7.6	7.5	8.6	8.3	8.7	7.3	91	98	95	98
		3	2.5	3.1	2.5	2.9	7.8	8.2	7.3	8.0	6.5	7.3	7.6	8.9	5	0	22	77
		4	2.2	2.6	2.9	3.3	7.7	7.9	8.2	8.3	8.2	9.6	8.9	9.0	50	5	86	100
ff99SB-ILDN	1	4.8	5.3	5.9	5.4	8.9	9.2	9.2	8.3	6.1	5.5	4.0	3.6	47	2	1	30	
	2	3.8	4.3	4.6	5.3	8.6	9.3	9.4	9.1	6.1	5.8	5.4	5.4	75	97	90	40	
	3	3.0	5.9	3.9	5.9	8.2	10.0	8.4	9.8	5.9	6.5	3.7	4.7	51	16	2	0	
	4	3.5	3.5	5.9	6.8	7.9	8.0	9.3	10.8	3.5	3.0	4.6	3.2	97	86	0	0	

Solvent	Force Field	Simulation	Secondary structure
Water	ff03	1	
		2	
		3	
		4	
	ff99SB	1	
		2	
		3	
		4	
	charmm22/CMAP	1	
		2	
		3	
		4	
	OPLS-AA/L	1	
		2	
		3	
		4	
30%TFE	ff03	1	
		2	
		3	
		4	
	ff99SB	1	
		2	
		3	
		4	

**Figure 5.2:** Secondary structure in the course of the simulations. The secondary structure were assigned by DSSP algorithm.<sup>250</sup> The secondary structure elements are coded by colors:  $\alpha$ -helix – blue,  $3_{10}$ -helix – gray, turn – yellow,  $\beta$ -bridge – black, bend – red.



**Figure 5.3:** Correlation between C $\alpha$ -RMSD and contacts between tryptophane and arginine. Presence of Trp-Arg contact does not guarantee low C $\alpha$ -RMSD value and *vice versa*.



## 6 Optimization of force field parameters for 2,2,2-trifluoroethanol

### 6.1 Motivation

2,2,2-trifluoroethanol (TFE) is often used as a cosolvent in experimental studies of peptides and proteins. Buffers containing TFE (up to 50% v/v) usually stabilize the  $\alpha$ -helices or induce their formation. Additionally, TFE improves solubility of peptides and proteins and thus facilitates experiments in solutions.<sup>71</sup> A mechanism how TFE stabilizes peptides and proteins has not been completely elucidated. The proposed hypothesis expect either direct or indirect effects on polypeptide in TFE/water mixture. The direct effects involve preferential binding on surface of the molecule, local increase of TFE concentration in vicinity of the protein and stabilizing hydrogen bonding of backbone due to the lower effective dielectric constant of the mixed solvent.<sup>251,252</sup> The indirect mechanism influences the thermodynamics of folding by chaotropic effect on solvation layers of folded and unfolded states.<sup>253</sup> Furthermore, the activity of TFE can be pronounced in both cases by transient self-aggregation of TFE molecules.<sup>254</sup>

We observed an inability of the standard GAFF model of TFE to stabilize structure of retro Trp-cage construct in simulations as reported in the previous chapter. The almost identical GAFF model was later introduced and tested with positive effects on stability of helical peptides.<sup>255,256</sup> However, no properties of TFE/water mixtures were reported by the authors although the good calibration of TFE–water interactions is the necessary prerequisite for reliable computational model. The conclusions of the earlier studies by Chitra and Smith suggest that the common TFE models often fail to describe the pure liquid TFE phase as well as their mixtures with water.<sup>257,258</sup> Qualitatively different behavior of TFE/water mixtures was recently observed also for different combination of water models.<sup>259</sup>

The microscopic structure of liquid mixtures can be analyzed in framework of Kirkwood–Buff theory.<sup>260</sup> The unique feature of this approach is the fact, that the corresponding values of Kirkwood–Buff integrals (KBI) can be evaluated from MD simulations as well as calculated from experimental thermodynamic quantities. Kirkwood–Buff theory therefore represents a valuable link between theory and experiment.<sup>261</sup>

The Kirkwood–Buff integrals can be obtained from simulations by two methods. The first—classical approach employs the definition of KBI for components of binary mixture via radial distribution functions:

$$G_{ij} = \int_0^\infty 4\pi r^2 (g_{ij}(r) - 1) dr, \quad (6.1)$$

where  $g_{ij}$  is the radial distribution function between particles  $i$  and  $j$ . The alternative way how to calculate KBI uses the definition through particle fluctuations in an embedded region<sup>262</sup> characterized by volume  $V$  and linear size  $L_s$ :

$$G_{ij}(L_s) = V_s \left( \frac{\langle N_i N_j \rangle - \langle N_i \rangle \langle N_j \rangle}{\langle N_i \rangle \langle N_j \rangle} - \frac{\delta_{ij}}{\langle N_i \rangle} \right), \quad (6.2)$$

The final KBI can be extrapolated for infinite  $L_s$  using the relation:

$$G_{ij}(L_s) = G_{ij} + \frac{const}{L_s}. \quad (6.3)$$

We decided to optimize TFE model based on GAFF in order to better reproduce liquid state properties and more importantly—the properties of TFE/water mixtures represented by Kirkwood–Buff integrals (*submitted*, Appendix F). We expect that such carefully optimized force field parameters will be able to provide also more realistic picture of ternary TFE/water/protein complexes and could shed a light on the atomistic mechanism of stabilizing effects of TFE.

Because the precise calibration depends also critically on the water model we have chosen TIP4P/Ew<sup>115</sup> and TIP4P/2005<sup>111</sup> as the references. Both models provide superior description of water properties in comparison to the common 3-site models. Although TIP4P/Ew and TIP4P/2005 are not accommodated by the amber force fields and their usage is spurious they did not show overall performance worse than TIP3P.<sup>151,263</sup> Moreover, the TIP4P/Ew is intended to replace obsolete TIP3P model in the newly developed versions of amber force field.<sup>116,157</sup>

## 6.2 Parametrization

For the sake of compatibility with the amber family of force field the initial model was adopted from GAFF and the partial charges from R.E.D. database.<sup>264</sup> These charges were calculated as highly reproducible multi-conformational RESP fit.<sup>265</sup>

The parametrization were scheduled to 3 stages.

Firstly, the parameters suitable for optimization were selected. We introduced perturbations to different partial charges and Lennard-Jones parameters of atoms and tested their impact on properties of liquid TFE in short simulations, namely—density, heat of vaporization, self-diffusion coefficient and population of *trans* conformer. Finally, we identified 5 parameters for further optimization. The final set included Lennard-Jones parameters ( $\sigma$  and  $\epsilon$ ) of fluorine, partial charges on fluorine and hydrogen in hydroxyl group and a parameter of C–C–O–H torsion.

Secondly, the different combinations of parameters were tested exhaustively in simulations of pure TFE liquid. We developed a protocol for simultaneous optimization of all 5 parameters. The dependence of 4 optimized properties was iteratively fitted in analytical forms and evaluated by an objective function. The analytic fit allowed us to predict the values of parameters in the best agreement with experiment and verify them subsequently in

simulations. Moreover, the exhaustive protocol guaranteed that the optimal combination of parameters could be found and provided a valuable view of their ambiguity.

In addition to the classical amber force field parametrization strategy we employed more precise approach to calculate vaporization enthalpy. This property is usually calculated as a difference between average energy of the molecule in gas phase  $\langle U \rangle_{gas}$  and in the liquid  $\langle U \rangle_{liq}/N$ :

$$\Delta H_{vap} = \langle U \rangle_{gas} - \langle U \rangle_{liq}/N + RT. \quad (6.4)$$

Since the partial atomic charges of amber force fields mimics condense phase charge distribution, they are not transferable between phases. The energy costs for this pre-polarization can be estimated and the affected properties (enthalpy of vaporization, free energy of hydration) can be corrected by an additional term.<sup>266</sup> Equivalent corrections were used in parametrization of both TIP4P/Ew and TIP4P/2005 models.

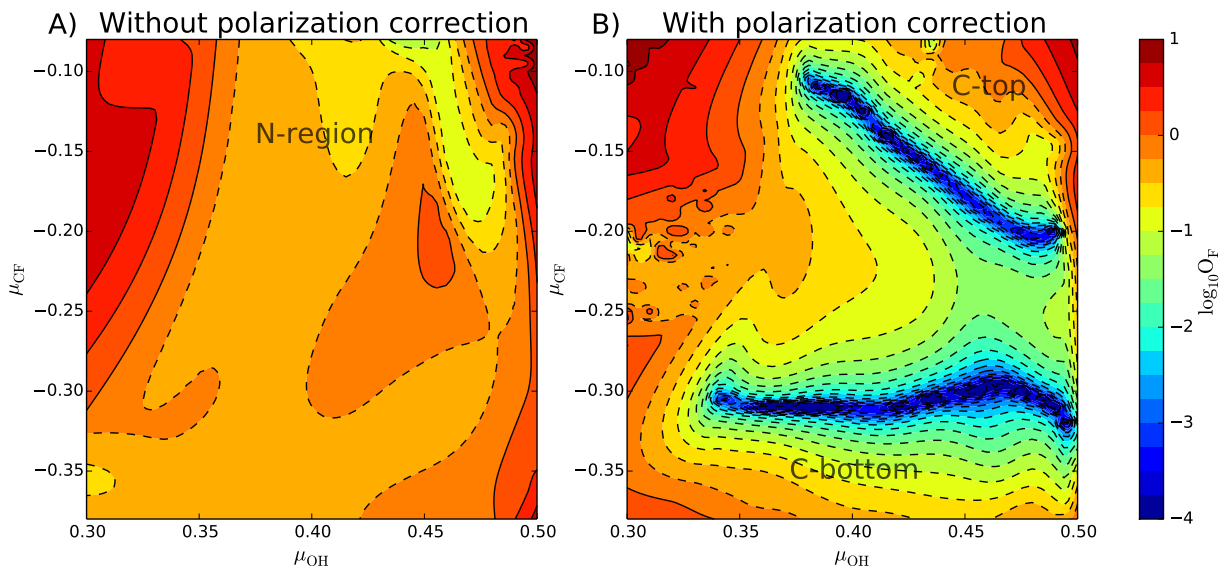
Lastly, the parameters describing sufficiently pure liquid phase of TFE were tested in simulation of mixtures with TIP4P/Ew and TIP4P/2005 water models. The Kirkwood–Buff integrals were used for comparison with experiment because of their sensitivity on the respective microscopic distribution of components in the mixture and their self-aggregation tendencies. However, the correct evaluation of the KBIs requires long simulations of large boxes. Therefore it could not be used routinely for development of parameters but only to discriminate parameters incompatible with water models.

### 6.3 Results

The systematic search in 5-dimensional space of TFE parameters for optimization required approximately 2500 simulations of liquid TFE and allowed us to map values of the selected force field parameters to the liquid state properties. Both ways to calculate vaporization enthalpy led to parameters that were able to describe simultaneously all optimized properties. However, if the polarization costs were considered, ambiguous and more precise solutions were found as illustrated in Fig 6.1. This conclusions could not be clearly drawn without the extensive protocol we employed. Obviously, there were no unique parameters which could be chosen preferentially. Therefore 28 candidate parameters from all promising regions of parameter space (named as N-region, C-top and C-bottom, see Fig. 6.1) were selected and tested in simulation of mixtures with water.

TFE/water mixtures modeled by candidate TFE parameters differed significantly in their properties. Although TFE is fully miscible with water in all concentrations in experiments, several parameters from N-, C-top and C-bottom region showed the opposite behavior. This TFE models with low partial charge on hydrogen atom manifested separation of both liquids and no meaningful KBIs could be calculated from such simulations. The clear trends were observed—the higher polarity of O–H bonds, the lower values of  $G_{TT}$  Kirkwood–Buff integrals which indicated lower self-aggregation of TFE molecules. The visual demonstration of self-aggregation is provided in Fig. 6.2.

Nevertheless, the values of calculated  $G_{TT}$  comparable to the experimental counterparts were found only for parameters from C-bottom region with the most polar partial charges.



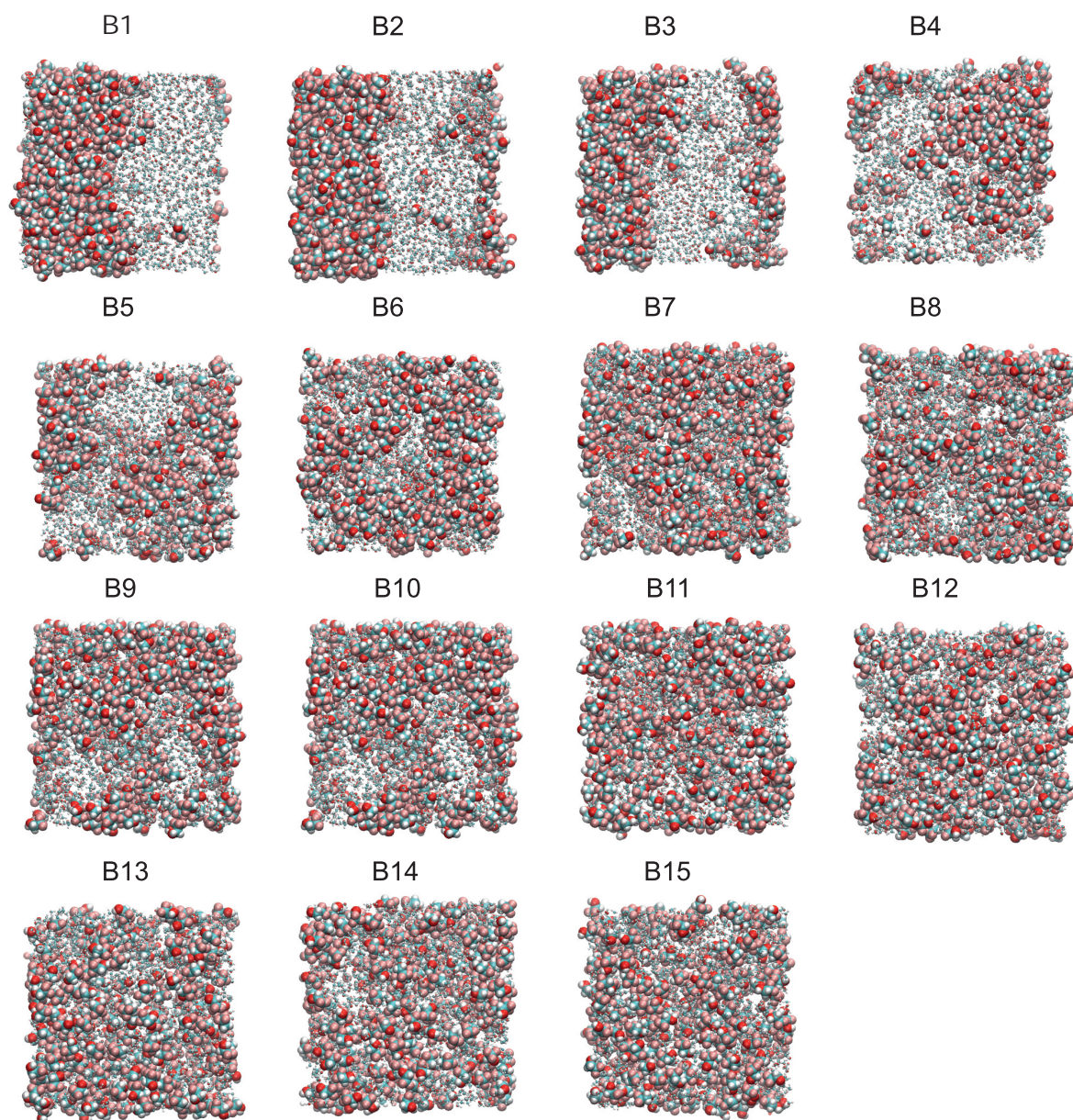
**Figure 6.1:** Exploration of TFE parameter space.

For sake of clarity the 5-dimensional parameter space were reduced to 2 dimensions. The panel A and B show values of objective function using partial charge on hydrogen ( $\mu_{OH}$ ) and fluorine ( $\mu_{CF}$ ) as parameters. The values of other three parameters were optimized for the best agreement with experimental data, i.e. the lowest value of objective function. Alternative calculation of vaporization enthalpy produced completely different results (compare panel A and B).

These 5 candidates were further investigated in details. The additional simulations of TFE/water mixtures in different ratios of both components were conducted in order to check the experimental trends. Almost all parameters exaggerated the  $G_{TT}$  for lower concentration of TFE but the overall trend were captured correctly. This test selected the final set of parameters (dubbed as B15) that best reproduced the experimental KBIs of mixtures with both water models (see Fig. 6.3).

The properties of final model B15 were thoroughly characterized. The pure liquid TFE properties were reproduced excellently. Not only these involved in parametrization but also shear viscosity, coefficient of thermal expansion, isothermal compressibility, static dielectric constant and free energy of hydration were predicted by TFE model in good agreement with experimental data (see Tab. 6.1).

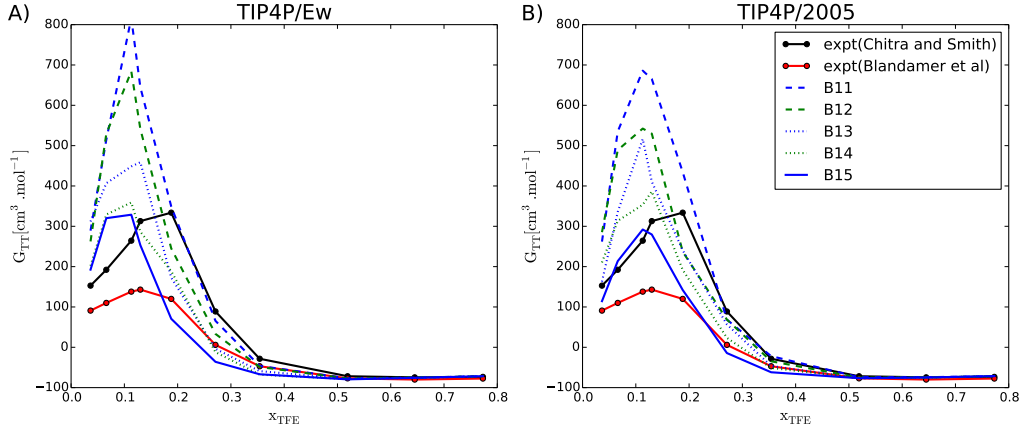
Calibration of TFE model to realistic mixture properties with 2 advanced water models (TIP4P/Ew and TIP4P/2005) showed clearly the importance of polarization correction in course of parametrization process. The implicit correction to polarization cost in gas phase significantly improved reproduction of liquid properties of TFE. More importantly, it allowed to balance TFE–TFE, TFE–water and water–water interactions and thus reduce the self-aggregation of TFE molecules at level supported by experimental data.



**Figure 6.2:** Self-aggregation of TFE molecules.

The different TFE force field parameters manifested distinct self-aggregation tendencies. The complete separation of both liquid phases was observed for models B1-B3. Models B4-B15 produced TFE clusters of decreasing size.

---



**Figure 6.3:** Kirkwood–Boff integral  $G_{TT}$  as function of TFE molar ratio. Values of  $G_{TT}$  were calculated for 5 candidate parameters (B11–B15) and both TIP4P/Ew (A) and TIP4P/2005 (B) water models. The clear trend can be observed for the polarity of the TFE model (increasing continuously from B11 to B15) and KBI  $G_{TT}$ . The experimental values were adopted from Ref. 261 and 267.

**Table 6.1:** The properties of pure liquid TFE at 298.15K and 100 kPa. Comparison of the predicted and experimental values.

The density  $\rho$ , self-diffusion coefficient  $D$ , enthalpy of vaporization  $\Delta H_{vap}$ , ratio of *trans* conformer in liquid, coefficient of thermal expansion  $\alpha$ , isothermal compressibility  $\kappa_T$ , shear viscosity  $\eta$ , static dielectric constant  $\epsilon$ , average dipole moment of the molecule  $\langle \mu \rangle$  and free energy of hydration  $\Delta G_{solv}$  are compared for predicted and experimental data.

property	simulation	expt <sup>a</sup>
$\rho$ [kg.m <sup>-3</sup> ]	1382.5	1382.4
$D$ [10 <sup>9</sup> m <sup>2</sup> .s <sup>-1</sup> ]	0.70	0.68
$\Delta H_{vap}$ [kJ.mol <sup>-1</sup> ]	42.9(47.5) <sup>b</sup>	42.9
% <i>trans</i>	39.9	40
$\alpha$ [10 <sup>3</sup> K <sup>-1</sup> ]	1.27	1.28
$\kappa_T$ [GPa <sup>-1</sup> ]	1.32	1.23
$\eta$ [10 <sup>3</sup> kg.m <sup>-1</sup> .s <sup>-1</sup> ]	1.52	1.72
$\epsilon_r$	34	27
$\langle \mu \rangle$ [Debye]	3.69	2.46
$\Delta G_{solv}$ [kJ.mol <sup>-1</sup> ]	-16.8 (-21.4) <sup>b,c</sup>	-18.02
$\Delta G_{solv}$ [kJ.mol <sup>-1</sup> ]	-16.6 (-21.2) <sup>b,d</sup>	-18.02

<sup>a</sup>References to the original experimental studies are provided in the enclosed manuscript.

<sup>b</sup>The value in parenthesis represent a version without correction on polarization costs in gas phase.

<sup>c</sup>Obtained with TIP4P/Ew water model.

<sup>d</sup>Obtained with TIP4P/2005 water model.

## 7 Concluding remarks

This thesis deals with several topics highly relevant for simulations of peptides and miniproteins in water and mixtures with cosolvents. Here I would like to conclude the most important results.

- The force fields involved in this study are not able to reproduce fine intrinsic properties of amino acids. The propensities obtained from simulations reflected rather artifacts of non-physical 1–4 electrostatic interactions and different charge distributions used in force fields.
- Conditions of thermal and chemical denaturation generate different conformational ensembles of model peptides which are neither random nor uniform and shifted from the reference state under physiological condition. The chemical shifts were shown to be less sensitive to the local backbone conformations and therefore they might provide incorrect picture of denatured or unfolded states.
- The gyration- and inertia-tensor based collective coordinates for metadynamics can be efficiently used for extensive sampling in bias exchange metadynamics. This collective coordinates were capable of exhaustive sampling of alanine polypeptides and succeeded in reproducible folding of Trp-cage miniprotein.
- Reverse sequence of Trp-cage does not fold in water but acquire a structure in 30% solution of 2,2,2-trifluoroethanol (TFE). The resulting structure resembles the structure of Trp-cage but possesses different packing amino acids in the core. Robetta and PEP-FOLD tools succeeded in correct prediction of helix in the structure but failed to model native interactions within the core. The stability of miniprotein was interpreted differently upon a choice of a force field. However, the attempt to mimic closely the experimental conditions by standard GAFF model of TFE led to behavior incompatible with experimental observations.
- We exhaustively optimized force field parameters of 2,2,2-trifluoroethanol for better description of pure liquid properties and realistic behavior in TFE/water mixtures. We demonstrated that the improvement over the standard GAFF parameters were achieved by implicit polarization correction treating the polarization costs of molecule in gas phase.

## List of Abbreviations

B3LYP	Becke–three-parameter–Lee–Yang–Parr Density Functional
CV	Collective Variable (Coordinate)
GAFF	Generalized Amber Force Field
HF	Hartree–Fock Method
KBI	Kirkwood–Buff Integral
LMP2	Local MP2 Method
MD	Molecular dynamics
MP2	Second-order Møller-Plesset Method
NMA	N-methylacetamide
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
RESP	Restrained Electrostatic Potential Method
RIMP2	Resolution of the Identity approximated MP2 Method
RMSD	Root Mean Square Deviation
TFE	2,2,2-trifluoroethanol



## List of Figures

1.1	Current view on intrinsically disordered proteins. . . . .	11
1.2	Force field genealogy. . . . .	16
1.3	Comparison of force fields against NMR experimental data. . . . .	24
1.4	Force fields getting better. . . . .	25
3.1	Definition of backbone conformers. . . . .	35
3.2	Relative population of backbone conformers. . . . .	36
3.3	Propensities for $\chi_1$ torsion. . . . .	37
3.4	Trends in the hybrid force field (HYB), FF99SB and FF03. . . . .	38
3.5	The correlation between experimental ppII content and predictions of individual force fields. . . . .	39
3.6	Conformational preferences of amino acids in AAXAA host peptide. . . . .	41
3.7	$^1\text{H}_\text{N}$ and $^{15}\text{N}_\text{H}$ chemical shifts for the guest X residues. . . . .	42
4.1	Free energy profiles of alanine peptides (Ala <sub>3-12</sub> ). . . . .	47
4.2	Superposition of Trp-cage structures. . . . .	48
4.3	Distribution of folded molecules in simulations. . . . .	49
5.1	The experimental structure of retro Trp-cage (A), Trp cage (B) and retro Trp-cage models (C, D). . . . .	52
5.2	Secondary structure in the course of the simulations. . . . .	55
5.3	Correlation between $C_\alpha$ -RMSD and contacts between tryptophane and arginine. . . . .	56
6.1	Exploration of TFE parameter space. . . . .	60
6.2	Self-aggregation of TFE molecules. . . . .	61
6.3	Kirkwood–Buff integral $G_{TT}$ as function of TFE molar ratio. . . . .	62

## List of Tables

4.1	Collective coordinates involved in bias exchange protocols. . . . .	46
5.1	Result of retro Trp-cage simulations. . . . .	54
6.1	The properties of pure liquid TFE at 298.15K and 100 kPa. Comparison of the predicted and experimental values. . . . .	62

## References

- [1] Wright, P. E.; and Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology* **1999**, *293*, 321–31.
- [2] Dill, K. a.; and MacCallum, J. L. The protein-folding problem, 50 years on. *Science (New York, N.Y.)* **2012**, *338*, 1042–6.
- [3] Sosnick, T. R.; and Barrick, D. The folding of single domain proteins—have we reached a consensus? *Current opinion in structural biology* **2011**, *21*, 12–24.
- [4] Daggett, V.; and Fersht, A. R. Is there a unifying mechanism for protein folding? *Trends in biochemical sciences* **2003**, *28*, 18–25.
- [5] Tsytlonok, M.; and Itzhaki, L. S. The how's and why's of protein folding intermediates. *Archives of biochemistry and biophysics* **2013**, *531*, 14–23.
- [6] Robic, S.; Guzman-Casado, M.; Sanchez-Ruiz, J. M.; and Marqusee, S. Role of residual structure in the unfolded state of a thermophilic protein. *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, 11345–9.
- [7] Hartl, F. U.; Bracher, A.; and Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **2011**, *475*, 324–32.
- [8] McMorran, L. M.; Brockwell, D. J.; and Radford, S. E. Mechanistic studies of the biogenesis and folding of outer membrane proteins in vitro and in vivo: What have we learned to date? *Archives of biochemistry and biophysics* **2014**,
- [9] Braselmann, E.; Chaney, J. L.; and Clark, P. L. Folding the proteome. *Trends in biochemical sciences* **2013**, *38*, 337–44.
- [10] Wirth, A. J.; and Gruebele, M. Quinary protein structure and the consequences of crowding in living cells: leaving the test-tube behind. *BioEssays : news and reviews in molecular, cellular and developmental biology* **2013**, *35*, 984–93.
- [11] Hingorani, K. S.; and Gierasch, L. M. Comparing protein folding in vitro and in vivo: foldability meets the fitness challenge. *Current opinion in structural biology* **2014**, *24*, 81–90.
- [12] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (New York, N.Y.)* **1973**, *181*, 223–30.
- [13] Levinthal, C. Are There Pathways For Protein Folding. *Journal De Chimie Physique Et De Physico-Chimie Biologique* **1968**, *65*, 44–&.
- [14] Kim, P. S.; and Baldwin, R. L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annual Review Of Biochemistry* **1982**, *51*, 459–489.
- [15] Barrick, D. What have we learned from the studies of two-state folders, and what are the unanswered questions about two-state protein folding? *Physical biology* **2009**, *6*, 015001.
- [16] Wetlaufer, D. B. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proceedings of the National Academy of Sciences* **1973**, *70*, 697–701.
- [17] Onuchic, J. N.; Luthey-Schulten, Z.; and Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry* **1997**, *48*, 545–600.
- [18] Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; and Muñoz, V. Experimental identification of downhill protein folding. *Science (New York, N.Y.)* **2002**, *298*, 2191–5.
- [19] Carvalho, F. a.; Martins, I. C.; and Santos, N. C. Atomic force microscopy and force spectroscopy on the assessment of protein folding and functionality. *Archives of biochemistry and biophysics* **2013**, *531*, 116–27.
- [20] Rizzuti, B.; and Daggett, V. Using simulations to provide the framework for experimental protein folding studies. *Archives of biochemistry and biophysics* **2013**, *531*, 128–35.
- [21] Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; and Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annual Review of Biophysics* **2012**, *41*, 429–452.
- [22] Piana, S.; Klepeis, J. L.; and Shaw, D. E. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current opinion in structural biology* **2014**, *24*, 98–105.
- [23] Piana, S.; Lindorff-Larsen, K.; and Shaw, D. E. Protein folding kinetics and thermodynamics from

## References

- atomistic simulation. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109*, 17845–50.
- [24] Tompa, P. Intrinsically unstructured proteins. *Trends in biochemical sciences* **2002**, *27*, 527–33.
- [25] Fink, A. L. Natively unfolded proteins. *Current opinion in structural biology* **2005**, *15*, 35–41.
- [26] Uversky, V. N. Intrinsically disordered proteins from A to Z. *The international journal of biochemistry & cell biology* **2011**, *43*, 1090–103.
- [27] He, B.; Wang, K.; Liu, Y.; Xue, B.; Uversky, V. N.; and Dunker, a. K. Predicting intrinsic disorder in proteins: an overview. *Cell research* **2009**, *19*, 929–49.
- [28] Uversky, V. N. Unusual biophysics of intrinsically disordered proteins. *Biochimica et biophysica acta* **2013**, *1834*, 932–51.
- [29] Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS letters* **2005**, *579*, 3346–54.
- [30] Dyson, H. J.; and Wright, P. E. Coupling of folding and binding for unstructured proteins. *Current opinion in structural biology* **2002**, *12*, 54–60.
- [31] Tompa, P.; Szász, C.; and Buday, L. Structural disorder throws new light on moonlighting. *Trends in biochemical sciences* **2005**, *30*, 484–9.
- [32] Cumberworth, A.; Lamour, G.; Babu, M. M.; and Gsponer, J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *The Biochemical journal* **2013**, *454*, 361–9.
- [33] Hazy, E.; and Tompa, P. Limitations of induced folding in molecular recognition by intrinsically disordered proteins. *Chemphyschem : a European journal of chemical physics and physical chemistry* **2009**, *10*, 1415–9.
- [34] Jensen, M. R. b.; Zweckstetter, M.; Huang, J.-R.; and Blackledge, M. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chemical reviews* **2014**,
- [35] Kjaergaard, M.; and Poulsen, F. M. Disordered proteins studied by chemical shifts. *Progress in nuclear magnetic resonance spectroscopy* **2012**, *60*, 42–51.
- [36] Jensen, M. R. b.; Markwick, P. R. L.; Meier, S.; Griesinger, C.; Zweckstetter, M.; Grzesiek, S.; Bernadó, P.; and Blackledge, M. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure (London, England : 1993)* **2009**, *17*, 1169–85.
- [37] Kosol, S.; Contreras-Martos, S.; Cedeño, C.; and Tompa, P. Structural characterization of intrinsically disordered proteins by NMR spectroscopy. *Molecules (Basel, Switzerland)* **2013**, *18*, 10802–28.
- [38] Mittag, T.; and Forman-Kay, J. D. Atomic-level characterization of disordered protein ensembles. *Current opinion in structural biology* **2007**, *17*, 3–14.
- [39] Fisher, C. K.; and Stultz, C. M. Constructing ensembles for intrinsically disordered proteins. *Current opinion in structural biology* **2011**, *21*, 426–31.
- [40] Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; and Shaw, D. E. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society* **2012**, *134*, 3787–91.
- [41] Best, R. B.; Buchete, N.-V.; and Hummer, G. Are current molecular dynamics force fields too helical? *Biophysical journal* **2008**, *95*, L07–9.
- [42] Camilloni, C.; Cavalli, A.; and Vendruscolo, M. Replica-Averaged Metadynamics. *Journal of Chemical Theory and Computation* **2013**, *9*, 5610–5617.
- [43] Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; and Phillips, D. C. 3-Dimensional Model Of The Myoglobin Molecule Obtained By X-Ray Analysis. *Nature* **1958**, *181*, 662–666.
- [44] Schmidt, A.; Teeter, M.; Weckert, E.; and Lamzin, V. S. Crystal structure of small protein crambin at 0.48Å resolution. *Acta Crystallographica Section F* **2011**, *67*, 424–428.
- [45] Williamson, M. P.; Havel, T. F.; and Wüthrich, K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by <sup>1</sup>H nuclear magnetic resonance and distance geometry. *Journal of Molecular Biology* **1985**, *182*, 295–315.
- [46] Torchia, D. a. Dynamics of biomolecules from picoseconds to seconds at atomic resolution. *Journal of magnetic resonance (San Diego, Calif. : 1997)* **2011**, *212*, 1–10.
- [47] Tochio, H. Watching protein structure at work in living cells using NMR spectroscopy. *Current opinion in chemical biology* **2012**, *16*, 609–13.
- [48] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; and Tasumi, M. The Protein Data Bank. A Computer-Based Archival

## References

- File for Macromolecular Structures. *European Journal of Biochemistry* **1977**, *80*, 319–324.
- [49] Murzin, A. G.; Brenner, S. E.; Hubbard, T.; and Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* **1995**, *247*, 536–540.
- [50] Orengo, C. a.; Michie, a. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; and Thornton, J. M. CATH—a hierarchic classification of protein domain structures. *Structure (London, England : 1993)* **1997**, *5*, 1093–108.
- [51] Laskowski, R. a.; Watson, J. D.; and Thornton, J. M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic acids research* **2005**, *33*, W89–93.
- [52] Huber, R.; and Bennett, W. S. Functional significance of flexibility in proteins. *Biopolymers* **1983**, *22*, 261–79.
- [53] Gerstein, M.; Lesk, a. M.; and Chothia, C. Structural mechanisms for domain movements in proteins. *Biochemistry* **1994**, *33*, 6739–49.
- [54] Frauenfelder, H.; Sligar, S.; and Wolynes, P. The energy landscapes and motions of proteins. *Science* **1991**, *254*, 1598–1603.
- [55] Teilum, K.; Olsen, J. G.; and Kragelund, B. B. Functional aspects of protein flexibility. *Cellular and molecular life sciences : CMLS* **2009**, *66*, 2231–47.
- [56] Popovych, N.; Sun, S.; Ebright, R. H.; and Kalodimos, C. G. Dynamically driven protein allostery. *Nature structural & molecular biology* **2006**, *13*, 831–8.
- [57] van der Kamp, M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R. D.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S.; Bromley, D.; Beck, D. A. C.; and Daggett, V. Dynameomics: a comprehensive database of protein dynamics. *Structure (London, England : 1993)* **2010**, *18*, 423–35.
- [58] Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Pérez, A.; Camps, J.; Hospital, A.; Gelpí, J. L.; and Orozco, M. A consensus view of protein dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104*, 796–801.
- [59] Benson, N. C.; and Daggett, V. Dynameomics: large-scale assessment of native protein flexibility. *Protein science : a publication of the Protein Society* **2008**, *17*, 2038–50.
- [60] Ángyán, A. F.; and Gáspári, Z. Ensemble-based interpretations of NMR structural data to describe protein internal dynamics. *Molecules (Basel, Switzerland)* **2013**, *18*, 10548–67.
- [61] Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; and Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *Journal of the American Chemical Society* **2007**, *129*, 5656–64.
- [62] Lange, O. F.; Lakomek, N.-A.; Farès, C.; Schröder, G. F.; Walter, K. F. a.; Becker, S.; Meiler, J.; Grubmüller, H.; Griesinger, C.; and de Groot, B. L. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science (New York, N.Y.)* **2008**, *320*, 1471–5.
- [63] Motlagh, H. N.; Li, J.; Thompson, E. B.; and Hilser, V. J. Interplay between allostery and intrinsic disorder in an ensemble. *Biochemical Society transactions* **2012**, *40*, 975–80.
- [64] Motlagh, H. N.; Wrabl, J. O.; Li, J.; and Hilser, V. J. The ensemble nature of allostery. *Nature* **2014**, *508*, 331–9.
- [65] Becktel, W. J.; and Schellman, J. A. Protein stability curves. *Biopolymers* **1987**, *26*, 1859–77.
- [66] Canchi, D. R.; and García, A. E. Cosolvent effects on protein stability. *Annual review of physical chemistry* **2013**, *64*, 273–93.
- [67] Tanford, C.; Kawahara, K.; and Lapanje, S. Proteins as Random Coils. I. Intrinsic Viscosities and Sedimentation Coefficients in Concentrated Guanidine Hydrochloride. *Journal of the American Chemical Society* **1967**, *89*, 729–736.
- [68] Dill, K. a.; and Shortle, D. Denatured states of proteins. *Annual review of biochemistry* **1991**, *60*, 795–825.
- [69] Ohgushi, M.; and Wada, a. 'Molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS letters* **1983**, *164*, 21–4.
- [70] Yancey, P. H.; Clark, M. E.; Hand, S. C.; Bowlus, R. D.; and Somero, G. N. Living with water stress: evolution of osmolyte systems. *Science (New York, N.Y.)* **1982**, *217*, 1214–22.
- [71] Buck, M. Trifluoroethanol and colleagues: cosolvents come of age. Recent studies with peptides and proteins. *Quarterly reviews of biophysics* **1998**, *31*, 297–355.
- [72] Tanford, C. Isothermal Unfolding of Globular Proteins in Aqueous Urea Solutions. *Journal of the American Chemical Society* **1964**, *86*, 2050–2059.
- [73] Bolen, D. W.; and Rose, G. D. Structure and energetics of the hydrogen-bonded backbone in protein

## References

- folding. *Annual review of biochemistry* **2008**, *77*, 339–62.
- [74] Flory, P. J. Thermodynamics of High Polymer Solutions. *The Journal of Chemical Physics* **1942**, *10*.
- [75] Timasheff, S. N. Protein-solvent preferential interactions, protein hydration, and the modulation of biochemical reactions by solvent components. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 9721–6.
- [76] Daggett, V. Molecular dynamics simulations of the protein unfolding/folding reaction. *Accounts of chemical research* **2002**, *35*, 422–9.
- [77] Heyda, J.; Kozisek, M.; Bednářová, L.; Thompson, G.; Konvalinka, J.; Vondrasek, J.; and Jungwirth, P. Urea and guanidinium induced denaturation of a Trp-cage miniprotein. *The journal of physical chemistry. B* **2011**, *115*, 8910–24.
- [78] Schneck, E.; Horinek, D.; and Netz, R. R. Insight into the molecular mechanisms of protein stabilizing osmolytes from global force-field variations. *The journal of physical chemistry. B* **2013**, *117*, 8310–21.
- [79] Shaw, a.; and Bott, R. Engineering enzymes for stability. *Current opinion in structural biology* **1996**, *6*, 546–50.
- [80] Levitt, M.; and Lifson, S. Refinement of protein conformations using a macromolecular energy minimization procedure. *Journal of molecular biology* **1969**, *46*, 269–79.
- [81] McCammon, J. A.; Gelin, B. R.; and Karplus, M. Dynamics of folded proteins. *NATURE* **1977**, *267*, 585–590.
- [82] Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; and Ren, P. The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. *Journal of chemical theory and computation* **2013**, *9*, 4046–4063.
- [83] Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; and Wang, J. Polarization effects in molecular mechanical force fields. *Journal of physics. Condensed matter : an Institute of Physics journal* **2009**, *21*, 333102.
- [84] Shaw, D. E. et al. Millisecond-Scale Molecular Dynamics Simulations on Anton. PROCEEDINGS OF THE CONFERENCE ON HIGH PERFORMANCE COMPUTING NETWORKING, STORAGE AND ANALYSIS. 345 E 47TH ST, NEW YORK, NY 10017 USA, 2009.
- [85] MacKerell, A. In *COMPUTATIONAL BIOCHEMISTRY AND BIOPHYSICS*; Becker, O., MacKerell, A., Roux, B., and Watanabe, M., Eds.; MARCEL DEKKER, 270 MADISON AVE, NEW YORK, NY 10016 USA: NEW YORK, 2001; Chapter Atomistic, pp 7–38.
- [86] Ponder, J. W.; and Case, D. A. Force fields for protein simulations. *Advances in protein chemistry* **2003**, *66*, 27–85.
- [87] Lifson, S.; and Warshel, A. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *The Journal of Chemical Physics* **1968**, *49*.
- [88] Allinger, N. L. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *Journal of the American Chemical Society* **1977**, *99*, 8127–8134.
- [89] Hagler, a. T.; Huler, E.; and Lifson, S. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of the American Chemical Society* **1974**, *96*, 5319–27.
- [90] Momany, F. a.; McGuire, R. F.; Burgess, a. W.; and Scheraga, H. a. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *The Journal of Physical Chemistry* **1975**, *79*, 2361–2381.
- [91] Nemethy, G.; Pottle, M. S.; and Scheraga, H. A. Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *The Journal of Physical Chemistry* **1983**, *87*, 1883–1887.
- [92] Weiner, P. K.; and Kollman, P. A. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry* **1981**, *2*, 287–303.
- [93] Case, D. a.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; and Woods, R. J. The Amber biomolecular simulation programs. *Journal of computational chemistry* **2005**, *26*, 1668–88.
- [94] Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; and Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* **1984**, *106*, 765–784.

## References

- [95] Blaney, J. M.; Weiner, P. K.; Dearing, A.; Kollman, P. A.; Jorgensen, E. C.; Oatley, S. J.; Burridge, J. M.; and Blake, C. C. F. Molecular mechanics simulation of protein-ligand interactions: binding of thyroid hormone analogs to prealbumin. *Journal of the American Chemical Society* **1982**, *104*, 6424–6434.
- [96] Singh, U. C.; and Kollman, P. A. An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **1984**, *5*, 129–145.
- [97] Jorgensen, W. L. Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *Journal of the American Chemical Society* **1981**, *103*, 335–340.
- [98] Weiner, S. J.; Kollman, P. a.; Nguyen, D. T.; and Case, D. a. An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry* **1986**, *7*, 230–252.
- [99] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; and Kollman, P. a. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.
- [100] Bayly, C. I.; Cieplak, P.; Cornell, W.; and Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993**, *97*, 10269–10280.
- [101] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; and Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*.
- [102] Jorgensen, W. L.; and Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **1988**, *110*, 1657–1666.
- [103] Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; and Pohorille, A. In *Computer Simulation of Biomolecular Systems SE - 2*; Gunsteren, W., Weiner, P., and Wilkinson, A., Eds.; Computer Simulations of Biomolecular Systems; Springer Netherlands, 1997; Vol. 3; pp 83–96.
- [104] Wang, J.; Cieplak, P.; and Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **2000**, *21*, 1049–1074.
- [105] Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; and Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry* **2003**, *24*, 1999–2012.
- [106] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. a.; and Case, D. a. Development and testing of a general amber force field. *Journal of computational chemistry* **2004**, *25*, 1157–74.
- [107] Sorin, E. J.; and Pande, V. S. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophysical journal* **2005**, *88*, 2472–93.
- [108] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; and Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–25.
- [109] Best, R.; and Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix- Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004–9015.
- [110] Best, R. B.; and Mittal, J. Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *The journal of physical chemistry. B* **2010**, *114*, 14916–23.
- [111] Abascal, J. L. F.; and Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *The Journal of chemical physics* **2005**, *123*, 234505.
- [112] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; and Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–8.
- [113] Li, D.-W.; and Brüschweiler, R. NMR-based protein potentials. *Angewandte Chemie (International ed. in English)* **2010**, *49*, 6778–80.
- [114] Nerenberg, P. S.; and Head-Gordon, T. Optimizing Protein-Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides. *Journal of Chemical Theory and Computation* **2011**, *7*, 1220–1230.
- [115] Horn, H. W.; Swope, W. C.; Pitner, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; and Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-

## References

- Ew. *The Journal of chemical physics* **2004**, *120*, 9665–78.
- [116] Cerutti, D. S.; Rice, J. E.; Swope, W. C.; and Case, D. a. Derivation of fixed partial charges for amino acids accommodating a specific water model and implicit polarization. *The journal of physical chemistry. B* **2013**, *117*, 2328–38.
- [117] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; and Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **1983**, *4*, 187–217.
- [118] Brooks, B. R. et al. CHARMM: the biomolecular simulation program. *Journal of computational chemistry* **2009**, *30*, 1545–614.
- [119] Nilsson, L.; and Karplus, M. Empirical energy functions for energy minimization and dynamics of nucleic acids. *Journal of Computational Chemistry* **1986**, *7*, 591–616.
- [120] Neria, E.; Fischer, S.; and Karplus, M. Simulation of activation free energies in molecular systems. *The Journal of Chemical Physics* **1996**, *105*, 1902.
- [121] Gelin, B. R.; and Karplus, M. Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochemistry* **1979**, *18*, 1256–68.
- [122] MacKerell, a. D. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B* **1998**, *102*, 3586–3616.
- [123] MacKerell, A. D.; and Banavali, N. K. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *Journal of Computational Chemistry* **2000**, *21*, 105–120.
- [124] Mackerell, A. D.; Feig, M.; and Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of computational chemistry* **2004**, *25*, 1400–15.
- [125] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; and MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone  $\phi$ ,  $\psi$  and Side-Chain  $\chi_1$  and  $\chi_2$  Dihedral Angles. *Journal of Chemical Theory and Computation* **2012**, *8*, 3257–3273.
- [126] Huang, J.; and MacKerell, A. D. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of computational chemistry* **2013**, *34*, 2135–45.
- [127] Piana, S.; Lindorff-Larsen, K.; and Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophysical journal* **2011**, *100*, L47–9.
- [128] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; and Mackerell, A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–90.
- [129] Vanommeslaeghe, K.; and MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *Journal of chemical information and modeling* **2012**, *52*, 3144–54.
- [130] Jorgensen, W.; Madura, J.; and Swenson, C. Optimized intermolecular potential functions for liquid hydrocarbons. *Journal of the American Chemical Society* **1984**, *106*, 6638–6646.
- [131] Jorgensen, W. L.; and Swenson, C. J. Optimized intermolecular potential functions for amides and peptides. Hydration of amides. *Journal of the American Chemical Society* **1985**, *107*, 1489–1496.
- [132] Jorgensen, W. L.; Maxwell, D. S.; and Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.
- [133] Kaminski, G. a.; Friesner, R. a.; Tirado-Rives, J.; and Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *The Journal of Physical Chemistry B* **2001**, *105*, 6474–6487.
- [134] Jiang, F.; Zhou, C.-y.; and Wu, Y.-d. Residue-Specific Force Field Based on the Protein Coil Library. RSFF1: Modification of OPLS-AA/L. *The journal of physical chemistry. B* **2014**,
- [135] Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; and van Gunsteren, W. F. The GROMOS Biomolecular Simulation Program Package. *The Journal of Physical Chemistry A* **1999**, *103*, 3596–3607.
- [136] Dunfield, L. G.; Burgess, A. W.; and Scheraga, H. A. Energy parameters in polypeptides. 8. Empirical potential energy algorithm for the conformational analysis of large molecules. *The Journal of Physical Chemistry* **1978**, *82*, 2609–2616.



## References

- [137] Berendsen, H.J.C.; Postma, J.P.M. ; van Gunsteren, W. . H. J. In *Intermolecular Forces*; Pullman, B., Ed.; 1981; p 331.
- [138] Daura, X.; Mark, A. E.; and Van Gunsteren, W. F. Parametrization of aliphatic CH<sub>n</sub> united atoms of GROMOS96 force field. *Journal of Computational Chemistry* **1998**, *19*, 535–547.
- [139] Schuler, L. D.; Daura, X.; and van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry* **2001**, *22*, 1205–1218.
- [140] Oostenbrink, C.; Villa, A.; Mark, A. E.; and van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry* **2004**, *25*, 1656–76.
- [141] Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; and van Gunsteren, W. F. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *European biophysics journal : EBJ* **2011**, *40*, 843–56.
- [142] Reif, M. M.; Hünenberger, P. H.; and Oostenbrink, C. New Interaction Parameters for Charged Amino Acid Side Chains in the GROMOS Force Field. *Journal of Chemical Theory and Computation* **2012**, *8*, 3705–3723.
- [143] Levitt, M.; Hirshberg, M.; Sharon, R.; and Daggett, V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Computer physics communications* **1995**, *91*, 215–231.
- [144] Levitt, M. Molecular dynamics of native protein. *Journal of Molecular Biology* **1983**, *168*, 595–617.
- [145] Levitt, M.; Hirshberg, M.; Sharon, R.; Laidig, K. E.; and Daggett, V. Calibration and Testing of a Water Model for Simulation of the Molecular Dynamics of Proteins and Nucleic Acids in Solution. *The Journal of Physical Chemistry B* **1997**, *101*, 5051–5061.
- [146] Price, D. J.; and Brooks, C. L. Modern protein force fields behave comparably in molecular dynamics simulations. *Journal of computational chemistry* **2002**, *23*, 1045–57.
- [147] Kosov, D. S.; and Stock, G. Conformational Dynamics of Trialanine in Water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS Force Fields to NMR and Infrared Experiments. *The Journal of Physical Chemistry B* **2003**, *107*, 5064–5073.
- [148] Gnanakaran, S.; and García, A. E. Helix-coil transition of alanine peptides in water: force field dependence on the folded and unfolded structures. *Proteins* **2005**, *59*, 773–82.
- [149] Vymětal, J.; and Vondrášek, J. Metadynamics as a tool for mapping the conformational and free-energy space of peptides—the alanine dipeptide case study. *The journal of physical chemistry. B* **2010**, *114*, 5632–42.
- [150] Vymětal, J.; and Vondrášek, J. Critical Assessment of Current Force Fields. Short Peptide Test Case. *Journal of Chemical Theory and Computation* **2013**, *9*, 441–451.
- [151] Beauchamp, K. a.; Lin, Y.-S.; Das, R.; and Pande, V. S. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of Chemical Theory and Computation* **2012**, *8*, 1409–1414.
- [152] Best, R. B.; and Mittal, J. Balance between alpha and beta Structures in Ab Initio Protein Folding. *The journal of physical chemistry. B* **2010**,
- [153] Mittal, J.; and Best, R. B. Tackling force-field bias in protein folding simulations: folding of Villin HP35 and Pin WW domains in explicit water. *Biophysical journal* **2010**, *99*, L26–8.
- [154] Cino, E. a.; Choy, W.-Y.; and Karttunen, M. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *Journal of chemical theory and computation* **2012**, *8*, 2725–2740.
- [155] Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; and Shaw, D. E. Systematic validation of protein force fields against experimental data. *PloS one* **2012**, *7*, e32131.
- [156] Vymětal, J.; and Vondrášek, J. The DF-LCCSD(T0) correction of the  $\phi/\psi$  force field dihedral parameters significantly influences the free energy profile of the alanine dipeptide. *Chemical Physics Letters* **2011**, *503*, 301–304.
- [157] Nerenberg, P. S.; Jo, B.; So, C.; Tripathy, A.; and Head-Gordon, T. Optimizing solute-water van der waals interactions to reproduce solvation free energies. *The journal of physical chemistry. B* **2012**, *116*, 4524–34.
- [158] Laio, A.; and Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 12562–6.
- [159] Laio, A.; and Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. 2008.
- [160] Leone, V.; Marinelli, F.; Carloni, P.; and Parrinello, M. Targeting biomolecular flexibility with

## References

- metadynamics. *Current opinion in structural biology* **2010**, *20*, 148–54.
- [161] Barducci, A.; Bonomi, M.; and Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 826–843.
- [162] Sutto, L.; Marsili, S.; and Gervasio, F. L. New advances in metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 771–779.
- [163] Fiorin, G.; Klein, M. L.; and Hémin, J. Using collective variables to drive molecular dynamics simulations. *Molecular Physics* **2013**, *111*, 3345–3362.
- [164] Iannuzzi, M.; Laio, A.; and Parrinello, M. Efficient Exploration of Reactive Potential Energy Surfaces Using Car-Parrinello Molecular Dynamics. *Physical Review Letters* **2003**, *90*, 238–302.
- [165] Stirling, A.; Iannuzzi, M.; Laio, A.; and Parrinello, M. Azulene-to-naphthalene rearrangement: the Car-Parrinello metadynamics method explores various reaction mechanisms. *Chemphyschem : a European journal of chemical physics and physical chemistry* **2004**, *5*, 1558–68.
- [166] Ensing, B.; De Vivo, M.; Liu, Z.; Moore, P.; and Klein, M. L. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Accounts of chemical research* **2006**, *39*, 73–81.
- [167] van der Vaart, A. Simulation of conformational transitions. *Theoretical Chemistry Accounts* **2006**, *116*, 183–193.
- [168] Vymětal, J.; and Vondrášek, J. Gyration- and inertia-tensor-based collective coordinates for metadynamics. Application on the conformational behavior of polyalanine peptides and Trp-cage folding. *The journal of physical chemistry. A* **2011**, *115*, 11455–65.
- [169] Gervasio, F. L.; Laio, A.; and Parrinello, M. Flexible docking in solution using metadynamics. *Journal of the American Chemical Society* **2005**, *127*, 2600–7.
- [170] Limongelli, V.; Bonomi, M.; and Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 6358–63.
- [171] Fiorin, G.; Pastore, a.; Carloni, P.; and Parrinello, M. Using metadynamics to understand the mechanism of calmodulin/target recognition at atomic detail. *Biophysical journal* **2006**, *91*, 2768–77.
- [172] Barducci, A.; Bonomi, M.; Prakash, M. K.; and Parrinello, M. Free-energy landscape of protein oligomerization from atomistic simulations. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, E4708–13.
- [173] Bussi, G.; Gervasio, F. L.; Laio, A.; and Parrinello, M. Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *Journal of the American Chemical Society* **2006**, *128*, 13435–41.
- [174] Piana, S.; and Laio, A. A Bias-Exchange Approach to Protein Folding. *The Journal of Physical Chemistry B* **2007**, *111*, 4553–4559.
- [175] Babin, V.; Roland, C.; Darden, T. a.; and Sagui, C. The free energy landscape of small peptides as obtained from metadynamics with umbrella sampling corrections. *The Journal of chemical physics* **2006**, *125*, 204909.
- [176] Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V.; Gervasio, F. L.; and Parrinello, M. Metadynamics simulation of prion protein: beta-structure stability and the early stages of misfolding. *Journal of the American Chemical Society* **2006**, *128*, 2705–10.
- [177] Martoňák, R.; Laio, A.; and Parrinello, M. Predicting Crystal Structures: The Parrinello-Rahman Method Revisited. *Physical Review Letters* **2003**, *90*, 075503.
- [178] Raiteri, P.; Martoňák, R.; and Parrinello, M. Exploring Polymorphism: The Case of Benzene13. *Angewandte Chemie International Edition* **2005**, *44*, 3769–3773.
- [179] Martoňák, R.; Laio, A.; Bernasconi, M.; Ceriani, C.; Raiteri, P.; Zipoli, F.; and Parrinello, M. Simulation of structural phase transitions by metadynamics. *Zeitschrift für Kristallographie* **2005**, *220*, 489–498.
- [180] Prestipino, S.; and Giaquinta, P. V. Liquid-solid coexistence via the metadynamics approach. *The Journal of chemical physics* **2008**, *128*, 114707.
- [181] Valsson, O.; and Parrinello, M. Thermodynamical Description of a Quasi-First-Order Phase Transition from the Well-Tempered Ensemble. *Journal of Chemical Theory and Computation* **2013**, *9*, 5267–5276.
- [182] Deighan, M.; and Pfaendtner, J. Exhaustively sampling peptide adsorption with metadynamics. *Langmuir : the ACS journal of surfaces and colloids* **2013**, *29*, 7999–8009.
- [183] Ilott, A. J.; Palucha, S.; Hodgkinson, P.; and Wilson, M. R. Well-tempered metadynamics as a tool for characterizing multi-component, crystalline molecular machines. *The journal of physical chemistry. B* **2013**, *117*, 12286–95.

## References

- [184] Bussi, G.; Laio, A.; and Parrinello, M. Equilibrium Free Energies from Nonequilibrium Metadynamics. *Physical Review Letters* **2006**, *96*, 10–13.
- [185] Wang, F.; and Landau, D. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Physical Review Letters* **2001**, *86*, 2050–2053.
- [186] Junghans, C.; Perez, D.; and Vogel, T. Molecular Dynamics in the Multicanonical Ensemble: Equivalence of Wang-Landau Sampling, Statistical Temperature Molecular Dynamics, and Metadynamics. *Journal of Chemical Theory and Computation* **2014**, *10*, 1843–1847.
- [187] Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; and Parrinello, M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *The journal of physical chemistry. B* **2006**, *110*, 3533–9.
- [188] Barducci, A.; Bussi, G.; and Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **2008**, *100*, 20603.
- [189] Spiwok, V.; Lipovova, P.; and Kralova, B. Metadynamics in Essential Coordinates: Free Energy Simulation of Conformational Changes. *The Journal of Physical Chemistry B* **2007**, *111*, 3073–3076.
- [190] Branduardi, D.; Gervasio, F. L.; and Parrinello, M. From A to B in free energy space. *The Journal of chemical physics* **2007**, *126*, 054103.
- [191] Bonomi, M.; and Parrinello, M. Enhanced Sampling in the Well-Tempered Ensemble. *Physical Review Letters* **2010**, *104*, 190601.
- [192] Branduardi, D.; Bussi, G.; and Parrinello, M. Metadynamics with Adaptive Gaussians. *Journal of Chemical Theory and Computation* **2012**, *8*, 2247–2254.
- [193] Laio, A.; Rodriguez-Forteza, A.; Gervasio, F. L.; Ceccarelli, M.; and Parrinello, M. Assessing the accuracy of metadynamics. *The journal of physical chemistry. B* **2005**, *109*, 6714–21.
- [194] Abrams, C.; and Bussi, G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2013**, *16*, 163–199.
- [195] Bonomi, M.; Barducci, A.; and Parrinello, M. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *Journal of computational chemistry* **2009**, *30*, 1615–21.
- [196] Barducci, A.; Bonomi, M.; and Parrinello, M. Linking well-tempered metadynamics simulations with experiments. *Biophysical journal* **2010**, *98*, L44–6.
- [197] Shi, Z.; Chen, K.; Liu, Z.; and Kallenbach, N. R. Conformation of the backbone in unfolded proteins. *Chemical reviews* **2006**, *106*, 1877–97.
- [198] Schweitzer-Stenner, R. Conformational propensities and residual structures in unfolded peptides and proteins. *Molecular bioSystems* **2012**, *8*, 122–33.
- [199] Eker, F.; Griebenow, K.; Cao, X.; Nafie, L.; and Schweitzer-Stenner, R. Preferred peptide backbone conformations in the unfolded state revealed by the structure analysis of alanine-based (AXA) tripeptides in aqueous solution. *Proceedings of the National Academy of Sciences* **2004**, *101*, 10054.
- [200] Pappu, R. V.; and Rose, G. D. A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Society* **2002**, 2437–2455.
- [201] Beck, D. a. C.; Alonso, D. O. V.; Inoyama, D.; and Daggett, V. The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 12259–64.
- [202] Schweitzer-Stenner, R.; Hagarman, A.; Measey, T. J.; Mathieu, D.; and Schwalbe, H. Intrinsic propensities of amino acid residues in GxG peptides inferred from amide I' band profiles and NMR scalar coupling constants. *Journal of the American Chemical Society* **2010**, *132*, 540–51.
- [203] Hagarman, A.; Mathieu, D.; Toal, S.; Measey, T. J.; Schwalbe, H.; and Schweitzer-Stenner, R. Amino acids with hydrogen-bonding side chains have an intrinsic tendency to sample various turn conformations in aqueous solution. *Chemistry (Weinheim an der Bergstrasse, Germany)* **2011**, *17*, 6789–97.
- [204] Brant, D. A.; and Flory, P. J. The Configuration of Random Polypeptide Chains. II. Theory. *Journal of the American Chemical Society* **1965**, *87*, 2791–2800.
- [205] Shi, Z.; Olson, C. A.; Rose, G. D.; Baldwin, R. L.; and Kallenbach, N. R. Polyproline II structure in a sequence of seven alanine residues. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 9190–5.
- [206] Zagrovic, B.; Lipfert, J.; Sorin, E. J.; Millett, I. S.; van Gunsteren, W. F.; Doniach, S.; and Pande, V. S. Unusual compactness of a polyproline type II structure. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 11698–703.
- [207] Makowska, J.; Rodziewicz-Motowidło, S.; Bagińska, K.; Vila, J. a.; Liwo, A.; Chmurzyński, L.; and

## References

- Scheraga, H. a. Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded peptides and proteins. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 1744–9.
- [208] Makowska, J.; Rodziewicz-Motowidlo, S.; Baginska, K.; Makowski, M.; Vila, J. a.; Liwo, A.; Chmurzynski, L.; and Scheraga, H. a. Further evidence for the absence of polyproline II stretch in the XAO peptide. *Biophysical journal* **2007**, *92*, 2904–17.
- [209] Chen, K.; Liu, Z.; Zhou, C.; Bracken, W. C.; and Kallenbach, N. R. Spin relaxation enhancement confirms dominance of extended conformations in short alanine peptides. *Angewandte Chemie (International ed. in English)* **2007**, *46*, 9036–9.
- [210] Schweitzer-Stenner, R.; and Measey, T. J. The alanine-rich XAO peptide adopts a heterogeneous population, including turn-like and polyproline II conformations. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104*, 6649–54.
- [211] Shi, Z.; Chen, K.; Liu, Z.; Ng, A.; Bracken, W. C.; and Kallenbach, N. R. Polyproline II propensities from GGXGG peptides reveal an anticorrelation with beta-sheet scales. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 17964–8.
- [212] Grdadolnik, J.; Mohacek-Grosev, V.; Baldwin, R. L.; and Avbelj, F. Populations of the three major backbone conformations in 19 amino acid dipeptides. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108*, 1794–8.
- [213] Brown, A. M.; and Zondlo, N. J. A propensity scale for type II polyproline helices (PPII): aromatic amino acids in proline-rich sequences strongly disfavor PPII due to proline-aromatic interactions. *Biochemistry* **2012**, *51*, 5041–51.
- [214] Whittington, S. J.; Chellgren, B. W.; Hermann, V. M.; and Creamer, T. P. Urea promotes polyproline II helix formation: implications for protein denatured states. *Biochemistry* **2005**, *44*, 6269–75.
- [215] Elam, W. A.; Schrank, T. P.; Campagnolo, A. J.; and Hilser, V. J. Temperature and urea have opposing impacts on polyproline II conformational bias. *Biochemistry* **2013**, *52*, 949–58.
- [216] Tiffany, M. L.; and Krimm, S. Circular dichroism of poly-L-proline in an unordered conformation. *Biopolymers* **1968**, *6*, 1767–1770.
- [217] Dukor, R. K.; and Keiderling, T. A. Reassessment of the random coil conformation: Vibrational CD study of proline oligopeptides and related polypeptides. *Biopolymers* **1991**, *31*, 1747–1761.
- [218] Oh, K.-I.; Jung, Y.-S.; Hwang, G.-S.; and Cho, M. Conformational distributions of denatured and unstructured proteins are similar to those of 20 x 20 blocked dipeptides. *Journal of biomolecular NMR* **2012**, *53*, 25–41.
- [219] Fitzkee, N. C.; and Rose, G. D. Reassessing random-coil statistics in unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 12497–502.
- [220] Chou, P. Y.; and Fasman, G. D. Conformational parameters for amino acids in helical,  $\beta$ -sheet, and random coil regions calculated from proteins. *Biochemistry* **1974**, *13*, 211–222.
- [221] Jha, A.; Colubri, A.; Zaman, M.; Koide, S.; Sosnick, T.; and Freed, K. Helix, Sheet, and Polyproline II Frequencies and Strong Nearest Neighbor Effects in a Restricted Coil Library. *Biochemistry* **2005**, *44*, 9691–9702.
- [222] Avbelj, F.; Grdadolnik, S. G.; Grdadolnik, J.; and Baldwin, R. L. Intrinsic backbone preferences are fully present in blocked amino acids. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 1272–7.
- [223] Avbelj, F.; and Baldwin, R. L. Origin of the neighboring residue effect on peptide backbone conformation. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 10967–72.
- [224] Jung, Y.-s.; Oh, K.-i.; Hwang, G.-s.; and Cho, M. Neighboring Residue Effects in Terminally Blocked Dipeptides: Implications for Residual Secondary Structures in Intrinsically Unfolded/Disordered Proteins. *Chirality* **2014**, *361*.
- [225] Feig, M. Is Alanine Dipeptide a Good Model for Representing the Torsional Preferences of Protein Backbones? *Journal of Chemical Theory and Computation* **2008**, *4*, 1555–1564.
- [226] Cruz, V. L.; Ramos, J.; and Martinez-Salazar, J. Assessment of the intrinsic conformational preferences of dipeptide amino acids in aqueous solution by combined umbrella sampling/MBAR statistics. A comparison with experimental results. *The journal of physical chemistry. B* **2012**, *116*, 469–75.
- [227] Han, B.; Liu, Y.; Ginzinger, S. W.; and Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *Journal of biomolecular NMR* **2011**, *50*, 43–57.
- [228] Cho, S. S.; Levy, Y.; and Wolynes, P. G. P versus Q: structural reaction coordinates capture protein folding on smooth landscapes. *Proceedings of the National Academy of Sciences of the United States*

## References

- of America* **2006**, *103*, 586–91.
- [229] Das, P.; Moll, M.; Stamati, H.; Kaviraki, L. E.; and Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 9885–90.
- [230] Beck, D. a. C.; and Daggett, V. A one-dimensional reaction coordinate for identification of transition states from explicit solvent P(fold)-like calculations. *Biophysical journal* **2007**, *93*, 3382–91.
- [231] Toofanny, R. D.; Jonsson, A. L.; and Daggett, V. A comprehensive multidimensional-embedded, one-dimensional reaction coordinate for protein unfolding/folding. *Biophysical journal* **2010**, *98*, 2671–81.
- [232] Biarnés, X.; Pietrucci, F.; Marinelli, F.; and Laio, A. METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. *Computer Physics Communications* **2012**, *183*, 203–211.
- [233] Šolc, K.; and Stockmayer, W. H. Shape of a random-flight chain. *Journal Of Chemical Physics* **1971**, *54*, 2756–&.
- [234] Minato, T.; and Hatano, A. Distribution function and principal components for a polymer chain with excluded volume. *Macromolecules* **1981**, *14*, 1035–1038.
- [235] Theodorou, D. N.; and Suter, U. W. Shape of unperturbed linear polymers: polypropylene. *Macromolecules* **1985**, *18*, 1206–1214.
- [236] Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R.; and Others, PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009**, *180*, 1961–1972.
- [237] Guptasarma, P. Reversal of peptide backbone direction may result in the mirroring of protein structure. *FEBS letters* **1992**, *310*, 205–10.
- [238] Lacroix, E.; Viguera, a. R.; and Serrano, L. Reading protein sequences backwards. *Folding & design* **1998**, *3*, 79–85.
- [239] Olszewski, K. a.; Kolinski, a.; and Skolnick, J. Does a backwardly read protein sequence have a unique native state? *Protein engineering* **1996**, *9*, 5–14.
- [240] Witte, K.; Skolnick, J.; and Wong, C.-h. A Synthetic Retrotransition (Backward Reading) Sequence of the Right-Handed Three-Helix Bundle Domain (10-53) of Protein A Shows Similarity in Conformation as Predicted by Computation. *Journal of the American Chemical Society* **1998**, *120*, 13042–13045.
- [241] Haack, T.; Sánchez, Y. M.; González, M. J.; and Giralt, E. Structural comparison in solution of a native and retro peptide derived from the third helix of Staphylococcus aureus protein A, domain B: retro peptides, a useful tool for the discrimination of helix stabilization factors dependent on the peptide chain o. *Journal of peptide science : an official publication of the European Peptide Society* **1997**, *3*, 299–313.
- [242] Pan, P. K.; Zheng, Z. F.; Lyu, P. C.; and Huang, P. C. Why reversing the sequence of the alpha domain of human metallothionein-2 does not change its metal-binding and folding characteristics. *European journal of biochemistry / FEBS* **1999**, *266*, 33–9.
- [243] Neidigh, J. W.; Fesinmeyer, R. M.; and Andersen, N. H. Designing a 20-residue protein. *Nature structural biology* **2002**, *9*, 425–30.
- [244] Qiu, L.; Pabit, S. a.; Roitberg, A. E.; and Hagen, S. J. Smaller and faster: the 20-residue Trp-cage protein folds in 4 micros. *Journal of the American Chemical Society* **2002**, *124*, 12952–3.
- [245] Biedermannova, L.; E Riley, K.; Berka, K.; Hobza, P.; and Vondrasek, J. Another role of proline: stabilization interactions in proteins and protein complexes concerning proline and tryptophane. *Physical chemistry chemical physics : PCCP* **2008**, *10*, 6350–9.
- [246] Barua, B.; Lin, J. C.; Williams, V. D.; Kummeler, P.; Neidigh, J. W.; and Andersen, N. H. The Trp-cage: optimizing the stability of a globular miniprotein. *Protein engineering, design & selection : PEDS* **2008**, *21*, 171–85.
- [247] Compiani, M.; and Capriotti, E. Computational and theoretical methods for protein folding. *Biochemistry* **2013**, *52*, 8601–24.
- [248] Kim, D. E.; Chivian, D.; and Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic acids research* **2004**, *32*, W526–31.
- [249] Maupetit, J.; Derreumaux, P.; and Tuffery, P. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic acids research* **2009**, *37*, W498–503.
- [250] Kabsch, W.; and Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
- [251] Cammers-Goodwin, A.; Allen, T. J.; Oslick, S. L.; McClure, K. F.; Lee, J. H.; and Kemp, D. S.

## References

- Mechanism of Stabilization of Helical Conformations of Polypeptides by Water Containing Trifluoroethanol. *Journal of the American Chemical Society* **1996**, *118*, 3082–3090.
- [252] Luo, P.; and Baldwin, R. L. Mechanism of helix induction by trifluoroethanol: a framework for extrapolating the helix-forming properties of peptides from trifluoroethanol/water mixtures back to water. *Biochemistry* **1997**, *36*, 8413–21.
- [253] Walgers, R.; Lee, T. C.; and Cammers-Goodwin, A. An Indirect Chaotropic Mechanism for the Stabilization of Helix Conformation of Peptides in Aqueous Trifluoroethanol and Hexafluoro-2-propanol. *Journal of the American Chemical Society* **1998**, *120*, 5073–5079.
- [254] Hong, D.-p.; Hoshino, M.; Kuboi, R.; and Goto, Y. Clustering of Fluorine-Substituted Alcohols as a Factor Responsible for Their Marked Effects on Proteins and Peptides. *Journal of the American Chemical Society* **1999**, *121*, 8427–8433.
- [255] Jia, X.; Zhang, J. Z. H.; and Mei, Y. Assessing the accuracy of the general AMBER force field for 2,2,2-trifluoroethanol as solvent. *Journal of molecular modeling* **2013**, *19*, 2355–61.
- [256] Guo, M.; and Mei, Y. Equilibrium and folding simulations of NS4B H2 in pure water and water/2,2,2-trifluoroethanol mixed solvent: examination of solvation models. *Journal of molecular modeling* **2013**, *19*, 3931–9.
- [257] Chitra, R.; and Smith, P. E. A comparison of the properties of 2,2,2-trifluoroethanol and 2,2,2-trifluoroethanol/water mixtures using different force fields. *The Journal of Chemical Physics* **2001**, *115*, 5521.
- [258] Chitra, R.; and Smith, P. E. Properties of 2,2,2-trifluoroethanol and water mixtures. *The Journal of Chemical Physics* **2001**, *114*, 426.
- [259] Gerig, J. T. Toward a molecular dynamics force field for simulations of 40% trifluoroethanol-water. *The journal of physical chemistry. B* **2014**, *118*, 1471–80.
- [260] Kirkwood, J. G.; and Buff, F. P. The Statistical Mechanical Theory of Solutions. I. *The Journal of Chemical Physics* **1951**, *19*.
- [261] Chitra, R.; and Smith, P. E. Molecular Association in Solution: A Kirkwood-Buff Analysis of Sodium Chloride, Ammonium Sulfate, Guanidinium Chloride, Urea, and 2,2,2-Trifluoroethanol in Water. *The Journal of Physical Chemistry B* **2002**, *106*, 1491–1500.
- [262] Schnell, S. K.; Liu, X.; Simon, J.-M.; Bardow, A.; Bedeaux, D.; Vlugt, T. J. H.; and Kjelstrup, S. Calculating thermodynamic properties from fluctuations at small scales. *The journal of physical chemistry. B* **2011**, *115*, 10911–8.
- [263] Shirts, M. R.; and Pande, V. S. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *The Journal of chemical physics* **2005**, *122*, 134508.
- [264] Dupradeau, F.-Y.; Cézard, C.; Lelong, R.; Stanislawiak, E.; Pêcher, J.; Delepine, J. C.; and Cieplak, P. R.E.D.D.B.: a database for RESP and ESP atomic charges, and force field libraries. *Nucleic acids research* **2008**, *36*, D360–7.
- [265] Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; and Cieplak, P. The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Physical chemistry chemical physics : PCCP* **2010**, *12*, 7821–39.
- [266] Swope, W. C.; Horn, H. W.; and Rice, J. E. Accounting for polarization cost when using fixed charge force fields. I. Method for computing energy. *The journal of physical chemistry. B* **2010**, *114*, 8621–30.
- [267] Blandamer, M. J.; Burgess, J.; Cooney, A.; Cowles, H. J.; Horn, I. M.; Martin, K. J.; Morcom, K. W.; and Warrick, P. Excess molar Gibbs energies of mixing of water and 1,1,1,3,3,3-hexafluoropropan-2-ol mixtures at 298.15 K. Comparison of thermodynamic properties and inverse Kirkwood-Buff integral functions for binary aqueous mixtures formed by ethanol, propan-2-ol, 2,2,2-trifluoroethanol, and water. *Journal of the Chemical Society, Faraday Transactions* **1990**, *86*, 2209.

## Appendices

- Appendix A:** Vymětal, J.; and Vondrášek, J. Critical Assessment of Current Force Fields. Short Peptide Test Case. *Journal of Chemical Theory and Computation* **2013**, *9*, 441–451.
- Appendix B:** Vymětal, J.; and Vondrášek, J. The DF-LCCSD(T0) correction of the  $\phi/\psi$  force field dihedral parameters significantly influences the free energy profile of alanine dipeptide. *Chemical Physics Letters* **2011**, *503*(4-6), 301-304.
- Appendix C:** Towse, C.-L.; Vymětal, J.; Vondrášek, J.; and Daggett, V. Potential for underestimating residual structure in denatured states and intrinsically disordered proteins (*submitted*)
- Appendix D:** Vymětal, J.; and Vondrášek, J. Gyration- and inertia-tensor-based collective coordinates for metadynamics. Application on the conformational behavior of polyalanine peptides and Trp-cage folding. *The journal of physical chemistry. A* **2011**, *115*, 11455–65.
- Appendix E:** Vymětal, J.; Reddy, B.S.; Černý, J.; Chaloupková, R.; Židek, L.; Sklenář, V.; and Vondrášek, J. Retro Operation on the Trp-cage Miniprotein Sequence Produces an Unstructured Molecule Capable of Folding Similar to the Original Only upon 2,2,2-trifluoroethanol Addition. (*submitted*)
- Appendix F:** Vymětal, J.; and Vondrášek, J. Parametrization of 2,2,2-trifluoroethanol based on Generalized Amber Force Field provides realistic agreement between experimental and calculated properties of pure liquid as well as water mixed solutions. (*submitted*)