University of Natural Resources and Life Sciences, Vienna

Department of Material Sciences

and Process Engineering

Institute of Molecular Modeling and Simulation

Optimization and Validation of force-field parameters for Molecular Dynamics Simulations

Submitted in partial fulfillment of the requirements for the degree

"Dr. nat. techn."

submitted by

Dipl. Ing. Matthias Diem

supervised by

Prof. Dr. Chris Oostenbrink

Vienna, Austria

February, 2021

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre eidesstattlich, dass ich die vorliegende Arbeit selbständig angefertigt, keine anderen als die angegebenen Hilfsmittel benutzt und alle aus ungedruckten Quellen, gedruckter Literatur oder aus dem Internet im Wortlaut oder im wesentlichen Inhalt übernommenen Formulierungen und Konzepte gemäß den Richtlinien wissenschaftlicher Arbeiten zitiert, durch Fußnoten gekennzeichnet bzw. mit genauer Quellenangabe kenntlich gemacht habe.

Wien, am 25.01.2021

Matthias Diem

Abstract

Molecular dynamics simulations are an invaluable tool to investigate biomolecular systems and deepen our understanding of dynamics on an atomistic level. To obtain reliable results from simulations it is key to have a well parameterized force field. In this work a new parameter set for the dihedral angles of the protein backbone is presented, that is parameterized on the properties of small blocked di-peptides and extensively validated on a set of 52 proteins. Furthermore, earlier parameterization approaches are investigated to determine if the use of different cut-off schemes and time saving techniques leads to inconsistent results. Repeating the experiments used for the parameterization showed that the use of a twin-range cut-off scheme does not result in a significant effect on the properties investigated in the simulation. Based on these results the effects of different cut-off schemes on protein simulations were statistically analysed and the variation observed in from them could be traced to different effective temperatures in the systems. A more stringent thermostat setting, or the use of a mixed cut-off scheme could help overcome these issues. This work provides valuable insights and parameters to push the accuracy of molecular dynamics simulations even further.

Zusammenfassung

Molekulardynamische Simulationen sind ein sehr nützliches Werkzeug, um biologische Systeme zu untersuchen und unser Verständnis von dynamischen Prozessen auf molekularer Ebene zu vertiefen. Um zuverlässige Ergebnisse von Simulationen zu erhalten sind genaue Kraftfeldparameter eine Grundvoraussetzung. In dieser Arbeit wird ein neues Parameterset für das Proteinrückgrat vorgestellt, welches auf Grundlage von kleinen chemisch geblockten Di-Aminosäuren parametrisiert wurde und mit einem extensiven Set, bestehend aus 52 Proteinen, validiert wurde. Weiters wurden frühere Parameter untersucht, ob die Verwendung von verschiedenen Cut-off-Schemata und zeitsparende Algorithmen einen signifikanten Einfluss auf die Parametrisierung hat. Nach der Wiederholung der Experimente zeigte sich, dass die Verwendung von zwei geteilten Cut-off Distanzschemata keinen Einfluss auf die untersuchten Größen hat. Basierend auf diesen Resultaten wurden die Effekte verschiedener Cut-off Schemata auf Proteinsimulationen statistisch untersucht. Die Abweichungen konnten auf Unterschiede in der effektiven Temperatur der Systeme zurückgeführt werden. Eine stringentere Temperaturkontrolle oder die Verwendung von gemischten Cut-off Schemata können helfen diese Probleme zu lindern. In dieser Arbeit werden wertvolle Erkenntnisse im Bereich der Cut-off Schemata und ein Parameterset, welches die Genauigkeit von molekulardynamischen Simulationen weiter verbessert, vorgestellt.

Acknowledgements

This thesis marks the end of a chapter in my life, I started as a bachelor student and became a scientist. Many people accompanied me on this journey and without them it would not have been such a great experience. First and foremost I want to thank my supervisor **Chris Oostenbrink** for discovering my talent as a computational chemist and giving me guidance along the way. Indeed it was quite a stretch from the start of my bachelor project in 2013 up to finishing my Phd in 2021, Chris was always there for me and helped me overcome all difficulties. I also want to thank **Christian Margreitter**, who helped me making my first steps in molecular dynamics simulations and introduced me the group back then, who would have guessed that it would last for more than 7 years. In those years many people joined the group and left again, unfortunately this section here is too short to thank all of them individually, but I want to explicitly thank **Drazen Petrov** for his scientific support and for motivating me in challenging phases. I especially want to thank **Bettina Lier** for always organizing group activities and I will remember her kind and inspiring personality. When it comes to "Lieblingswaldviertler" I will always remember **Christoph Öhlknecht** for showing me the world of photography and being calm as a rock when things heat up. One of the greatest experiences was to supervise **Jakob Dong Hua Liu**, it was amazing to see you going all the way from starting the master project to becoming a PhD student yourself. One person without whom this thesis would not have been possible is **Sonja Wit**, our secretary. Without her the bureaucratic challenges would not have been solvable and I also enjoyed our morning coffee breaks. Last but certainly not least I want to thank my family, my parents **Heidemarie** and **Reinhard** and my sister **Magdalena**, who are always there for me and supported me, no matter what.

Contents

List of Publications

This thesis led to the following publications:

Chapter 2

Diem M. and Oostenbrink C.

Hamiltonian Reweighing To Refine Protein Backbone Dihedral Angle Parameters in the GROMOS Force Field.

Accepted by *J. Chem. Inf. Model.* . **2020**; 60, 1, 279–288.

https://doi.org/10.1021/acs.jcim.9b01034

Chapter 3

Diem M. and Oostenbrink C.

The Effect of Using a Twin-Range Cutoff Scheme for Nonbonded Interactions: Implications for Force-Field Parametrization?

Accepted by *J. Chem. Theory Comput.*. **2020**;16, 10, 5985–5990.

https://doi.org/10.1021/acs.jctc.0c00509

Chapter4

Diem M. and Oostenbrink C.

The effect of different cutoff schemes in molecular simulations of proteins

Accepted by *J Comput. Chem.*. **2020**; 41: 2740– 2749.

https://doi.org/10.1002/jcc.26426

Additional publications

T. F. D. Silva, D. Vila-Viçosa, P. B. P. S. Reis, et al.
The Impact of Using Single Atomistic Long-Range Cutoff Schemes with the GROMOS 54A7 Force Field
Accepted by *J. Chem. Theory Comput.*. **2018** 14 (11), 5823-5833
https://doi.org/10.1021/acs.jctc.8b00758

Schwaigerlehner, L., Mayrhofer, P., Diem, M. et al.
Germinality does not necessarily define mAb expression and thermal stability.
Accepted by *Appl Microbiol Biotechnol* 103, 7505–7518 **2019**
https://doi.org/10.1007/s00253-019-09998-3

A. Turupcu, M. Diem, L. J. Smith, C. Oostenbrink
Structural Aspects of the O-glycosylation Linkage in Glycopeptides via MD Simulations and Comparison with NMR Experiments
Accepted by *ChemPhysChem* **2019**, 20, 1527.
https://doi.org/10.1002/cphc.201900079

Yang, S., Diem, M., Liu, J.D.H. et al.
Cellular levels and molecular dynamics simulations of estragole DNA adducts point at inefficient repair resulting from limited distortion of the double-stranded DNA helix
Accepted by *Arch Toxico*l 94, 1349–1365 **2020**.
https://doi.org/10.1007/s00204-020-02695-5

S. Yang, J. D.H. Liu, M. Diem et al.
Molecular Dynamics and In Vitro Quantification of Safrole DNA Adducts Reveal DNA Adduct Persistence Due to Limited DNA Distortion Resulting in Inefficient Repair
Accepted by *Chem. Res. Toxicol.* **2020** 33 (9), 2298-2309
https://doi.org/10.1021/acs.chemrestox.0c00097

Chapter 1

Introduction

Introduction

Mankind was always driven by curiosity, from the discovery of fire to quantum computers. Thousands of years ago scientist like Aristotle, Socrates and Pythagoras founded the basis of modern Science. The amount of knowledge has increased dramatically since then, but the very basic principle of how to generate it never changed. Therefore, scientists across the ages observed nature, drew their conclusion and tested their hypothesis to validate the results. This straightforward workflow of generating knowledge is applicable up to this day. But the sheer amount of knowledge makes it impossible for one person to be a universal genius of all fields of science. So, despite the similarity in the generation of knowledge the different fields of science diverge in recent centuries with arguably Gottfried Wilhelm Leibnitz or Johann Wolfgang Goethe being the last universal geniuses of our time. Nowadays however the scientific community seems to come closer together again, combining different fields and drive scientific progress along those interfaces to deepen the understanding of the world around us [1].

Molecular dynamics simulations

Molecular dynamics (MD) simulations are a field of science, that is located at the interface of chemistry, biology, physics and computer sciences. What makes it difficult to classify, on the other hand makes it fascinating to study. The high resolution and the dynamic matter of the field allow to gain insight in many chemical systems. MD simulations rely on accurate structural representations of the systems simulated, these typically are obtained experimentally by nuclear magnetic resonance spectroscopy or X-ray crystallography, or structure predictions (e.g. homology modelling, alphaFold[2]). Once the positions are obtained random, initial velocities, sampled from a Maxwell Boltzmann distribution, are assigned and the system is propagated through time. This is needed since the experimental or statistical approaches don't contain any information on velocities. So, it comes down to calculating the new positions x using veloc-

ities v at time t, whith a timestep δt as shown in eq 1.1.

$$x(t + \Delta t) = x(t) + v(t + \frac{1}{2}\Delta t) * \Delta t \tag{1.1}$$

$$v(t + \frac{1}{2}\Delta t) = v(t - \frac{1}{2}\Delta t) + a(t) + \Delta t \tag{1.2}$$

The leapfrog algorithm[3] alternates between updating positions and velocities, latter are calculated using the previous velocities and multiplying them with the acceleration a of the time interval data (eq 1.2). This is achieved by applying Newtons second law of motion (eq 1.3) which states that the force F equals mass times acceleration. The force is calculated by the negative partial derivative of the potential energy U. The name leapfrog comes from the fact that one can imagine the propagation through time by imagining two frogs leaping over each other.

$$F = -\frac{\partial U}{\partial x} = m * a \rightarrow a = \frac{F}{m} \tag{1.3}$$

Force field

As mentioned before to calculate the force of a particle it is crucial to know it's potential energy. The potential energy is the sum of all energy contibutions of the system (eq 1.6). It is typically split in two parts, which are bonded and non bonded energies. Equations in this section are written in a typical functional form, in the GROMOS force field a slightly different functional form is used.

$$U^{bonded} = U^{bonds} + U^{angles} + U^{dihedral} + U^{improper} \tag{1.4}$$

$$U^{non-bonded} = U^{vdW} + U^{Coulomb} \tag{1.5}$$

$$U^{total} = U^{bonded} + U^{non-bonded} \tag{1.6}$$

Bonded

Bonded terms are considered all contributions which are propagated along the bonds of the molecule. The most basic are two atoms connected at a defined

distance. The potential energy bond term, as calculated in eq 1.7 , is calculated as the squared difference of the actual bond length d and the ideal bond length d_{type} multiplied by the force constant k halved. Since disturbed bond lengths are engergetically very unfavorable the force arising from deviations in the bond length are very high. The fact that bond vibrations are decoupled from other degrees of freedom in simulation allows for their removal altogether using algorithms such as SHAKE [4] or LINCS [5]. Since bond vibrations are the fastest motions in the system, removing them allows the use of bigger timesteps (2 fs vs. 0.5 fs).

$$V^{bonds} = \sum_{bonds} \frac{1}{2} K_{type(i)} (d_i - d^0_{type(i)})^2 \tag{1.7}$$

The next type of bonded interactions are angles, which are described by three atoms and the embedded angle. The functional form describing the potential energy term for angles is very similar to bonds. The difference between the ideal and the real angle are calculated and squared, but the halfed force constant which it is multiplied with is much lower than in the case of bond the bond terms.

$$V^{angles} = \sum_{angles} \frac{1}{2} K_{type(i)} (\theta_i - \theta^0_{type(i)})^2 \tag{1.8}$$

Four sequential atoms can form a dihedral angle, this interaction term mimics the effect of electron density clouds around the central bond. It describes the energy of the two plains formed by the first and last three atoms of the dihedral which depending of the angle can be more (higher energy) or less (lower energy) favorable. The functional form to calculate the potential energy of a dihedral is depicted in eq 1.9, it uses the cosine of the multiplicity m (a measure of how many maxima and minima the energy landscape has) times the angle φ, and adds a shift parameter δ wich shifts the curve and is multiplied by a force constant k_{type}. To achieve more complex curvatures multiple of these potential can be used on the same four atoms. The implications on the protein

backbone are descibled in the following chapter.

$$V^{torsions} = \sum_{torsion} K_{type(i)}[1 + \cos(m_{type(i)}\varphi_i - \delta_{type(i)})] \tag{1.9}$$

Dihedrals can also be used to mimic the behaviour of π-hybrid orbitals, e.g. in benzene. To achieve a flat ring structure so called improper dihedrals are applied to avoid out of plain bending of atoms. The functional form is equal to those of the normal dihedral angles.

Non bonded

Non-polar van der Waals interactions in simulations account for the transient dipoles in molecules that induce each other and can be either attractive and in close distance due to Pauli's exclusion principle repulsive. In MD simulation the Lennard-Jones (LJ) potential energy term is used to describe this effect. In eq 1.10 there is one term r^{-12} which accounts for the repulsion at very close distance of two atoms and the r^{-6} term accounting for the attraction. Sigma gives the zero energy distance and epsilon the depth of the energy well. The energy is very low and decays very fast, r^{-6}. As opposed to the Coulomb energy term, vdW interactions can also occur between non-polar atoms and the specific terms depends on the combiantion of both atom types.

$$V^{vdW}(r^N) = \sum_{pairs} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.10}$$

The second non-bonded energy term is the coulomb energy term, it is calculated between polar atoms. Since in MD simulations all charges are fixed, the effect of charge-charge attraction and repulsion is accounted for by the Coulomb interaction term. The effect two charged atoms have on each other is described in eq 1.11, q_i and q_j are the two charges, wich lead to an attractive or repulsive interaction, depending on their signs. The product of the two charges is divided by their distance to each other r_{ij} and mulitplied by the inverse of Epsilon zero (di-electric permitivitiy of vaccum) times Epsilon r (relative di-electric permitivity of the medium) times 4 π. These interactions decay with 1/r which is very long range and therefore usually bonded atoms are grouped

in neutral charge-groups since the dipole-dipole interaction decay with r^{-3}. A more detailed discussion can be found in chapters 3 and 4. In the GROMOS force fields, a cutoff is typically used, with a reaction-field contribution added to the Coulomb interaction, to account for the effect of a homogeneous medium outside the cutoff sphere (see eq 1.11 below). R_{RF} and C_{RF} are the parameters of the reaction field [6].

$$V^{Coulomb}(r^N) = \sum_{pairs} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1} \left[\frac{1}{r_{ij}} - \frac{C_{RF} r^2}{2R_{RF}^3} - \frac{1 - \frac{1}{2}C_{RF}}{R_{RF}} \right] \tag{1.11}$$

Cut-off

Since non bonded interactions are very time consuming to calculate, there are several ways to treat them in a MD simulation. One of the possibilities is straight cut-off truncation, where the energies abruptly go to zero at a certain cutoff distance and vacuum is considered outside. There are very few systems, where this is sensible and using periodic boundary conditions it turned out to be very error prone[7, 8, 9, 10, 11]. To avoid the errors from straight cutoff truncation a mean field approach is widely used. Where beyond a certain cutoff distance a homogeneous environment is assumed. To smoothen the transition from the calculated interactions to the homogeneous field shifting functions are used [10, 6, 12]. A more detailed discussion of cutoff schemes and their impact on protein simulations is discussed in chapters 3 and 4.

Lattice summation

Another way to treat the non bonded interactions are Lattice summation approaches, where an infinite pseudo crystal is assumed. These schemes are based on Ewalds work on crystals [13]. The point charges are convoluted using charge shaping functions and the interactions split into a real and a reciprocal space contribution. The real space contribution term has a smaller range then the cutoff so it can by calculated by summing up the pairs. The reciprocal space contribution is long ranged and periodic, it's smoothness lets it converge

very quickly. For particle-particle particle-mesh (P3M) [14] and particle-mesh Ewald (PME) [15, 16] grid based fast Fourier transformation is used to sum up the reciprocal space contributions.

Parameterization

Finding precise and accurate parameters for biomolecular systems, one has to consider all of the aspects mentioned above, because the better the model, the more accurate the predictions it can make.Typically biological systems are split up into smaller subsystems to parameterize, this allows for a more accurate parameter search and tranferability between similar chemical groups. One way to obtain reliable parameters is by using ab-initio QM calculations and fitting the parameters iteratively to those e.g. dihedral torsion profiles. This approach is universally applicable but unfortionately lacks information from the surrounding of the system. Therefore GROMOS force fields use experimental data, e.g. J-values, hydration free energies,… , to parameterize against [17]. The focus in this work is on dihedral angles, which play a crucial role in simulations of biomolecular systems since they are important parameters for the structure of macro molecules. J-coupling values obtained from NMR experments can be used to parameterize the ϕ dihedral angle. Using the Karplus relation [18] eq. 1.12 the dihedral angle can be converted into J coupling constants. Despite the known inaccuracy of 1-2 Hz [19] of this conversion allows a very detailed and direct insight into the protein structure.

$$\mathrm{J}(\theta) = \mathrm{A} * \cos^2\theta + \mathrm{B} * \cos\theta + \mathrm{C} \tag{1.12}$$

Since this type of experiments is not feasable for the ψ dihedral angle, a more indirect approach was used here and secondary strucure propensities obained for Raman- and infrared spectrocopy were fitted via a Ramachandran plot to the distribution obtained in simualtions. These procedures are described in more detail in chapter 2.

A brief history of GROMOS force fields

Force fields and parameterization have come a long way from the first simulations of a protein by ammon et al. [20] in 1977 up to this day. The first GROMOS force field (26C1) [21] dates back to 1981, it consisted of 26 atom types used in amino acids and a heme group. Two years later in 1983 eleven new atom types ware released (37C2) alongside new 1-3 neighbouring Lennard-Jones interactions (37C4). Finally after some refinement in 1987 the GROMOS87 force field was released.

This version lasted for almost 10 years till GROMOS96 [22] in 1996 wich was a major revision of the force field. In this version the dihedral back bone parameter were redefined to make it more rigid. Furthermore additional atom types were introduced and the atom mass types were seperated from the LJ types. In the subsequent version the LJ parameters of the aliphatic atom types were reparameterized to fit enthalpy of vaporisation. Furthermore the C12 parameter of the SPC water was adapted for the interactions with non polar atoms (43A1). In 2000 new dihedral angle parameters for the linear alkanes were introduced to represent accurate trans – gauche ratios (43A2) [23]. In subsequent force field versions, 45A3 [24] and 45A4 [25, 26] the LJ parameters were again updated to fit experimental desity also for cyclic, branched and longer n-alkanes. Version 45A4 also brought an update to the DNA and RNA parameters since the were too floppy before and Watson-Crick hydrogen bonds were underrepresented. The parameterization on thermodynamic properities was continued in the 53A5 and 53A6 [27] version of the force field. There a set of pure liquids and small organic compounds was used fiting the densities, heat of vaporization. Furthermore the the solvation free energies of amino acid side chain analogues in cyclohexane and in water (53A6) were used for fitting the partial charges. The last major updates to the GROMOS force field date back to 2011 and 2012, where dihedral angle potential energy terms were reparameterized to a set of high resolution crystal structures and the LJ parameters of Na^+ and Cl^- were updated to fit experimental free energy of hydration data (54A7) [28]. In the most recent version of the GROMOS froce field, 54A8 [29, 30], the charged amino acid side chains were reparameterized and solvation free energies of

polyatomic ions were matched to experiments.

Proteins

One of the most studied biomolecular systems these day in academia and industry are proteins. They play a very big role in everyday life of living organisms. Because of this a precise representation of these macromolecules is essential to every molecular dynamics force field. Proteins consist of single building blocks, amino acids, sequentially connected and forming structural elements. Combining these sturctures to macro structures of many proteins allows them to fullfill chemical and physical tasks. So it is essential to have an precise description and accurate force field parameters to simulate their behaviour. In this thesis an extensive set of 52 different proteins from different species is used to validate new parameter sets and to observe the influence of certain simulation settings. The set described by Setz et al.[31, 32] contains 39 structures obtained from X-ray crystallography and 13 obtained for NMR experiments. The size ranges from less than 20 amino acids to several hundred and the protein properties are vastly different. This diversity allows to avoid biased results and allows insights into a broad range of proteins at the same time with statistical methods.

Aims

The overall aim of this thesis is to make molecular dynamics simulations a more accurate and robust tool to understand the dynamics and structure of biomolecules. For this we set out to improve specific force field parameters and to test the effect of simulation settings during the parameterization.

In previous work, it became clear that the GROMOS parameters for the amino-acid backbone dihedral angles do not represent the structural preferences of the smallest peptides very well. As parameterizations of the GROMOS force field rely on reproducing experimental data for small molecules, we aimed to expand this philosophy to the backbone parameters as well. The aim is to define a set of parameters that is appropriate for both folded and unfolded pro-

teins, such that the folding behaviour can be captured and an appropriate force field for intrinsically disordered proteins is obtained.

The choice of simulation settings during the parameterization procedure is highly relevant for the applicability domain of force fields. Recently, some doubts have been voiced about the appropriateness of the use of the twin-range cutoff scheme in GROMOS parameterizations. A second aim of the thesis is to explore the effect of different choices for the treatment of electrostatic interactions on the parameterization of the force field and on the behaviour of proteins in molecular simulations.

For both aims, it is very important to quantify the statistical relevance of any changes observed between two sets of simulations. As molecular dynamics is a stochastic process, it is important to distinguish differences in ensemble averages that are the result of a true effect from those that can be expected from natural fluctuations. For this reason, a large set of protein structures was used, which were simulated in triplicates. The statistical framework to determine significant differences was an important aspect of the research in this thesis.

Outline

For the backbone dihedral angle parameters it was shown that the set used in current versions of the GROMOS force field fails to represent accurate dynamics of small peptides and even on the protein level, some secondary structure propensities are over or under represented. In chapter 2, mathematical optimization techniques and extensive validation will be used to obtain a more reliable set of protein backbone dihedral angle parameters.

In chapter 3 the effects of using different cutoff types and distances on the parameterization of the GROMOS force field are investigated. Since recently concerns were voiced upon the validity of the parameterization of the force fields. Therefore all simulations used in the validation of the parameterization will be repeated and a sound statistical analysis will be performed.

Taking up on chapter 3, in chapter 4 the effects of theses cutoff schemes in protein systems will be investigated and compared. Therefore a large set of 52 proteins will be used and extensive structural analysis will be performed.

References

[1] B. Russell. *A History of Western Philosophy - And Its Connection with Political and Social Circumstances from the Earliest Times to the Present Day*. Simon and Schuster, 1964.

[2] J. Jumper et al. *High Accuracy Protein Structure Prediction Using Deep Learning*. Accessed: 2021-01-06. 2021. URL: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

[3] R.W. Hockney and J.W. Eastwood. *Computer simulation using particles*. 1988.

[4] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes". In: *J. Comp. Phys.* 23.3 (Mar. 1977), pp. 327–341. DOI: 10.1016/0021-9991(77)90098-5.

[5] Berk Hess et al. "LINCS: A linear constraint solver for molecular simulations". In: *Journal of Computational Chemistry* 18.12 (Sept. 1997), pp. 1463–1472.

[6] I. G. Tironi et al. "A generalized reaction field method for molecular dynamics simulations". In: *J. Chem. Phys.* 102.13 (Apr. 1995), pp. 5451–5459. DOI: 10.1063/1.469273.

[7] M. M. Reif et al. "Molecular dynamics simulations of a reversibly folding β-heptapeptide in methanol: Influence of the treatment of long-range electrostatic interactions". In: *J. Phys. Chem. B* 113.10 (Mar. 2009), pp. 3112–3128. DOI: 10.1021/jp807421a.

[8] H. Schreiber and O. Steinhauser. "Cutoff Size Does Strongly Influence Molecular Dynamics Results on Solvated Polypeptides". In: *Biochem.* 31.25 (Feb. 1992), pp. 5856–5860. DOI: 10.1021/bi00140a022.

[9] E. Spohr. "Effect of electrostatic boundary conditions and system size on the interfacial properties of water and aqueous solutions". In: *J. Chem. Phys.* 107.16 (Oct. 1997), pp. 6342–6348. DOI: 10.1063/1.474295.

[10] P. H. Hünenberger and W. F. Van Gunsteren. “Alternative schemes for the inclusion of a reaction-field correction into molecular dynamics simulations: Influence on the simulated energetic, structural, and dielectric properties of liquid water”. In: *J. Chem. Phys.* 108.15 (Apr. 1998), pp. 6117–6134. DOI: 10.1063/1.476022.

[11] C. L. Brooks. “The influence of long-range force truncation on the thermodynamics of aqueous ionic solutions”. In: *J. Chem. Phys.* 86.9 (May 1987), pp. 5156–5162. DOI: 10.1063/1.452636.

[12] J. A. Barker and R. O. Watts. “Monte carlo studies of the dielectric properties of water-like models”. In: *Mol. Phys.* 26.3 (1973), pp. 789–792. DOI: 10.1080/00268977300102101.

[13] P. P. Ewald. “Die Berechnung optischer und elektrostatischer Gitterpotentiale”. In: *Ann. Phys.* 369.3 (1921), pp. 253–287. DOI: https://doi.org/10.1002/andp.19213690304.

[14] J. W. Eastwood, R. W. Hockney, and D. N. Lawrence. “P3M3DP-The three-dimensional periodic particle-particle/ particle-mesh program”. In: *Compu. Phys. Comm.* 19.2 (Apr. 1980), pp. 215–261. ISSN: 00104655. DOI: 10.1016/0010-4655(80)90052-1.

[15] T. Darden, D. York, and L. Pedersen. “Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems”. In: *J. Chem. Phys.* 98.12 (June 1993), pp. 10089–10092. DOI: 10.1063/1.464397.

[16] U. Essmann et al. “A smooth particle mesh Ewald method”. In: *J. Chem. Phys.* 103.19 (Nov. 1995), pp. 8577–8593. DOI: 10.1063/1.470117.

[17] W. F. Van Gunsteren et al. “Biomolecular modeling: Goals, problems, perspectives”. In: *Angew. - Int. Ed.* 45.25 (2006), pp. 4064–4092. DOI: 10.1002/anie.200502655.

[18] M. Karplus. “Contact electron-spin coupling of nuclear magnetic moments”. In: *J. Chem. Phys.* 30.1 (Jan. 1959), pp. 11–15. DOI: 10.1063/1.1729860.

[19] W. F. Van Gunsteren et al. “Validation of Molecular Simulation: An Overview of Issues”. In: *Angew. - Int. Ed.* 57.4 (Jan. 2018), pp. 884–902. DOI: 10.1002/anie.201702945.

[20] J. A. McCammon, B. R. Gelin, and M. Karplus. “Dynamics of folded proteins”. In: *Nature* 267.5612 (1977), pp. 585–590. DOI: 10.1038/267585a0.

[21] *Biomolecular Simulation - The GROMOS Software*. http://gromos.net/. Accessed: 2021-01-06.

[22] Xavier Daura et al. “On the sensitivity of MD trajectories to changes in water-protein interaction parameters: The potato carboxypeptidase inhibitor in water as a test case for the GROMOS force field”. In: *Proteins: Structure, Function, and Bioinformatics* 25.1 (1996), pp. 89–103. DOI: https://doi.org/10.1002/1097-0134(199605)25:1.

[23] X. Daura, A. E. Mark, and W. F. Van Gunsteren. “Parametrization of aliphatic CHn united atoms of GROMOS96 force field”. In: *J. Comp. Chem.* 19.5 (Apr. 1998), pp. 535–547.

[24] L. D. Schuler, X. Daura, and W. F. Van Gunsteren. “An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase”. In: *J. Comp. Chem.* 22.11 (Aug. 2001), pp. 1205–1218. DOI: 10.1002/jcc.1078.

[25] T. A. Soares et al. “An improved nucleic acid parameter set for the GROMOS force field”. In: *J. Comp. Chem.* 26.7 (May 2005), pp. 725–737. DOI: 10.1002/jcc.20193. URL: http://doi.wiley.com/10.1002/jcc.20193.

[26] R. D. Lins and P. H. Hünenberger. “A new GROMOS force field for hexopyranose-based carbohydrates”. In: *J. Comp. Chem.* 26.13 (Oct. 2005), pp. 1400–1412. DOI: 10.1002/jcc.20275.

[27] C. Oostenbrink et al. “A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6”. In: *J. Comp. Chem.* 25.13 (Oct. 2004), pp. 1656–1676. DOI: 10.1002/jcc.20090.

[28] N. Schmid et al. “Definition and testing of the GROMOS force-field versions 54A7 and 54B7”. In: *Eur. Biophys. J.* 40.7 (July 2011), pp. 843–856. DOI: 10.1007/s00249-011-0700-9.

[29] M. M. Reif, P. H. Hünenberger, and C. Oostenbrink. “New interaction parameters for charged amino acid side chains in the GROMOS force field”. In: *J. Chem. Theo. Comp.* 8.10 (Oct. 2012), pp. 3705–3723. DOI: 10.1021/ct300156h.

[30] M. M. Reif, M. Winger, and C. Oostenbrink. “Testing of the GROMOS force-field parameter set 54A8: Structural properties of electrolyte solutions, lipid bilayers, and proteins”. In: *J. Chem. Theo. Comp.* 9.2 (Feb. 2013), pp. 1247–1264. DOI: 10.1021/ct300874c.

[31] Setz M. *Molecular dynamics simulations of biomolecules: from validation to application*. Vienna: Dissertation, BOKU, 2018, p. 285. DOI: https://permalink.obvsg.at/AC15159105.

[32] M. Stroet et al. “Challenges associated with the validation of protein force fields based on structural criteria.” In: *submitted to J Chem Inf Model* (2020).

Chapter 2

Hamiltonian Reweighing To Refine Protein Backbone Dihedral Angle Parameters in the GROMOS Force Field

This work was previously published in

Diem M. and Oostenbrink C.

Hamiltonian Reweighing To Refine Protein Backbone Dihedral Angle Parameters in the GROMOS Force Field.

*J Chem. Model. Inf.. **2020**; 60 (1), 279-288*

https://doi.org/10.1021/acs.jcim.9b01034

Abstract

Molecular dynamics simulations of proteins depend critically on the underlying force field, which may be parameterized against experimental data or high-quality quantum calculations. Here, we develop search algorithms, based on Monte Carlo and steepest descent calculations to optimize the backbone dihedral angle parameters from a single reference simulation. We apply these tools to improve the agreement between simulations of single, capped amino acids and experimentaly determined J-values and secondary structure propensities of these molecules. The parameters are further refined based on simulations of a set of seven proteins and finally validated in simulations on a large set of 52 protein structures. Improvements in the dihedral angle distributions are observed and structural propensities of the proteins are reproduced very well. Overall, the GROMOS 54A8_bb parameter set forms a improvement to previous parameter sets, both for small molecules and for protein simulations.

Introduction

In order to describe the structure and dynamics of proteins using computational methods, it is essential to accurately capture the interactions within the protein backbone. Here, we aim to enhance the agreement of small peptide simulations with experimentally obtained quantities, while at the same time maintain structural stability of the protein dynamics, when using the GROMOS force field. In the GROMOS force field the backbone dihedral energy terms were typically parameterized to be in agreement with quantum-mechanical calculations or experimental observations [1, 2, 3]. The most recent update of the official GROMOS force field dihedral angle parameters was described in parameter set 54A7 [4]. The strategy was to fit the ϕ and ψ angle distributions as obtained in simulations to distributions of a set of high resolution crystal structures. In the most recent parameter set 54A8, the non-bonded interaction terms of the charged side chains were reparameterized [5]. Other molecular dynamics force fields, like CHARMM and AMBER introduced the CMAP energy term to improve their peptide dihedral angle sampling. The CMAP potential energy term con-

tribution is a numerical term based on quantum mechanical calculations, of dipeptides, which are added to the other contributing potential energy terms [6, 7, 8]. Very recently an updated set of backbone CMAP parameters was reported for the AMBER force field (ff19SB). These were parameterized against quantum mechanical calculations for the individual amino acids [**Tian2019**]. We have recently shown, that differences between force fields can only be assessed simulating a large number of diverse proteins [9, 10]. In the context of this work an evaluation of various GROMOS parameter sets was performed and some discrepancies between dihedral angle distributions observed in molecular simulations and those derived from high quality crystallographic or solution protein structures were identified. We have recently described alternative peptide backbone dihedral angle parameters for use in the GROMOS force field [11, 12]. In this work a brute force approach was used, reweighing ensemble averages from reference simulations to a large number ($\sim 10^5$) of alternative parameter sets and selecting those parameters that reproduce experimental data best. Here, we refine this approach in several points. First, while the parameterization is primarily performed against experimental data on the capped amino acids, a validation cycle is inserted by assessing the behaviour of new parameters in simulations of seven protein structures. The final parameter set is validated against simulations of a large set of 52 proteins [9, 10]. Second, the parameterization procedure is more focused, using a Monte Carlo and Steepest Descent approach in parameter space, rather than a brute-force generation of many potential parameter sets. Third, an additional degree of freedom is used, allowing phase shifts to take any value in the range [-180°,180°]. In previous GROMOS force fields, the phase shift of torsional potential energy terms were restricted to values of 0° or 180°. Finally, the current set of parameters was derived in the context of the 54A8 parameter set, ensuring consistency with the most recent GROMOS force field.

Methods

In a force field, the change in potential energy upon rotation of a dihedral angle is determined by the non-bonded interactions of the substituents connected

to the rotating bond. As charges and Lennard-Jones parameters are fixed in a classical force field, an additional potential energy term is added to refine the potential energy profile along the dihedral angle. Equation 2.1 is used to calculate the potential energy of a dihedral angle, ϕ, contributing to the sum of all potential energies. This equation allows for three tunable parameters for each potential energy term: the force constant k^{Φ}, the multiplicity m^{Φ} and the phase shift ϕ^{0}. In current parameterization efforts, we allow the force constant k^{Φ} to take any value from 0.00 kJ/mol to 5.00 kJ/mol, the multiplicity m^{Φ} to take a value of 1,2,3 or 6 and the phase shift ϕ^{0} to take any value -180° to +180°. Furthermore, a sum of maximally two potential energy terms for each dihedral angle was used to model the final shape of the potential energy curve.

$$V^{\Phi}_{dihed} = k^{\Phi}[1 + \cos(m^{\Phi}\phi - \phi^{0})] \tag{2.1}$$

Small peptides

All single amino acids with an acetylated N-terminus and an N-methylated C-terminus, as proposed by Avbelj et al. [13], were simulated for 200 ns using the GROMOS11[14] software package and the GROMOS 54A8 force field[15]. All sidechains were described as appropriate for a pH of 7: Arg and Lys were protonated, Asp and Glu deprotonated. For His the neutral form, protonated at N_{δ} was taken. The small peptides were solvated using the SPC water model [16] and a minimum distance to the box edge of 1.4 nm. To equilibrate the systems a five step protocol was used, where the temperature was elevated at constant volume from 60 K to 300 K and at every step, a simulation time of 20 ps was applied. The production simulation was performed at 300 K and 1 atm pressure, using a weak coupling thermostat with two baths, one for the solvent and one for the solute, with a relaxation time of 0.1 ps and a weak coupling barostat with isothermal compressibility of $4.575 * 10^{-4}$(kJ mol^{-1}nm^{-3})$^{-1}$ and a relaxation time of 0.5 ps [17]. In order to maintain the bond distances at the energy minimum the SHAKE algorithm [18] was applied. The simulation timestep used was 2 fs, for the non-bonded interactions a triple range scheme was used with short range cutoff at 0.8 nm and an intermediate range cutoff at 1.4 nm, the interactions for the inner cutoff range were computed every step and the pairlist

and interactions for the intermediate range were updated every five step s[19]. Outside the cutoff a reaction field contribution was calculated assuming a homogeneous medium with a relative dielectric constant of 61, as apropriate for the SPC model [20].

Comparison to experimental quantities

In order to compare simulated properties with experimental values the timeseries of ϕ and ψ dihedral angles was extracted using the GROMOS++ program tser. Furthermore the timeseries of J-values was calculated using the Karplus relation with the parameter set by Pardi et al [21]. It should be pointed out that the Karplus relation leads to an uncertainty of 1-2 Hz [22]. These J-values were compared to the experimentally measured J coupling constants obtained by Avbelj [13]. Additionally the time series of ϕ and ψ dihedral angles were used to classify the molecular conformations in α-helix, β-sheet and poly-proline II ($\mathrm{PP_{II}}$) region, according to figure S1 in the supporting information, using the DISICL program [23]. The occurence of these conformations can be compared to the experimental IR- and Raman spectroscopie data by Grdadolnik et al. [24]. In general, the experimental quantities measured represent an ensemble average and are furthermore underlying an uncertainty themself.

Hamiltonian reweighing, Monte Carlo and steepest descent

Since simulating all potentially relevant parameter combinations is not feasible, Hamiltonian reweighing [25, 26], equation 2.2 was used to predict an ensemble average of a certain quantity Q, by using the timeseries of the Hamiltonian and the timeseries of that quantity Q [27, 11].

$$< \mathrm{Q} >_{\mathrm{mod}} = \frac{< \mathrm{Q} e^{-\frac{\Delta V}{k_B T}} >_{54A8}}{< e^{-\frac{\Delta V}{k_B T}} >_{54A8}} \tag{2.2}$$

Here a ΔV is the change in potential energy, due to modification in the force field, $\mathrm{k_B}$ is the Boltzmann constant, T the absolute temperature and the angular brackets indicate an ensemble average. This equation can be applied directly to the instantaneous J-values, yielding <J>$^{\mathrm{l}}_{\mathrm{mod}}$ for amino acid l, or to the

occurence of secondary structures, yielding the propensites P^l_α , P^l_β and $P^l_{P_{II}}$. The ensemble averages for a specific set of force field parameters can be scored using equation 2.3, by summing the absolute deviation to the experimental data, where the weights were set to w_J = 1 Hz^{-1} and w_α = w_β = $w_{P_{II}}$ = 1:

$$D_{tot} = \sum_l d_l = \sum_l (w_J \cdot | < J >^l - J^{l0}| + w_\alpha \cdot |P^l_\alpha - P^{l0}_\alpha| + w_\beta \cdot |P^l_\beta - P^{l0}_\beta| + w_{P_{II}} \cdot |P^l_{P_{II}} - P^{l0}_{P_{II}}|) \tag{2.3}$$

To sample the relevant regions in the parameter space, Hamiltonian reweighing was combined with a Monte Carlo search scheme. At every step one of the twelve input parameters (two times two sets of k, m and ϕ^0/ψ^0) was randomly modified and the appropriate ensemble averages of the J-coupling value, and the secondary structure propensities were predicted for all amino acids. A Metropolis acceptance critereon was used to avoid getting stuck in local minima, using for the acceptance probability eq 2.4. To allow for a broad sampling, while finding solutions in better agreement with the experimental target values the value of β used in the Metropolis acceptance criterion is set to N_l/10 with N_l being the number of amino acids.

$$P_{acc} = \min(1, \exp^{-\Delta D_{tot} \cdot \beta}) \tag{2.4}$$

Furthermore a steepest descent optimization algorithm was used to further optimize preselected parameter combinations. The optimization was simultaneously carried out for the force contants and the phase shifts, but not the multiplicities. Equation 2.5 shows the target function used to optimize the parameter set.

$$I = \sum_l [w_J \cdot (< J^l > - J^{l0})^2 + w_\alpha \cdot (P^l_\alpha - P^{l0}_\alpha)^2 + w_\beta \cdot (P^l_\beta - P^{l0}_\beta)^2 + w_{P_{II}} \cdot (P^l_{P_{II}} - P^{l0}_{P_{II}})^2] \tag{2.5}$$

In equation 2.6 the derivative of the target function with respect to the first

term of the force constant is shown.

$$\frac{\partial I}{\partial k_1} = \sum_l \Big[2 \cdot w_J \cdot (< J^l > - J^{l0})^2 \cdot \frac{\partial < J^l >}{\partial k_1} + 2 \cdot w_\alpha \cdot (P_\alpha^l - P_\alpha^{l0})^2 \cdot \frac{\partial P_\alpha^l}{\partial k_1} + 2 \cdot w_\beta \cdot (P_\beta^l - P_\beta^{l0})^2 \cdot \frac{\partial P_\beta^l}{\partial k_1} + 2 \cdot w_{P_{II}} \cdot (P_{P_{II}}^l - P_{P_{II}}^{l0})^2 \cdot \frac{\partial P_{P_{II}}^l}{\partial k_1} \Big] \tag{2.6}$$

With similar terms for all force constants and all phase shifts in the parameter set. As an example we write the individual derivative of the J-value with respect to the first force constant (eq. 2.7), by taking the derivative of equation 2.2 with the derivatives $\frac{\partial \Delta V}{\partial k_i}$ being readily available from equation 2.1 and the time series of the dihedral angles.

$$\begin{aligned} \frac{\partial \langle J^l \rangle_{mod}}{\partial k_1} &= \frac{\partial}{\partial k_1} \cdot \frac{\langle J e^{-\frac{\Delta V}{k_B T}} \rangle}{\langle e^{-\frac{\Delta V}{k_B T}} \rangle_{54A8}} \\ &= \frac{\langle -\frac{1}{k_B T} \cdot \frac{\partial \Delta V}{\partial k_1} \cdot J e^{-\frac{\Delta V}{k_B T}} \rangle_{54A8}}{\langle e^{-\frac{\Delta V}{k_B T}} \rangle_{54A8}} - \\ &\quad \frac{\langle J e^{-\frac{\Delta V}{k_B T}} \rangle_{54A8}}{\langle e^{-\frac{\Delta V}{k_B T}} \rangle_{54A8}} \cdot \frac{\langle -\frac{1}{k_B T} \cdot \frac{\partial \Delta V}{\partial k_1} \cdot e^{-\frac{\Delta V}{k_B T}} \rangle_{54A8}}{\langle e^{-\frac{\Delta V}{k_B T}} \rangle_{54A8}} \\ &= -\frac{1}{k_B T} \cdot \left[\langle \frac{\partial \Delta V}{\partial k_1} J \rangle_{mod} - \langle J \rangle_{mod} \cdot \langle \frac{\partial \Delta V}{\partial k_1} \rangle_{mod} \right] \end{aligned} \tag{2.7}$$

Using equations 2.5-2.7, a steepest descent optimization of *I* could be implemented,[28] using a stepsize of 0.001 kJ/mol for the force constants and of 0.001 degrees for the phase shifts.

Glycine, alanine, proline

For glycine a different approach was used to optimize the backbone dihedral angle potential energy term [11]. Since glycine allows, due to its small size, for a high amount of flexibility, the experimental data of Avbelj on the propensity for α, β or PP_{II} conformations is not complete [29, 13]. The potential energy terms should allow for a relatively homogeneous distribution of ϕ and ψ angles. Therefore the time resolved Ramachandran plot of the single amino acid simulations was split in a grid of 20 times 20 degrees and the time series of occurence in these bins was monitored. The occurence of conformations ac-

cording to these bins was reweighed alongside, using equation 2.2 to obtain a prediction for the glycine Ramachandran plot. The root-mean-square fluctuations over the bins were minimized as a target function in order to find the parameter set corresponding to the most evenly distributed Ramachandran plot.

Similar to glycine, the side chain of alanine is very small, a single methyl group. Following our previous work [11] and in agreement with preliminary calculations, it appears necessary to optimize the backbone dihedral angles for this amino acid seperately from the remaining amino acids, here referred to as common amino acids.

Proline is due to its structure a conformationally very restricted amino acid. As a result of the side chain and the backbone forming a five-membered ring, it lacks the flexibilty around the ϕ backbone angle. Therefore and for the lack of experimental values available, we decided not to include proline in the optimization process, but to apply the obtained common amino acid parameters in subsequent simulations.

Validation of parameters in proteins

As will be outlined in the results section, the Monte Carlo parameter search on the single amino acid backbone dihedral angle parameters still leads to a large number of possible parameter sets. For selected parameter sets, simulations were performed on seven proteins, of which the experimental structure was determined by NMR. These proteins are indicated by a dagger in table 2.1.

The final parameter set was subsequently validated using a set of 52 proteins (table 2.1, described by Setz et al.[9, 10]). The set consists of 13 NMR and 39 X-ray diffraction structures. Each of the 52 proteins was simulated for 15 ns in 3 replicates. The GROMOS11 software was used to simulate the different systems, the settings were the same as for the single amino acid systems, except that 0.15 M of NaCl was added to the simulation box. The results were compared to the most recent versions of the GROMOS force field, namely 53A6 and 54A8. To assess the stability of the protein simulations the root-mean-square deviation (RMSD) was calculated and corrected for the protein size, following Carugo and Pongor [30]. This means, that the protein backbone is normalized to a size of 100 residues. The root-mean-square deviation of pro-

X-ray (39)	1A19 1AKI, 1AMM, 1EW4, 1FAZ, 1FL0, 1MJC, 1NG6, 1PGB, 1QK8, 1SHG, 1T2I, 1TUA, 1TVQ, 1UBI, 1UCS, 1ULR, 1UXZ, 1YU5, 1ZLM, 1ZVG, 2CWR, 2GKT, 2J8B, 2NLS, 2PND, 2PNE, 2PPO, 2PTH, 2RB8, 2WLW, 2YXF, 3CA7, 3E7U†, 3EYE ,3WP5, 4LFQ†, 4MHP, 4RWU
NMR (13)	1AEY, 1AFI, 1BTA*, 1D3Z*, 1E8L*, 1IT5, 1MVG*, 1QQV*, 2AF8, 2CZN, 2GB1, 2OVN*†, 3CI2*

Table 2.1: PDB-codes of the 52 proteins used to test the backbone dihedral angle parameters. X-ray and NMR indicate the origin of the starting structure, either obtained from NMR-experiments or X-ray crystallography. Those marked using an asterisk refer to the smaller subset, used in initial tests of new dihedral angle parameters. PDB-codes marked with a dagger refer to proteins smaller or equal to 40 residues.

teins smaller than 40 residues was not normalized, since the method was not verified for such small proteins. These proteins are indicated by a dagger in table 2.1. Further the radius of gyration was calculated and compared to the experimental structure. To observe the impact of the different force fields on the proteins secondary structure elements, the DSSP algorithm by Kabsch and Sanders [31] was used. To compare the different proteins, the difference in secondary stucture content compared to the experimentally determined structure was calculated [24, 13]. NOE violations and average J-value deviations were calculated for all protein structures derived from NMR experiments. The backbone dihedral angle distributions were computed and compared to the distributions derived from the experimental structures.

To asses the significance of the impact of the force field on the simulation metrics (e.g.: RMSD, RGYR, secondary structure elements,..) a multivariate multilevel analysis was conducted as described in Setz et al. [9, 10].

Results and Discussion

Reparameterization Procedure

To optimize the backbone dihedral angle parameters 300000 Monte Carlo steps were performed, once using the experimental values of 17 common amino acids as target function and once to optimize for alanine. Out of these parameter combinations the 1000 best, based on the target function in eq. 3 were preselected. The Monte Carlo scheme yielded many different reasonable solutions

with a similar score. Figure 2.1 shows a heatmap of an overlay of the 1000 ϕ-dihedral angle potential energy terms leading to the best agreement with the experimental data of the common amino acids. In view of the uncertainty of the experimental data, it does not make much sense to select the very best performing parameter set, but rather each of these parameter sets are viable candidates. We decided to select potential energy functions with a low potential energy barrier height, to allow for flexibility in the protein and to avoid forcing the protein in unfavorable conformations. After testing several proposed parameter sets on the subset of 7 proteins, it became apparent that solutions very similar to the 54A8 parameter set performed very well in terms of the normalized ψ dihedral angle frequency as compared to the experimentally observed structures. Therefore we chose to select only ϕ potential energy terms for the common amino acids from the 1000 best ranking Monte Carlo parameter sets. We aimed to enhance occurence in the region from -130° to -90° which corresponds mainly to the β region. For alanine the exact opposite applied, here the normalized frequency of occurrence of the ϕ dihedral angle, as obtained by the 54A8 parameter set accurately matched the distribution determined by experiment, but we identified an artificial shoulder in the normalized frequency of occurrence of the ψ angle. Consequently the parameters for the alanine ψ angle were selected from the Monte Carlo solutions for alanine. Here, the best solutions were obtained when both potential energy terms had a multiplicity of two, allowing us to reduce the parameter set to a single term in ψ for alanine.

The combinations of ϕ and ψ dihedral angle parameters were subsequently refined in the force constants and phase shifts by using the steepest descent method, until D_{tot} was less than 0.75 N_l. We then performed new simulations of 100 ns for each of the amino acids using the optimized parameter sets. Table S2 shows a comparison of the reweighed quantities, using eq. 2 and the corresponding quantities. The values are in very good agreement, indicating overlap in phase-space was reasonable and the application of equation 2 appropriate. Figure 2.2 illustrates the breakdown of d_l of the single amino acids when using the 54A8 parameter set and the newly optimized parameters, from now on referred to as the 54A8_bb set. Figure S3 shows a more detailed breakdown of

the individual contributions to d_ι.

Note that in these figures, for glycine only the J-value is given. For this amino acid the bin reweighing approach was used in order to find an improved potential energy function. A Monte Carlo search was performed to find the 1000 parameter sets which were in closest agreement with the experimentally determined J-values. The bin reweighing approach was applied to all of these parameter sets and the parameter set with the smallest root mean square deviation between the individual bins was selected as 54A8_bb glycine parameter set.

See figure S2 in the supporting information for the Ramachandran plot of glycine using 54A8 and 54A8_bb. For all amino acids in figure 2.2 and figure S3 the agreement with the experimental data has improved. Remaining discrepancies are within the uncertainty of the experimental data or the Karplus curve for J-values. The experimental uncertainty for these kind of measurements is typically very low, around 0.5 Hz. However conversion to the dihedral angles observed in simulations via the Karplus curve to calculated estimates can introduce up to 1-2 Hz of uncertainty [22, 32]. The agreement between the experimental and calculated J-values can be quantified by computing the χ^2 values, which also take the uncertainty of the model into account [**Tian2019**, 33, 34]. Using an estimated value of 0.91 Hz for σ [34], we computed the χ^2 values in table S3 and obtained values that are below 1 for all amino acids, indicating a deviation that is smaller than σ. These values are comparable to the agreement that was recently described by Tian et al. for the Amber ff19SB+OPC force field [**Tian2019**].

Still the J-values only provide direct guidance for the ϕ-dihedral anlge in the backbone of the amino acids. The secondary propensities derived form Raman spectroscopy are furthermore representative of both ϕ and ψ dihedral angles. While the population of helical conformations was already largely within 10% for the 54A8 parameter set, deviations for β and PP_{II} regions were larger and improved throughout the 54A8_bb simulations (figure 3). These observations follow from the fact that the potential energy term of the ϕ-dihedral angle changed most (figure 3).

Overall largest deviations remain for Asn and Asp for which a deviation of

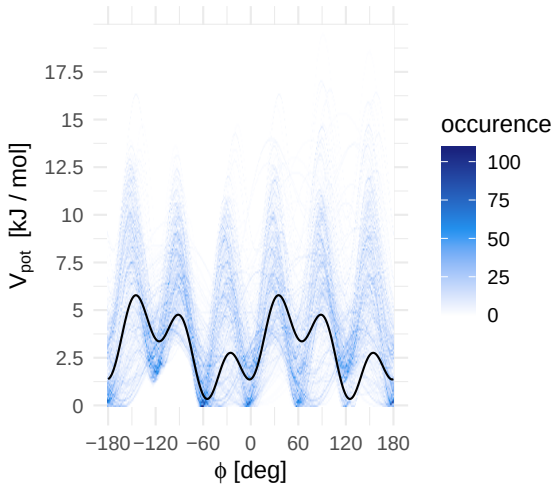

Figure 2.1: Binned distribution of the best 1000 parameter sets for the ϕ dihedral angle of the common amino acids obtained by the Monte Carlo scheme, with an overlay of the final potential energy curve choosen.

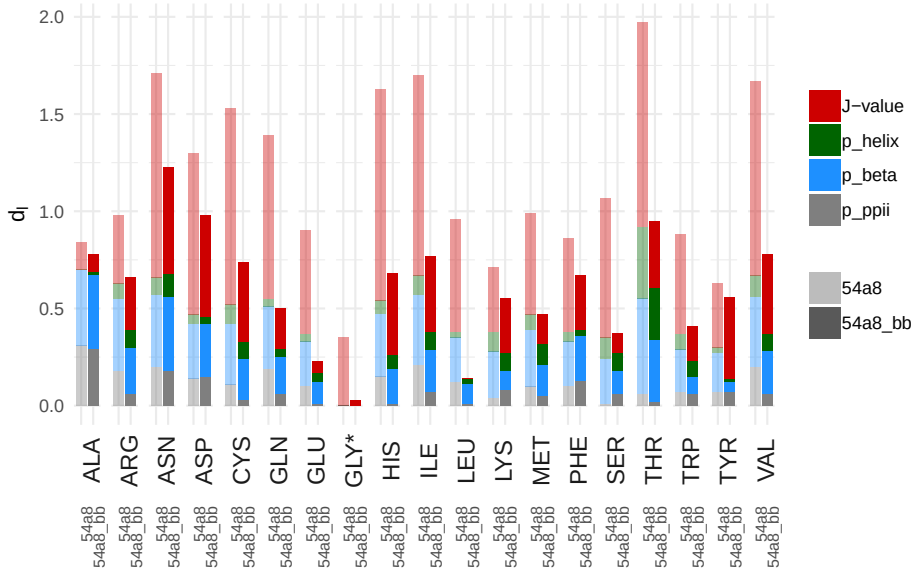

Figure 2.2: Cumulative deviation in the J-value and secondary structure propensities of the blocked amino acids, as computed using equation 2.3. * for glycine only the J-value is compared. See figure S3 in supporting information for a comparison of the deviations for the individual properties.

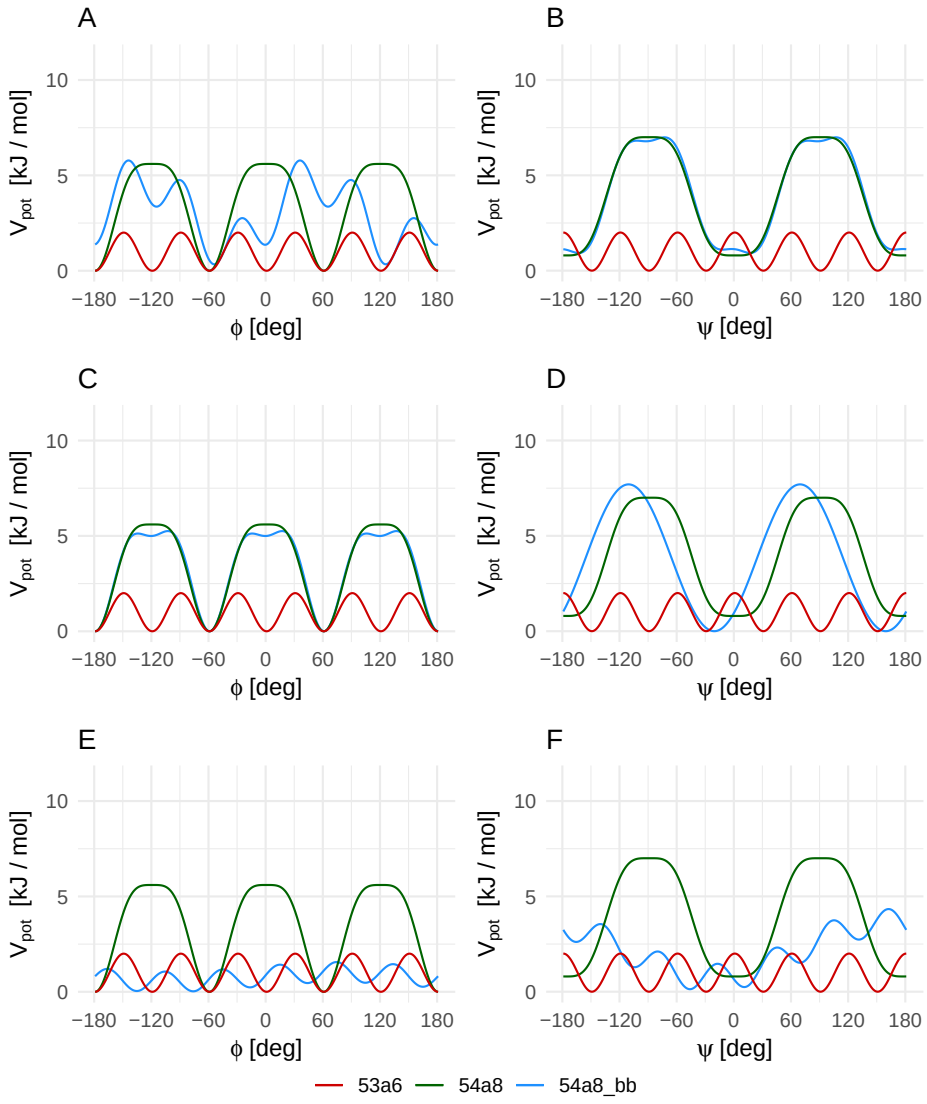

Figure 2.3: Backbone potential energy terms, as used in different GROMOS parameter sets. (A,B): ϕ,ψ potential energy terms used for the common amino acids, (C,D): used for alanine, (E,F): used for glycine

	ϕ			ψ		
	k^{ϕ} $[\frac{kJ}{mol}]$	ϕ^0 [°]	m^{ϕ}	k^{ψ} $[\frac{kJ}{mol}]$	ψ^0 [°]	m^{ψ}
53A6	1.00	180	6	1.00	0	6
54A8	2.80	0	3	3.50	180	2
	0.70	180	6	0.40	0	6
54A8_bb	1.75	104	2	3.40	180	2
common	1.31	-164	6	0.56	0	6
54A8_bb	2.50	2	3	3.85	137	2
ala	0.92	180	6			
54A8_bb	0.25	77	1	0.73	80	6
glycine	0.53	80	6	1.44	-110	1

Table 2.2: ϕ and ψ dihedral angle parameter sets used in different parameter sets.

about 0.5 Hz in the J-value is combined with a relative large underrepresentation of the β-propensity. Strikingly, agreement of the very similar amino acids Gln and Glu is among the best of all amino acids. Note the marked improvement for the β-branched amino acids, Ile, Thr and Val, when going from the 54A8 to the 54A8_bb parameter set. In our previous work a separate set of parameters was proposed for these amino acids [11], but the inclusion of the phase shifts in the optimization procedure has made this separation absolete, possibly at the cost of a remaining overrepresentation of the helical conformation for Thr.

The final 54A8_bb parameter set is described in table 2.2 and figure 2.3. It consists of distinct parameter sets for the common amino acids, alanine and glycine. For proline no corresponding experimental data is available to allow for a seperate optimization, so the same parameters are used as for the common amino acids. The highest barrier in the new parameter set amounts to 7.7 kJ/mol for the ψ dihedral angle of Ala. As for the common ψ and the alanine ϕ potential the resemblance to the 54A8 potential energy is still striking. Note that the potential energy for the ϕ dihedral angle of the common amino acids roughly follows the shape of the 54A8 parameters in the relevant interval [-180°, -30°]. A striking modification, however is the reintroduction of the local energy minimum around 120°, as was previously seen for the 53A6 parameter set. Similarly the glycine potential energy terms seem to be more similar to the 53A6 version of the force field. As is clear from figure S2, the high barrier

in 54A8 severely hampered the sampling of ϕ and ψ in 54A8. Next, we performed an extensive validation process with 52 proteins and three replicates each. While the simulation time per simulation is moderate (15 ns), the set of simulations is far more extensive than any force-field validation sets reported in literature. In Figure 2.4 the probability density functions of occurrence for the ϕ, ψ angles, when using the 54A8 and the 54A8_bb parameter set is depicted. One can observe a more distinct sampling of the β region corresponding to the local minimum in the common ψ potential energy at 120°. Futhermore, the α region peak is slightly more intense than before. In figure 2.5, the one-dimensional distributions of ϕ and ψ are compared to the distributions observed in the experimental structures of our set of proteins. We compare simulations performed with 53A6, 54A8 and 54A8_bb parameters. It can be seen that 54A8_bb simulations strike a balance between the 53A6 and 54A8 parameter sets. The artificial shoulder in the ψ-distribution at 90° that was observed for 53A6 was removed in the 54A8 and 54A8_bb parameter sets. This shoulder led to a strong preference for α-structures in 53A6. On the other hand, the ψ-distribution of 54A8 showed an underrepresentation of the β-region around 120°. The 54A8_bb parameters lead to an increased sampling of this region and match the relative populations of the β ($\psi \sim$ -120°) and combined PP_{II} α ($\psi \sim$ -60°) better. Figure S4 shows the same data, represented as potentials of mean force (PMF) to allow for an energetic comparison of the observed frequencies. We also provide the differences between the PMFs derived from experimental data and from the various simulations. Time series of the root-mean-square deviations with respect to the experimental structure for all simulations using the 54A8 and the 54A8_bb parameter sets are provided in figures S6 to S9. The normalized average values over the last 5 ns are represented as box-plots in figure 6, supplemented with similar plots for the changes in the radius of gyration and for the agreement with NMR observables. For the $RMSD_{100}$ the values are very comparable for the 53A6, 54A8 and 54A8_bb parameter sets with a small improvement for some of the outliers. Only two simulations out of 156 show an RMSD, which is higher than 0.5 nm, when using the 54A8_bb set. The deviation in the radius of gyration is less than 5% for the majority of the proteins in all three parameter sets. For the twelve proteins for which NMR data is available, the average NOE

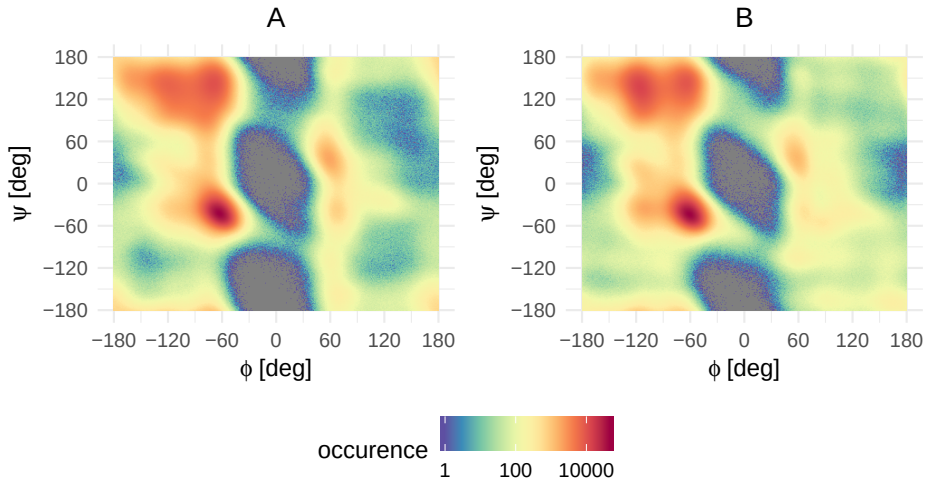

Figure 2.4: Probability density functions of ϕ, ψ, as obtained in protein simulations of the 54A8 parameter set (A) and the 54A8_bb parameter set (B). See figure S2 for the corresponding figures for glycine and figure S5 for pre-proline residues.

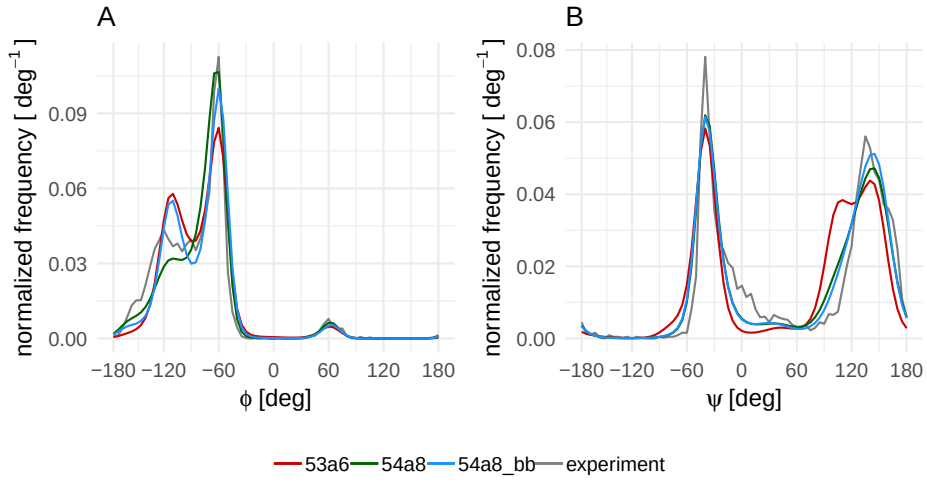

Figure 2.5: Comparison of the normalized frequency of occurence of the ϕ and ψ backbone angles as observed in experimentally observed structures (experiment) and in MD simulations using different parameter sets.

violations pooled over all 3 replicates seem very comparable and also there are no outliers. For the average deviation to the J values the 54A8_bb parameter set performed slightly better than the previous force fields, which is likely the result of the parameterization against the J-values for the single amino acids in solution (figure 2.2). The occurence of secondary structure elements is shown in figure 2.7. The deviations with respect to the experimental structures are very akin across the three force fields with slight improvements for the 54A8_bb force field in all secondary structure elements. In general the deviations are on a rather small scale, only up to $\pm$ 5 % for the vast majority of the proteins. A notable exception is the protein with PDB code 2OVN, which is a small 17 residue peptide, that shows transitions from α- to 3_{10}- and π-helical conformations in all parameter sets. The deviations are large (up to 20 %), falling off the scale of figure 2.7. Note, that in spite of these larger deviations, it does not show exceptionally large NOE violations (average violation of 0.039 nm) or deviations in the J-value (average deviation of 1.3 Hz). Earlier work showed that the NMR data allows for a wider range of conformations than reported in the NMR bundle of structures [35].

The statistical analysis showed that the investigated metrics behave very similar to the 54A8 force field. Except for the β-strand none of the measured quantities shows signifcant differences. P-values on the differences between the measures based on a multivariate multilevel analysis are shown in table S1 in the supporting information.

Conclusion

In the current work, we have described a reparameterization of the backbone dihedral angles in the GROMOS force fields. Following the philosophy of the GROMOS force field, the parameterization was initially performed against experimental data on small molecules, the capped amino acids. They were subsequently refined against simulations of a smaller set of proteins and validated in 156 simulations of 52 distinct proteins. The newly derived dihedral angle parameters fullfil the experimental data for small molecules well, considering the uncertainty in the experimental data and their processing into structural

J-values and secondary structure propensities. The potential energy profiles show similarities to both the 53A6 and the 54A8 parameters sets and accordingly strike a balance between these force fields in simulations of a large set of proteins, improving the overall agreement in the dihedral angle distributions. Secondary structure propensities and structural parameters improved slightly with the newly derived parameter set. One exciting question that remains to be tested is the behaviour of these parameters in simulations of intrinsically disordered proteins. As the side-chains in the GROMOS force fields are parameterized against experimental solvation free-energies and the backbone parameters perform well for individual amino acids, there is no reason to suspect an overly strong preference for (mis)folded states of disordered proteins. This question will be adressed in future work.

Supplementary Material

The supporting information contains the Ramachandran plot used to classify the secodary structure (figure S1), a comparison of glycine in the 54A8 and 54A8_bb parameter set (figure S2), a breakdown of figure 2 in the individual contributions (figure S3). Further more figure 5 plotted as potential of mean force (figure S4), the ramachandran plot of experimental and simultated pre-proline values (figure S5), the comparsion of RMSDs of all 52 protein systems for the 54A8 and 54A8_bb force field (figure S6-S9), the results of the statistical analysis in table S1, the comparison for the reweighted and the simulated properties of the single amino acid systems (table S2) and in table S3 the χ^2 values between the experimental and simulated values.

The supporting information for this chapter can be found under:
https://doi.org/10.1021/acs.jcim.9b01034

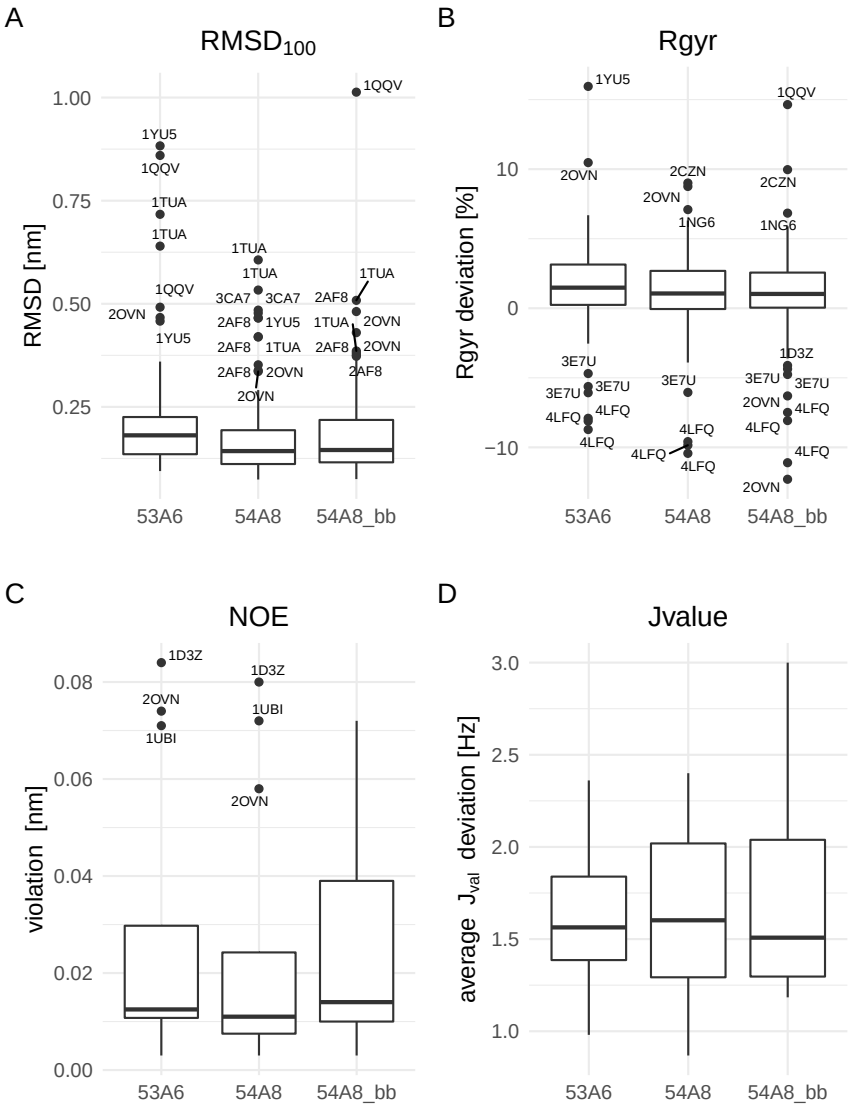
A
RMSD100
RMSD [nm]
1.00
0.75
0.50
0.25
53A6
54A8
54A8_bb
B
Rgyr
Rgyr deviation [%]
10
0
−10
C
NOE
violation [nm]
0.08
0.06
0.04
0.02
0.00
D
Jvalue
average Jval deviation [Hz]
3.0
2.5
2.0
1.5
1.0

Figure 2.6:

Figure 2.6: Boxplots representing the deviation of three independent simulations of every protein with respect to the initial, experimentally derived structure. A) Normalized root-mean-square deviations ($RMSD_{100}$) with respect to the initial structure. B) Deviation in terms of the radius of gyration as percentage change from the initial structure. C) Average NOE violations pooled over all 3 replicates, compared to experimentally determined upper bounds. D) Average deviation in the J-value as compared to experimental values. Solid horizontal lines represent the average values of three independent simulations of 52 (A,B) or 13 (C,D) protein systems. Boxes represent the first and third quartiles, whiskers extend to the extremes of the distribution with a maximum of 1.5 times the inter-quartile range. Individual outlying simulations are marked with dots and labeled with the PDB-code.

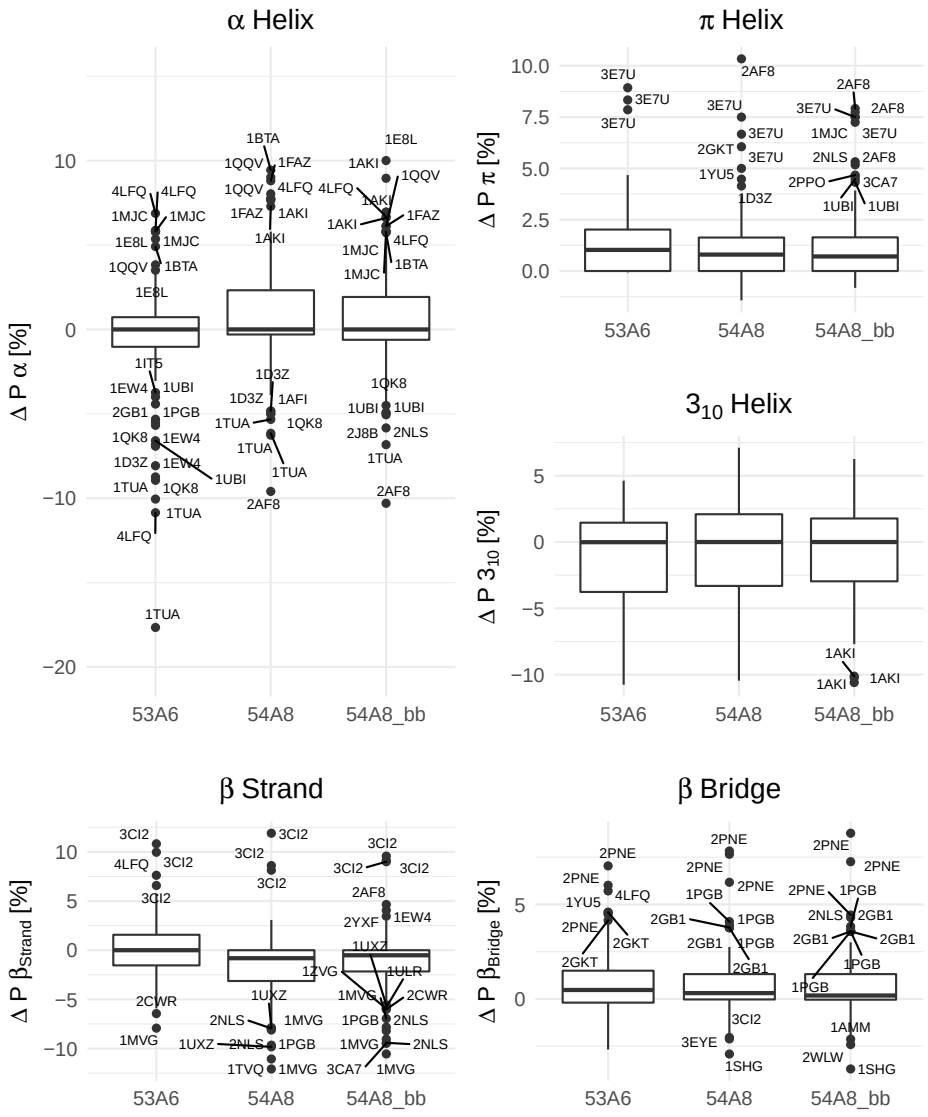

Figure 2.7:

Figure 2.7: Boxplots of the average deviation ΔP of the occurence of secondary structure elements in three independent simulations of 52 proteins as percentage change from the assignment in the experimentally determined structure, following the DSSP classification. The box indicates the upper and lower quartile of the deviation with the whiskers extend to the extremes of the distribution, with a maximum of 1.5 times the inter-quartile range. Individual outliers are represented with dots and are labeled with the PDB-code of the simulation. 2OVN falls of the scale of the α-Helix panel in all three parameter sets.

References

[1] C. Oostenbrink et al. “A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6”. In: *J. Comp. Chem.* 25.13 (Oct. 2004), pp. 1656–1676. DOI: 10.1002/jcc.20090.

[2] L. D. Schuler and W. F. Van Gunsteren. “On the Choice of Dihedral Angle Potential Energy Functions for n -Alkanes”. In: *Mol Sim* 25.5 (Oct. 2000), pp. 301–319. DOI: 10.1080/08927020008024504.

[3] T. A. Soares et al. “An improved nucleic acid parameter set for the GROMOS force field”. In: *J. Comp. Chem.* 26.7 (May 2005), pp. 725–737. DOI: 10.1002/jcc.20193. URL: http://doi.wiley.com/10.1002/jcc.20193.

[4] N. Schmid et al. “Definition and testing of the GROMOS force-field versions 54A7 and 54B7”. In: *Eur. Biophys. J.* 40.7 (July 2011), pp. 843–856. DOI: 10.1007/s00249-011-0700-9.

[5] M. M. Reif, M. Winger, and C. Oostenbrink. “Testing of the GROMOS force-field parameter set 54A8: Structural properties of electrolyte solutions, lipid bilayers, and proteins”. In: *J. Chem. Theo. Comp.* 9.2 (Feb. 2013), pp. 1247–1264. DOI: 10.1021/ct300874c.

[6] M. Buck et al. “Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme.” In: *Biophys J* 90.4 (Feb. 2006), pp. L36–8. DOI: 10.1529/biophysj.105.078154.

[7] A. Perez et al. “Grid-based backbone correction to the ff12SB protein force field for implicit-solvent simulations.” In: *J Chem Theory Comput* 11.10 (Oct. 2015), pp. 4770–9. DOI: 10.1021/acs.jctc.5b00662.

[8] A. D. Mackerell, M. Feig, and C. L. Brooks. “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”. In: *J Comput Chem* 25.11 (Aug. 2004), pp. 1400–1415. DOI: 10.1002/jcc.20065.

[9] Setz M. *Molecular dynamics simulations of biomolecules: from validation to application*. Vienna: Dissertation, BOKU, 2018, p. 285. DOI: https://permalink.obvsg.at/AC15159105.

[10] M. Stroet et al. “Challenges associated with the validation of protein force fields based on structural criteria.” In: *submitted to J Chem Inf Model* (2020).

[11] C. Margreitter and C. Oostenbrink. “Optimization of Protein Backbone Dihedral Angles by Means of Hamiltonian Reweighting”. In: *J Chem Inf Model* 56.9 (Sept. 2016), pp. 1823–1834. DOI: 10.1021/acs.jcim.6b00399.

[12] C. Margreitter and C. Oostenbrink. “Correction to Optimization of Protein Backbone Dihedral Angles by Means of Hamiltonian Reweighting”. In: *J Chem Inf Model* 58.8 (Aug. 2018), pp. 1716–1720. DOI: 10.1021/acs.jcim.8b00470.

[13] F. Avbelj et al. “Intrinsic backbone preferences are fully present in blocked amino acids.” In: *Proc Nat Acad Sci USA* 103.5 (Jan. 2006), pp. 1272–7. DOI: 10.1073/pnas.0510420103.

[14] N. Schmid et al. “Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation”. In: *Comp Phys Commun* 183.4 (Apr. 2012), pp. 890–903. DOI: 10.1016/j.cpc.2011.12.014.

[15] M. M. Reif, P. H. Hünenberger, and C. Oostenbrink. “New interaction parameters for charged amino acid side chains in the GROMOS force field”. In: *J. Chem. Theo. Comp.* 8.10 (Oct. 2012), pp. 3705–3723. DOI: 10.1021/ct300156h.

[16] W. J. Meath et al. “Intermolecular Forces”. In: *Intermolecular Forces*. Ed. by Bernard Pullman. Vol. 14. May. Dordrecht: Springer Netherlands, 1981. Chap. Interactions, p. 101. DOI: 10.1007/978-94-015-7658-1.

[17] H. J. C. Berendsen et al. “Molecular dynamics with coupling to an external bath”. In: *J Chem Phys* 81.8 (Oct. 1984), pp. 3684–3690. DOI: 10.1063/1.448118.

[18] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. “Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes”. In: *J Comput Phys* 23.3 (Mar. 1977), pp. 327–341. DOI: 10.1016/0021-9991(77)90098-5.

[19] I. G. Tironi et al. “A generalized reaction field method for molecular dynamics simulations”. In: *J. Chem. Phys.* 102.13 (Apr. 1995), pp. 5451–5459. DOI: 10.1063/1.469273.

[20] T. N. Heinz, W. F. van Gunsteren, and P. H. Hünenberger. “Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations”. In: *J Chem Phys* 115.3 (July 2001), p. 1125. DOI: 10.1063/1.1379764.

[21] A. Pardi, M. Billeter, and K. Wüthrich. “Calibration of the angular dependence of the amide proton-Cα proton coupling constants, 3JHNα, in a globular protein: Use of 3JHNα for identification of helical secondary structure”. In: *J Mol Biol* 180.3 (Dec. 1984), pp. 741–751. DOI: 10.1016/0022-2836(84)90035-4.

[22] W. F. Van Gunsteren et al. “Validation of Molecular Simulation: An Overview of Issues”. In: *Angew. - Int. Ed.* 57.4 (Jan. 2018), pp. 884–902. DOI: 10.1002/anie.201702945.

[23] G. Nagy and C. Oostenbrink. “Dihedral-Based Segment Identification and Classification of Biopolymers I: Proteins”. In: *J Chem Inf Model* 54.1 (Jan. 2014), pp. 266–277. DOI: 10.1021/ci400541d.

[24] J. Grdadolnik et al. “Populations of the three major backbone conformations in 19 amino acid dipeptides.” In: *Proc Nat Acad Sci USA* 108.5 (Feb. 2011), pp. 1794–8. DOI: 10.1073/pnas.1017317108.

[25] R. W. Zwanzig. “High Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases”. In: *J Chem Phys* 22.8 (Aug. 1954), pp. 1420–1426. DOI: 10.1063/1.1740409.

[26] G.M. Torrie and J.P. Valleau. “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling”. In: *J Comput Phys* 23.2 (Feb. 1977), pp. 187–199. DOI: 10.1016/0021-9991(77)90121-8.

[27] Z. Lin, C. Oostenbrink, and W. F. van Gunsteren. “On the use of one-step perturbation to investigate the dependence of NOE-derived atom-atom distance bound violations of peptides upon a variation of force-field parameters”. In: *Eur Biophys J* 43.2-3 (Mar. 2014), pp. 113–119. DOI: 10.1007/s00249-014-0943-3.

[28] M. Pechlaner, M. M. Reif, and C. Oostenbrink. “Reparametrisation of united-atom amine solvation in the GROMOS force field”. In: *Mol Phys* 115.9-12 (June 2017), pp. 1144–1154. DOI: 10.1080/00268976.2016.1255797.

[29] B. K Ho and R. Brasseur. “The Ramachandran plots of glycine and pre-proline”. In: *BMC Struct Bio* 5.1 (Aug. 2005), p. 14. DOI: 10.1186/1472-6807-5-14.

[30] O. Carugo. “How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared”. In: *J Appl Cryst* 36.1 (Feb. 2003), pp. 125–128. DOI: 10.1107/S0021889802020502.

[31] W. Kabsch and C. Sander. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. In: *Biopolymers* 22.12 (1983), pp. 2577–2637. DOI: 10.1002/bip.360221211. arXiv: 83/122577-6 [0006-3525].

[32] J. R. Steiner D.and Allison, A. P. Eichenberger, and W. F. van Gunsteren. “On the calculation of 3J$\alpha\beta$-coupling constants for side chains in proteins”. In: *Journal of Biomolecular NMR* 53.3 (July 2012), pp. 223–246. DOI: 10.1007/s10858-012-9634-5.

[33] R. B. Best, N.-V. Buchete, and G. Hummer. “Are Current Molecular Dynamics Force Fields too Helical?” In: *Biophys J* 95.1 (2008), pp. L07–L09. DOI: https://doi.org/10.1529/biophysj.108.132696.

[34] L Wickstrom, A Okur, and C Simmerling. “Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data”. In: *Biophys J* 97.3 (2009), pp. 853–856. DOI: https://doi.org/10.1016/j.bpj.2009.04.063.

[35] J. Dolenc et al. “Methods of NMR structure refinement: molecular dynamics simulations improve the agreement with measured NMR data of a C-terminal peptide of GCN4-p1”. In: *J Biomol NMR* 47.3 (July 2010), pp. 221–235. DOI: 10.1007/s10858-010-9425-9.

Chapter 3

The effect of using a twin-range cut-off scheme for non-bonded interactions: Implications for force-field parameterization?

This work was previously published in

Diem M. and Oostenbrink C.

The effect of using a twin-range cut-off scheme for non-bonded interactions: Implications for force-field parameterization?

J. Chem. Theory Comput. ***2020*** *16, 10, 5985–5990*

https://doi.org/10.1021/acs.jctc.0c00509

Abstract

Recently, concerns have been voiced concerning the validity of the GROMOS force fields, being parameterized using a twin-range cut-off scheme, in which longer-ranged non-bonded forces and energies are updated less frequently than shorter-ranged ones. Here we demonstrate that the influence of such a scheme on the thermodynamic, structural and dynamic properties used in the parameterization of the GROMOS force fields is minor. We find root-mean-square differences of maximally 0.5 kJ/mol for the solvation free energy and heat of vaporization and of maximally 0.4% for the density. Slightly larger differences are observed when switching from a group-based to an atom-based cut-off scheme. In cases where the twin-range cut-off scheme does result in minor differences compared to a single range cut-off these are well within the deviation from the experimentally measured values.

Main

Recently, concerns have been voiced concerning the validity of the GROMOS force fields for biomolecular simulation.[1] In particular, when using the GROMOS force field in the popular molecular dynamics program GROMACS (version 2019.3 and newer), the following warning is issued: "The GROMOS force fields have been parametrized with a physically incorrect multiple-time-stepping scheme for a twin-range cut-off. When used with a single-range cut-off (or a correct Trotter multiple-time-stepping scheme), physical properties, such as the density, might differ from the intended values. Check if molecules in your system are affected by such issues before proceeding." Here, we address the question if this warning is warranted and if the use of this time-saving technique affected the parameterization of the GROMOS force field in a relevant way.

The use of force fields to describe the interaction energy between atoms and molecules in molecular dynamics simulation is a commonly used approach and the validation of such simulations is a crucial matter.[2] Apart from the accuracy of the force field, a wide range of simulation parameter settings will influence the outcome of a simulation. Ideally, the outcome of a simulation using

a given force field should be independent of the parameter settings that were used to parameterize the force field. Unfortunately, this situation is extremely difficult, if not impossible, to reach. It is certainly not currently possible as atomistic simulations do not reach macroscopic space and time scales. This means that, by necessity, a range of approximations, including treating electrostatic interactions as either lattice-sum or cut-off with reaction-field, are required. For this reason, it is generally prudent to use simulation parameter settings that are similar to those used to derive the force field parameters.

The GROMOS force field parameters have been derived using a twin-range (TR) cut-off scheme for the non-bonded interactions as a simulation-time saving technique. Typically, a pair-list is generated after a fixed time interval (e.g. 10 fs). The forces and energies up to a short-range cut-off (typically 0.8 nm) are computed at every time-step according to this pair-list. At pair-list updates, also after the chosen time interval, forces and energies up to a long-range cut-off (typically 1.4 nm) are computed and kept constant in between updates [3]. The discontinuity this introduces to the forces will lead to additional noise in the properties of the system, which will be small if the longer-ranged non-bonded forces can be expected to change little in between the updates. This approximation was introduced at a time when computational power was limited with tests at the time suggesting the increase in computational efficiency outweighed any possible loss in accuracy. We note that the choice of a specific time-step, van der Waals cut-off, shifting function, the use of constraints or the choice of a particular level of numerical precision or solvation model in a simulation are all based on a similar balance between computational efficiency and accuracy [4]. However, as the current GROMOS force field was parameterized using this approximation it is reasonable to ask if this approximation affected the parameterization to such an extent that if pairlist and non-bonded forces were computed at every time step (single range, SR) the results of the simulation would no-longer be valid.

Before presenting our findings we note that there has been a number of recent reports suggesting that the results obtained in simulations performed using the twin-range scheme and simulations performed using a single range cut-off were essentially identical. No significant differences were observed in

the area per lipid of a membrane, when the only difference in the simulation parameter settings was the use of a SR versus a TR cut-off scheme [5]. Similarly, the radius of gyration of a polyamidoamine dendrimer and the structural properties of a protein or a membrane in constant pH simulations were indistinguishable when using SR versus TR cut-offs [6]. However, differences were found when using an atomistic (AT) cut-off rather than a charge-group (CG) based cut-off for non-bonded forces [6]. Also, the potential of mean force between small solutes and carbon nanotube model systems were not significantly affected by the choice of SR versus TR cut-offs [7]. None of these studies directly addressed the question of whether the types of properties used in the parameterization of the GROMOS force field were affected by the use of a twin-range cut-off scheme.

In a related and very extensive study, Gonsalves et al. examined the effect of various simulation parameter settings on the thermodynamic and transport properties of pure liquids [8]. Significant differences were observed, depending on the exact choice of parameters governing the computation of non-bonded interaction energies. However, a direct comparison of SR and TR using otherwise identical parameter settings was not considered.

The non-bonded parameters of the GROMOS force fields were parameterized against thermodynamic properties of small molecules. Specifically, the density and heat of vaporization of pure liquids and the solvation free energy of amino acid side-chain analogues in water and in cyclohexane [9, 10, 11]. Here, the hydration free energies for the neutral amino-acid side-chain analogues used in the parameterization of the GROMOS force field are re-calculated using SR and TR cut-off schemes. In addition, the effect of the cut-off scheme on key properties of a set of five liquids of different polarity and molecular complexity is also examined. All calculations were performed using both an atomistic (AT) and a charge-group based (CG) cut-off, to further clarify whether any of the key assumptions made during the parameterization of the force field affects the outcome of the simulations. The GROMOS force fields were initially parameterized using a charge-group cut-off scheme. This was used as it reduces the artefacts in the forces due to atoms at the cut-off distance. The main question we aim to address is whether the GROMOS force field parameters would have

been affected, if a SR cut-off was used instead of the TR cut-off.

To address this question four sets of simulations were performed, two with a SR cut-off scheme, two with a TR cut-off scheme. In both cases either a charge-group based (CG) cut-off and pair list or an atomistic (AT) cut-off and pair list were used. All simulations were performed using GROMOS11 [12]. Note that other simulation packages may not support all four combinations of cut-off schemes.

Analogous to our previous work, the amino acid side-chain analogues were generated by breaking the C_α – C_β bond in the naturally occurring neutral amino acids and adjusting the parameters of the united atom Cβ carbon [10]. The molecules were placed in a rectangular periodic box and solvated in SPC water [13]. The systems were then simulated at a constant reference temperature of 298.15 K and a reference pressure of 1 atm, using the weak-coupling scheme [14]. The coupling times were τ_T = 0.1 ps and τ_P = 0.5 ps. The isothermal compressibility was set to 4.575 x 10^{-4} (kJ mol^{-1} nm^{-3})$^{-1}$. Separate temperature baths were used for the solute and solvent. Bond-lengths were constraint using the SHAKE algorithm, with a relative geometric accuracy of 10^{-4} [15]. A time step of 2 fs was used in the leap-frog algorithm[16] to integrate the equations of motion. For the TR cut-off scheme, a short-range cut-off of 0.8 nm and a long-range cut-off of 1.4 nm was used with updates of longer-ranged forces and energies and the pair list every 10 fs (5 steps). The SR cut-off scheme involved a single cut-off of 1.4 nm, with an update of pair list and forces and energies at every step. A reaction-field contribution to the energies and forces was added to the electrostatic interactions to account for a homogeneous medium with a relative dielectric constant of 61 [17] beyond the 1.4 nm cut-off [18].

Solvation free energies were calculated by alchemically removing the non-bonded interactions of the amino acid side-chain analogues using a λ-coupling parameter approach and thermodynamic integration [20]. Simulations were performed at 21 equally spaced λ-values. At each λ-value, the systems were equilibrated for 100 ps and data was collected for 1 ns. Every set of simulations was performed three times, using different random number seeds for the

Table 3.1: Solvation free enthalpies of the amino-acid side-chain analogues in SPC water using different non-bonded interaction cut-off schemes. CG: charge-group based cut-off scheme with reaction field. AT: atom-based cut-off scheme with reaction field. TR: twin-range cut-off scheme. SR: single-range cut-off scheme. Averages over three independent sets of simulations are reported with errors reported as standard deviations. [a] reference[10]; [b] reference[9];[c] reference[19];

amino acid	ΔG_{solv} CG/SR [kJ/mol]	ΔG_{solv} CG/TR [kJ/mol]	ΔG_{solv} AT/SR [kJ/mol]	ΔG_{solv} AT/TR [kJ/mol]	ΔG_{solv} lit. [kJ/mol]	ΔG_{solv} exp. [c] [kJ/mol]
ALA	9.2 $\pm$ 0.0	9.1 $\pm$ 0.2	9.2 $\pm$ 0.1	9.1 $\pm$ 0.1	6.2[b]	8.4
ARG	-45.8 $\pm$ 0.0	-46.8 $\pm$ 0.3	-43.9 $\pm$ 0.1	-44.4 $\pm$ 0.1	-46.1[a]	-45.7
ASN	-40.3 $\pm$ 0.1	-40.5 $\pm$ 0.1	-39.0 $\pm$ 0.1	-39.5 $\pm$ 0.2	-40.6[a]	-40.6
ASP	-25.8 $\pm$ 0.1	-26.0 $\pm$ 0.2	-24.9 $\pm$ 0.2	-25.7 $\pm$ 0.7	-30.6[a]	-28.0
CYS	-6.8 $\pm$ 0.1	-6.9 $\pm$ 0.1	-6.5 $\pm$ 0.1	-6.9 $\pm$ 0.1	-4.7[a]	-5.2
GLU	-28.2 $\pm$ 0.1	-28.6 $\pm$ 0.1	-27.7 $\pm$ 0.3	-27.8 $\pm$ 0.2	-27.2[a]	-27.0
GLN	-39.1 $\pm$ 0.1	-39.4 $\pm$ 0.3	-37.7 $\pm$ 0.1	-38.1 $\pm$ 0.3	-38.5[a]	-39.4
HIS	-43.8 $\pm$ 0.3	-44.7 $\pm$ 0.1	-43.3 $\pm$ 0.0	-43.8 $\pm$ 0.1	-42.7[a]	-42.9
ILE	9.5 $\pm$ 0.2	9.3 $\pm$ 0.3	9.9 $\pm$ 0.2	9.8 $\pm$ 0.3	8.7[b]	8.7
LEU	10.6 $\pm$ 0.0	10.2 $\pm$ 0.1	10.8 $\pm$ 0.1	10.4 $\pm$ 0.2	10.4[b]	9.7
LYS	-18.7 $\pm$ 0.4	-19.1 $\pm$ 0.4	-17.7 $\pm$ 0.3	-18.4 $\pm$ 0.5	-17.3[a]	18.3
MET	-7.8 $\pm$ 0.2	-8.2 $\pm$ 0.2	-7.0 $\pm$ 0.3	-7.3 $\pm$ 0.2	-6.8[a]	-6.2
PHE	-0.8 $\pm$ 0.1	-1.4 $\pm$ 0.2	-0.6 $\pm$ 0.3	-0.9 $\pm$ 0.1	0.0[a]	-3.1
SER	-23.3 $\pm$ 0.2	-23.3 $\pm$ 0.1	-22.5 $\pm$ 0.2	-22.7 $\pm$ 0.1	-22.1[a]	-21.2
THR	-21.3 $\pm$ 0.2	-21.6 $\pm$ 0.1	-20.5 $\pm$ 0.2	-20.7 $\pm$ 0.1	-20.0[a]	-20.5
TRP	-24.3 $\pm$ 0.1	-25.1 $\pm$ 0.4	-23.5 $\pm$ 0.2	-24.4 $\pm$ 0.1	-25.7[a]	-24.7
TYR	-24.8 $\pm$ 0.1	-25.6 $\pm$ 0.3	-23.7 $\pm$ 0.3	-24.4 $\pm$ 0.1	-25.5[a]	-26.6
VAL	8.5 $\pm$ 0.2	8.0 $\pm$ 0.1	8.6 $\pm$ 0.2	8.3 $\pm$ 0.1	8.6[b]	8.2

Table 3.2: Root-mean-square difference of the hydration free enthalpies between the different columns of table 3.1 (upper half) and average errors (column – row) between these columns (lower half). All values in kJ/mol.

	CG/SR	CG/TR	AT/SR	AT/TR	lit.	exp
CG/SR		0.53	0.85	0.52	1.66	1.26
CG/TR	0.44		1.32	0.92	1.73	1.28
AT/SR	-0.71	-1.15		0.47	1.88	1.54
AT/TR	-0.29	-0.74	0.42		1.67	1.28
lit.	0.07	-0.38	0.77	0.36		1.24
exp.	0.10	-0.35	0.80	0.39	-0.03	

generation of initial velocities. Error estimates are reported as standard deviations over the three independent estimates.

Five liquids of different polarity were also examined: hexane, butylamine (BAN), ethanol, dimethylsulfoxide (DMSO) and water. Each system consisted of 1000 molecules. The systems were simulated at constant pressure, using the same simulation parameter settings as above. The only differences were that the isothermal compressibility for the pressure scaling algorithm and the relative dielectric constant for the reaction-field contributions to long range interactions were set to values derived from experiment [6]. The translational and internal/rotational degrees of freedom were coupled to separate temperature baths. Three independent simulations of 10 ns each were performed, using different initial velocities. The last 5 ns of every simulation was used to analyze the data. Error estimates are reported as the standard error of the mean over the three simulations.

Simulations in the gas phase were performed by simulating a single amino-acid analogue or by placing the molecules at a distance of at least 50 nm of each other (no inter-molecular interactions) for multi-molecular systems. In these simulations a Langevin thermostat[21] involving stochastic forces was added to help the molecule to escape conformational local minima (friction coefficient γ = 91 ps^{-1}). Diffusion constants (D) were estimated from the Einstein-relation, rotational relaxation constants (τ_2) were estimated from the auto-correlation function of the second order Legendre polynomial of a molecular axis. To estimate the relative dielectric permittivity (ϵ), 5 additional simu-

Table 3.3: Selected properties for six pure liquids using different cut-off schemes. Literature data is taken from refs.[10, 9, 23] a)

property	cut-off	Hexane	Butylamine	Ethanol	DMSO	Water
ΔH_{vap}	CG/SR	31.8 ± 0.0	39.3 ± 0.0	45.5 ± 0.0	52.9 ± 0.0	44.3 ± 0.0
$(kJmol^{-1})$	CG/TR	31.8 ± 0.0	39.7 ± 0.0	45.2 ± 0.0	53.0 ± 0.0	44.1 ± 0.0
	AT/SR	31.8 ± 0.0	39.5 ± 0.0	45.8 ± 0.0	53.6 ± 0.0	44.2 ± 0.0
	AT/TR	31.3 ± 0.0	40.1 ± 0.0	46.2 ± 0.0	53.3 ± 0.0	44.0 ± 0.0
	lit.	31.6	30.1	44.3	52.9	43.7
	exp.	31.6	35.7	42.3	52.9	44.0
ρ	CG/SR	655.8 ± 0.0	734.1 ± 0.1	770.9 ± 0.0	1096.1 ± 0.1	972.1 ± 0.1
(kgm^{-3})	CG/TR	656.1 ± 0.0	738.4 ± 0.1	772.1 ± 0.0	1096.5 ± 0.1	972.1 ± 0.1
	AT/SR	656.2 ± 0.1	735.7 ± 0.1	772.3 ± 0.0	1103.0 ± 0.0	975.1 ± 0.0
	AT/TR	656.5 ± 0.1	742.3 ± 0.0	773.6 ± 0.1	1103.2 ± 0.1	974.9 ± 0.1
	lit.	655.0	745.0	778.0	1096.0	972.0
	exp.	655.0	737.0	785.0	1095.0	997.0
D	CG/SR	4.8 ± 0.2	2.4 ± 0.0	1.1 ± 0.0	1.1 ± 0.0	4.2 ± 0.2
$10^{-9}m^2s^{-1}$	CG/TR	4.9 ± 0.1	2.4 ± 0.1	1.1 ± 0.0	1.0 ± 0.1	4.1 ± 0.1
	AT/SR	4.9 ± 0.2	2.3 ± 0.1	1.1 ± 0.0	0.9 ± 0.0	4.0 ± 0.1
	AT/TR	4.8 ± 0.1	2.3 ± 0.1	1.1 ± 0.0	0.9 ± 0.1	3.8 ± 0.1
$\tau_2(ps)$	CG/SR	2.2 ± 0.0	1.7 ± 0.0	4.3 ± 0.0	2.7 ± 0.0	1.1 ± 0.0
	CG/TR	2.2 ± 0.0	1.9 ± 0.0	4.3 ± 0.0	2.7 ± 0.0	1.1 ± 0.0
	AT/SR	2.2 ± 0.0	1.8 ± 0.0	4.3 ± 0.0	2.9 ± 0.0	1.1 ± 0.0
	AT/TR	2.2 ± 0.0	2.1 ± 0.0	4.3 ± 0.1	2.9 ± 0.0	1.1 ± 0.0
ϵ_r	CG/SR	1	7	12	39	62
	CG/TR	1	7	12	39	63
	AT/SR	1	10	14	52	81
	AT/TR	1	12	15	54	80

a) Error estimates are calculated as the standard deviation of the mean of three independent simulations. Error estimates of 0.0 indicate errors smaller than 0.1.

lations were performed in the condensed phase for 0.5 ns under the influence of an external electric field, with sizes 0.00625, 0.0125, 0.025, 0.05, and 0.1 $e{\cdot}nm^{-2}$ [22].

Table 3.1 shows the average solvation free energies using the different cut-off schemes. For comparison the literature values from the original parametrizations [9, 10] are given, as well as the experimental data used in the parametrization [19]. Deviations between the literature values and the current values can be traced to longer simulation times and the use of more λ values in the thermodynamic integration in the current work. Absolute differences between the nonbonded cutoff schemes for TR vs SR and CG vs AT comparisons are given

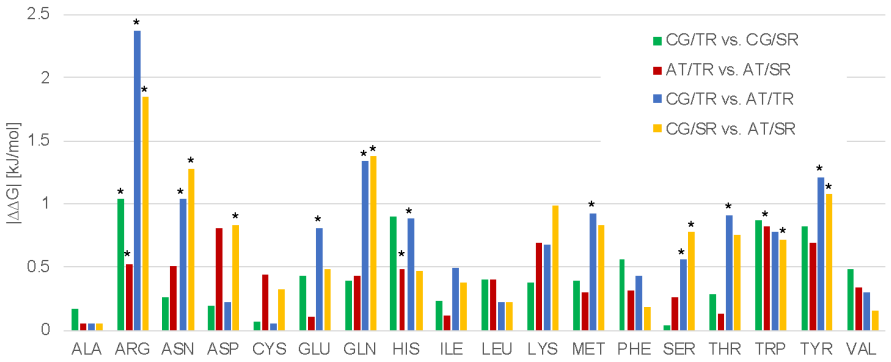

Figure 3.1: Absolute differences in solvation free energy for four different comparisons of non-bonded interaction cut-off schemes. CG: charge-group based cut-off scheme with reaction field. AT: atom-based cut-off scheme with reaction field. TR: twin-range cut-off scheme. SR: single-range cut-off scheme. Differences based on the averages (n = 3) in Table 3.1, with significant differences ($p < 0.01$) indicated by asterisks.

in Figure 3.1, along with an indication of significant differences ($p < 0.01$).

The differences between TR and SR are very small. The largest deviation is observed for the histidine side chain at 1.1 kJ mol^{-1} when using an group-based cut-off. The top half of Table 3.2 shows the root-mean-square difference for comparisons between each of the columns in table 3.1. A comparison between TR and SR amounts to 0.53 kJ mol^{-1} (CG) or 0.52 kJ mol^{-1} (AT). Using a two-tailed t-test, the observed differences are significant only for Arg when using a group-based cut-off and for Arg, His and Trp when using an atom-based cut-off. The differences between CG and AT are slightly larger, with a maximum difference of 2.4 kJ mol^{-1} for Arg when using a twin-range cut-off scheme. Root-mean-square differences amount to 0.85 kJ mol^{-1} (SR) and 0.92 kJ mol^{-1} (TR). The t-test reveals significant differences for seven (SR) and nine (TR) compounds. The lower half of Table 3.2 shows the average differences between the columns of Table 3.1. Comparison of these values to the RMSD-values in the upper half reveals that the small differences are largely systematic. Overall, the largest difference in the simulations is observed for CG/TR vs. AT/SR, with an RMSD of 1.32 kJ mol^{-1}, where the two effects (CG to AT and TR to SR) seem

to add up. For the comparison CG/SR vs. AT/TR the RMSD amounts to 0.52 kJ mol^{-1} and systematic effects partially cancel.

Table 3.3 shows selected properties for the pure liquids. Comparing the SR and TR schemes, the largest deviation in terms of the heat of vaporization amounts to 0.6 kJ mol^{-1} for butylamine using AT. The root-mean-square deviation over these five liquids amounts to 0.24 kJ mol^{-1} (CG) and 0.42 kJ mol^{-1} (AT). The densities of all liquids are rather similar. The largest deviation in a SR vs. TR comparison is seen for butylamine and amounts to 0.9% (AT). Root-mean-square deviations amount to 2.0 kg m^{-3} (CG) and 3.0 kg m^{-3} (AT) or 0.3% (CG) and 0.4% (AT). Differences regarding the comparison CG vs. AT are slightly larger with root-mean-square differences of 3.8 kg m^{-3} (TR) and 3.5 kg m^{-3} (SR). Figure 3.2 shows the radial distribution functions and dipole correlation functions for the simulations of the polar liquids. As observed before, differences are found when comparing CG and AT simulations [6, 24]. When comparing simulations performed with SR to those with TR, the curves for both these structural properties are indistinguishable. A detailed comparison for group-based and atom-based cut-off schemes shows that artifacts at the cut-off appear for atom-based cut-off schemes [6, 25]. Dynamic properties were computed for the pure liquids too. For the diffusion coefficient, τ_2 rotational relaxation time and the relative dielectric permittivity, the deviations between SR and TR are negligible. For the dielectric permittivity, the difference between AT and CG that was previously reported [6] is reproduced.

In conclusion, we have repeated the key simulations used in the parameterization of the GROMOS force fields for 18 amino acid side-chain analogues and five pure liquids of different polarities. We have shown that the effect of using a twin-range cut-off scheme as compared to a single-range cut-off on the thermodynamic, structural and dynamic properties is minor. All differences that were observed are well within the deviation from the experimentally measured values and are likely to be irrelevant in the vast majority of applications of biomolecular simulations.[25] Based on these results we can conclude that the parameterization of the GROMOS force fields has not been influenced by the use of the twin-range cut-off scheme to a large extent. While ideally all force

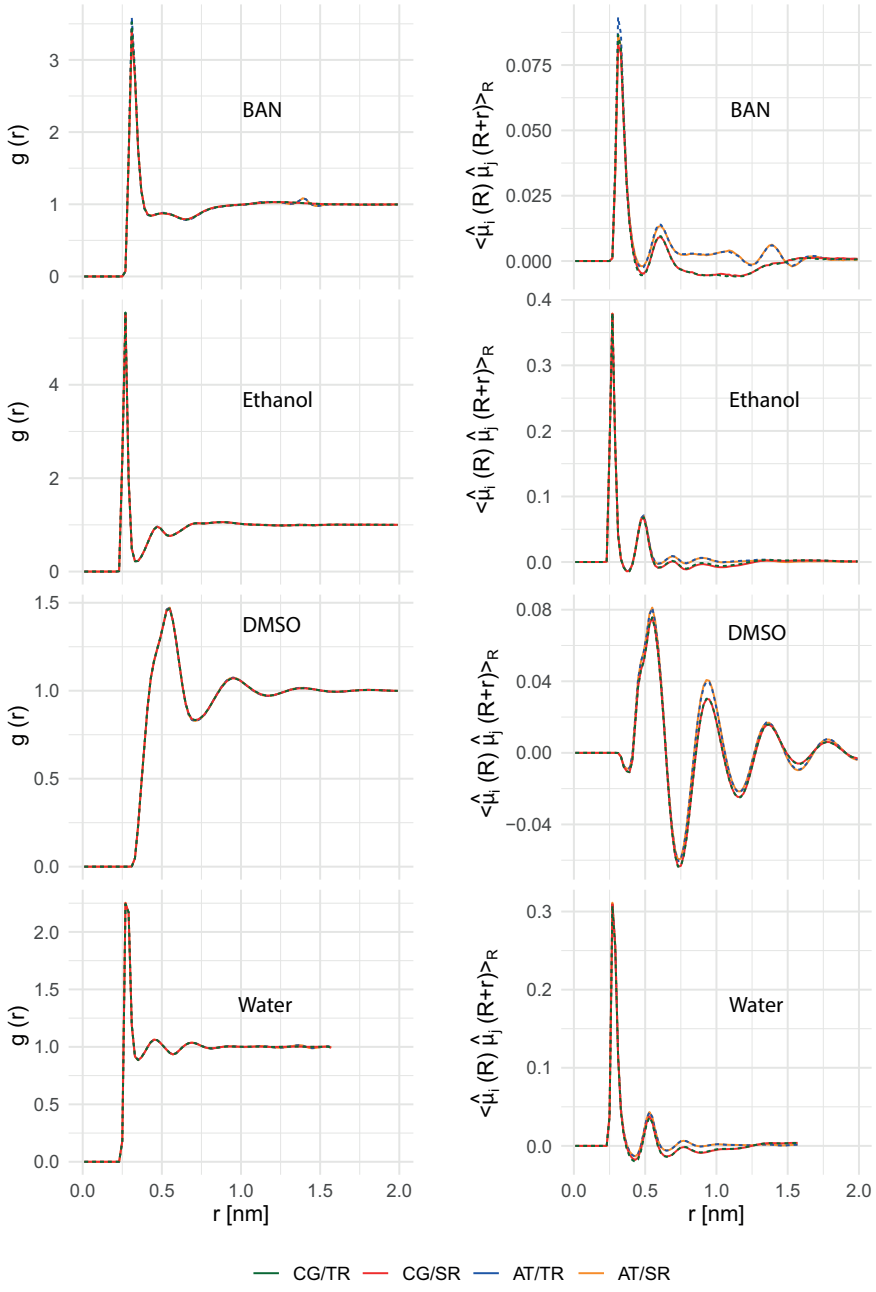

Figure 3.2: radial distribution functions, g(r), and dipole correlation functions, $\langle \hat{\mu}(R)\hat{\mu}(R+r)\rangle_R$ for the four polar liquids. Molecule positions taken as the position of the N-atom in butylamine (BAN), the O-atom in ethanol, the S-atom in DMSO and the O-atom in water.

fields should be parameterized without the use of any time saving approximations this has not been true in the past. At least in the case of the GROMOS family of force fields we can safely say that the use of a single as opposed to a twin-range cut-off has had no effect on the parameterization in terms of the agreement with experiment. Thus, in codes where a GROMOS-type twin-range is not implemented users should use a group-based, single-range cut-off to obtain equivalent results.

References

[1] B. Hess et al. “On The Importance of Accurate Algorithms for Reliable Molecular Dynamics Simulations”. In: *ChemRxiv* (Dec. 2019). DOI: 10.26434/chemrxiv.11474583.v1.

[2] W. F. Van Gunsteren et al. “Validation of Molecular Simulation: An Overview of Issues”. In: *Angew. - Int. Ed.* 57.4 (Jan. 2018), pp. 884–902. DOI: 10.1002/anie.201702945.

[3] H. J. C. Berendsen et al. “Simulations of Proteins in Water”. In: *Annals of the New York Academy of Sciences* 482.1 Computer Simu (Dec. 1986), pp. 269–286.

[4] W. F. van Gunsteren et al. “On the Effect of the Various Assumptions and Approximations used in Molecular Simulation on the Properties of Bio-Molecular Systems: A Review of Issues”. In: *submitted* (2020).

[5] S. Reißer et al. “Real Cost of Speed: The Effect of a Time-Saving Multiple-Time-Stepping Algorithm on the Accuracy of Molecular Dynamics Simulations”. In: *Journal of Chemical Theory and Computation* 13.6 (2017), pp. 2367–2372.

[6] F. D. Silva et al. “The Impact of Using Single Atomistic Long-Range Cutoff Schemes with the GROMOS 54A7 Force Field”. In: *Journal of Chemical Theory and Computation* 14.11 (2018), pp. 5823–5833.

[7] D.l Markthaler, S. Jakobtorweihen, and N. Hansen. “Lessons Learned from the Calculation of One-Dimensional Potentials of Mean Force [Article v1.0]”. In: *Living Journal of Computational Molecular Science* 1.1 (2019). DOI: 10.33011/livecoms.1.2.11073.

[8] Y. M. H. Gonçalves et al. “Influence of the Treatment of Nonbonded Interactions on the Thermodynamic and Transport Properties of Pure Liquids Calculated Using the 2016H66 Force Field”. In: *Journal of Chemical Theory and Computation* 15.3 (2019), pp. 1806–1826.

[9] L. D. Schuler, X. Daura, and W. F. Van Gunsteren. “An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase”.

In: *J. Comp. Chem.* 22.11 (Aug. 2001), pp. 1205–1218. DOI: 10.1002/jcc.1078.

[10] C. Oostenbrink et al. "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6". In: *J. Comp. Chem.* 25.13 (Oct. 2004), pp. 1656–1676. DOI: 10.1002/jcc.20090.

[11] M. M. Reif, P. H. Hünenberger, and C. Oostenbrink. "New interaction parameters for charged amino acid side chains in the GROMOS force field". In: *J. Chem. Theo. Comp.* 8.10 (Oct. 2012), pp. 3705–3723. DOI: 10.1021/ct300156h.

[12] N. Schmid et al. "Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation". In: *Comp Phys Commun* 183.4 (Apr. 2012), pp. 890–903. DOI: 10.1016/j.cpc.2011.12.014.

[13] W. J. Meath et al. "Intermolecular Forces". In: *Intermolecular Forces*. Ed. by Bernard Pullman. Vol. 14. May. Dordrecht: Springer Netherlands, 1981. Chap. Interactions, p. 101. DOI: 10.1007/978-94-015-7658-1.

[14] H. J. C. Berendsen et al. "Molecular dynamics with coupling to an external bath". In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3684–3690.

[15] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes". In: *J. Comp. Phys.* 23.3 (Mar. 1977), pp. 327–341. DOI: 10.1016/0021-9991(77)90098-5.

[16] Hockney RW. "The potential calculation and some applications." In: *Methods Comput Phys* 209 (1970), pp. 136–211.

[17] T. N. Heinz, W. F. van Gunsteren, and P. H. Hünenberger. "Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations". In: *J Chem Phys* 115.3 (July 2001), p. 1125. DOI: 10.1063/1.1379764.

[18] I. G. Tironi et al. "A generalized reaction field method for molecular dynamics simulations". In: *J. Chem. Phys.* 102.13 (Apr. 1995), pp. 5451–5459. DOI: 10.1063/1.469273.

[19] R. Wolfenden et al. “Affinities of Amino Acid Side Chains for Solvent Water”. In: *Biochemistry* 20.4 (1981), pp. 849–855. DOI: 10.1021/bi00507a030.

[20] J. G. Kirkwood. “Statistical mechanics of fluid mixtures”. In: *The Journal of Chemical Physics* 3.5 (1935), pp. 300–313. DOI: 10.1063/1.1749657.

[21] W. F. Van Gunsteren and H. J.C. Berendsen. “A Leap-Frog Algorithm for Stochastic Dynamics”. In: *Molecular Simulation* 1.3 (1988), pp. 173–185. DOI: 10.1080/08927028808080941.

[22] S. Riniker, A. Pitschna E. Kunz, and W. F. Van Gunsteren. “On the calculation of the dielectric permittivity and relaxation of molecular models in the liquid phase”. In: *Journal of Chemical Theory and Computation* 7.5 (May 2011), pp. 1469–1475. DOI: 10.1021/ct100610v.

[23] D. P. Geerke et al. “An effective force field for molecular dynamics simulations of dimethyl sulfoxide and dimethyl sulfoxide-water mixtures”. In: *Journal of Physical Chemistry B* 108.4 (Jan. 2004), pp. 1436–1445. DOI: 10.1021/jp035034i.

[24] P. H. Hünenberger and W. F. van Gunsteren. “Alternative schemes for the inclusion of a reaction-field correction into molecular dynamics simulations: Influence on the simulated energetic, structural, and dielectric properties of liquid water”. In: *The Journal of Chemical Physics* 108.15 (1998), pp. 6117–6134.

[25] M. Diem and C. Oostenbrink. “Hamiltonian Reweighing To Refine Protein Backbone Dihedral Angle Parameters in the GROMOS Force Field”. In: *Journal of Chemical Information and Modeling* 60.1 (2020), pp. 279–288.

Chapter 4

The effect of different cutoff schemes in molecular simulations of proteins

This work was previously published in

Diem M. and Oostenbrink C.

*The effect of different cutoff schemes in molecular simulations of proteins. J Comput Chem.. **2020**; 41: 2740– 2749.*

https://doi.org/10.1002/jcc.26426

Abstract

Molecular simulations of nanoscale systems invariably involve assumptions and approximations to describe the electrostatic interactions, which are long-ranged in nature. One approach is the use of cutoff schemes with a reaction-field contribution to account for the medium outside the cutoff scheme. Recent reports show that macroscopic properties may depend on the exact choice of cutoff schemes in modern day simulations. In this work a systematic analysis of the effects of different cutoff schemes was performed using a set of 52 proteins. We find no statistically significant differences between using a twin-range or a single-range cutoff scheme. Applying the cutoff based on charge groups or based on atomic positions, does lead to significant differences, which is traced to the cutoff noise for energies and forces. While group-based cutoff schemes show increased cutoff noise in the potential energy, applying an atomistic cutoff leads to artificial structure in the solvent at the cutoff distance. Carefully setting the temperature control, or using an atomistic cutoff for the solute and a group-based cutoff for the solvent significantly reduces the effects of the cutoff noise, without introducing structure in the solvent. This study aims to deepen the understanding of the implications different cutoffs have on molecular dynamics simulations.

Introduction

Molecular dynamics simulations are an invaluable tool to study the behaviour of proteins in aqueous solution in great detail. Nowadays time scales up to milliseconds can be simulated, which lead to new insights, that were not possible before [1]. Prolonged simulations possibly bring to light new challenges in the development of reliable force fields as well as effects of assumptions and approximations in algorithms that have been widely used [2, 3, 4]. The biggest part of computer time is used to identify and calculate the non-bonded interactions.

One way of treating the non-bonded interactions is based on lattice summation schemes. These methods make use of the commonly applied periodic boundary conditions and assume a periodic repetition of charges at an infinite

range [5, 6]. This assumption is challenged in systems, which are not perfectly periodic but should represent a dilute solution of biomolecules. While lattice summation methods are very commonly applied, these methods are not without artefacts. The induced periodicity leads to an underrepresentation of the electrostatic interactions (unlike charges at exactly half the box-length have no interaction). This underpolarisation has effects on the calculation of thermodynamic properties as well as the structures sampled in simulations. The effects on system properties have been described repeatedly [7, 8, 9, 10, 11].

Another way to treat electrostatic and van der Waals interactions are cutoff schemes, in which interactions are only computed up to a fixed atomic or molecular distance. Since a straight truncation leads to major artefacts,[12, 13] a reaction-field contribution combined with shifting or switching functions are used to ensure that the energy approaches zero at the cutoff distance. For the reaction-field contribution a continuous medium outside the cutoff-region is assumed [14, 15]. Given a box size that is larger than twice the cutoff distance, these approaches do not show artefacts due to periodicity. However, the neglect of molecular detail beyond the cutoff distance does affect the thermodynamics of the system in different ways, in particular for charged species [8, 9, 10]. A cutoff can be either imposed based on interatomic distances or by using charge-groups. In the latter case the molecular interactions between all atoms that are part of two charge-groups interact as long as the centres of the charge-groups are within the cutoff distance. The advantage of this approach is, that the definition of neutral charge groups reduces the vast majority of the electrostatic interactions to dipole-dipole interactions which have a shorter range than charge-charge interactions (r^{-3} vs. r^{-1}). For efficiency reasons, the electrostatic interactions within an atomistic or charge-group based cutoff scheme are typically calculated from a pairlist that is not necessarily updated at every timestep of the simulation. In addition, the GROMOS force fields that will be used in the current work were parameterised with a twin-range cutoff scheme. In this approach, a pairlist is calculated at specific time intervals (e.g. every 10 fs). Short-range interactions e.g. up to 0.8 nm are computed at every timestep from this pairlist. Upon pairlist construction, interactions up to a longer range cutoff (e.g. 1.4 nm) are also computed and kept constant between pairlist up-

dates. The twin-range cutoff scheme is a way to speed up simulations and allow for longer simulation timescales, but it also introduces discontinuities in the non-bonded energies and forces which leads to additional noise in the simulation. Therefore it is crucial to finetune the update intervals, an update every 10 fs was commonly seen to increase the efficiency in protein simulations without leading to significant differences in thermodynamic and structural properties [16, 17, 18].

Ideally, a force field should be independent of the simulation settings used at parameterization, but unfortunately using non-bonded interactions that are approximated by cutoff or lattice summation schemes, this is very hard if not impossible to achieve. Therefore it is recommended to use simulation settings similar to the ones that were used upon parameterization. To parameterize the GROMOS force field a twin-range, charge-group based cutoff scheme, combined with a reaction-field contribution was used. Recently, some discussion has come as to the validity of this approach [19]. Recent studies indeed show that different results are obtained when using alternative cutoff schemes for e.g. the area per lipid,[16] the radius of gyration of a dendrimer or constant pH simulations of membranes and proteins [17]. Also, the thermodynamic properties of small molecules may be affected[20]. We recently showed for small molecules, that these differences are not due to the use of the twin-range, but may be attributed to the use of lattice sum electrostatics, or the switch to an atomistic rather than group-based cutoff scheme[18]. In this study we aim to expand this analysis to a large number of simulations of proteins, such that statistically sound conclusions can be drawn with respect to any observed differences.

Methodology

The investigation is based on a set of 52 protein structures described by Setz et al. [21, 22] This set consists of 39 structures obtained by X-ray diffraction and 13 obtained by NMR experiments. Simulations were performed using the GROMOS11 software package and the GROMOS 54A8 force field [23, 24, 25]. The systems were solvated using the SPC water model and 0.15 M NaCl was

added to the simulation box. For the equilibration a 8 step protocol was used. In the first 6 steps the temperature was increased by 60 K at constant volume. At the same time harmonic position restraints were loosened by one order of magnitude from an initial force constant of 2.5 10^4 kJ mol^{-1} nm^{-2}. Step 7 was used to instantiate the roto-translational[26] constraints on the solute atoms and in the last step pressure coupling was applied at 1 atm. The equilibration took 160 ps in total, 20 ps at every step.

Unless stated differently, the weak-coupling scheme [27] with relaxation times of 0.1 ps and 0.5 ps was used to keep the temperature and pressure constant at 298.15 K and 1 atm with an estimated isothermal compressibility of 4.575 10^{-4} (kJ mol^{-1} $nm^{-3})^{-1}$. Solute and solvent were coupled to two separate temperature baths. The SHAKE [28] algorithm was used with a relative tolerance of 10^{-4} to keep the bond lengths constrained to their minimum-energy value, using a timestep of 2 fs. In this study we compare four different sets of simulations, that differ in the way the non-bonded interactions are calculated. In the first set of simulations, the non-bonded interactions were calculated using a group-based, twin-range cutoff scheme (CG/TR), with a short-range cutoff at 0.8 nm and a long-range cutoff at 1.4 nm. The short-range interactions were computed every timestep (2 fs) from a pairlist that was updated every 10 fs. The intermediate range interactions, up to the long-range cutoff were computed at pairlist updates and kept constant in between. A reaction-field contribution [15] was added to all electrostatic interactions to account for a homogeneous medium beyond the long-range cutoff with a relative dielectric constant of 61 [29]. In the second set of simulations, the frequency of the pairlist update and the calculation of intermediate-range interactions was set to every 2 fs, resulting in a single range pairlist scheme (CG/SR). In the third set of simulations, the cut-off was applied based on interatomic distances (AT/TR). For the fourth set of simulations, the protein was treated atomistic, by treating every atom of the solute as a separate charge-group, while the solvent was treated as in the charge-group simulations (solute-atomistic, SA/TR). Every protein system was simulated for all four cutoff schemes in triplicates for 15 ns, yielding in a total simulation time of around 10 μs.

For the simulations of pure SPC [30] water, a box of 1000 molecules was

simulated in analogy to the protein simulations. The isothermal compressibility was estimated at 7.51 10^{-4} (kJ mol^{-1} nm^{-3})$^{-1}$ and the relative dielectric constant of the reaction field was set to 78. The simulations were performed in triplicates for 10 ns each. Three different options for the cutoff were used, first an atomistic cutoff was used (AT), second a charge-group based cutoff scheme was used with the center of the charge-group being the center of geometry [CG(cog)] and in the third set of simulations the center of the charge-group was placed on the oxygen atom of the water [CG(OW)]. These simulations were performed with the TR cutoff scheme.

The analysis was performed on the last 5 ns of the simulation trajectory. Structural features were compared using the RMSD$_{100}$ proposed by Carugo and Pongor [31]. Attempting to correct for differently sized proteins the RMSD$_{100}$ normalizes the RMSD value to a protein of a 100 amino acid length. Hydrogen bond analysis was performed on the backbone of the protein. As geometric criterion an acceptor - donor distance below 0.25 nm and an acceptor - hydrogen - donor angle larger than 120° was applied. The solvent accessible surface area of the protein was split, by amino acid type, in a non polar (A,C,F,I,L,M,V,W,Y) and a polar (remaining residues) contribution. The radius of gyration is calculated according to equation 4.1 with $\mathrm{m_i}$ being the mass of atom i, $\mathrm{r_i}$ the position vector of every atom i and $\mathrm{r_{com}}$ as the position vector of the center of mass of all atoms. M is the total mass of the protein.

$$\mathrm{R_{gyr}} = \sqrt{\frac{1}{\mathrm{M}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{m_i}(\mathrm{r_i} - \mathrm{r_{com}})^2} \tag{4.1}$$

The occurence of secondary structure motives was assigned using the Dictionary of Secondary Structures of Proteins (DSSP), by Kabsch and Sander [32]. For the structures resolved by NMR experiments J-coupling constants and NOE intensities were also evaluated for the statistical comparison of the protein set. J-coupling constants were calculated via the related dihedral angle, using the empirical parameters for the Karplus relation proposed by Lindorff-Larsen [33]. Experimentally proposed NOE upper-bounds for inter-proton distances were compared to simulated distance averages, computed as $< \mathrm{r}^{-3} >^{-1/3}$ and using pseudoatom-corrections proposed by Wüthrich [34]. The technical

replicates of the simulation were pooled for this analysis. To investigate the structure of the solvent the radial distribution function (RDF) and the dipole-dipole orientation correlation function (DCF, $C(r)$, equation 4.2) was used with $\hat{\mu_i}$ the direction of the water dipole moment.

$$C(r) = \langle \hat{\mu_i}(R) \hat{\mu_j}(R + r) \rangle_R \tag{4.2}$$

To determine whether the variation of results obtained from different sets of simulations are significant, a mixed-model linear analysis was used as described in Setz et al [21, 22]. The p-values of the binary contrasts of the different metrics were adjusted using the Benjamini-Yekutieli correction for multiple testing [35].

Further investigations were conducted using the EGF domain of Spitz (PDB code: 3CA7). All different simulations were performed using 6 technical replicates and a simulation time of 15 ns. Apart from different cutoff schemes, a number of different reference temperatures were set for the temperature baths. Furthermore, two sets of simulations were conducted using a Nosé-Hoover chains thermostat with a chain length of 3. One set of simulations used particle-particle-particle-mesh (P3M) lattice summation to account for long-range electrostatics, using a real-space cutoff of 0.8 nm and a grid spacing of 0.12 nm. The data is represented as mean values with standard deviations over the last 5 ns of the simulation. To compare the individual means, a pairwise t-test was performed with a Holm-Bonferroni multiple testing correction [36].

Results and Discussion

Recent studies indicate that quantities obtained from molecular dynamics simulations depend on the treatment of the pairlist and the cutoff type [17, 16, 20]. While there is some debate that observed differences are due to the use of a twin-range cutoff scheme,[19] this does not follow from the data in table 4.1. No significant differences were observed for any of the analysed properties between the CG/TR and CG/SR sets of simulations.

property	CG/TR vs. CG/SR significance	p-value	CG/TR vs. AT/TR significance	p-value
$RMSD_{100}$	—	1	***	< 0.0001
Nr. of H-$bond_{backbone}$	—	0.3757	***	< 0.0001
$SASA_{polar}$	—	1	***	< 0.0001
Radius of gyration	—	1	**	0.0078
NOE violations[a]	—	1	*	0.0151
J-value[a]	—	1	—	0.8488
$SASA_{non-polar}$	—	0.14444	—	1
Occurrence of α-Helix	—	1	.	0.0780
Occurrence of π-Helix	—	1	—	0.5290
Occurrence of 3_{10}-Helix	—	1	—	1
Occurrence of β-Strand	—	1	—	1
Occurrence of β-Bridge	—	1	—	1

[a] NMR data is available for a subset of 13 proteins

Table 4.1: Statistical analysis on the significance of differences. P-values obtained from a multivariate multilevel analysis on 52 proteins with 3 replicates each.

On the other hand, significant differences are observed when comparing set CG/TR with set AT/TR for the $RMSD_{100}$, the number of backbone hydrogen bonds, the solvent accessible surface area of polar amino acid residues, the radius of gyration, the violations of NOE distances and the occurrence of α-helical structures. Figure 4.1 shows the $RMSD_{100}$ for all proteins. Simulations of 1ng6 show in general very high $RMSD_{100}$ values. This structure of a cytosolic protein of unknown function consists of two four-helix bundles with a relatively flexible linker. Interestingly, the use of our recently updated parameter set 54A8_bb significantly reduced the values of $RMSD_{100}$ [3]. For almost all proteins, the RMSD is higher in the case of the charge-group based cutoff scheme. The differences in RMSD could be traced to the temperature of the solvent and solute in both simulation sets. The cutoff noise in either simulations leads to deviations from the target temperature. The solvent and solute temperatures were always lower in the simulations that used an atomistic cutoff than the ones that used a group-based cutoff, as can be seen in Figure 4.2. The difference between the cutoff schemes was around 1.5 K for the solute degrees of freedom and only around 0.3 K for the solvent degrees of freedom. Though these differences are small they seem to affect the system and lead to significant differences in the properties indicated in Table 4.1.

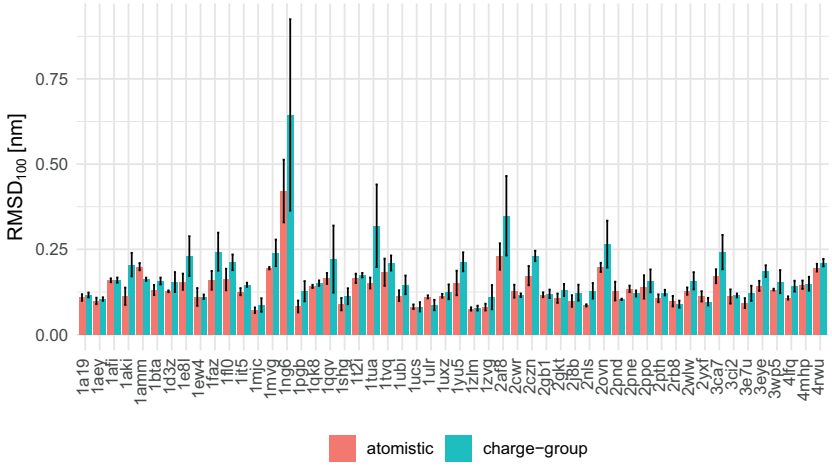

Figure 4.1: Average backbone $RMSD_{100}$ values of all proteins simulated. The error bars indicate the standard deviation from the mean of three replicates of one protein system.

To determine if these differences are specific for soluble, structured proteins, we also performed the same set of simulations for the unstructured pentapeptide Ala_5, see figure S1 in the supplementary material. While the solute temperatures are generally maintained better for such a small peptide, the deviations from the reference values are still smaller for the atomistic cutoff scheme. For the solvent, the deviations from the reference temperature are similar to the values in Figure 4.2. The SASA values for atomistic cutoffs are in general lower than for charge-group cutoffs and the radius of gyration and total number of backbone hydrogen bonds are very similar in all simulation settings.

To rationalise the differences we observed for the proteins, a simple one-dimensional system as described in Figure 4.3 was used to investigate the energies and forces at the cutoff. Two diatomic molecules with different charge distributions were placed at different intermolecular distances. Both molecules were 0.1 nm in size. The electrostatic interaction and force along the molecular axes were calculated using different cutoff schemes. The 'mixed' cutoff scheme is a combination of the atomistic and charge-group based schemes, where charges q1 and q2 are treated as atomistic and charges q3 and q4 as

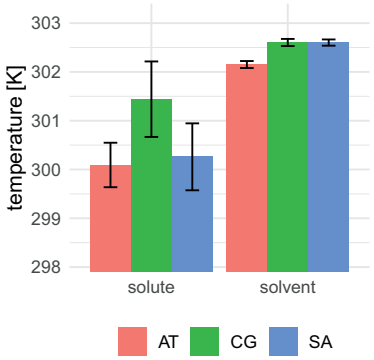

Figure 4.2: Average temperatures observed for solute and solvent degrees of freedom for atomistic (AT), charge-group based (CG) and solute atomistic (SA) cutoff schemes. The reference temperature was set to 289.15 K. The error bars indicate the standard deviation over all 156 simulations.

charge group. The interaction energy between two atoms i and j is calculated using:

$$V_{ij}^{el} = \frac{q_i q_j}{4\pi\epsilon_0}\left[\frac{1}{r_{ij}} + \frac{\frac{1}{2}C_{rf}r_{ij}^2}{R_{rf}^3} + \frac{1-\frac{1}{2}C_{rf}}{R_{rf}}\right] \tag{4.3}$$

where r_{ij} is the interatomic distance, C_{rf} is a reaction-field constant depending on the reaction-field dielectric constant and R_{rf} is the reaction-field cutoff distance [15]. The last distance-independent term, ensures that the electrostatic energy approaches zero when $r_{ij} = R_{rf}$.

In panels C-E of Figure 4.3, the energies occurring around the 1.4 nm cutoff were plotted, for dipole-dipole, dipole-charge and charge-charge interactions. This example shows that for an atomistic cutoff scheme, the overall energy goes to zero more smoothly than for charge-group based cutoffs. This can be explained from equation 4.3, which goes to zero if the interatomic distance $r_{ij} = R_{rf}$. However, in the group-based cutoff scheme, some atoms may no longer interact at distances shorter than R_{rf}, or still interact beyond this distance, leading to sudden jumps in the electrostatic interaction energy between the molecules. These sudden changes lead to larger cutoff noise, and hence demand more heat exchange with the temperature baths to maintain the temperature at the target value. Indeed, in previous work, we observed that the difference between AT and CG becomes smaller when using larger cutoffs, as the size of the energy jumps diminishes [17]. For the forces in panels F-

A

type	q1	q2	q3	q4
dipole–dipole	0.25	−0.25	0.5	−0.5
dipole–charge	0.25	−0.25	0.5	0.5
charge–charge	0.5	0.5	0.5	0.5

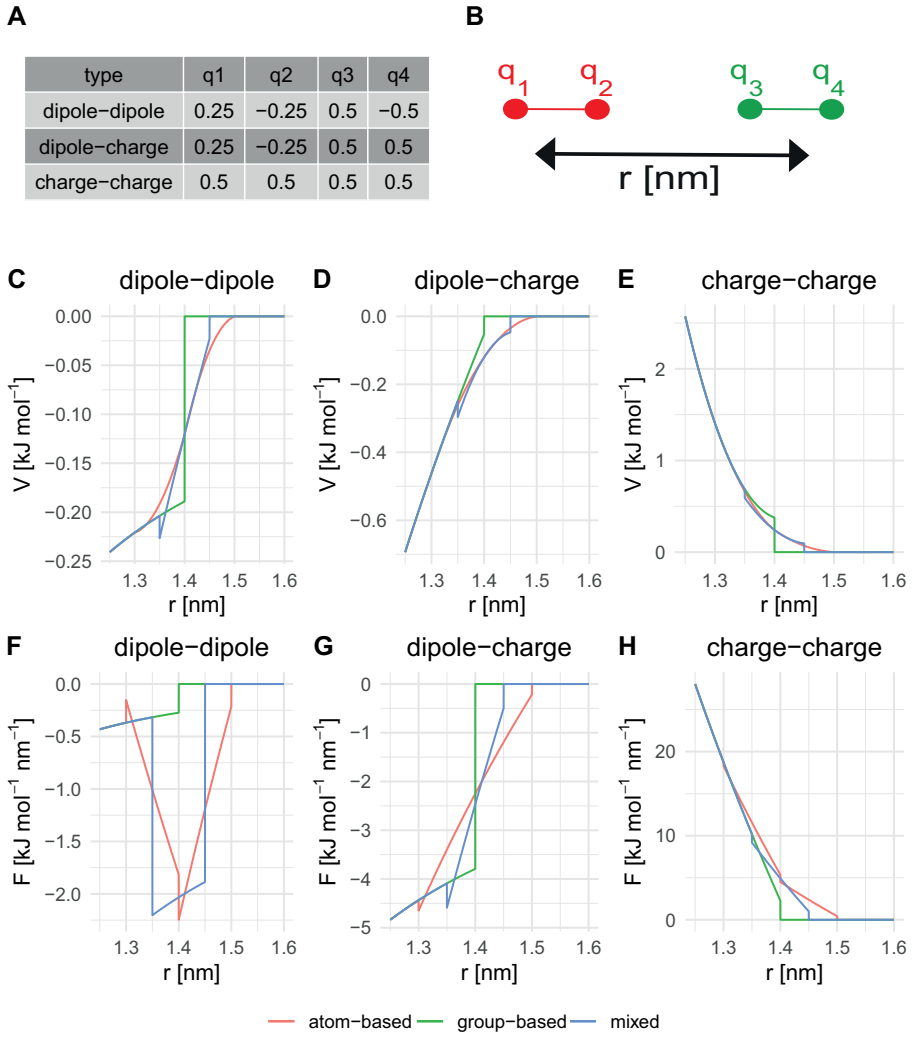

Figure 4.3: Simple one-dimensional system (panel B) to analyse the cutoff effect in detail. Two diatomic molecules are translated along the x-axis and molecular interaction energies and forces are computed, for charges according to panel A. Interaction energy in the cut-off region for dipole-dipole (C), dipole-charge (D), charge-charge (E) system. The same for the forces in the lower panels (F)-(H).

H of Figure 4.3, however, the dipole-dipole interaction leads to irregular spikes around the cutoff for the atomistic cutoff scheme. At distances where some atoms no longer interact, the molecular interaction changes to a dipole-charge or charge-charge interaction, with different slopes in the energy profile, and hence different forces. As the two molecules move further apart, the forces fluctuate strongly. The blue line of the mixed-cutoff scheme approximates the smooth energy profile of the atomistic cutoff scheme, and also shows the artificial spikes in the dipole-dipole forces.

The effect of the irregular forces in the atomistic cutoff scheme around the cutoff can be seen by analysing the radial distribution functions (RDF) and dipole correlation functions (DCF) for a box of 1000 SPC water molecules (Figure 4.4). The close-up of the RDF shows an artificial structure around the cutoff region for the simulations using an atomistic cutoff scheme. For the dipole correlation function a slight anti-correlation can be observed for the charge-group case, as was observed previously[13, 17, 18]. Different centers of the charge-group do not seem to have a major influence on the RDF and the DCF [compare CG(cog) and CG(OW)]. To ensure that this observation is not a peculiarity of the SPC water model, we have performed AT and CG(OW) simulations of the TIP4P water model, and find very similar artefacts around the cutoff (Figure S2 in supplementary material).

Following up on the mixed cutoff scheme in Figure 4.3, the 52 proteins were simulated using a cutoff scheme in which the protein atoms were treated as individual groups, while the solvent was treated using a group-based cutoff [CG(OW)]. Table 4.2 shows the differences in the monitored protein quantities. It can be observed that the solute atomistic cutoff set leads to the proteins behaving comparably to the atomistic case, except for the $\mathrm{SASA}_{\mathrm{polar}}$ which seems to be governed by the water being treated as charge-group. This is in agreement with the observations in Figure 4.3, where the mixed cutoff scheme is most similar to the atomistic scheme. Also for the temperatures in Figure 4.2, the solute behaves similar to the atomic case and the solvent similar to the charge-group case.

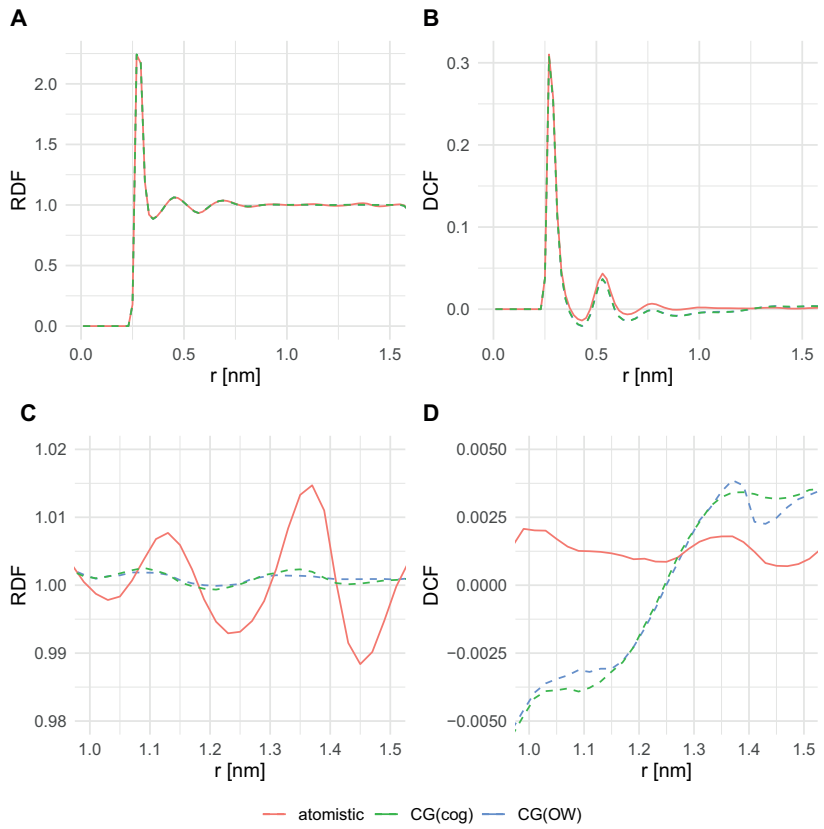

Figure 4.4: Radial distribution function of water oxygen atoms (A) and dipole correlations function for water (B). Panels (C) and (D) zoom in to the region around the cutoff (1.4 nm).

	AT/TR vs. SA/SR		CG/TR vs. SA/TR	
property	significance	p-value	significance	p-value
$RMSD_{100}$	—	1	***	< 0.0001
Nr. of H-$bond_{backbone}$	—	0.1382	**	0.0016
$SASA_{polar}$	***	< 0.0001	—	1
Radius of gyration	—	1	—	0.1382
NOE violations[a]	—	1	.	0.0151
J-value[a]	—	1	—	1
$SASA_{non-polar}$	—	0.5494	—	0.2303
Occurrence of α-Helix	—	1	**	0.0075
Occurrence of π-Helix	—	1	—	1
Occurrence of 3_{10}-Helix	—	1	—	0.8488
Occurrence of β-Strand	—	1	—	1
Occurrence of β-Bridge	—	1	—	1

[a] NMR data is available for a subset of 13 proteins

Table 4.2: Statistical analysis on the significance of differences. P-values obtained from a multivariate multilevel analysis on 52 proteins with 3 replicates each.

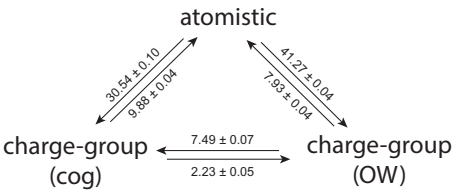

Figure 4.5: Change in non-bonded electrostatic energy in kJ mol^{-1} upon reanalysis of trajectories of pure water simulations using an atomistic cut-off (AT), a charge-group based cut-off using the center of geometry CG(cog) and using a charge-group based cut-off at the oxygen of the water molecule CG(OW).

Next, we turn our attention to the energetic differences between the different cutoff schemes. The potential energy was recalculated for the configurations that were obtained from simulations with one cutoff scheme, applying an alternative cutoff scheme. Figure 4.5 shows the resulting change in potential energy for the simulations of pure water. All values in this figure are positive, which follows from the fact that configurations are generated that are most favorable for the cutoff scheme used in the simulation. However, there is an asymmetry in the values. As can be seen in this picture the difference in energy going from an atomistic simulation to a charge-group based cutoff scheme is much more unfavourable than vice versa. This is due to the fact that in the atomistic case a higher water density is artificially observed before the cutoff. Furthermore, water molecules at the cutoff will orient themselves such that unfavourable interactions are placed out of the cutoff. Reintroducing these in a group-based recalculation subsequently leads to unfavourable interactions. This is in line with the differences in density and orientations at the cutoff as seen in Figure 4.4. Similarly, the much smaller differences between the CG(cog) and CG(OW) may be explained by the difference between the green and blue curves of the DCF at exactly 1.4 nm. Using the oxygen atom as the centre of the water molecules, leads to a slightly larger positive correlation just before the cutoff, followed by a slight drop in the correlation beyond the cutoff.

A similar recalculation of the potential energy was performed for the protein simulations. Figure 4.6 shows the change in energy from all three different

cutoff schemes reanalysed using the other schemes, separated into protein-protein, protein-solvent and solvent-solvent contributions. All contributions were normalised with respect to the number of atoms prior to averaging over the proteins. Again, the difference in energy seems always unfavourable, but statistical significance is only reached for few energy terms and simulation settings. The most pronounced difference in terms of energy is seen in the solvent-solvent interactions when recalculating a simulation that was performed with an atomistic cutoff to a (solvent) group-based cutoff. This is in line with the larger values for similar changes in Figure 4.5. We interpret this such, that the added structure in the solvent that is observed in the RDFs for atomistic cutoff simulations is also relevant in the protein simulations and should be avoided.

To test if the changes that are observed between atomistic and group-based cutoff schemes can be compensated by different settings of the temperature baths, the EGF domain of Spitz (PDB-Id: 3ca7) was simulated using eight different simulation settings. The settings are outlined in Table 4.3 and include the use of lower reference temperatures, the use of a Nosé-Hoover chains thermostat and the use of P3M for the long-range electrostatics. Every individual parameter set was simulated in sixtuples. The actual average temperatures observed in the simulations are also listed in this table. The P3M simulations show, that a complete removal of the cutoff noise, reduces the solute temperature close to the target, while the solvent temperature remains high. This suggests that the noise in the solvent is mainly due to another source, possibly related to the use of distance constraints. Figure 4.7 shows the effects for the properties for which significant differences were observed in Table 4.1. For the RMSD_{100} significant differences can be seen comparing the charge-group based cutoff to almost every other simulation setting. This confirms that a more exact temperature control can indeed reduce the RMSD_{100} values. For the radius of gyration there were no significant differences observed and for the number of hydrogen bonds the differences are only between the atomistic and charge-group based cutoff schemes. For $\mathrm{SASA}_{\mathrm{pol}}$ significant differences can be seen between the atomistic and all other simulated sets, except for the

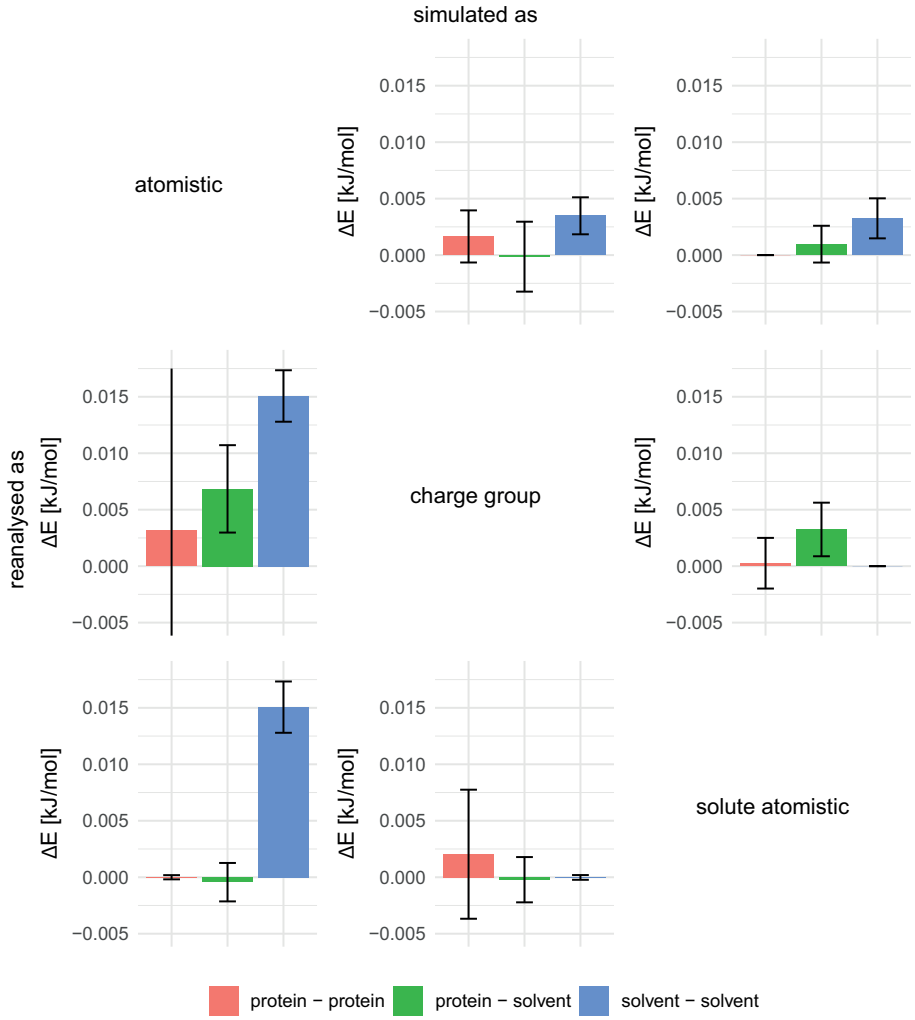

Figure 4.6: Differences in the non-bonded electrostatic energy upon reanalysis of protein trajectories. The protein-protein and solvent-solvent energies were normalized by the number of atoms in the respective sets (N_a^{solute} and $N_a^{solvent}$, respectively) and the protein-solvent interactions were normalized by $(N_a^{solute} \cdot N_a^{solvent})^{\frac{1}{2}}$.

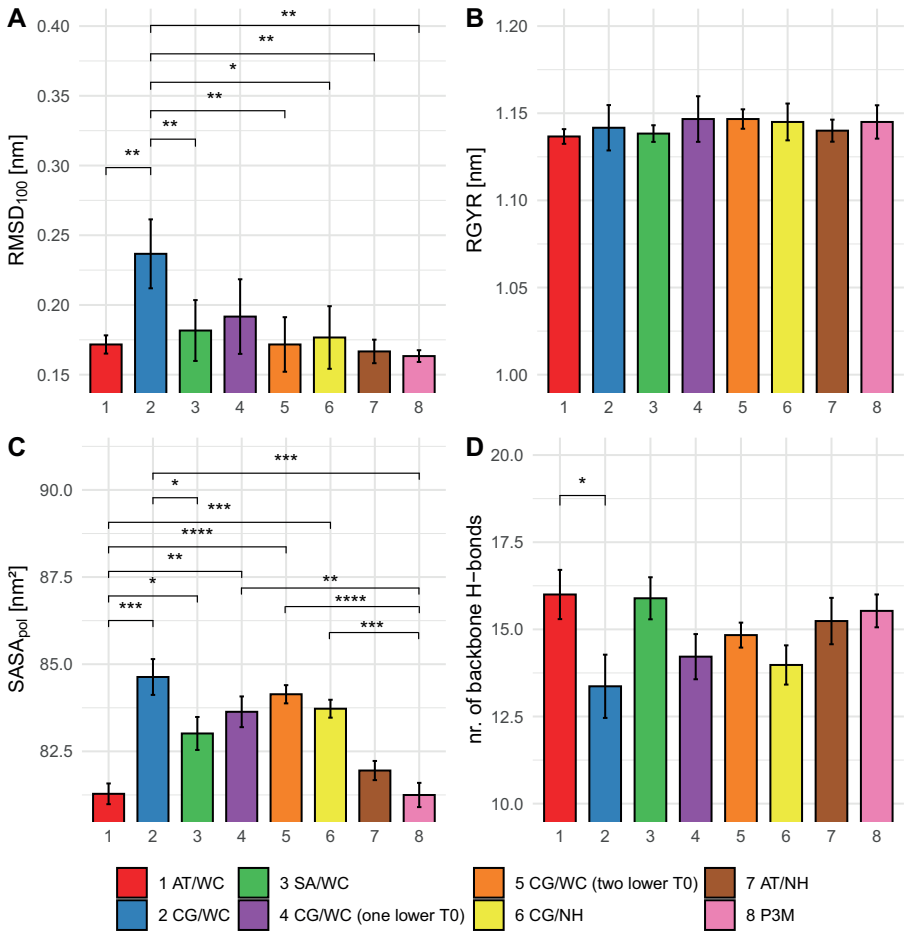

Figure 4.7: Simulations of Spitz EGF using seven different temperature settings (Table 4.3), with average RMSD_{100} in panel A, the radius of gyration in panel B, the $\mathrm{SASA}_{\mathrm{pol}}$ in panel C and the number of backbone hydrogen bonds in panel D.

atomistic simulations performed using a Nose-Hoover thermostat and the P3M simulations.

Figure 4.8 shows the water-water radial distribution function in the Spitz simulations. The overall downward trend in panel B can be explained from the fact that the protein occupies a considerable volume in the simulation box. The curves for the charge-group and atomistic cut-off schemes and temperature

nr.	cutoff	thermostat	reference solute temperature [K]	observed solute temperature [K]	reference solvent temperature [K]	observed solvent temperature [K]
1	AT	WC	298.15	299.42	298.15	302.27
2	CG	WC	298.15	301.75	298.15	302.63
3	SA	WC	298.15	299.76	298.15	302.69
4	CG	WC	295.85	299.28	295.85	300.33
5	CG	WC	296.02	299.79	297.75	302.27
6	CG	NH	298.15	299.01	298.15	300.28
7	AT	NH	298.15	298.34	298.15	300.13
8	P3M	NH	298.15	299.39	298.15	302.23

Table 4.3: Simulation settings for the additional sets of simulations of the EGF protein. Cutoff schemes used are atomistic (AT), charge-group based (CG), solute-atomistic (SA) or particle-particle-particle-mesh (P3M). Thermostats refer to weak-coupling (WC) or Nosé-Hoover chains (NH). In simulation sets 4 and 5, the reference temperatures were reduced to obtain observed temperatures closer to the target (set 4) or to the AT setup (set 5).

settings can be clearly distinguished. As expected, the artificial structure at the cutoff for AT simulations persists in the protein simulation, the effect of a more precise temperature control is minor, with the blue curve (CG/TR) slightly above the other CG curves. The SA scheme is indistinguishable from the CG schemes with a more precise solute temperature, in spite of the higher solvent temperature (Table 4.3). These data suggest that a close look at the temperature control of simulations remains an important check for any biomolecular simulation.

Conclusion

We described simulations of 52 protein systems, using four different cutoff and pairlist schemes. No significant differences were observed for any of the analysed properties when comparing the twin-range cutoff scheme to a single-range cutoff scheme. However, the choice of the entities to which the cutoff is applied (atomistic vs. group-based) does have a significant influence on some of the molecular properties. Investigations on pure water simulations show that using an atomistic cutoff leads to artificial structure in the water at the cut-

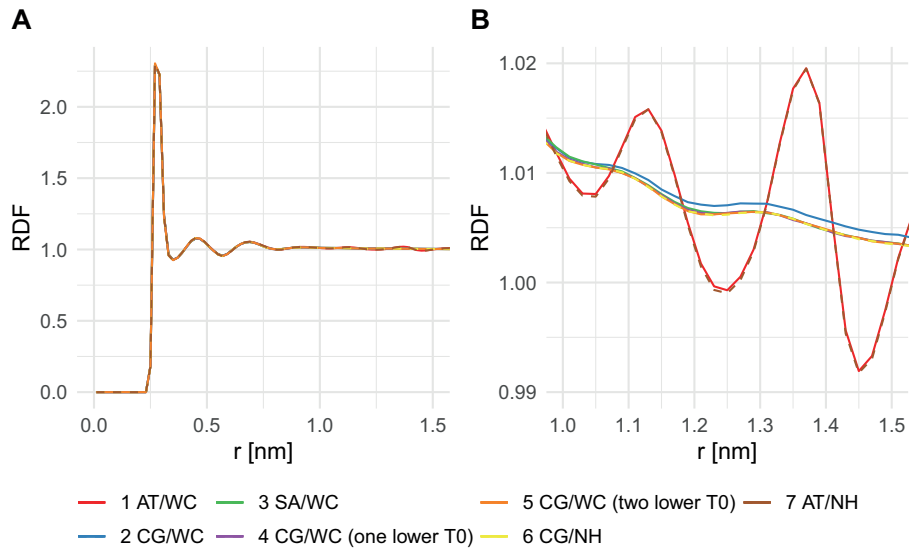

Figure 4.8: Radial distribution function of the solute obtained in the simulations of Spitz EGF in panel A and the zoomed representation in panel B. The dashed lines represent simulations where a Nose-Hoover thermostat was used.

off, whereas the water-dipoles seem to be slightly anticorrelated in the charge-group case. Re-analysis of simulations with alternative cutoff schemes suggests, that these structural effects also propagate into the energetics of the solvent in protein-water simulations. Simulations of the Spitz EGF protein suggest that proper control of the effective simulation temperature can remove the observed differences in the analysed properties. A solute-atomistic simulation scheme seems to have the same effect, leading to less noise in the protein degrees of freedom, while still avoiding the artificial structure of the solvent at the cutoff. This approach has the added advantage that the speed-up of using group-based water molecules can be maintained. Overall, we conclude that while the cutoff noise may be less with an atomistic cutoff, due to smoother energy curves, this comes as the expense of artificial structure in the solvent, due to irregular forces at the cutoff. A solute-atomistic cutoff scheme or simply a close look at the settings of the temperature baths is sufficient to control the charge-group based cutoff noise.

Supporting Information

Figure S1 of the supporting information shows Simulations of ALA5 using five different cutoff settings, with average temperature of the solute (light colours) and solvent (dark colours) in panel A, the radius of gyration in panel B, the solvent accessible surface area (SASA) in panel C and the number of backbone hydrogen bonds in panel D. And figure S2, the radial distribution function of water oxygen atoms (A) and dipole correlations function for water (B), comparing SPC and TIP4P water models. Panels (C) and (D) zoom in to the region around the cutoff (1.4 nm).

The supporting information for this chapter can be found under https://doi.org/10.1002/jcc.26426

References

[1] K. Lindorff-Larsen et al. “Picosecond to Millisecond Structural Dynamics in Human Ubiquitin”. In: *The Journal of Physical Chemistry B* 120.33 (2016), pp. 8313–8320.

[2] S. Riniker. “Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview”. In: *Journal of Chemical Information and Modeling* 58.3 (2018), pp. 565–578.

[3] M. Diem and C. Oostenbrink. “Hamiltonian Reweighing To Refine Protein Backbone Dihedral Angle Parameters in the GROMOS Force Field”. In: *Journal of Chemical Information and Modeling* 60.1 (2020), pp. 279–288.

[4] W. F. van Gunsteren et al. “On the Effect of the Various Assumptions and Approximations used in Molecular Simulation on the Properties of Bio-Molecular Systems: A Review of Issues”. In: *submitted* (2020).

[5] P. P. Ewald. “Die Berechnung optischer und elektrostatischer Gitterpotentiale”. In: *Ann. Phys.* 369.3 (1921), pp. 253–287. DOI: https://doi.org/10.1002/andp.19213690304.

[6] B. A. Luty et al. “A Comparison of Particle-Particle, Particle-Mesh and Ewald Methods for Calculating Electrostatic Interactions in Periodic Molecular Systems”. In: *Molecular Simulation* 14.1 (1994), pp. 11–20.

[7] W. Weber, P. H. Hünenberger, and J. A. McCammon. “Molecular Dynamics Simulations of a Polyalanine Octapeptide under Ewald Boundary Conditions: Influence of Artificial Periodicity on Peptide Conformation”. In: *The Journal of Physical Chemistry B* 104.15 (2000), pp. 3668–3675.

[8] M. A. Kastenholz and P. H. Hünenberger. “Influence of Artificial Periodicity and Ionic Strength in Molecular Dynamics Simulations of Charged Biomolecules Employing Lattice-Sum Methods”. In: *The Journal of Physical Chemistry B* 108.2 (2004), pp. 774–788.

[9] M. A. Kastenholz and P. H. Hünenberger. “Computation of methodology-independent ionic solvation free energies from molecular simulations. I. The electrostatic potential in molecular liquids”. In: *The Journal of Chemical Physics* 124.12 (Mar. 2006), p. 124106.

[10] M. M. Reif and C. Oostenbrink. “Net charge changes in the calculation of relative ligand-binding free energies via classical atomistic molecular dynamics simulation”. In: *Journal of Computational Chemistry* 35.3 (Nov. 2013), pp. 227–243.

[11] M M Reif and Chris Oostenbrink. “Toward the correction of effective electrostatic forces in explicit-solvent molecular dynamics simulations: restraints on solvent-generated electrostatic potential and solvent polarization”. In: *Theoretical Chemistry Accounts* 134.2 (2015), p. 2.

[12] S. E. Feller et al. “Effect of Electrostatic Force Truncation on Interfacial and Transport Properties of Water”. In: *The Journal of Physical Chemistry* 100.42 (1996), pp. 17011–17020.

[13] P. H. Hünenberger and W. F. van Gunsteren. “Alternative schemes for the inclusion of a reaction-field correction into molecular dynamics simulations: Influence on the simulated energetic, structural, and dielectric properties of liquid water”. In: *The Journal of Chemical Physics* 108.15 (1998), pp. 6117–6134.

[14] M. Neumann. “Dipole moment fluctuation formulas in computer simulations of polar systems”. In: *Molecular Physics* 50.4 (1983), pp. 841–858.

[15] I. G. Tironi et al. “A generalized reaction field method for molecular dynamics simulations”. In: *J. Chem. Phys.* 102.13 (Apr. 1995), pp. 5451–5459. DOI: 10.1063/1.469273.

[16] S. Reißer et al. “Real Cost of Speed: The Effect of a Time-Saving Multiple-Time-Stepping Algorithm on the Accuracy of Molecular Dynamics Simulations”. In: *Journal of Chemical Theory and Computation* 13.6 (2017), pp. 2367–2372.

[17] F. D. Silva et al. “The Impact of Using Single Atomistic Long-Range Cutoff Schemes with the GROMOS 54A7 Force Field”. In: *Journal of Chemical Theory and Computation* 14.11 (2018), pp. 5823–5833.

[18] M. Diem and C. Oostenbrink. “The effect of using a twin-range cut-off scheme for non-bonded interactions: Implications for force-field parameterization?” In: *Journal of Chemical Theory and Computation* 16 (2020), online, doi: 10.1021/acs.jctc.0c00509.

[19] B. Hess et al. “On The Importance of Accurate Algorithms for Reliable Molecular Dynamics Simulations”. In: *ChemRxiv* (Dec. 2019). DOI: 10.26434/chemrxiv.11474583.v1.

[20] Y. M. H. Gonçalves et al. “Influence of the Treatment of Nonbonded Interactions on the Thermodynamic and Transport Properties of Pure Liquids Calculated Using the 2016H66 Force Field”. In: *Journal of Chemical Theory and Computation* 15.3 (2019), pp. 1806–1826.

[21] Setz M. *Molecular dynamics simulations of biomolecules: from validation to application*. Vienna: Dissertation, BOKU, 2018, p. 285. DOI: https://permalink.obvsg.at/AC15159105.

[22] M. Stroet et al. “Challenges associated with the validation of protein force fields based on structural criteria.” In: *submitted to J Chem Inf Model* (2020).

[23] N. Schmid et al. “Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation”. In: *Comp Phys Commun* 183.4 (Apr. 2012), pp. 890–903. DOI: 10.1016/j.cpc.2011.12.014.

[24] M. M. Reif, P. H. Hünenberger, and C. Oostenbrink. “New interaction parameters for charged amino acid side chains in the GROMOS force field”. In: *J. Chem. Theo. Comp.* 8.10 (Oct. 2012), pp. 3705–3723. DOI: 10.1021/ct300156h.

[25] M. M. Reif, M. Winger, and C. Oostenbrink. “Testing of the GROMOS force-field parameter set 54A8: Structural properties of electrolyte solutions, lipid bilayers, and proteins”. In: *J. Chem. Theo. Comp.* 9.2 (Feb. 2013), pp. 1247–1264. DOI: 10.1021/ct300874c.

[26] A. Amadei et al. “Molecular dynamics simulations with constrained roto-translational motions: Theoretical basis and statistical mechanical consistency”. In: *The Journal of Chemical Physics* 112.1 (2000), pp. 9–23.

[27] H. J. C. Berendsen et al. “Molecular dynamics with coupling to an external bath”. In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3684–3690.

[28] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. “Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes”. In: *J. Comp. Phys.* 23.3 (Mar. 1977), pp. 327–341. DOI: 10.1016/0021-9991(77)90098-5.

[29] T. N. Heinz, W. F. van Gunsteren, and P. H. Hünenberger. “Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations”. In: *J Chem Phys* 115.3 (July 2001), p. 1125. DOI: 10.1063/1.1379764.

[30] J. Hermans et al. “A consistent empirical potential for water–protein interactions”. In: *Biopolymers* 23.8 (1984), pp. 1513–1518.

[31] O. Carugo and S. Pongor. “A normalized root-mean-spuare distance for comparing protein three-dimensional structures”. In: *Protein Science* 10.7 (2001), pp. 1470–1473.

[32] Wolfgang Kabsch and Christian Sander. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. In: *Biopolymers* 22.12 (Dec. 1983), pp. 2577–2637.

[33] K. Lindorff-Larsen, R. B. Best, and M. Vendruscolo. “Interpreting Dynamically-Averaged Scalar Couplings in Proteins”. In: *Journal of Biomolecular NMR* 32.4 (Aug. 2005), pp. 273–280.

[34] K. Wüthrich, M. Billeter, and W. Braun. “Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-proton distance constraints with nuclear magnetic resonance”. In: *Journal of Molecular Biology* 169.4 (Oct. 1983), pp. 949–961.

[35] Y. Benjamini and D. Yekutieli. “The control of the false discovery rate in multiple testing under dependency”. In: *The Annals of Statistics* 29.4 (2001), pp. 1165–1188.

[36] S. Holm. “A Simple Sequentially Rejective Multiple Test Procedure”. In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70.

Chapter 5

Final Conclusions & Outlook

In an ideal world simulation results would be independent of the force field used to simulate the system, but at the current state of the field this is unfortunately not entirely true. In this thesis various aspects and effects of force fields on simulations of biomolecular systems are analysed, discussed and if necessary updated, in order to obtain more reliable and robust results from simulations of biomolecular systems.

To achieve this goal in chapter 2 an update of the protein backbone dihedral angle parameter set is performed. By using experimental data of small blocked di-peptides as target values for mathematical and statistical optimization schemes a whole new description of the φ and ψ angles in the protein backbone was found for all amino acids. Allowing the shift of the potential energy curve to take up any value from -180° to180° increased the flexibility of this approach. The large scale validation performed on a set of 52 protein systems gave a very detailed insight in the features of the new system and made sure that the new 54A8_bb parameters for the dihedral angle backbone pose a reasonable update over the old ones. The new set features a better ratio of the $\mathrm{PP_{II}}$ region and the β region in the φ coordinate and the removal of an artificial shoulder around 90° in the ψ coordinate of the Ramachandran plot.

The infinite nature of Coulomb interactions poses a tough challenge in molecular dynamics simulations ever since, their calculation and the generation of the pair-list consumes most of the computational resources. Over time several clever schemes were developed to allow for more efficient calculations of these interactions, in chapter 3 recent concerns are addressed, that these influence the outcome of a simulation in a significant way. By repeating key experiments performed to parameterize the GROMOS force fields, no significant impact of using a single or twin range cut-off scheme could be observed, the deviations are within the experimental error.

Picking up the results of chapter 3 in chapter 4 the impacts of different cut-off schemes on proteins in molecular dynamics simulations are investigated. The results from chapter 3, that single-range and twin-range cut-off schemes do not influence the simulation results obtained, could be confirmed. But the large set of proteins showed significant differences depending on cut-off choice. After tracing the root of differences down to slightly different temperatures, a de-

tailed investigation of a simple 1-dimensional system of 4 charges gave insight in the dynamics at the cut-off region. Since the charge group-based cut-off on the one hand leads to a higher cut-off noise, the atomistic cut-off on the other hand the leads to artificial structure of the water around the cut-off. Unfortunately, both effects are not desirable, one proposed fix is the use of a charge-group based cut-off for water in the system and an atomistic for the solute, removing the artificial water structure and decrease the temperature difference. But in order to get the most reliable results the use of a more stringent temperature control is advised and tested in systems of the epidermal growth factor receptor of spitz.

MD simulations are on the one hand limited by an infinite sampling and on the other hand by imperfections in the accuracy of the force field. In a nutshell the work performed in this thesis is another small step towards better molecular dynamics force fields and more reliable simulations. The future of force field development will critically depend on proper validation. The increase in computational power will enable the use of bigger sets of biomolecules with sufficient replicates to validate updates in the parameters and to determine the accuracy and performance of the force field overall. With the gain of importance of intrinsically disordered proteins and the trend to simulate larger and more complex systems in general the demands force fields have to fulfil will rise. This development will call for more automatized ways of parameter search, as described in chapter 2, where data from wet-lab experiments and ab-initio quantum mechanical calculations will be combined to find better parameter sets. With the increase of computational power and the development of more refined algorithms it will be possible to compute sufficiently large systems with large cut-offs that only have a very limited effect on the simulation. The overall goal should be to obtain the same simulation results independent of the force field used.

Curriculum Vitae

Name: Matthias Diem
Nationality: Austrian
Email: matthias.diem@gmail.com

Areas of specialisation

Biomolecular Simulation; Force Field Parameterization; Statistics;

Technical skills

Bash; Python; R; Gromos; Gromacs; Pymol; VMD; MOE; Linux; Windows; Tensorflow;

Education

2017 *MSc* in Biotechnology (emphasis on Bioinformatics), University of Natural Resources and Life Sciences, Vienna, Austria

2015 *BSc* in Food Science and Biotechnology, University of Natural Resources and Life Sciences, Vienna, Austria

Research experience

2017 - 2020 Graduate Researcher: Optimization and Validation of force-field parameters for molecular dynamics simulations; Advisor: Chris Oostenbrink

2015 - 2017 Master Thesis: The effect of changes in the environment on the stability and surface properties of proteins. Studied by in silico methods; Advisor: Chris Oostenbrink

2013 - 2015 Bachelor Thesis: The effects of post-translational modifications on helical structure elements of peptides. A computational study; Advisor: Chris Oostenbrink

Conferences

2019 Evolving protein backbone parameters using Hamiltonian reweighing;
Algorithms, Biomolecules and Computers,
Mar 17, 2019, Vienna, Austria (Talk)

2019 Some like it hot - comparing atomistic and charge-group cut-off schemes;
29th Intl. BIOMOS Symposium on Biomolecular Simulation,
Sep 11-13, 2019, Ausserberg, Switzerland (Talk)

2018 Evolving protein backbone parameters using Hamiltonian reweighing;
28th Intl. BIOMOS Symposium on Biomolecular Simulation,
Sep 12-14, 2018, Ausserberg, Schweiz (Talk)

2018 Optimising protein backbone parameters combining Hamiltonian reweighing and elaborate search schemes;
14th Greta Pifat Mrzljak International School of Biophysics ,
Aug 23- Sep 1, 2018, Split, Croatia (Poster)

2018 The effect of changes in the environment on the structural stability of model proteins;
CECAM workshop: Proteins in realistic environments,
May 23-25, 2018, Stuttgart, Germany (Poster)

Teaching

2017 - 2020 Teaching Assistant. Modeling and Simulation of Biomolecules.
University of Natural Ressources and Life Science, Vienna, Austria