Kumulative
Habilitationsschrift über

Biomolecular Simulations

From mechanics of a blood protein to peptides in isolation to molecular structure sampling

zur Erlangung der *venia legendi*
für Biochemie und Theoretische Chemie
am Fachbereich Biologie, Chemie, Pharmazie der
Freien Universität Berlin

vorgelegt von
Dr. Carsten Baldauf
im Juni 2016

Fritz-Haber-Institut
der Max-Planck-Gesellschaft, Berlin

Acknowledgements

I had and still have the honor of working with excellent scientists. At **Universität Leipzig**, I have learned a lot from Prof. Hans-Jörg Hofmann and Dr. Robert Günther.
While working at **BIOTEC Dresden**, I made my first independent steps as a postdoctoral researcher under the supervision of Dr. Mayte Pisabarro. I enjoyed working and socializing with Mandy Erlitz and Dr. Jens Lättig. My next stop was **MPG-CAS PICB Shanghai** where I got to know Dr. Wolfram Stacklies, Prof. Scott Edwards, Marko Briesemann, Prof. Fei Xia, and Dr. Senbo Xiao. I am glad that I still meet all of them occasionally. I want to especially thank Prof. Frauke Gräter and Prof. Reinhard Schneppenheim for scientific discussion and support.
I paid a visit to Prof. Alfredo Alexander-Katz's lab at **MIT** and got to know Prof. Matthias Schneider. We were talking a lot in the lab and elsewhere. Thanks to that I became part of the **SHENC group** and have the honor of working with great colleagues.
Then I moved to Berlin and joined the Theory department of **Fritz Haber Institute**. I am grateful to Prof. Matthias Scheffler for providing an excellent and challenging research environment. I learned from Prof. Volker Blum (now Duke University) what it means to take a second, third, fourth, *etc.* look at your own work and after a while I even understood that it makes sense. I am glad that we still work together. In the Bio Group and beyond, I met excellent scientists and friendly humans, among others I want to thank Dr. Matti Ropo, Dr. Franziska Schubert, Dr. Mariana Rossi, Dr. Mateusz Marianski, Adriana Supady, Markus Schneider, Teresa Ingram, Arvid Ihrig, Dr. Lydia Nemec, Dr. Mathis Gruber, Dr. Viktor Atalla, and Dr. Luca Ghiringhelli. I thank Dr. Gert von Helden (FHI) and Prof. Kevin Pagel (FU Berlin) for the long-standing collaboration and for fruitful (*sic!*) discussions. At some point I started to work with the group of Prof. Pavel Hobza at **IOCB Prague**. I am still impressed by the friendly atmosphere in this excellent research group. I am grateful to Dr. Adam Pecina for keeping his enthusiasm about our joint project. I found support of my academic career at **Freie Universität Berlin** from Prof. Beate Koksch who gave me the opportunity to teach a course in bioorganic chemistry. I am also extremely grateful to Prof. Beate Paulus, first of all for the challenge of the lecture *Atombau und chemische Bindung*, but also for her support with the habilitation process.
Two colleagues have been with me throughout my scientific life: Dr. René Meier (Leipzig) and Prof. Daniel Merkle (Odense). Thank you for your support and friendship.

Berlin, June 2016 C. B.

Abstract

My research interest are biomolecular structure and dynamics. In essence, I want to contribute to an understanding of how the machinery of life works. Proteins and peptides play a central role in most of life's manifestations. I study their properties at different levels of detail and with varying levels of theory.

The molecular mechanics of the blood protein von Willebrand factor (VWF) and the regulation of its function by mechanical stimuli may serve as an entry point. Here we see a gigantic protein at work that acts as a shear-flow sensor in hemostasis. At the level of protein domains, I investigate the effect of tensile force on protein structure. Partial unfolding or interface opening as a response to a stretching force regulates VWFs blood-clotting activity.

In order to get a deeper insight into the rules that shape proteins, I study their structural building blocks, i.e. comparably short sequences that form helices, strands, or turns, in isolation. The empirical force fields that are standardly used to simulate large systems, for example the VWF simulations, fail here and I resort to the first principles of density-functional theory (DFT). Especially by comparison to ion mobility spectrometry and gas-phase infrared spectroscopy one can validate the accuracy of a simulation approach. Furthermore, the combination of computational and experimental infrared spectroscopy can be used for structure elucidation of peptides in the gas phase. By that I can take a close and unperturbed look at the interactions that shape polypeptides.

The first step in order to predict molecular properties is often the prediction of a molecular structure or the structure of a complex of two or more molecules. This represents a high-dimensional search problem. I discuss this on the example of molecular docking, where I present a search method to predict protein-ligand complexes and where we assessed the accuracy of commonly-used empirical energy function in comparison to semi-empirical quantum mechanics. Furthermore, I present a systematic first-principles based search across chemical space for amino acids and their complexes with divalent cations and a genetic algorithm implementation to perform DFT-based structure searches for medium-sized bioorganic molecules.

Zusammenfassung

Das Hauptaugenmerk meiner Arbeit liegt auf der Untersuchung der Struktur und Dynamik von Biomolekülen, ich möchte also mit meiner Forschung zum Verständnis der Prozesse in lebenden Organismen beitragen. Proteine und Peptide spielen hier eine zentrale Rolle, da sie an faktisch allen Lebensäußerungen beteiligt sind. Ich untersuche ihre Eigenschaften mit verschiedenen Methoden der Computer-gestützten Chemie.

Von Willebrand Faktor (VWF) ist das größte extrazelluläre Protein im menschlichen Körper und nimmt eine Schlüsselstellung in der Hämostase ein. VWF fungiert unter anderem als Scherflusssensor: hinreichend hohe Scherraten strecken das Riesenmolekül. Ich untersuche mit Hilfen von Kraftfeld-basierten Molekulardynamiksimulationen wie die resultierende Streckkraft entlang der Proteinkette nun die Blutgerinnungsaktivität des VWF durch teilweise Entfaltung von speziellen Domänen bzw. durch das Eröffnen von Wechselwirkungen zwischen benachbarten Domänen reguliert.

Zur Erforschung der grundlegenden Regeln der Protein- und Peptidstrukturbildung untersuche ich Sekundärstrukturbausteine, also kurze Sequenzen die entweder Helices, Bänder oder Umkehrschleifen bilden, in der Gasphase. Empirische Kraftfelder, wie sie zur Simulation von Proteinen häufig genutzt werden, versagen hier und ich verwende stattdessen *first principles* Methoden, vor allem Dichtefunktionaltheorie (DFT). Besonders die Möglichkeit zum Vergleich mit experimentellen Daten aus Ionenmobilitäts-Spektrometrie oder Gasphasen-Infrarotspektroskopie erlaubt es die Genauigkeit der Simulationstechniken kritisch zu beurteilen. Darüberhinaus ergibt sich aus der Kombination von theoretischer und experimenteller Spektroskopie eine Methode zur Strukturaufklärung die es mir ermöglicht die intrinsische, ungestörte Strukturbildung von Polypeptide zu untersuchen.

Der erste Schritt zur Berechnung von Eigenschaften ist zu meist die Vorhersage der Struktur eines Moleküls oder eines molekularen Komplexes. Ich diskutiere Möglichkeiten zur Lösung dieses hochdimensionalen Suchproblems unter anderem am Beispiel der Vorhersage von Protein-Ligand-Komplexen durch *molecular docking*, der Suche nach strukturellen Trends unter den proteinogenen Aminosäuren auf Basis einer *first principles* basierten Struktursuche und anhand der Implementierung einer globalen DFT Struktursuche mit Hilfe genetischer Algorithmen für bioorganische Moleküle.

Contents

1 Introduction and overview

The molecular machinery of life works at length and time scales that are hard to grasp as they differ so much from our normally experienced environment [1]. When taking a macroscopic view on matter, things may often appear solid and seem to essentially not change over years, while in the nanoscopic world of biomolecules everything appears to be in constant movement [2]. However, this seemingly chaotic scenario follows rules and by that enables the manifestation, progression, and evolution of life as we know it. The work that I summarize here deals with limited efforts towards an understanding of the basics of structure formation and dynamics of biomolecules. My tools are molecular simulations at different accuracy levels of description, from the atomistic molecular mechanics viewpoint of force fields to more elaborate electronic structure theory calculations by means of density-functional theory or high-level quantum chemistry.

In order to keep this introduction concise, I decided not to repeat the introductory remarks and references that the reader will find in the respective sections of the attached publications. Instead, this chapter gives a short overview of my work. The chapters that follow the introduction summarize three broader research areas of mine represented by selected publications:

Chapter 2 introduces important aspects of the mechano-regulation of the multimeric blood protein von Willebrand factor (VWF): (i) VWF function is dependent on the multimer length. The protease ADAMTS13 cleaves a binding site that is only accessible upon mechanical activation by partial unfolding of the VWF A2 domain [CB1]. (ii) In this unfolding process, the extended peptide chain is under tension. We investigate the isomerization of the prolyl *cis*/*trans* peptide bonds under stretching force as a possible re-folding timer [CB2]. (iii) By the same stretching force along the polypeptide chain, the interactions between neighboring domains can be broken. The interaction between the neighboring A1 and A2 domains auto-inhibits the interaction between VWF-A1 and the platelet receptor GPIb in a force-sensitive fashion [CB3].

Chapter 3 deals with examples of peptide foldamer structure formation studied in a col-

laborative effort by gas-phase spectroscopy and molecular simulations. The aim is a detailed understanding of the principles that govern structure formation of the basic building blocks, i.e. secondary structure elements like helices, strands, and turns, of the large proteins that were studied in Chapter 2. Theoretical structure/energy predictions in conjunction with gas-phase experiments enable on the one hand the unambiguous structure assignment and on the other a rigorous assessment of the accuracy of the applied methods. In that spirit, I included three manuscripts in Chapter 3: (i) a recent review that I have written together with Mariana Rossi (EPF Lausanne) that introduces the experiments and simulation approaches [CB4], (ii) a comparative study of helix forming α and β peptides in the gas phase [CB5], and (iii) a study of the impact of Li^+ and Na^+ cations on the structure and dynamics of short prototypical turn-forming peptides [CB6].

Chapter 4 discusses the sampling and presentation of potential-energy surfaces of molecules and molecular complexes with different approaches and based on different flavors of energy functions. I include four manuscripts from the area of global structure search in Chapter 4 that describe (i) the implementation the bio-inspired particle-swarm optimization search method for molecular docking [CB7], (ii) an assessment of the accuracy of commonly used scoring function in molecular docking in comparison to semi-empirical quantum mechanics [CB8], (iii) an implementation of a genetic algorithm as a global search technique for the use with first-principles methods [CB9], and (iv) an application of global structure search to investigate a region of chemical space, namely the proteinogenic amino acids in isolation or interacting with a divalent cation [CB10].

1.1 Biomolecules and how to describe their energetics

There are three main classes of biomolecular oligomers and polymers, namely nucleic acids (Figure 1.1A), peptides and proteins (see Figure 1.1B), and carbohydrates (Figure 1.1C). In the following, each of these classes will be briefly introduced. Due to my research interest, the reader will notice a bias towards gas-phase investigations. The investigation of molecules in isolation allows for a detailed look at their intrinsic properties and serves as a references point to, for example, estimate the impact of solute-solvent interactions.

Nucleic acids are carriers of genetic information and act as coenzymes in biochemical reactions. Furthermore, they might also have played a key-role in chemical evolution at the postulated RNA-world stage [3, 4, 5]. Indeed, nucleic acids can act as catalysts, information storage, and as an energy source. In todays living organisms, a sequence of nucleotides in deoxyribose nucleic acids (DNA) can be transcribed into ribose nucleic acids (RNA) that then serves as template for the stepwise linkage of the amino acids into a peptide or protein chain. This process, and flow of information, is known as central dogma of molecular biology. Nucleic acids feature a sugar-phosphate backbone with nucleobases

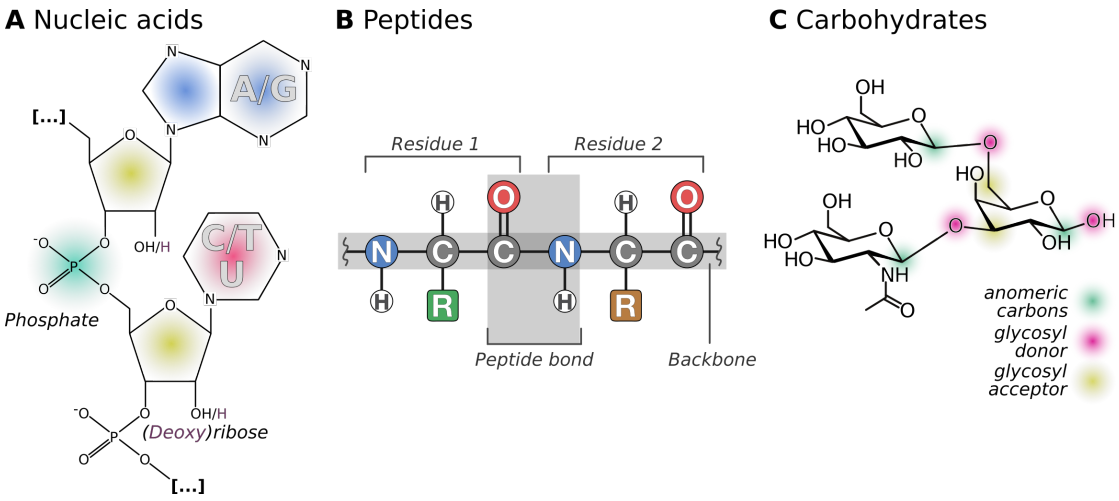


Figure 1.1 – Schematic chemical structures of the three dominant classes of biopolymers: **A**) nucleic acids, **B**) peptides, and **C**) carbohydrates.

connected to the (deoxy)ribose moieties (see Figure 1.1A for a pictorial representation of the different groups). Structure formation is mainly triggered by stacking of base pairs and by intermolecular hydrogen bonding between specific pairs of bases (base pairing) in case of DNA or intramolecular base pairing in case of RNA. Gas-phase studies allow to decipher the basics of these interactions in great detail, a recent review by Abi-Ghanem and Gabelica [6] can serve as entry point to the literature about nucleic acids in the gas phase.

Peptides and proteins make up the machinery of life and are involved in essentially all of its manifestations, from comparably small signaling peptides to gigantic protein complexes. A peptide or protein is a linear chain (oligomer) of amino acids (residues) that are linked by so-called peptide bonds (see Figure 1.1B). Besides the amino and carboxy groups that form the peptide bonds, the different amino acids carry a side chain '**R**' with differing chemical functionality. The sequence of the different amino acids that are linked to form a peptide or protein is called primary structure. Based on the length of such a sequence, shorter oligomers are called peptides, while oligomers beyond a certain length (from about 50 amino acids on) are called proteins. Secondary structure formation occurs at the level of peptides (Figure 1.2) and is mainly dependent on the conformational properties of the monomers and backbone hydrogen bonding. In larger oligomers, i.e. in proteins, side chain interactions and packing gain importance and govern tertiary structure formation. Christian Anfinsen postulated that, at least for small proteins, the native structure is fully encoded in the amino acid sequence [7]. For a functional protein, under its natural solvent, pH, and temperature conditions, the native state is uniquely stable, is robust with respect to small perturbations of the environmental conditions, and must be kinetically accessible. The latter means that the free energy path from an unfolded state to the native state must be downhill in energy and without too high barriers. Gentle treatment during ionization kinetically traps them in the solution state and allows for their interrogation under clean-room conditions.

Recent reviews introduce the field with a focus on peptides [CB4] or on large proteins and assemblies thereof [8].

Polymeric carbohydrates serve as nutrition and energy source or as structural scaffolds. Complex carbohydrates, a.k.a. **glycans**, can be linked to proteins where they promote folding and function as recognition tags [9, 10]. The monomeric building blocks are connected by glycosidic bonds (see Figure 1.1C) to form polymeric or complex carbohydrates. Diversity here stems not only from the available about 20 different monosaccharide units. In contrast to the backbones of peptides or nucleic acids, carbohydrates are not necessarily composed as linear chains. The building blocks have one donor (the anomeric C) but multiple acceptors for glycosidic bonds, such that branched structures can be realized. In addition, due to chirality, glycosidic bonds can be formed in two chiral forms: the α and β enantiomers. These features result in a diversity of possible topologies of carbohydrates that surpasses the number of possible sequences in nucleic acids and peptides by orders of magnitude, even with relatively small numbers of building blocks [11]. The significant conformational degrees of freedom are rotations around the single bonds of the glycosidic linkages and the puckerings of the monosaccharide rings [12].

A central focus of my work has been on the **secondary structure formation and dynamics in peptides**. The respective secondary structure elements, i.e. helices, pleated-sheets, and turns (Figure 1.2) [13], form already in peptides of a few to some tens amino acids of length. Turns are non-periodic motifs, while helices and sheets are regarded as periodic, in the sense that a repeating unit can be defined, allowing for a characterization based on pairs of torsional angles. Helix nomenclature is based on the periodic hydrogen bonding patterns between close non-nearest neighbor building blocks. This will be discussed in Section 1.3, see also Figure 1.7 there. The most common types, the α and the 3_{10} helices, are characterized by H bonds between residue i to $i+4$ and residue i to $i+3$, respectively. Sheets are H bonded extended strands that are characterized as parallel and anti-parallel depending on the relative orientation of their peptide chains. Finally, turns are important for higher-order structures as they reverse the propagation direction of sheets and helices, so that compact structures can be formed [14, 15, 16]. Turns do not necessarily form H bonds, yet H bonds are a common feature among them. The most common are β turns, which cause a 180^o change in the propagation direction.

Computational studies allow us to investigate structure and dynamics of peptides and proteins, nucleic acids, carbohydrates *etc.* in atomistic and electronic detail and can provide a link to biochemical experiments and biophysical measurements. Practical **biomolecular simulations** aiming at the prediction of experimental observables have to balance:

- the required accuracy of the underlying description of the potential energy to predict,
- the eventually immense computational costs of simulating systems of biomolecular size (hundreds to thousands to ten thousands of atoms), and

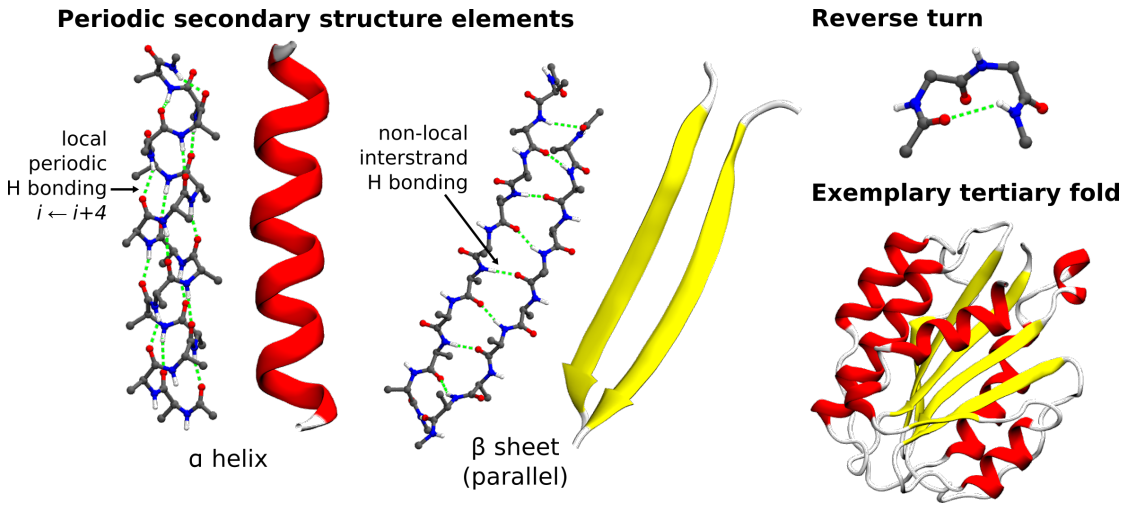


Figure 1.2 – Overview over the structure levels of proteins: periodic and non-periodic secondary structure elements, and an example of a tertiary protein fold. The three-dimensional structure examples are taken from PDB-ID 3PPY [CB11].

- the necessary numbers of energy and force evaluation to sample a energy surface during conformational searches or molecular dynamics simulations until convergence.

Every practically applicable **simulation approximates reality**, consequently the resulting simulation-derived physical observables deviate from the results of an experimental measurement. But also experimental values come with uncertainties that stem from the techniques themselves or from problems with preparing samples or controlling the environment. On the theory side, a description of the molecular potential-energy surface (PES) is basis to almost all molecular simulation techniques. Molecular mechanics rely on a classical mechanics description of the molecule and offer atomistic resolution. First-principles based electronic structure theory methods are based on fundamental physical laws and constants. Besides atomic coordinates and number of electrons, no further input is required.

A computationally efficient, but rather approximate, level of description is the use of **molecular mechanics**. The total energy of a molecules is here described by additive contributions of bonded and non-bonded interactions. Bonded interactions model chemical bonds and are based on the connectivity of the molecular structure. They are represented by harmonic potentials for bond lengths (2 atoms), bond angles (3 atoms), and torsional angles (4 atoms). Non-bonded interactions are electrostatic interactions of (partially) charged atoms that are modeled by Coulomb's law and van-der-Waals interactions that are accounted for by Lennard-Jones potentials. The combination of the formulation of the molecular mechanics equation with the parameters and atom typings is called a **force field**. Popular force fields are, for example, OPLS-AA [17], Amber99sb [18], and Charmm22 [19, 20]. Force-field based simulations allow for pushing the boundaries when it comes to system sizes or time scales. However, some clear limitations have become obvious in recent years, mainly due to the

limited achievable accuracy of the underlying parametrization and formulation. There exist examples where it has been shown that simulations can diverge when running for long times [21] and that different standard protein force fields yield completely different kinetic and dynamic data [22, 23]. Nevertheless, force field simulations offer an illustrative qualitative glimpse at the dynamic behavior of the molecules that make up living organisms.

In principle, **electronic-structure theory** offers an empiricism- and parameter-free description. To that end, the ultimate aim of most computational chemistry approaches is to solve the time-independent, non-relativistic **Schrödinger equation**. That can be used to estimate ground-state properties of systems in which relativistic effects can be neglected.

$$H\Psi_n = E_n\Psi_n$$

The wave function $\Psi_n(x_i, R_I)$ is a function of the coordinates of all electrons i and their spin (x_i) and of the position of the nuclei R_I. The wave functions Ψ_n at the states n would give us access to all the information we are interested, most prominently the energy E_0 of the ground state $n = 0$. The Hamilton operator of this eigenvalue equation contains a kinetic part dealing with the movements of all involved particles (electrons and nuclei) and a potential part that describes the attractive interactions between nuclei and electrons as well as the repulsive interactions among the nuclei and electrons, respectively.

The mass difference between the involved particles, nuclei and electrons, is high. That allows for an extreme interpretation of the relative movements of both types of particles: electrons move fast in a static arrangement of the atomic nuclei. The approximation of "clamped nuclei" allows for the formulation of an electronic Hamiltonian H_{elec} that acts on the electronic wave function Ψ_{elec}, here, the nuclear coordinates R_I only enter as parameters, not as variables anymore. This separation of electronic and nuclear degrees of freedom is known as **Born-Oppenheimer approximation** [24]. The total energy of the electronic system is now defined as the sum of E_{elec}, *i.e.* the eigenvalue of H_{elec} with Ψ_{elec}, and the energy of the nuclear repulsion E_{nucl}.

The calculation of the ground state energy for a molecule seems now a simple task. The only input required are the positions of the involved atoms and the number of electrons. The remaining parts, for example the operator for the electronic kinetic energy, are independent of the systems that is under investigation and all properties of interest could be derived by applying the respective operator to the wave function. Unfortunately, there is no way to directly analytically solve Schrödinger's equation for systems with practical relevance due to the many-body problem of the electron interactions.

Instead, approximate solution must be sought. One way is to approach the exact solution for the ground state, which is defined by the ground state wave function Ψ_0 and has the ground state energy E_0. To that end, the **variational principle** is employed: computing the energy E_{trial} with a wave function Ψ_{trial} with the Hamilton operator for a given system will always

give an upper bound for the observable. As a consequence, the best possible guess with regards to yielding the lowest value for E_{trial} for the trial wave function Ψ_{trial} is identical to the ground-state wave function Ψ_0 of the system.

Approaching the exact solution by applying the variational principle would require to search through the space of all possible wave functions – an approximation is necessary here. In the **Hartree-Fock method**, a reasonable subset of physically meaningful wave functions is defined [25, 26, 27]. The **Slater determinant** Φ_{SD} approximates the wave function that describes the behavior of the N electrons of the system. Φ_{SD} is the anti-symmetrized product of N one-electron wave functions: spin orbitals $\chi_i(x_i)$ that are composed of a spatial orbital $\phi(r)$ and one of the two possible spin functions α or β.

Wave function based methods rely on the Hartree-Fock approximation that the wave function of a systems can be written as Slater determinant of one-electron wave functions. In the Hartree-Fock method, a single electron moves in an average electrostatic potential of all other electrons and the instantaneous part of the electron-electron interaction is neglected. The missing correlation energy contribution can be attributed for by post-Hartree-Fock methods like Møller-Plesset perturbation theory [28] and coupled-cluster theory [29].

Density-functional theory (DFT) represents an alternative approach to wave function theory as it seeks to predict the ground-state properties from the electron density. Foundation of DFT is the Hohenberg-Kohn theorem [30] that states:

1. There exists a bijective relation between the ground-state wave function and the ground-state electron density for the Hamiltonian of a given system. Consequently, any ground-state property can be formulated as a functional of the density of electrons.

2. The application of the variational principle now allows to find this ground-state electron density as it is minimizes the total energy.

The Kohn-Sham *Ansatz* [31] makes the idea by Hohenberg and Kohn practically usable. They assumed a set on non-interacting particles that have the same electron density and total energy as the realistic set of interacting particles. By that, the energy functional is essentially a sum of contributions, i.e. non-interacting kinetic energy, potential energy, Coulomb energy, and the exchange correlation energy. This formulation of the total energy by Kohn and Sham is exact and its solution based on the exact electron density would give the same result as the solution of the Schrödinger equation with the exact wave function. In practice however, the exchange-correlation (XC) functionals are unknown and have to be approximated. This gives rise to a large zoo of functionals that can be sorted according to what Perdew has named Jacob's ladder of DFT [32].

By choosing one of the above mentioned energy functions, a **potential-energy surface (PES)** is defined for a given molecule: for each combination of nuclear coordinates a total

energy value can be computed. In summary, this results in a high-dimensional and rather rugged landscape with multiple points of interest, especially the low-energy minima. The discussion of accuracy of density-functional approximations and other methods to compute potential energy of biomolecules is a central part of the research that I am involved in. This is exemplary being discussed in Chapters 3 and 4 and in some of publications that I have (co)authored, e.g. [CB4, CB6, CB12, CB10].

The force field based simulation of proteins, like the ones on von Willebrand factor in Section 1.2, are usually the starting point for a biochemist in the field of computational chemistry. In order to take a closer and more accurate look on the structure and dynamics, the systems have get smaller as the methods get more accurate and computationally more demanding. Such studies on peptides and peptide foldamers are discussed in Section 1.3. Computational studies in biochemistry often involve the prediction of structure of molecules or molecular complexes. There structure sampling methods like the ones discussed in Section 1.4 are being used.

1.2 The molecular mechanics of the blood protein von Willebrand factor

In the late nineteenth century, Erik Adolf von Willebrand, a Finnish physician, started to practice on the Åland islands. He discovered a bleeding disorder that he called an inherited pseudo-hemophilia [33]. Later, the syndrome was named after him as von Willebrand disease (VWD). VWD is the most common inherited bleeding disorder and it took several decades until the molecular origin of the disease was identified. Only in the 1970s the protein von Willebrand factor (VWF) was identified as the coagulation factor linked to VWD [34, 35]. The different manifestations of VWD are caused by functional defects of VWF itself as well as by reduced expression or total absence of VWF. Besides its role as cause of VWD, VWF is linked to inflammation as well as to pathological blood-clot formation in stroke scenarios.

VWF is a giant glycoprotein that is released by endothelial cells into human blood. Mature monomeric subunits of VWF are 2,050 amino acids long and are connected to multimer chains of lengths of up to 100 monomers. The domain structure of the monomer is shown in Figure 1.3. In blood vessel endothelial cells, pre-pro VWF is a product of protein biosynthesis and is subsequently transferred to the endoplasmatic reticulum (ER, Figure 1.4). In the ER, the signaling peptide is cleaved-off and C terminal dimerization is facilitated via the formation of covalent disulfide bonds between the CK domains of pro-VWF. While transferring to the Golgi and post-Golgi apparatus, the dimers multimerize via N terminal disulfide-bond formation and simultaneous propeptide cleavage. VWF multimers of masses between 800 kDa and 20,000 kDa are stored in Weibel-Palade bodies and are eventually released to the blood stream [36].

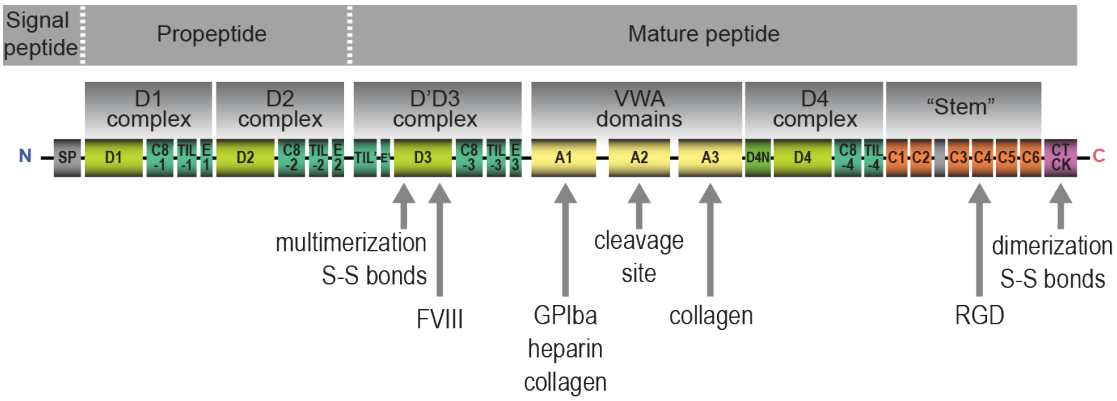


Figure 1.3 – The von Willebrand factor (VWF) domain structure.

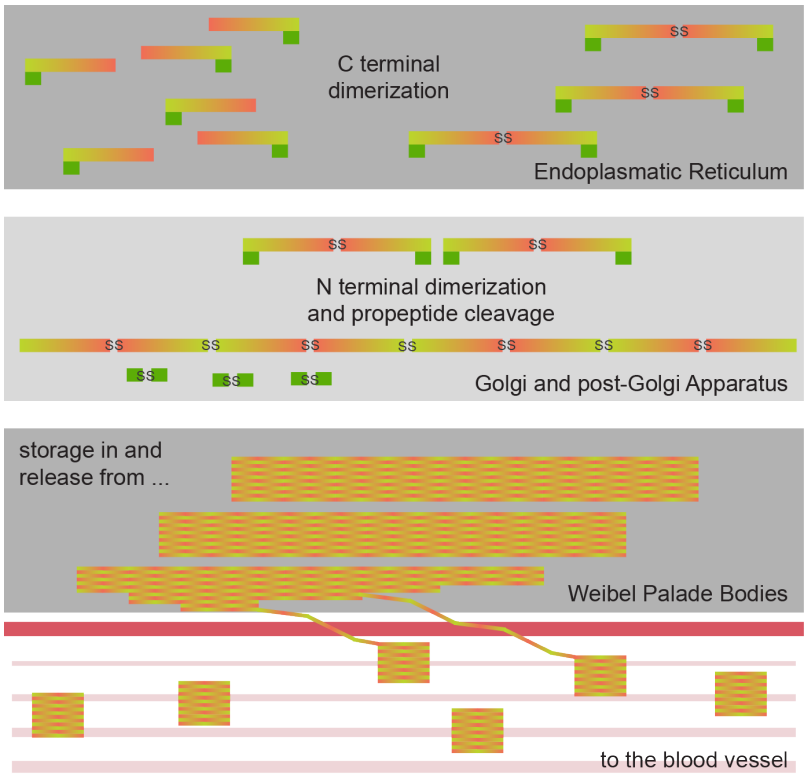


Figure 1.4 – The von Willebrand factor (VWF) biosynthesis.

In a collaborative project that is part of the DFG-funded Research Unit SHENC,[1] we were able to identify protein disulfide isomerase PDIA1 as the enzyme that catalyzes the C terminal linkage of VWF monomers to dimers in the ER [CB13]. In conjunction to the experimental investigations, we contributed with protein docking and molecular dynamics simulations to the understanding of the mechanism of VWF dimerization: PDIA1 initiates the dimerization by forming two disulfide bonds Cys2771-2773' and Cys2771'-2773 between the CK domains of the two monomers. Subsequently, the third bond Cys2811-2811' is formed, presumably, to protect the first two bonds from reduction, thereby rendering the dimerization irreversible.

Key property of VWF is its sensitivity to the shear flow that results from the parabolic distribution of flow velocities in an approximately cylindric blood vessel with flow velocities being maximal in the center and going to zero at the vessel wall [35, 37, 38]. Special flow conditions, e.g. turbulent flow, can result from branching vessels, pathological stenoses, or at sites of vessel rupture. At low shear rates, VWF is wound-up to a globular shape due to specific and unspecific intra-chain interactions. With increasing shear rate, VWF starts to form tethers and, with even higher shear rates acting upon it, fully untangles to an extended chain [39, 40]. The shear force is still acting on the molecule and is translated into an extensional force along the macromolecule. This force acts as trigger for several of the functions that VWF participates in and some of them will be discussed in the following. Two aspects are crucial for the shear-flow sensing properties of VWF: (i) as a molecule of some ten micrometers in size, VWF large enough to sense shear and (ii) a second aspect, which appears to be often underestimated, is the impact of the heavy glycosylation of the macromolecule. The complex carbohydrates that are covalently linked to VWF are hydrophilic and, in contrast to a protein, do not collapse in water to a densely packed entity. Rather, these hydrophilic glycans stick out to the solvent and act as sails that have a strong impact on the hydrodynamic radius of VWF and its ability to sense shear flow.

We have studied several aspects of VWF acting in physiological and pathophysiological scenarios by means of molecular simulations. Of special importance here is the role of VWF to form blood clots at sites of vascular injury. At such a site, VWF binds (mediated by its A1 and A3 domains) to collagen, a main component of the extracellular matrix (ECM) [CB14]. The extended VWF chain is sticking out and recruits platelets by a specific interaction with glycoprotein Ibα, a receptor presented by platelets. VWF, the ECM, and platelets represent multi-valent binding partners. More and more of them are recruited during the growth of such a clot and the vascular injury is closed. Two aspects are critical here for the regulation of this process: the initiation of clot formation to close the injury and the inhibition of clot growth to avoid closure of the blood vessel. Both aspects are mechano-regulated via the ability of VWF to act as force sensor.

Experiments have shown that the binding of VWF to platelets is triggered by mechanical stimulation through shear flow. Based on that we have developed the hypothesis that under

[1] DFG Research Unit FOR 1543: http://www.shenc.de

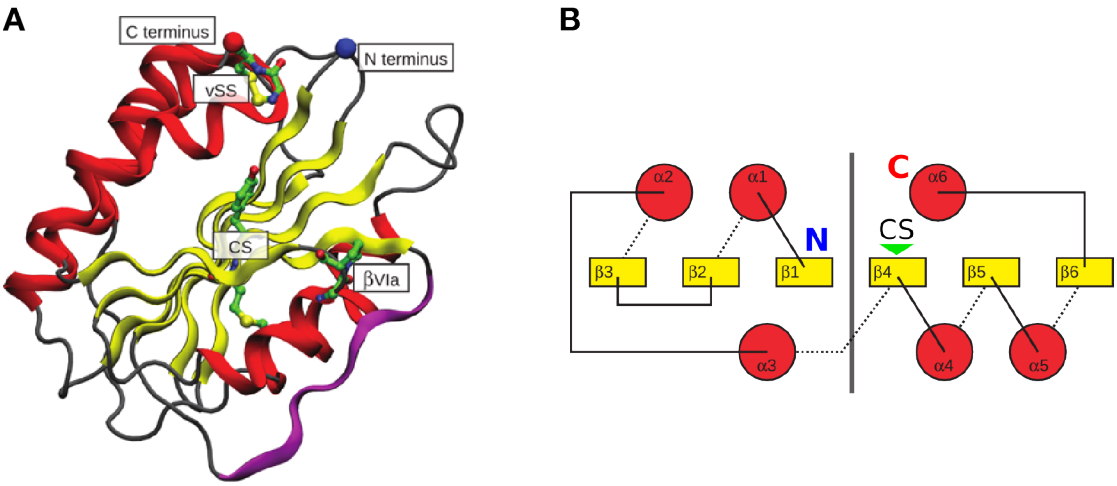


Figure 1.5 – Structural features (**A**) and topology of secondary structure elements (**B**) of the VWF A2 domain.

unperturbed conditions, with no force acting, the binding site for GPIbα at the A1 domain of VWF is occluded by the directly neighboring A2 domain [CB3]. In order to test this possibility, we performed unbiased molecular dynamics simulations as well as protein-protein docking simulations. Indeed, a number of stable poses were predicted that suggest that such competitive binding of A2 to A1 is possible. Under shear stress, when a stretching force acts on the macromolecular chain of VWF, the two domains are pulled apart from each other and the GPIbα binding site of A1 is presented. Now the interaction between VWF-A1 and platelet-GPIbα is possible. In order to test the hypothesis of VWF auto-inhibiting its binding to GPIbα under low-shear conditions, collaborators performed binding studies of wild-type VWF and ΔA2 VWF (i.e. VWF without the A2 domain). They could show that the ΔA2 variant, in comparison to the wild type, required little to no mechanical activation to form aggregates.

The A2 domain has a further function, it carries a cleavage site for the plasma protease ADAMTS13. However, this cleavage site is hidden in the center of the domain and is thus not accessible for the cleaving enzyme. Only when a stretching force is applied along the protein chain, the A2 domain partially unfolds [41, CB1]. As a consequence, VWF under above-critical shear is activated for both, cleavage and clotting. Besides the hidden cleavage site, the A2 domain has several additional interesting features (see Figure 1.5A), for example a vicinal disulfide bridge at the end of the C terminal helix that sets it apart from the flanking and otherwise highly homologous A1 and A3 domain. In the latter, the disulfide bridge spans the whole sequence of the respective domain and prohibits any force induced unfolding. Furthermore, one of the loops connecting the secondary structure elements in the C terminal half of the A2 domain adopts the somewhat uncommon βVIa turn. This type of turn is characterized by a *cis* peptide bond. A stretching force that rests on an extended peptide chain acts on the *cis* bond and triggers the isomerization to a *trans* peptide bond [CB2].

Overall, the architecture of the A2 domain is peculiar: the regularity of the α-β alternation that is expected for a Rossman fold is broken in case of the VWF A2 domain. The secondary structure elements of the N-terminal half are oriented as follows (see Figure 1.5B): β1-α2-β2-β3-α2-α3. The C terminus however is perfectly regular: β4-α4-β5-α5-β6-α6. The alteration in the C terminal half and the resulting knot (see scheme in Figure 1.5B) renders this part of A2 stable against unfolding. In the future I would like to understand how, once the mechanical load is released, the intact N terminus acts as template and supports the refolding of the C terminus of A2 by simulation.

In order to gain a deeper understanding of peptide structure formation, we study shorter peptides in more detail. Some aspects of that are introduced in the next section.

1.3 Peptide foldamers in the gas phase

Structure formation of natural peptides and peptide foldamers has been a longstanding research topic of mine [CB15, CB16]. This work is motivated by the importance of secondary structure elements, for example helices, as recognition elements in protein-protein interactions. The applicability of, in particular, natural peptides to design modulators of protein-protein interactions is hindered by their rapid metabolization and limited selectivity for alternative binding partners [42]. Already seemingly trivial modifications like the inversion of the chirality at the Cα can overcome the enzymatic susceptibility of peptides. A promising route to peptides with improved bioavailability and structure formation properties are homologous peptides. In comparison to the native α peptides, homologous peptides feature an increased backbone length of their building blocks. Or in other words, homologous peptides are composed of β, γ, or δ amino acids, see Figure 1.6. The groundbreaking synthetic work in this field came from the groups of Gellman [43, 44, 45] and Seebach [46, 47].

Previous work of mine focused on the investigation of the helix formation propensities of such peptides by an automatic approach to characterize structures using first-principles methods. A thorough survey of the field can be found in a review to which I have contributed [CB15]. Helical hydrogen bonding is not limited to the patterns that are known and preferred in natural α peptides, i.e. the α and 3_{10} helices. Instead, hydrogen bonding can occur in multiple patterns in forward or backward direction relative to the sequence direction (Figure 1.7A). The preference for a specific helix type can for example be triggered by side chain substitution patterns [CB17] or by backbone modifications [CB18].

Mixed or β helices (Figure 1.7B) represent a particular type of secondary structure that represents an overall helical fold combined with an H-bonding pattern that reminds of β strand/sheet structures (see Figure 1.2). A known naturally occurring example of this structure type is the peptide antibiotic gramicidin A embedded in a membrane [48]. We have shown that such structure types are intrinsically preferred over the alternative, α- or

3_{10}-helix like, structures and only penalized in polar environments [CB19].

Combinations of different non-natural building block types further enrich the possibilities of structure formation and allow for the isosteric replacement of natural α peptide sequences [CB20, CB21]. Especially the feature of β/γ peptides to represent an isosteric replacement of α peptides has initiated a fruitful collaboration with experimentalists, the group of Beate Koksch at FU Berlin [CB22, CB23, CB24, CB25, CB26].

The first-principles calculations of peptides and homologous peptides are often performed on isolated molecules in the gas phase. The relevance of such studies is highlighted in an article by Franziska Schubert *et al.* [CB12]. There, the conformational space of two large polyalanine peptides that differ in the N-terminal *versus* C-terminal localization of a protonated lysine residue is investigated. This alternative placement of the positive charge relative to the sequence has dramatic effects on the structure formation of the peptides. The C-terminal placement triggers perfect helix formation with a clear and steep folding funnel due to a favorable charge-dipole interactions. In the contrary, the N-terminal localization of the charge results in diverse globular structures that are close in energy. Such case is of course a challenge for a theoretical description as already small systematic errors in the energy function lead to inconsistencies in the predictions. In that work, due to the careful comparison to experimental data, we were able to assess and push the current limitations of DFT-based structure prediction for biomolecules. Another aspect besides the energy function, namely the search algorithm itself, is the subject of the following section.

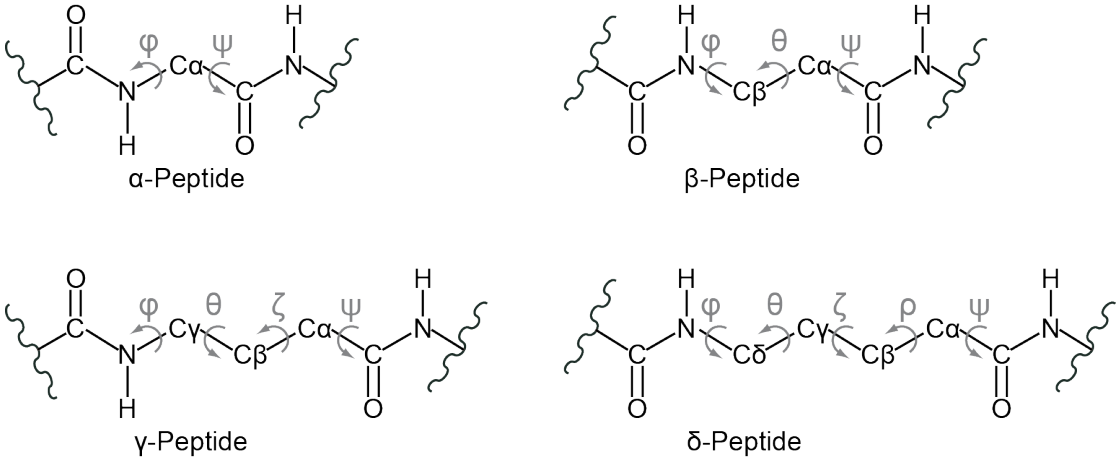


Figure 1.6 – Chemical formulas of α, β, γ and δ peptides. The backbone torsion angles are highlighted in grey. Only the peptide main chain atoms are shown, aliphatic hydrogens and side chains are not shown for clarity.

1.4 Sampling biomolecular potential-energy landscapes

Simulations have the potential to be faster and less costly than experiments. In principle, simulations represent an ideal way to compute properties like binding energies, diverse types of spectra, catalytic activity, and many more. These computed observables can then be compared to experimental measurements and by that simulation can serve as a microscope to investigate biochemical processes at atomistic and even electronic resolution. In addition, *in silico* approaches can map uncharted territory of chemical or materials space or explain under-resolved experiments. The first step towards the prediction of properties or physical observables is usually the search for the three-dimensional structure of a given molecule.

The Born-Oppenheimer approximation yields the definition of the potential-energy surface (PES) of a molecule: the potential energy as a function of the nuclear degrees of freedom. On

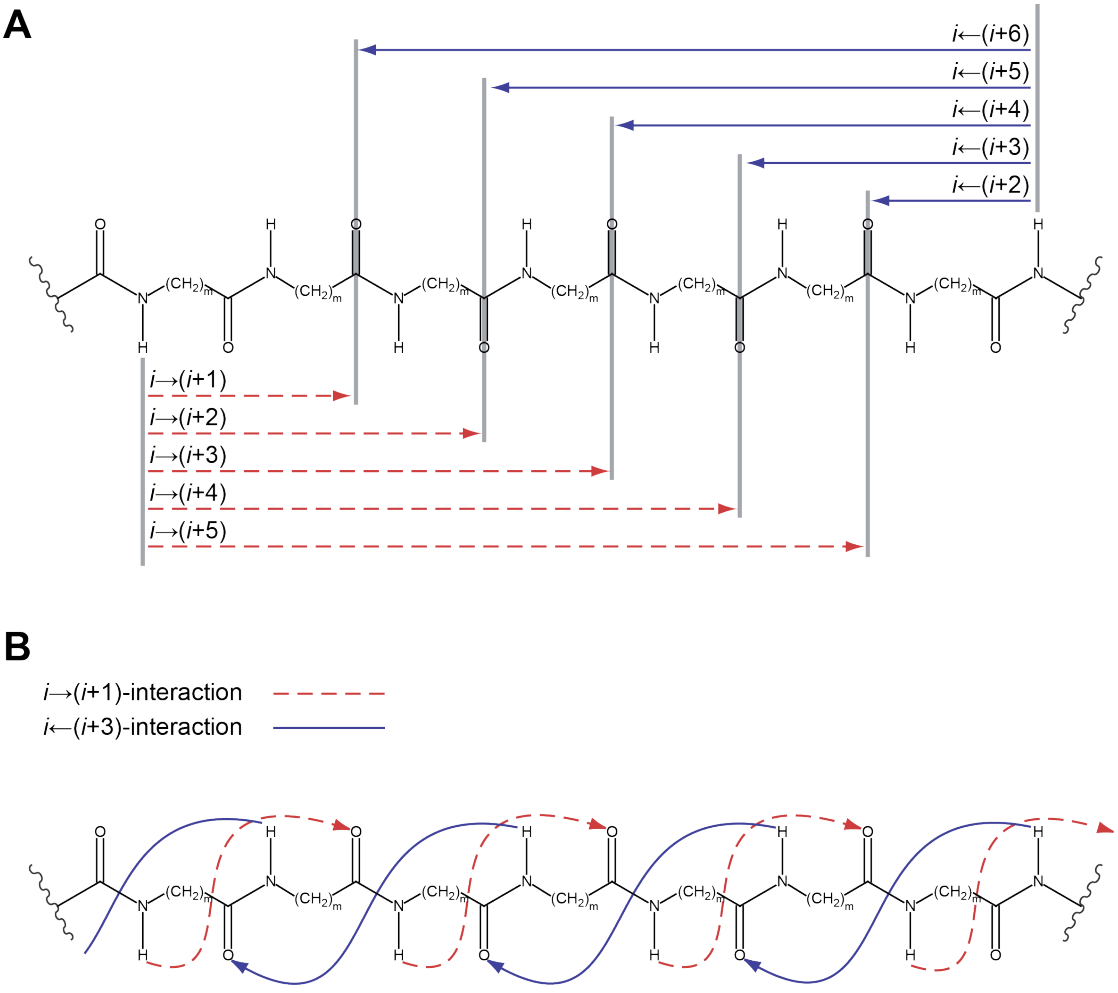


Figure 1.7 – H-bonding types for helices in homologous peptide foldamers. **A** Unidirectional helical H-bonding patterns with hydrogen bonds pointing backward (blue) or forward (red) relative to the sequence direction. **B** H-bonding pattern for mixed or β helices [CB17, CB19].

the PES, points of interest are the global minimum, low-energy local minima, and transition states between them. When searching for minima – and some examples will be discussed in the following, a few considerations have to be made beforehand according to the choice of the coordinate system, the type of energy function, and the applied search method. A good search method has to fulfill several requirements [49, 50]. For example, it should:

- be able to find the global minimum,
- quickly locate and leave local minima,
- focus on overall PES structure and explore conformational space as fast as possible,
- not blindly jump, but use accumulated knowledge to avoid complete enumeration as well as revisiting of known regions of structure space.

Systematic searches

As stated above, the choice of the coordinate system is critical for the way we can search a PES. In order to represent a molecule in Cartesian coordinates, the position in space of each atom is given by a coordinate in x, y, and z direction. In molecules, an intuitive and successful choice are internal coordinates that neglect external motion (rotation and translation). Internal coordinates are build-up starting from a seed atom, all other atoms are related to it by bond lengths, bond angles, and torsion angles. Bond lengths and angles deviate only slightly from average values for the different minima of a given molecule. Torsion angles on the other hand are sufficient to describe the different possible conformations well. Consequently, systematic searches of the structure space of molecules are often based on the discretization of the torsion angles and a complete enumeration of the resulting possible combinations. A classical example is shown in Figure 1.8: The conformational space of the acetylated and amino-methylated amino acid alanine (the so-called alanine dipeptide) can be described by the two backbone torsion angles ϕ and ψ. For all possible combinations of the discretized torsion angles a constrained relaxation with the torsion angles kept constant can be performed. Plotting now potential energy as a function of the two torsion angles results in a Ramachandran plot. This representation was first proposed by Ramachandran, Ramakrishnan, and Sasisekharan in 1963 [52] based on simple steric repulsion of the involved atoms. Later, first-principles calculations were performed at the Hartree-Fock level, for example by Head-Gordon in 1991 [53] or by Rommel-Möhle in 1993 [51]. Since then, the Ramachandran plot of the the alanine dipeptide has been continuously refined at ever higher levels of theory. The energetically favored structures are well visible: for example the global minimum, the C_7^{eq} conformer, and the second stable C_5 conformer in the second quadrant. In such representation, also the high barriers separating the basins can be seen. In the meantime, a multitude of such studies has been performed for many of the proteinogenic amino acids. However, the data is highly diverse, the individual structure searches and first-principles calculations differ in many parameters. A comparison across

chemical space however is only possible based on a rigorously consistent search approach setup for all different amino acids. An example is a search on the conformational trends and the impact of divalent cations for twenty proteinogenic amino acids and dipeptides [CB10].

The dimensionality of the search space can easily be increased by increasing the number of rotatable bonds. This is shown for the example of homologous peptides in Figure 1.6. Already with three torsional degrees of freedom, for example in the β amino acid in Figure 1.6, a straightforward visualization of the structure space becomes complicated. Further increasing the numbers of degrees of freedom leads to numbers of structures to consider that make complete systematic enumerations impossible, for example for one of the oligomeric molecules that are exemplary sketched in Figure 1.7. In order to investigate such structures in a systematic way, the actual volume of conformational space that is sampled has to be reduced in a clever way. In case of the homologous peptide oligomers that are used as an example here, the following criteria were applied:

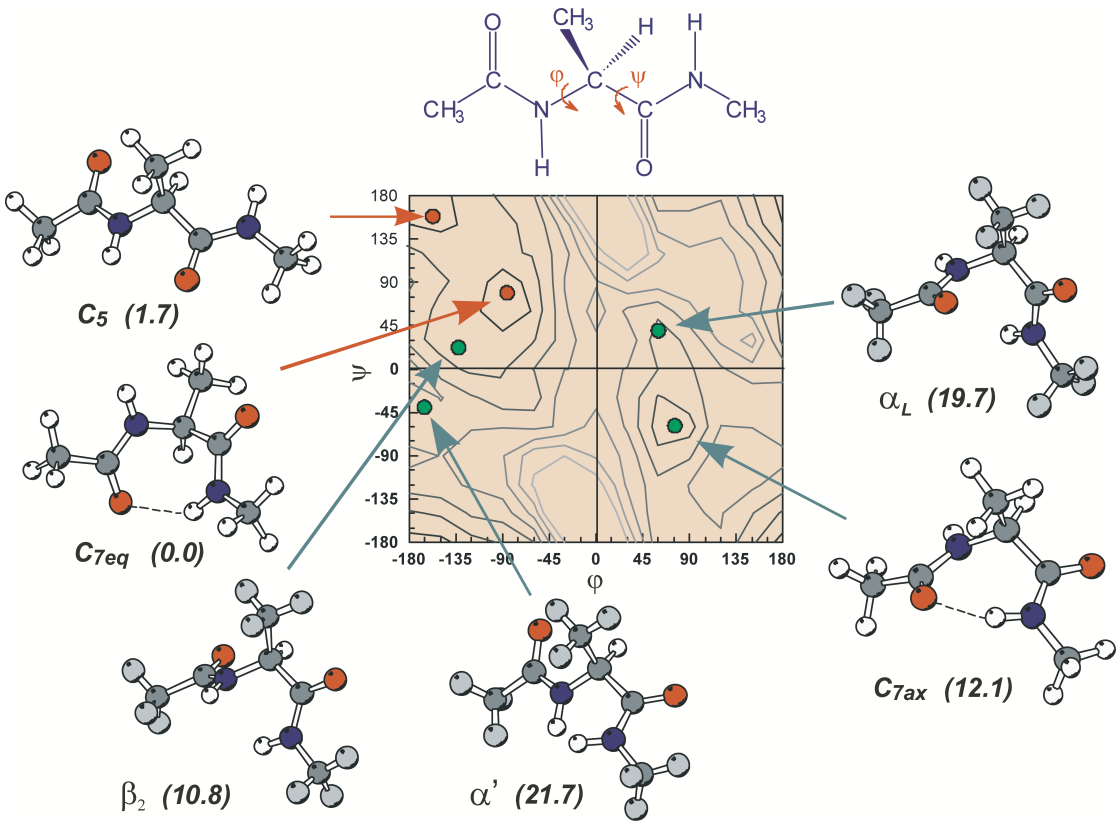


Figure 1.8 – Systematic evaluation of the conformational space of the alanine dipeptide. By plotting potential energy as a function of the torsion angles ϕ and ψ, a first principles Ramachandran plot can be obtained. Numbers in parentheses are HF/6-31G* relative energies in kJ/mol. The plot is work by Rommel-Möhle [51] and was published before in [CB15].

- Only periodic structures are considered. Periodic here refers to repeating conformations of the building blocks, in short, the torsion angle patterns of all monomeric subunits must be the same.
- Only *sensible* structures, *i.e.* structures without atom clashes, are accepted.
- Only structures that feature helical hydrogen bonds (see Figure 1.7) are considered.

The input structures that remain after filtering with these criteria are then subjected to first-principles geometry optimizations. This criteria-based reduction of search space was successful for multiple types of homologous peptides (see references cited in [CB15]), but a weak point remains: by that approach we gain no information about alternative structures that are potentially more stable, but do not fit to the applied selection criteria for input structures. In this very case, we are on the safe side: experiments confirm the assumption of helical structures, see [CB15] and references cited therein. Besides chemical intuition and structural considerations, experimental constraints could eventually be incorporated in such a search as well.

Minima hopping and basin hopping

The basic idea of minima and basin hopping is the reduction of the PES to so-called attraction basins. An attraction basin is a region of the potential-energy surface where a geometry optimization from each point leads to the same minimum [49]. The PES sampling is realized by alternating steps of generating input structures and local relaxations. Barriers that flank the current minimum are overcome by projections out of the current minimum that generate new input structures. Possible ways to generate these new geometries are for example: (i) Monte-Carlo moves [54, 55, 56], (ii) short molecular dynamics trajectories [57], and (iii) projections along normal modes [58]. Structure searches employing basin or minima hopping usually involve immense amounts of energy function calls and force evaluations, see for example the case of a large silicon crystal unit cell in the paper by Goedecker [57]. While using computationally less costly empirical energy functions, this is not an issue for moderately sized systems. However when describing the PES using computationally demanding first principles methods or when investigating systems with more and more degrees of freedom, this can render such searches intractable.

Replica-exchange molecular dynamics

The use of replica-exchange molecular dynamics (REMD) offers an unbiased, straightforward and easy to implement way for structure searching [59, 60]. Within an ensemble of molecular dynamics trajectories that run in parallel at different temperatures, exchange attempts based on the Metropolis criterion [61] facilitate traversing through a range of temperatures. The idea is to overcome barriers at high temperature and to freeze-out structures at low

temperature. With current computational resources, force-field based REMD searches can very well be performed for reasonably large systems and reach μs to ms time scales, see for example references [CB5] and [CB12]. For the biomolecular systems that are of interest here, first principles based REMD [62] can only be performed as a refinement step in the ps time range [CB5, CB10, CB12].

Random structure search

A simple yet effective way to reduce the computational overhead in the creation of new candidate structures for local optimization is *ab initio* random structure search (AIRSS) advocated for by Pickard and Needs [63]. Randomness is an important ingredient that ensures the success of the method, deep (global) minima often have huge attraction basins and thus "random hits often". But also chemical intuition or experiment-derived knowledge can be incorporated. For example, only *sensible* structure guesses are further processed. That means that structures with, for example, too short bond lengths are skipped. And indeed, it seems that the fraction of the PES with too short bonds contains little to no minima. Once a low energy basin is located, *shake* steps in the spirit of the projections that are used in minima or basin hopping can be used to overcome barriers. Based on the funnel hypothesis it is expected that low-energy basins are located near to low-energy basins in structure space. All in all, one can say that AIRSS utilizes randomness within boundaries that are based on chemical knowledge and intuition.

Genetic algorithm searches

Genetic algorithms (GA) belong to the family of evolutionary algorithms (EA) or to an even broader group of bio-inspired and population based search techniques [64]. The quest for the globally optimal structure is reformulated to the Darwinian survival of the fittest concept and applied to a population of, at first, randomly generated structure guesses (individuals). During the course of the global optimization, genetic operations (crossing over and mutation) are being used to evolve the population over generations. As a concession to the peculiarities of chemical structure search, usually local optimization steps are performed for each newly generated individual in the population, see for example reference [65]. As such, the Darwinian evolutionary concept is superseded by Lamarck's idea in such algorithms. An implementation of a GA for structure search with first-principles methods has been realized by Adriana Supady [CB9] and is described in Section 4.3.[2]

Particle-swarm optimization

The bio-inspired search technique particle-swarm optimization (PSO) mimics the social behavior of animals, for example, bird flocking or fish schooling [66]. This technique was

[2] Source code available from: https://github.com/adrianasupady/fafoom

successfully applied [67, 68] to molecular docking, an *in silico* drug design technique that aims to predict the pose of a molecular ligand in the binding pocket of a protein [69, 70]. PSO is well suited to tackle the continuous search space of protein ligand interaction. The search for the global minimum starts with a population of random solutions; the search for optima is facilitated by updating generations, making the swarm virtually fly through the search space. The best position in search space so far (best solution achieved) is tracked for the individual particle as well as for the whole swarm. With the change of generations of the swarm, the particles are accelerated toward these best solutions. We have implemented such an algorithm in the molecular docking program Paradocks [CB7].[3]

Stochastic surface walking

The above mentioned search techniques like basin hopping, AIRSS, or GA search focus on finding low-energy minima and neglect the connecting paths between them. In the contrary, the stochastic surface walking (SSW) aims at finding local and global minima while also maintaining information about the paths connecting them [71]. To achieve that, in SSW a complicated PES is explored by repeating a three-step process consisting of climbing, relaxation, and Metropolis Monte Carlo. The climbing is implemented by modifying the PES: between two points on the PES, a minimum and a higher-energy structure for example, biasing Gaussian potentials are placed to stepwise move the structure uphill in energy. After a while, all biasing potentials are deleted and a local relaxation is performed. The acceptance of the new local minimum is based on a Metropolis criterion. By design, the resulting SSW trajectories contain information about transition paths as well as about minima.

[3] Source code available from: https://github.com/cbaldauf/paradocks

Publications by C. Baldauf

[CB1] C. Baldauf, R. Schneppenheim, W. Stacklies, T. Obser, A. Pieconka, S. Schneppenheim, U. Budde, J. Zhou, and F. Gräter: Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis. *J. Thromb. Haemost.*, 7(12) 2096–2105, 2009.

[CB2] J. Chen, S. A. Edwards, F. Gräter, and C. Baldauf: On the *cis* to *trans* isomerization of prolyl-peptide bonds under tension. *J. Phys. Chem. B*, 116(31) 9346–9351, 2012.

[CB3] C. Aponte-Santamaría, V. Huck, S. Posch, A. K. Bronowska, S. Grässle, M. A. Brehm, T. Obser, R. Schneppenheim, P. Hinterdorfer, S. W. Schneider, C. Baldauf, and F. Gräter: Force-sensitive autoinhibition of the von Willebrand factor is mediated by interdomain interactions. *Biophys. J.*, 108(9) 2312–2321, 2015.

[CB4] C. Baldauf and M. Rossi: Going clean: Structure and dynamics of peptides in the gas phase and paths to solvation. *J. Phys.: Condens. Matter*, 27(49) 493002, 2015.

[CB5] F. Schubert, K. Pagel, M. Rossi, S. Warnke, M. Salwiczek, B. Koksch, G. von Helden, V. Blum, C. Baldauf, and M. Scheffler: Native like helices in a specially designed β peptide in the gas phase. *Phys. Chem. Chem. Phys.*, 17(7) 5376–5385, 2015.

[CB6] C. Baldauf, K. Pagel, S. Warnke, G. von Helden, B. Koksch, V. Blum, and M. Scheffler: How cations change peptide structure. *Chem. Eur. J.*, 19(34) 11224–11234, 2013.

[CB7] R. Meier, M. Pippel, F. Brandt, W. Sippl, and C. Baldauf: Paradocks: A framework for molecular docking with population-based metaheuristics. *J. Chem. Inf. Model.*, 50(5) 879–889, 2010.

[CB8] A. Pecina, R. Meier, J. Fanfrlík, M. Lepšík, J. Rezác, P. Hobza, and C. Baldauf: The SQM / COSMO filter: reliable native pose identification based on the quantum-mechanical description of protein-ligand interactions and implicit COSMO solvation. *Chem. Commun.*, 52(16) 3312–3315, 2016.

[CB9] A. Supady, V. Blum, and C. Baldauf: First-principles molecular structure search with a genetic algorithm. *J. Chem. Inf. Model.*, 55(11) 2338–2348, 2015.

[CB10] M. Ropo, M. Schneider, C. Baldauf, and V. Blum: First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data*, 3 160009, 2016.

[CB11] M. Zhou, X. Dong, C. Baldauf, H. Chen, Y. Zhou, T. A. Springer, X. Luo, C. Zhong, F. Gräter, and J. Ding: A novel calcium-binding site of von Willebrand factor A2 domain regulates its cleavage by ADAMTS13. *Blood*, 117(17) 4623–4631, 2011.

[CB12] F. Schubert, M. Rossi, C. Baldauf, K. Pagel, S. Warnke, G. von Helden, F. Filsinger, P. Kupser, G. Meijer, M. Salwiczek, B. Koksch, M. Scheffler, and V. Blum: Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala_{19}-Lys + H^+ *vs.* Ac-Lys-Ala_{19} + H^+ and the current reach of DFT. *Phys. Chem. Chem. Phys.*, 17(11) 7373–7385, 2015.

[CB13] S. Lippok, K. Kolšek, A. Löf, D. Eggert, W. Vanderlinden, J. P. Müller, G. König, T. Obser, K. Röhrs, S. Schneppenheim, U. Budde, C. Baldauf, C. Aponte-Santamaría, F. Gräter, R. Schneppenheim, J. O. Rädler, and M. A. Brehm: von Willebrand factor is dimerized by protein disulfide isomerase. *Blood*, 127(9) 1183–1191, 2016.

[CB14] S. Posch, C. Aponte-Santamaría, R. Schwarzl, A. Karner, M. Radtke, F. Gräter, T. Obser, G. König, M. A. Brehm, H. J. Gruber, R. R. Netz, C. Baldauf, R. Schneppenheim, R. Tampé, and P. Hinterdorfer: Mutual A domain interactions in the force sensing protein von Willebrand factor. *J. Struct. Biol.*, *in press* , 2016.

[CB15] C. Baldauf and H.-J. Hofmann: *Ab initio* MO theory – An important tool in foldamer research: Prediction of helices in oligomers of ω amino acids. *Helv. Chim. Acta*, 95(12) 2348–2383, 2012.

[CB16] C. Baldauf, R. Günther, and H.-J. Hofmann: Helix formation and folding in γ-peptides and their vinylogues. *Helv. Chim. Acta*, 86(7) 2573–2588, 2003.

[CB17] C. Baldauf, R. Günther, and H.-J. Hofmann: Side-chain control of folding of the homologous α-, β-, and γ-peptides into "mixed" helices (β-helices). *Biopolymers*, 80(5) 675–687, 2005.

[CB18] C. Baldauf, R. Günther, and H.-J. Hofmann: Control of helix formation in vinylogous γ-peptides by (E)- and (Z)-double bonds: A way to ion channels and monomolecular nanotubes. *J. Org. Chem.*, 70(14) 5351–5361, 2005.

[CB19] C. Baldauf, R. Günther, and H.-J. Hofmann: Mixed helices – A general folding pattern in homologous peptides?. *Angew. Chem. Int. Ed.*, 43(12) 1594–1597, 2004.

[CB20] C. Baldauf, R. Günther, and H.-J. Hofmann: Helix formation in α,γ- and β,γ-hybrid peptides: Theoretical insights into mimicry of α- and β-peptides. *J. Org. Chem.*, 71(3) 1200–1208, 2006.

[CB21] G. V. M. Sharma, B. S. Babu, K. V. S. Ramakrishna, P. Nagendar, A. C. Kunwar, P. Schramm, C. Baldauf, and H.-J. Hofmann: Synthesis and structure of α/δ-hybrid peptides - Access to novel helix patterns in foldamers. *Chem. Eur. J.*, 15(22) 5552–5566, 2009.

[CB22] R. Rezaei Araghi, C. Jäckel, H. Cölfen, M. Salwiczek, A. Völkel, S. C. Wagner, S. Wieczorek, C. Baldauf, and B. Koksch: A β/γ motif to mimic α-helical turns in proteins. *ChemBioChem*, 11(3) 335–339, 2010.

[CB23] R. Rezai Araghi, C. Baldauf, U. I. M. Gerling, C. D. Cadicamo, and B. Koksch: A systematic study of fundamentals in α-helical coiled coil mimicry by alternating sequences of β- and γ-amino acids. *Amino Acids*, 41(3) 733–742, 2011.

[CB24] E. K. Nyakatura, R. Rezaei Araghi, J. Mortier, S. Wieczorek, C. Baldauf, G. Wolber, and B. Koksch: An unusual interstrand H-bond stabilizes the heteroassembly of helical $\alpha\beta\gamma$-chimeras with natural peptides. *ACS Chem. Biol.*, 9(3) 613–616, 2014.

[CB25] E. K. Nyakatura, J. Mortier, V. S. Radtke, S. Wieczorek, R. Rezaei Araghi, C. Baldauf, G. Wolber, and B. Koksch: β- and γ-Amino acids at α-helical interfaces: Toward the formation of highly stable foldameric coiled coils. *ACS Med. Chem. Lett.*, 5(12) 1300–1303, 2014.

[CB26] J. Mortier, E. K. Nyakatura, O. Reimann, S. Huhmann, J. O. Daldrop, C. Baldauf, G. Wolber, M. S. Miettinen, and B. Koksch: Coiled-coils in phage display screening: Insight into exceptional selectivity provided by molecular dynamics. *J. Chem. Inf. Model.*, 55(3) 495–500, 2015.

A complete list of my publications can be found at the end of this thesis.

Bibliography

[1] D.S. Goodsell: Bionanotechnology: Lessons from Nature. Wiley, 2004.

[2] R. P. Feynman: There's plenty of room at the bottom. *Engineering and Science*, 23(5) 22–36, 1960.

[3] W. Gilbert: Origin of life: The RNA world. *Nature*, 319(6055) 618–618, 1986.

[4] L. E. Orgel: Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.*, 39(2) 99–123, 2004.

[5] J. E. Wilusz, H. Sunwoo, and D. L. Spector: Long noncoding RNAs: Functional surprises from the RNA world. *Genes Dev.*, 23(13) 1494–1504, 2009.

[6] J. Abi-Ghanem and V. Gabelica: Nucleic acid ion structures in the gas phase. *Phys. Chem. Chem. Phys.*, 16(39) 21204–21218, 2014.

[7] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White: The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.*, 47 1309–1314, 1961.

[8] J. Marcoux and C.V. Robinson: Twenty years of gas phase structural biology. *Structure*, 21(9) 1541–1550, 2013.

[9] R. Daniels, B. Kurowski, A. E. Johnson, and D. N. Hebert: N-linked glycans direct the cotranslational folding pathway of influenza hemagglutinin. *Mol. Cell*, 11(1) 79–90, 2003.

[10] A. Helenius and M. Aebi: Roles of N-linked glycans in the endoplasmic reticulum. *Annu. Rev. Biochem.*, 73(1) 1019–1049, 2004.

[11] D. B. Werz, R. Ranzinger, S. Herget, A. Adibekian, C.-W. von der Lieth, and P. H. Seeberger: Exploring the structural diversity of mammalian carbohydrates ("Glycospace") by statistical databank analysis. *ACS Chem. Biol.*, 2(10) 685–691, 2007.

[12] D. Cremer and J. A. Pople: General definition of ring puckering coordinates. *J. Am. Chem. Soc.*, 97(6) 1354–1358, 1975.

[13] J. S. Richardson: The anatomy and taxonomy of protein structure URL: http://kinemage.biochem.duke.edu/teaching/anatax/index.html.

[14] B. L. Sibanda and J. M. Thornton: β-Hairpin families in globular proteins. *Nature*, 316(6024) 170–174, 1985.

[15] E. G. Hutchinson and J. M. Thornton: A revised set of potentials for β-turn formation in proteins. *Protein Sci.*, 3(12) 2207–2216, 1994.

[16] K. Möhle, M. Gußmann, and H.-J. Hofmann: Structural and energetic relations between β turns. *J. Comput. Chem.*, 18(11) 1415–1430, 1997.

[17] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives: Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45) 11225–11236, 1996.

[18] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling: Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3) 712–725, 2006.

[19] 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. *J. Comput. Chem.*, 25(11) 1400–1415, 2004.

[20] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus: All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18) 3586–3616, 1998.

[21] J. K. Weber, R. L. Jack, C. R. Schwantes, and V. S. Pande: Dynamical phase transitions reveal amyloid-like states on protein folding landscapes. *Biophys. J.*, 107(4) 974–982, 2014.

[22] S. Piana, K. Lindorff-Larsen, and D. E. Shaw: How robust are protein folding simulations with respect to force field parameterization?. *Biophys. J.*, 100(9) L47–L49, 2011.

[23] F. Vitalini, A. S. J. S. Mey, F. Noé, and B. G. Keller: Dynamic properties of force fields. *J. Chem. Phys.*, 142(8) 084101, 2015.

[24] M. Born and R. Oppenheimer: Zur Quantentheorie der Molekeln. *Ann. Phys. (Berlin)*, 389(20) 457–484, 1927.

[25] D. R. Hartree: The wave mechanics of an atom with a non-Coulomb central field. Part II. Some results and discussion. *Math. Proc. Cambridge Phil. Soc.*, 24(1) 111–132, 1928.

[26] J. C. Slater: Note on Hartree's method. *Phys. Rev.*, 35(2) 210–211, 1930.

[27] V. Fock: Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Phys.*, 61(1-2) 126–148, 1930.

[28] C. Møller and M. S. Plesset: Note on an approximation treatment for many-electron systems. *Phys. Rev.*, 46(7) 618–622, 1934.

[29] J. Cízek: On the correlation problem in atomic and molecular Systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods. *J. Chem. Phys.*, 45(11) , 1966.

[30] P. Hohenberg and W. Kohn: Inhomogeneous electron gas. *Phys. Rev.*, 136 B864–B871, 1964.

[31] W. Kohn and L. J. Sham: Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140 A1133–A1138, 1965.

[32] J. P. Perdew and K. Schmidt: Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc.*, 577(1) 1–20, 2001.

[33] E. A. von Willebrand: Hereditary pseudohaemophilia. *Haemophilia*, 5(3) 223–231, 1999.

[34] J. E. Sadler: Biochemistry and genetics of von Willebrand factor. *Annu. Rev. Biochem.*, 67(1) 395–424, 1998.

[35] T. A. Springer: Von Willebrand factor, Jedi knight of the bloodstream. *Blood*, 124(9) 1412–1425, 2014.

[36] U. Budde and R. Schneppenheim: von Willebrand-Syndrom und von Willebrand-Faktor - Aktuelle Aspekte der Diagnostik und Therapie. Uni-Med Verlag AG, 2010.

[37] G.J. Tortora and B.H. Derrickson: Principles of anatomy and physiology. John Wiley & Sons, 2011.

[38] X. Zhang, K. Halvorsen, C.-Z. Zhang, W. P. Wong, and T. A. Springer: Mechanoenzymatic cleavage of the ultralarge vascular protein von Willebrand factor. *Science*, 324(5932) 1330–1334, 2009.

[39] A. Alexander-Katz, M. F. Schneider, S. W. Schneider, A. Wixforth, and R. R. Netz: Shear-flow-induced unfolding of polymeric globules. *Phys. Rev. Lett.*, 97 138101, 2006.

[40] S. W. Schneider, S. Nuschele, A. Wixforth, C. Gorzelanny, A. Alexander-Katz, R. R. Netz, and M. F. Schneider: Shear-induced unfolding triggers adhesion of von Willebrand factor fibers. *Proc. Natl. Acad. Sci. U.S.A.*, 104(19) 7899–7903, 2007.

[41] Q. Zhang, Y.-F. Zhou, C.-Z. Zhang, X. Zhang, C. Lu, and T. A. Springer: Structural specializations of A2, a force-sensing domain in the ultralarge vascular protein von Willebrand factor. *Proc. Natl. Acad. Sci. U.S.A.*, 106(23) 9226–9231, 2009.

[42] A. Giannis and T. Kolter: Peptidomimetics for receptor ligands - Discovery, development, and medical perspectives. *Angew. Chem. Int. Ed.*, 32(9) 1244–1267, 1993.

[43] D. H. Appella, L. A. Christianson, I. L. Karle, D. R. Powell, and S. H. Gellman: β-Peptide foldamers: Robust helix formation in a new family of β-amino acid oligomers. *J. Am. Chem. Soc.*, 118(51) 13071, 1996.

[44] S. H. Gellman: Foldamers: A Manifesto. *Acc. Chem. Res.*, 31(4) 173, 1998.

[45] D. H. Appella, L. A. Christianson, D. A. Klein, M. R. Richards, D. R. Powell, and S. H. Gellman: Synthesis and structural characterization of helix-forming β-peptides: *trans*-2-aminocyclopentanecarboxylic acid oligomers. *J. Am. Chem. Soc.*, 121(33) 7574, 1999.

[46] D. Seebach, P. E. Ciceri, M. Overhand, B. Jaun, D. Rigo, L. Oberer, U. Hommel, R. Amstutz, and H. Widmer: Probing the Helical Secondary Structure of Short-Chain β-Peptides. *Helv. Chim. Acta*, 79(8) 2043, 1996.

[47] D. Seebach, M. Overhand, F. N. M. Kühnle, B. Martinoni, L. Oberer, U. Hommel, and H. Widmer: β-Peptides: Synthesis by Arndt-Eistert homologation with concomitant peptide coupling. Structure determination by NMR and CD spectroscopy and by X-ray crystallography. Helical secondary structure of a β-hexapeptide in solution and its stability towards pepsin. *Helv. Chim. Acta*, 79(4) 913, 1996.

[48] R. R. Ketchem, K. -C. Lee, S. Huo, and T. A. Cross: Macromolecular structural elucidation with solid-state NMR-derived orientational constraints. *J. Biomol. NMR*, 8(1) 1–14, 1996.

[49] D. J. Wales and H. A. Scheraga: Global optimization of clusters, crystals, and biomolecules. *Science*, 285(5432) 1368–1372, 1999.

[50] B. Hartke: Global optimization. *WIREs Comput. Mol. Sci.*, 1(6) 879–887, 2011.

[51] K. Rommel-Möhle and H.-J. Hofmann: Conformation dynamics in peptides: Quantum chemical calculations and molecular dynamics simulations on N-acetylalanyl-N'-methylamide. *J. Molec. Struct.: THEOCHEM*, 285(2) 211–219, 1993.

[52] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan: Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7(1) 95–99, 1963.

[53] T. Head-Gordon, M. Head-Gordon, M. J. Frisch, C. L. Brooks III, and J. A. Pople: Theoretical study of blocked glycine and alanine peptide analogs. *J. Am. Chem. Soc.*, 113(16) 5989–5997, 1991.

[54] D. J. Wales and J. P. K. Doye: Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A*, 101(28) 5111–5116, 1997.

[55] J. P. K. Doye, D. J. Wales, and M. A. Miller: Thermodynamics and the global optimization of Lennard-Jones clusters. *J. Chem. Phys.*, 109(19) 8143–8153, 1998.

[56] J. P. K. Doye and D. J. Wales: Thermodynamics of global optimization. *Phys. Rev. Lett.*, 80(7) 1357–1360, 1998.

[57] S. Goedecker: Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.*, 120(21) 9911–9917, 2004.

[58] R. V. Pappu, R. K. Hart, and J. W. Ponder: Analysis and application of potential energy smoothing and search methods for global optimization. *J. Phys. Chem. B*, 102(48) 9725–9742, 1998.

[59] R. H. Swendsen and J.-S. Wang: Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57 2607–2609, 1986.

[60] Y. Sugita and Y. Okamoto: Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2) 141–151, 1999.

[61] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller: Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6) 1087–1092, 1953.

[62] E. C. Beret, L. M. Ghiringhelli, and M. Scheffler: Free gold clusters: Beyond the static, monostructure description. *Faraday Discuss.*, 152 153–167, 2011.

[63] C. J. Pickard and R. J. Needs: *Ab initio* random structure searching. *J. Phys.: Condens. Matter*, 23(5) 053201, 2011.

[64] D. E. Goldberg: Genetic algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc., 1989.

[65] N. Nair and J. M. Goodman: Genetic algorithms in conformational analysis. *J. Chem. Inf. Comput. Sci.*, 38(2) 317–320, 1998.

[66] J. Kennedy and R. Eberhart: Particle swarm optimization. *Proc. IEEE Int. Conf. Neural Networks*, 4 1942–1948, 1995.

[67] S. Janson and D. Merkle: A new multi-objective particle swarm optimization algorithm using clustering applied to automated docking. *Lecture Notes Comp. Sci.*, 3636(*Hybrid Metaheuristics*) 128–141, 2005.

[68] V. Namasivayam and R. Günther: PSO@Autodock: A fast flexible molecular docking program based on swarm intelligence. *Chem. Biol. Drug Design*, 70(6) 475–484, 2007.

[69] J. Bajorath: Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.*, 1(11) 882–894, 2002.

[70] G. Klebe: Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today*, 11(13–14) 580–594, 2006.

[71] C. Shang and Z.-P. Liu: Stochastic surface walking method for structure prediction and pathway searching. *J. Chem. Theory Comput.*, 9(3) 1838–1845, 2013.

2 The molecular mechanics of the blood protein von Willebrand factor

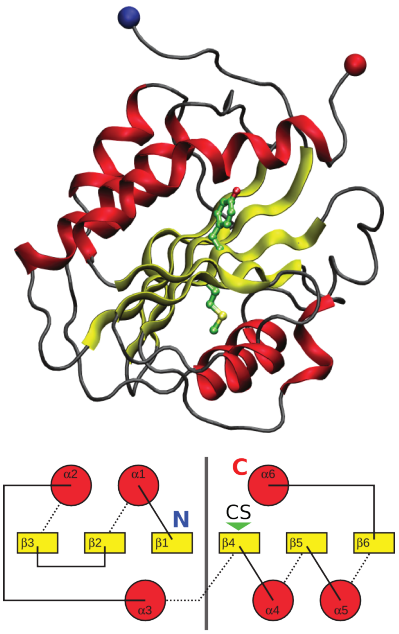
2.1 Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis

Baldauf, C.; Schneppenheim, R.; Stacklies, W.; Obser, T.; Pieconka, A.; Schneppenheim, S.; Budde, U.; Gräter, F.

Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis

J. Thromb. Haemost. **2009** (7), 2096-2105.

DOI: 10.1111/j.1538-7836.2009.03640.x



ORIGINAL ARTICLE

Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis

C. BALDAUF,*†[1] R. SCHNEPPENHEIM,‡[1] W. STACKLIES,*† T. OBSER,‡ A. PIECONKA,§ S. SCHNEPPENHEIM,§ U. BUDDE,§ J. ZHOU*† and F. GRÄTER*†¶

CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China; †EML Research, Heidelberg; ‡Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg; §Coagulation Laboratory, AescuLabor Hamburg, Hamburg; and ¶Max-Planck-Institute for Metals Research, Stuttgart, Germany

To cite this article: Baldauf C, Schneppenheim R, Stacklies W, Obser T, Pieconka A, Schneppenheim S, Budde U, Zhou J, Gräter F. Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis. *J Thromb Haemost* 2009; **7**: 2096–105.

Summary. *Background:* To avoid pathological platelet aggregation by von Willebrand factor (VWF), VWF multimers are regulated in size and reactivity for adhesion by ADAMTS13-mediated proteolysis in a shear flow dependent manner. *Objective and methods:* We examined whether tensile stress in VWF under shear flow activates the VWF A2 domain for cleavage by ADAMTS13 using molecular dynamics simulations. We generated a full length mutant VWF featuring a homologous disulfide bond in A2 (N1493C and C1670S), in an attempt to lock A2 against unfolding. *Results:* We indeed observed stepwise unfolding of A2 and exposure of its deeply buried ADAMTS13 cleavage site. Interestingly, disulfide bonds in the adjacent and highly homologous VWF A1 and A3 domains obstruct their mechanical unfolding. We find this mutant A2 (N1493C and C1670S) to feature ADAMTS13-resistant behavior *in vitro*. *Conclusions:* Our results yield molecular-detail evidence for the force-sensing function of VWF A2, by revealing how tension in VWF due to shear flow selectively exposes the A2 proteolysis site to ADAMTS13 for cleavage while keeping the folded remainder of A2 intact and functional. We find the unconventional 'knotted' Rossmann fold of A2 to be the key to this mechanical response, tailored for regulating VWF size and activity. Based on our model we discuss the pathomechanism of some natural mutations in the VWF A2 domain that significantly increase the cleavage by ADAMTS13 without shearing or chemical denaturation, and provide with the cleavage-activated A2 conformation a structural basis for the design of inhibitors for VWF type 2 diseases.

Keywords: ADAMTS13, force-probe molecular dynamics, Rossmann fold, shear flow, ultra-large von Willebrand factor.

Correspondence: Carsten Baldauf, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, 200031 Shanghai, China.
Tel.: +86 2154920475; fax: +86 2154920451.
E-mail: carsten@picb.ac.cn.

Reinhard Schneppenheim, Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, Martinistraβe 52, D-20246 Hamburg, Germany.
Tel.: +49 40428034270; fax: +49 40428034601.
E-mail: schneppenheim@uke.de.

Frauke Gräter, EML Research gGmbH – Villa Bosch, Schloss-Wolfsbrunnenweg 33, D-69118 Heidelberg, Germany.
Tel.: +49 6221533267; fax: +49 6221533298.
E-mail: frauke.graeter@eml-r.villa-bosch.de

[1]These authors contributed equally to this work.

Received 25 May 2009, accepted 16 September 2009

Introduction

von Willebrand factor (VWF) is a huge multimeric protein found in blood plasma. VWF mediates the adhesion of platelets to the sub-endothelial connective tissue and is the key protein in primary hemostasis in arterial vessels and the microcirculation [1,2]. Monomeric VWF is synthesized in megakaryocytes and endothelial cells. After transfer from the cytosol to the endoplasmatic reticulum, VWF matures by C terminal dimerization (disulfide bonds between CK domains) and N terminal multimerization (disulfide bonds between D3 domains) while being transferred through Golgi and post-Golgi apparatus. Finally stored in endothelial Weibel–Palade bodies and platelet α-granules, VWF is up to 100 monomers long and highly glycosylated [3]. Multimers are released from storage organelles by adequate stimuli.

The VWF multimers released from storage are particularly rich in ultra-large VWF (ULVWF). These highly active forms get rapidly yet only partially cleaved by the protease ADAMTS13 at the cleavage site Tyr1605–Met1606 within the A2 domain [4,5]. ADAMTS13 is a zinc-containing metalloprotease from the ADAMS/ADAMTS family. Shear stress in blood vessels has been shown to drive VWF multimers

into an elongated conformation with increased activity for adsorption to the blood vessel surface, a mechanism to stop bleeding after mechanical injury [6,7]. Mechanical forces due to shear flow regulate selective cleavage of ULVWF and thereby their size distribution [8,9]. If this size regulation fails, ULVWF accumulates and results in phenotypic manifestation of thrombotic thrombocytopenic purpura (TTP) [10]. In contrast, reduced VWF concentration, functional deficits, or complete absence of VWF results in the different types of von Willebrand disease (VWD) [11], the most common inherited bleeding disorder in humans. While the shear stress-induced adhesion and cleavage have been demonstrated in detail *in vitro*, the underlying molecular mechanism of shear-induced activation of VWF for ADAMTS13 cleavage is currently unknown.

Structural information in atomic detail for VWF is scarce. A single VWF is a multi-domain protein featuring a multitude of functionalities (Fig. 1A). The central A domain triplet is pivotal for adhesion and clotting, featuring binding sites for collagen (A1, A3) and the platelet receptor glycoprotein Ib (GPIb, A1), and the ADAMTS13 cleavage site (A2). A1 and A3 have been shown by X-ray crystallography [12,13] and A2 by homology modeling [14] and X-ray crystallography [15] to adopt a Rossmann α/β-fold. The ADAMTS13 cleavage site in A2 appears to be buried, suggesting that forces in stretched VWF multimers induce unfolding and exposure [16]. In recent experiments, described by the groups of Springer, Wong and co-workers [17], unfolding of the A2 domain by optical tweezers and subsequent cleavage by ADAMTS13 was observed.

We here investigate the unfolding and activation mechanism of A2 for ADAMTS13 cleavage under force by molecular simulations. By applying force distribution analysis, a method previously introduced by our group [18], we reveal how the atypical Rossmann fold topology of the VWF A2 domain senses mechanical force by selectively exposing and activating the ADAMTS13 cleavage site. We compare a homology model of the VWF A2 domain with the crystal structure 3GXB and discuss their special structural features. Furthermore, we predict and analyze, based on a homology model of the VWF A2 domain, the impact of mutations stabilizing the A2 domain by introducing a disulfide bond into VWF A2, in analogy with A1 and A3. We demonstrate this mutant VWF to be resistant against ADAMTS13 *in vitro*. Our results show VWF A2 domain unfolding as a response to shear stress to be the essential event in VWF size regulation.

Materials and methods

Homology modeling and in-silico mutation

The sequences of the VWF A domains have a residue identity of 20–25%. Based on multiple sequence alignments and structural alignments we created a homology model of the VWF A2 domain (residues 1488–1676 of human VWF) and

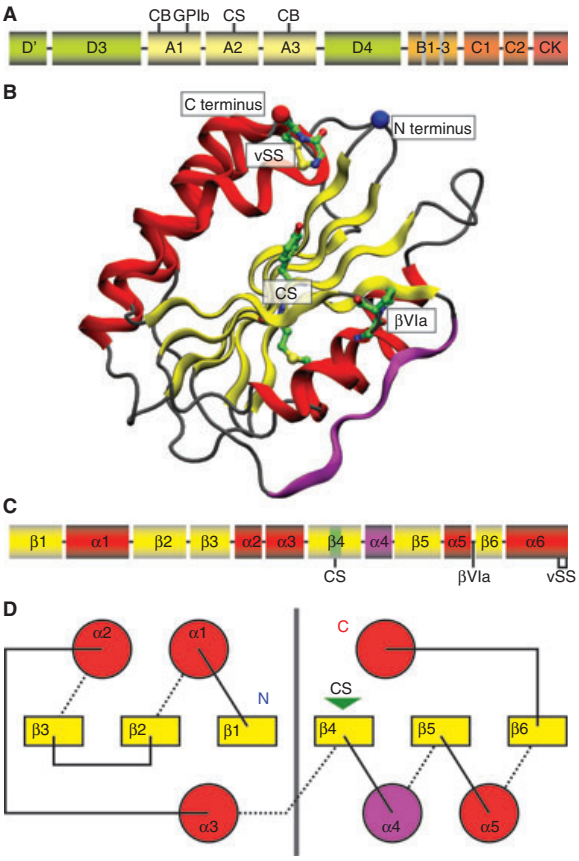


Fig. 1. (A) Domain organization of the VWF with collagen binding sites (CB) in domains A1 and A3, a glycoprotein Ib (GPIb) binding site in A3, and the ADAMTS13 cleavage site (CS) in A2. (B) Structure of the VWF A2 domain (PDB: 3GXB) in cartoon representation, the cleavage site (CS), the βVIa turn (βVIa), and the N terminal vicinal disulfide bridge (vSS) are highlighted in green; α-helices are red, β-sheets are yellow, and the α4-less loop is pink. (C) Secondary structure organization; β and α denote β-strand and α-helix, respectively. (D) The schematic sketch of the spatial secondary structure orientation shows the classical Rossmann fold of the C terminal half of the A2 domain with the cleavage site (CS, green marker) while the N terminal half shows a 'knotted' Rossmann fold with significantly higher stability under force.

the mutant A2 domain (N1493C and C1670S) from a human VWF A1 X-ray structure (PDB: 1AUQ). Details of the homology modeling performed with MOE (2007.9, Chemical Computing Group CCG, Montreal, Canada) can be found in the Supporting Information.

Based on the model of the A2 domain, the A2 double mutant N1493C/C1670S was generated. A disulfide bridge was introduced between the termini by the N1493C mutation, enabling a link between C1493 and C1669. To maintain a constant content of cysteine residues, known to be beneficial for protein expression (see below), a second mutation C1670S was introduced. Both models were validated by molecular dynamics (MD) simulation (Fig. S3B,C). The model is available as Supporting Information or from the authors.

Molecular dynamics simulation

All simulations and part of the analysis were carried out with the Gromacs suite of programs (version 3.3.1) [19]. The OPLS all-atom force field was used for the proteins [20] solvated in dodecahedral boxes with at least 7500 TIP4p water molecules [21], and periodic boundary conditions were applied. The typical protonation states at pH 7 were chosen for ionizable groups of the peptide. The necessary amount of counter-ions (Cl^- and Na^+) was added to ensure a neutral system. A temperature of 300 K and a pressure of 1 bar were assumed. The wild-type and mutant A2 models were simulated three times each for 30 ns and with different seeds for the initial velocity generation. Force-probe MD simulations, each ~26 ns in length, were performed two times independently on the 3GXB X-ray structure (residues 1495–1671) and three times independently on a truncated VWF A2 model (residues 1492–1670). Harmonic springs, attached to the terminal Cα atoms, with spring constants of 500 kJ $(\text{mol nm}^2)^{-1}$, were moved away from each other with a velocity of 1.25 nm ns^{-1}. To restrict the system size along the pulling direction, after partial unfolding the residues 1637–1671 of A2 (3GXB) and residues 1636–1670 of the A2 homology model were removed, water was added to the system, and the force-probe MD simulations were continued.

For FDA, two starting systems were taken from snapshots of the unfolding trajectory. Already unfolded parts, starting from Glu1652 for the first and from Ser1613 for the second system, were removed. Constant force of 10 and 100 pN, respectively, for the relaxed and stretched state, was applied in opposing direction to both termini. Each of the two systems was equilibrated under the respective constant force for 30 ns. For both systems, the all-atom RMSD to the starting structure remained below 0.35 nm for both pulling forces (Fig. S1), indicating that the system is able to bear the mechanical stress within this time scale without rupture. In the following, 10 simulations for the folded and 20 simulations for the unfolded state were performed for 30 ns and 15 ns each, starting with different random velocities. LINCS [22] and a time step of 2 fs were used for the folded state, whereas no constraints and a time step of 1 fs were used for the unfolded state. We used the FDA code [18] for Gromacs 4.0 [23] to write out forces F_{ij} between each atom pair i and j. Forces were averaged over the total simulation time of 300 ns per system, respectively, sufficient to obtain converged averages. Changes in forces, ΔF, are the differences in pair-wise forces between the systems pulled with 10 and 100 pN. Residue-wise forces F_{uv}^{res} were obtained by summing up forces F_{ij} for all pairs of atoms i and j in residues u and v, respectively. The absolute sum

$$\Delta F_u^{\text{res}} = \sum_v \left|\Delta F_{uv}^{\text{res}}\right|$$

reflects the changes in strain acting on a single residue and was used to color-code force distribution onto the protein backbone. Strain along the backbone was measured as the sum of all bonded interactions between adjacent residue pairs. As we use an approximation for angular and dihedral terms and solvent is not included in the FDA but in the simulations, changes in backbone forces indicate strain between two residues, but the values are not physically correct forces. Further simulation details can be found in the Supporting Information.

VWF engineering and analysis

By *in vitro* mutagenesis of full length VWF we exchanged N1493 at the N terminal site for cysteine and C1670, one of two neighboring cysteines at the C terminal site of the A2 domain, for serine to allow creation of a cysteine bond in the A2 domain. In additional mutagenesis experiments we also eliminated the existing disulfide bonds in the A1 (C1271S/D1459C) and A3 domains (C1686S/S1873C). *In vitro* mutagenesis of full length VWF cDNA in the mammalian expression vector pcDNA 3.1 was performed with the quick change mutagenesis kit (Stratagene) using primers of 41–46 bp in length harboring the particular base exchange. Transfer of the cDNA transfection of 293 cells by means of liposomal transfer, cell culture conditions, and harvesting and preparing of recombinant VWF, was performed as described previously [24]. The ADAMTS13 assay was based on recombinant human ADAMTS13 (rhuADAMTS13), adjusted to 0.05 U mL^{-1} in Tris/HCl buffer (5 mM, pH 8.0). 100 μL of this solution were added to 200 μL conditioned media containing rhuVWF (80 U dL^{-1}) and incubated with 10 mM barium chloride. The aliquots were then dialyzed against buffer solution (1.5 M urea, 5 mM Tris/HCl at pH 8.0) and incubated at 37 °C for 5h. The reaction was stopped with EDTA (10 mM) [24,25]. ADAMTS13-proteolyzed mutant and wild-type VWF was also analyzed by polyacrylamide gel electrophoresis under reducing conditions [26]. VWF phenotypic characterization by VWF multimer analysis recorded by digital photo imaging (FluorChem 8000) was carried out according to previously published protocols [27–29].

Results and discussion

Force-induced unfolding of the A2 domain

In vivo, the VWF multimer size is regulated by ADAMTS13 depending on shear flow conditions. Shear flow elongates VWF and results in a tensile force propagating throughout all VWF domains including A2 in the stretched protein [6,7]. We examined tensile stress on the VWF A2 conformation by force-probe MD simulations where a pulling force is applied on the termini of A2 in opposite directions. Force profiles for two independent simulations are shown in Fig. 2. The initial conformation (snapshot 1, Fig. 2) is stepwise unfolded. Starting from the C terminus the secondary structure elements are sequentially peeled-off, namely α6, β6 and α5 to yield a first intermediate (snapshot 3, Fig. 2), followed by β5 and the α4-less loop [15] leading to exposure of the cleavage site (snapshot 4, Fig. 2). Overall, inter-β

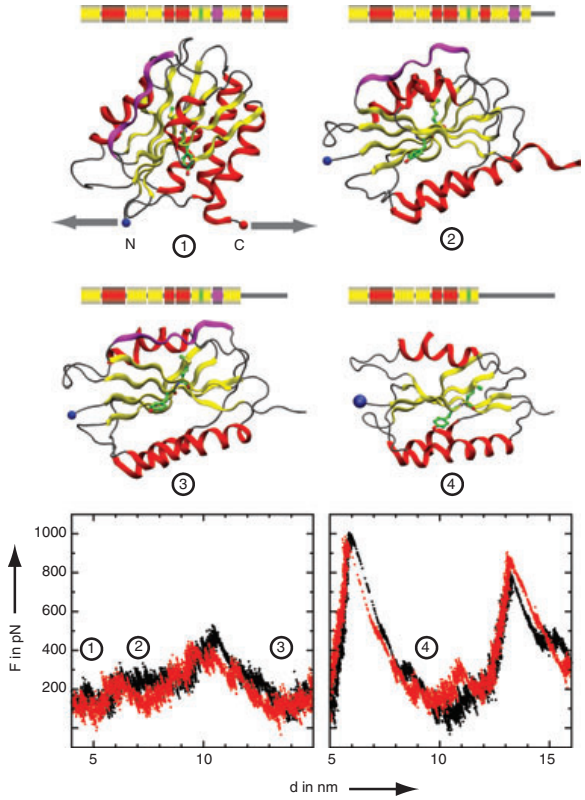


Fig. 2. The force profiles for two independent force-probe MD simulations of 3GXB are shown. After extending the protein chain to 15 nm, the simulations were continued with the unfolded C terminal part (sequence numbers 1637 and higher) being cut off. Selected snapshots are shown as cartoons; the cleavage site is shown in green; the fully unfolded C-terminal fragments in 2, 3 and 4 are omitted for clarity.

strand interactions show higher mechanical resistance than interactions involving helices. A short movie in the Supporting Information illustrates the sequential unfolding of VWF A2 under force.

In a very recent study, Wong, Springer and co-workers demonstrated the enforced activation of the VWF A2 domain for ADAMTS13 cleavage with a laser tweezers set-up [17]. They report a subset of their unfolding experiments to exhibit an intermediate state with a contour length of about 23 nm (40% of the length of the completely unfolded A2 domain with 58 ± 5 nm). Such an intermediate state corresponds well with the state 4 shown in Fig. 2. In this state 60 of 174 residues (35%) are unfolded at the C terminus, which results in an overall contour length of 24.6 nm, including the length of the intact N terminal part of the domain with 3 nm. Our force-probe simulations thus suggest that the experimentally detected unfolding intermediate is ready for cleavage by ADAMTS13. A study by De Cristofaro and co-workers is focused on the mechanism of ADAMTS13 catalysis. Their work with VWF73 (a truncated A2 domain covering VWF residues Asp1596–Arg1668) suggests as well that a partially unfolded state of VWF A2 is ready for ADAMTS13 cleavage [30].

Force distribution analysis of the VWF A domain

The stable N terminal β1 strand is locked to the center of the protein, keeping the protein core including the cleavage site largely intact (Fig. 1D), while the C terminal structural elements, being more responsive to the external force, are pulled out step by step until the cleavage site is accessible. This distinct response of the two halves of the domain is determined by the underlying topology of the VWF A-type domains. The C terminal part of the A2 domain represents a Rossmann fold,

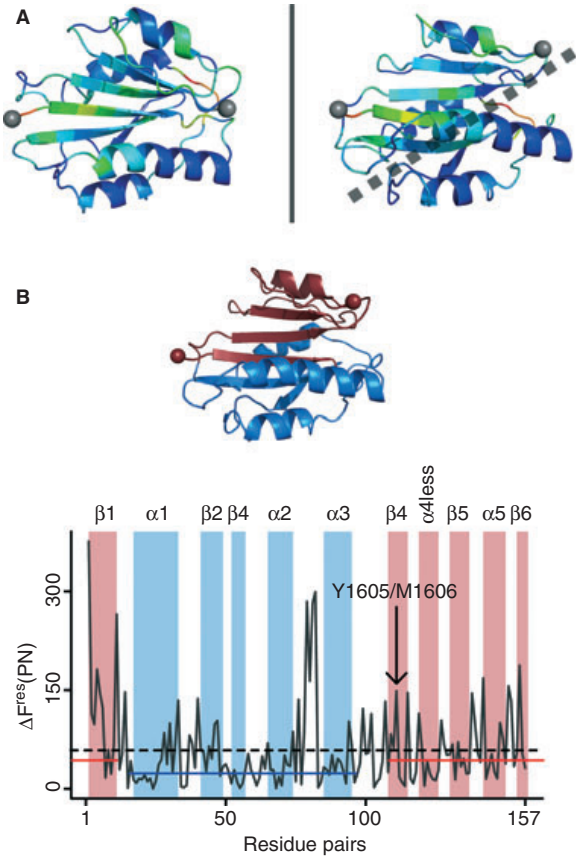


Fig. 3. Force distribution analysis (FDA) of the A2 domain (3GXB). (A) Cartoon representation of an A2 folded state (state 2 in Fig. 2) in two views. Changes in pair-wise forces, ΔF, are color-coded ranging from blue for ΔF = 0 to red for high ΔF. The external pulling force distributes along a direct path between termini, leaving the N-terminal part nearly unaffected (below the dotted line). (B) Strain along the backbone of the folded structure (solid gray line), measured in terms of changes in bonded interactions between residue pairs. Sequence positions of secondary structure elements (helices and strands) are highlighted by colored bars. The color-coding reflects the protein topology: structural elements that are unfolded in state 4 of Fig. 2 and strand β1 are red; structural elements that remain intact in state 4 of Fig. 2 are blue (compare with the cartoon representation above the plot). The dashed line represents the mean force over all bonded residue pairs. This mean force on structural elements α1, β2, β3, α2 and α3 (blue line) is lower than on structural elements β1, β4, α4less, β5, α5 and β6 (red line). High pair-wise forces around Thr1576 (between α2 and α3) are an artifact resulting from rare loop rearrangements.

with the characteristic sequential order of the secondary structure elements β4–α4 less–β5–α5–β6–α6, bridging each strand in the parallel β sheet alignment with an α helix (Fig. 1D). This sequential arrangement results in the stepwise unfolding under force. In contrast, the modified Rossmann fold of the N terminal half of the A2 domain prevents unfolding. Here, β strands are swapped such that β1, the strand directly subjected to the external force, is tightly embedded in the protein core (Fig. 1D), so as to form rupture-resistant interactions with adjacent strands β2 and β4.

We further validated the key role of this particular 'knotted' Rossmann topology for the mechanical response of A2 by force distribution analysis (FDA). FDA reveals the distribution of internal strain within a structure subjected to an external force by monitoring changes in pair-wise atomic forces ΔF [18]. We determined the strain distribution in an early unfolding intermediate of the VWF A2 domain in which the mechanically labile helix α6 is already unraveled (Fig. 2, snapshot 2). The tensile force mainly propagates through the part of the central β sheet formed by strands 1, 4, 5 and 6 of the domain (Fig. 3A), following a direct path between the two termini. Force distributes from the C terminal strand via strands β5 and β4 to the very center of the structure, strand β1, transferring the force out of the domain to the N terminus. From the pair-wise forces plotted in Fig. 3(B) it is also evident that substantial parts of the N terminal half are under low (sub-average) force as a direct result of the unconventional Rossmann fold. Significant high forces in this part of the protein are only observed at the loop directly attached to strand β1 (connection to helix α1) and at the loop connecting α1 and β2, that is in close proximity to the N terminus, the point of force application. Strand β1 virtually shields the tertiary structure formed by β1–α1–β2–β3–α2–α3 from force-induced unfolding. Accordingly, this area is under low strain, as evident by the cold coloring mapped on the structure in Fig. 3A. The shielded region is in the lower right of the dashed line. Interestingly, this architecture would allow the N terminal half of the A2 domain to carry out a particular – albeit currently unknown – function, even while the C terminus is unfolded and cleavage ready.

The cleavage site Tyr1605–Met1606 is the topological middle point of the folded and intact A2 domain and therewith protected from cleavage. Mechanical force induced a cleavage-ready unfolding intermediate as observed in our force-probe MD simulation (Fig. 2 state 4) as well as in experiments [17]. In order to investigate the impact of pulling forces on the cleavage site, the Tyr1605–Met1606 peptide bond, FDA was performed on the partially unfolded cleavage-ready A2 domain. We find the backbone between Tyr1605 and Met1606 to be under extra-ordinarily high strain, and strain distribution seems to be directed in a way to specifically target these residues, Fig. 4A. Analysis of inter-side chain forces, this is, forces that pairs of side chains exert on each other, reveals that a large part of the strain on the cleavage site results from neighboring residues located in the central β-strands. The force network in Fig. 4(B), mapping changes in forces between residue pairs as edges onto the structure, shows a clear polarization. There are almost no edges crossing a virtual plane that separates the cleavage site residues, resulting in a weakened peptide bond potentially mechanically activated for cleavage. This specific deflection of mechanical load onto the cleavage site is mainly realized by strong pair-wise interactions between Tyr1605 and Ala1500, Phe1501, Val1502 and Val1604 (Fig. 4C), and between Met1606 and Thr1608 (Fig. 4D), respectively. Thus, in addition to mere exposure to ADAMTS13, the Tyr1605–Met1606 proteolytic site in the VWF A2 unfolding intermediate is selectively tensed up due to an optimized force distribution. We therefore predict mutations in the local force network to attenuate the strain in the peptide bond and to consequently alter its susceptibility to ADAMTS13 cleavage.

Homology modeling of the VWF A2 domain

Until recently [15], no experimentally derived structural data on the VWF A2 domain was available. Thus, a homology model including residues 1488–1676 of human VWF was created. The model fully includes the very terminal sequences of A2, and thereby the site of mutagenesis for introducing a disulfide bond (see below). It is therefore more comprehensive but otherwise highly similar to a previous homology model that covers only the VWF residues 1496–1669 [14] and to the X-ray structure 3GXB covering residues 1495 to 1671 (Fig. S4) [15]. The rmsd (N, Cα, Cβ, C, O) between snapshots of our trajectories of the homology model and 3GXB is around 0.2 nm, and < 0.1 nm for the heavy atoms of the cleavage site. Also, the unfolding mechanism of A2 observed in our simulations largely agrees with the X-ray structure and the model (cf. Fig. 2 and Fig. S5). The predictability of the A2 structure, including differences in A1 and A3, implies that the tertiary fold is largely defined by the primary sequence of A2. However, our model also misses two features by which A2 differs from its highly related adjacent domains, as now revealed by the A2 crystal structure [15] (cf. Fig. S4):

First, the remarkable experimental finding of a vicinal disulfide bridge between the very terminal Cys residues 1669 and 1670 was not predicted in the homology model. Helix α6 is capped and rigidified by the C terminal vicinal disulfide bond, thereby mechanically stabilized and unfolding at once. With reduced Cys side chains as in our homology model, the helix unfolds stepwise in the force-probe simulations (Fig. S5). The presence of the helix cap, however, does not significantly change the rate limiting steps involving high forces for the rupture of the hydrogen network of the central β-sheet after the early helix α6 unfolding. Based on the high strain of the eight-membered disulfide-bonded ring and possible environmental changes, we hypothesize a redox-dependent regulation of the forced onset of A2 unfolding via the vicinal disulfide bond.

Second, the peptide bond between Trp1644 and Pro1645 is *cis* configured in the X-ray structure. The turn comprising this peptide bond is of type βVIa [31] in the X-ray structure, while a

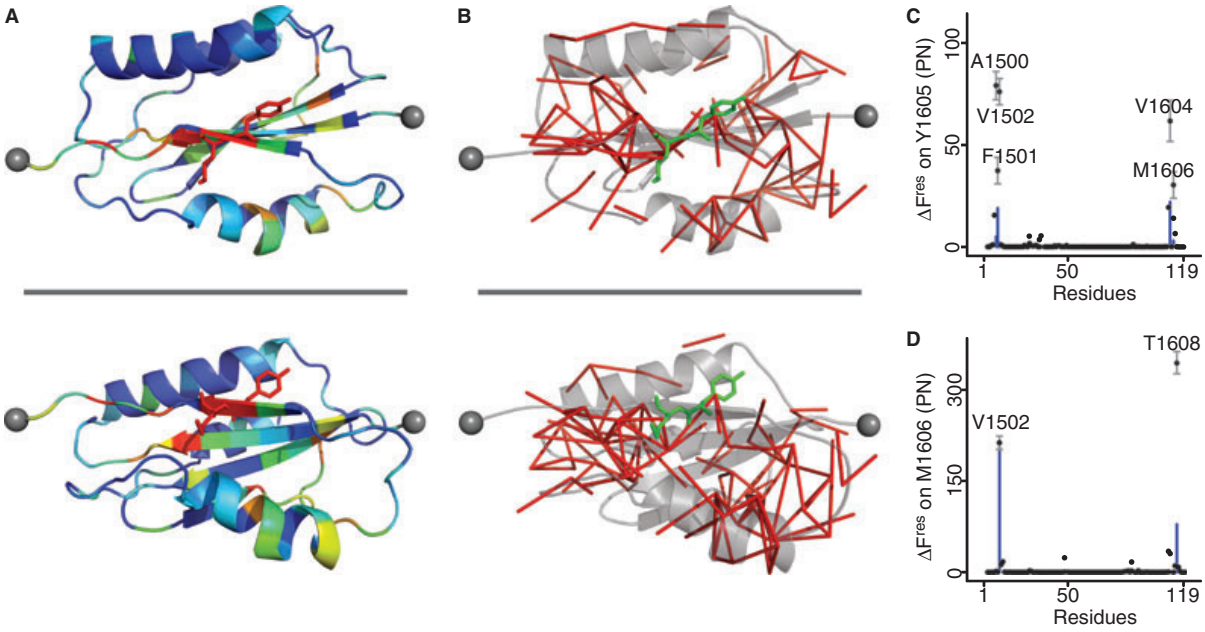


Fig. 4. Force distribution analysis (FDA) of the partially unfolded A2 domain (3GXB). (A) Cartoon representation of the cleavage-ready A2 unfolding intermediate (compare state 4 in Fig. 2). Backbone strain, measured in terms of changes in pair-wise forces, ΔF, is color coded onto the structure, with colors ranging from blue for $\Delta F = 0$ to red for high ΔF. The cleavage site (sticks) exhibits particularly high strain. Helices α1 to α4 are not under strain. (B) Graph-like representation of force distribution in the cleavage-ready A2 unfolding intermediate. Edges represent changes in forces between residue pairs that exceed a threshold of 30 pN. (C) Strain induced on the cleavage site residue Y1605 by adjacent residues. The plot shows changes in inter-residue forces ΔF^{res} for Y1605; the contribution of the Y1605 side-chain is plotted as straight blue lines. Standard errors are plotted as whiskers. (D) Same as (C) but for M1606.

βIII turn is predicted in our model instead (cf. Fig. S4). Our A2 model suggests that *cis* as well as *trans* configuration for this particular peptide bond is structurally feasible. Given the impact of *cis-trans* isomers on protein folding [32], their possible interconversion by mechanical forces [33], and the absence of Prolyl-*cis*/*trans*-isomerases in the extra-cellular space, they might play a regulatory role.

However, the actual physiological significance of both, the vicinal disulfide bridge and the *cis* or *trans* configuration of the peptide bond, remains to be identified for this particular case.

In vitro *mutagenesis and electrophoretic analysis*

Our unfolding simulations suggest A2 to be activated for ADAMTS13 cleavage under high shear flow conditions by exposing the cleavage site after partial unfolding of the C-terminal domain. A1 and A3 have highly similar amino acid sequences and three-dimensional structures, and thus would be expected to unfold along a similar mechanism. Examination of the 3D structures of the VWF A domains shows the existence of disulfide linkages between the termini of the A1 and A3 domains, respectively, but not for the A2 domain. The snipped sequence alignment in Fig. 5A illustrates that A1 and A3 feature two cysteine residues each at their N and C termini, allowing the formation of disulfide bridges. A2 has two vicinal cysteine residues at its C terminus (Fig. 5A) and none at the N terminus. The domain can unfold under force and is mechano-responsive. The cysteine hooks ensure the structural integrity of A1 and A3 under shear flow for specific interactions with collagen and GPIb as essential for VWF adhesion and aggregation, while allowing the selective force-induced unfolding of only the A2 domain for cleavage by ADAMTS13.

Based on the homology model of the VWF A2 domain, we designed the ADAMTS13-resistant VWF variant mutA2 (cf. Table 1 for nomenclature of all VWF variants) by introducing a cysteine at position N1493 to allow disulfide bond formation with residue C1669 at the A2 C-terminus in analogy with the A1 domain. The magnification in Fig. 5A illustrates the virtually perfect orientation of the side chains of residues N1493 and C1669. C1670 was changed to serine to generate maximal homology of A2 with A1 and A3 and to avoid the possibility of alternate disulfide bonding at the A2 carboxy terminal. A model of the mutant A2 domain was subjected to MD simulations to test the feasibility of disulfide bond formation and the domain's structural integrity upon mutation. The data are shown in the Supporting Information (Fig. S3C) and indicate tolerance of the mutations and conservation of the A2 structural features.

To confirm the generation of an artificially introduced A2 disulfide bond *in vitro*, we subjected A2 mutant full-length recombinant VWF (mutA2) to multimer analysis in

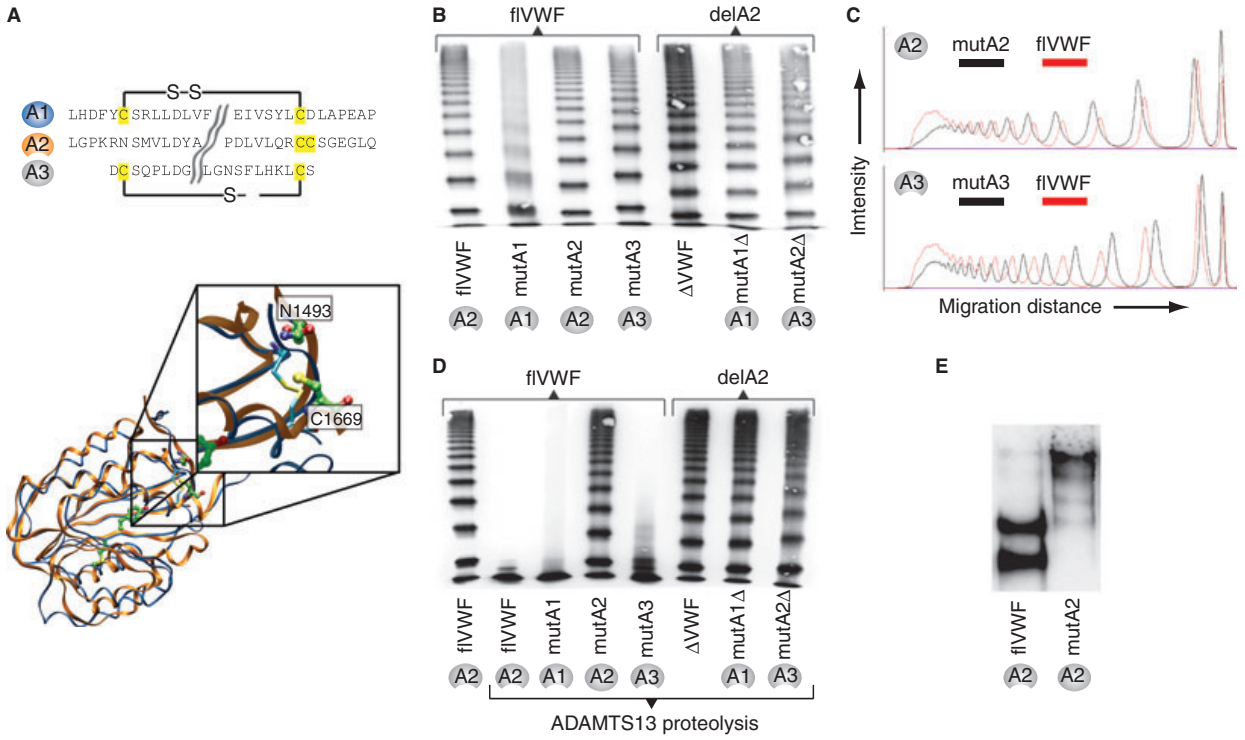


Fig. 5. (A) The A1 domain (blue ribbons and blue carbon atoms) is locked by a disulfide bond between the termini, while the A2 domain (orange ribbons and green atoms) lacks this feature. Mutation N1493C would allow disulfide bond formation (magnification). In the schematic sequence alignment of the VWF A domains only the N- and C-terminal sequences are shown (see Supporting Information for full alignment), cysteine residues are highlighted and disulfide bonds are shown as brackets. Closed circles indicate disulfide linkage of the domain termini; open circles indicate no disulfide linkage of the domain termini. (B) Electrophoretic multimer analysis of VWF variants. Downwards shifted bands indicate faster migration (mutA2) compared with flVWF, while upwards shifted bands indicate slower migration (mutA1 and mutA3). (C) Scan of selected lanes of the gel shown in Fig. 5(B): the shift of bands relative to flVWF multimers indicates faster migration of mutA2 and slower migration of mutA3. (D) ADAMTS13 proteolysis of VWF variants monitored by electrophoresis. Normal proteolysis is observed for flVWF, mutA1 and mutA3, whereas proteolysis is absent in all other variants including mutA2. (E) Full-length A2 domain mutant mutA2 in comparison with flVWF after ADAMTS13 treatment and reduction of disulfide bonds. Wild-type VWF is proteolyzed completely and displays the expected two proteolytic fragments after reduction, whereas mutA2 is not proteolysed.

Table 1 Overview of the expressed VWF variants, their sensitivity against ADAMTS13 and the relative migration speed of the multimers in electrophoresis

Name	Mutations	Cleavage by ADAMTS13	Migration relative to flVWF
Based on full-length VWF			
flVWF	None	+	–
mutA1	C1272S/D1459C	+	Slower
mutA2	N1493C/C1670S	–	Faster
mutA3	C1686S/S1873C	+	Slower
Based on VWF without A2 domain			
ΔVWF	None	–	–
mutA1Δ	C1272S/D1459C	–	–
mutA3Δ	C1686S/S1873C	–	–

comparison with flVWF. The mutA2 variant migrated faster than flVWF, suggesting a more compact structure with higher electrophoretic mobility (Fig. 5B,C). In contrast, removing the disulfide bridges in the A1 (C1272S and D1459C, mutA1) or A3 domain (C1686S and S1873S, mutA3) towards an open structure as in wild-type A2 resulted in a decrease of the electrophoretic mobility, both in the presence and absence of A2, respectively (Fig. 5B,C). These results support the assumption of the generation of a cysteine-bridge connection of the A2 N and C terminus analogous to A1 and A3. We then exposed all variants to ADAMTS13 and monitored proteolysis by multimer analysis. We could show that, in contrast to flVWF, ADAMTS13 proteolysis of mutA2 was completely absent, similar to the A2 domain deleted VWF variants ΔVWF, mutA1Δ and mutA3Δ (Fig. 5D). This was further confirmed by reduction of cysteine bonds by β-mercaptoethanol to exclude the possibility that the mutA2 variant was actually proteolysed but just held together by the created cysteine bonds (Fig. 5E). Opening the disulfide bond of the A1 and A3 domain by mutagenesis in A2 domain deleted VWF (variants mutA1D and mutA3D) did not result in proteolytic susceptibility of the respective domains, indicating that the homology of A1 and A3 to A2 is too low for substrate recognition by ADAMTS13 (Fig. 5D). A speculative physiological

effect of the loss of the termini-linking disulfide bonds in A1 and A3 domain would suggest a significantly reduced affinity of A1 to collagen and GPIb and of A3 to collagen due to the disturbed structural integrity, especially in blood flow.

Conclusions

We here show by simulations and *in vitro* mutagenesis how force-induced partial unfolding is required for ADAMTS13-mediated cleavage of VWF A2. The unfolding and activation mechanism of A2 can be abolished by a single mutation, N1493C, in analogy with the mechanism that presumably protects A1 and A3 from unfolding and loss of function. We find the C-terminal part of VWF A2 to be unraveled under force, suggesting ADAMTS13 to primarily recognize this partially unfolded domain rather than the native state of A2. This is in excellent agreement with recent *in vitro* studies on the interaction of VWF A2 with ADAMTS13 [34,35]. Our data also suggest that this ADAMTS13-susceptible unfolding intermediate corresponds to the intermediate very recently observed in A2 single molecule stretching experiments [17].

The force-sensing mechanism of the A2 domain provides an intriguing explanation for the size regulation of ULVWF: larger multimers involve higher pulling forces and therefore higher unfolding rates at a given shear flow. As a result, larger VWF is cleaved more readily. The forces required for the exposure of the cleavage site in A2 as observed here (up to 1000 pN) can be expected to be significantly larger than those inducing unfolding in *in vivo* conditions due to the short nanosecond time scale of the simulations within which the unfolding is forced to occur [36]. Under physiological conditions, cleavage will preferentially occur for the upper limit of VWF multimer sizes, and thus under flow conditions that lead to tensile forces beyond the 5–10 pN estimated for average VWF sizes [6] (A Alexander-Katz, personal communication, 2009). Indeed, the experimental evaluation of VWF A2 unfolding by optical tweezers suggests forces in the range of 10–15 pN [17].

The intermediates of mechanical unfolding of the VWF A2 domain observed here (and not a static intact equilibrium state) represent the substrate of ADAMTS13. These dynamics of the A2 domain during unfolding are prerequisite to explore the structural and functional determinants of A2 recognition by ADAMTS13. The gained knowledge (e.g. the actual structure of the partially unfolded A2 domain) can be used to design inhibitors of ADAMTS13 and can provide a route to drugs targeting enhanced VWF cleavage in blood.

VWF A2 mutations previously identified as causing von Willebrand disease type IIA due to an increased susceptibility to ADAMTS13 [24] cleavage can now be rationalized on the basis of our model. They can be expected to involve destabilization of the overall A2 structure by forcing charged groups into regions of hydrophobic packing (I1628T and G1629E), perturbing β-turn formation between the VWF A2 secondary structure elements β5 and α5 (G1631D), or by destabilizing A2 due to a drastic increase in spatial demand of the side chain (G1609R). Structural destabilization in turn facilitates A2 unfolding and cleavage site exposure to ADAMTS13.

While the C terminal part of the A2 domain follows a highly conserved unfolding pattern if subjected to tensile stress, the N terminal 'knotted' Rossmann fold remains completely intact even under high forces. We hypothesize that the second important function of A2, the proposed inhibition of the A1 GPIb interaction [37], which mediates the binding of VWF to platelets, is located at this force-resistant part of the domain. Thereby, as a consequence of the two distinct Rossmann topologies within the A2 domain, size regulation of VWF by ADAMTS13 does not affect platelet interaction. As a second consequence of the unconventional Rossmann fold, we find strain to internally propagate selectively to the ADAMTS13 cleavage site, bringing the peptide bond under tension. We hypothesize that this specific force-activation affects the catalytic activity of ADAMTS13, as a direct impact of the A2 mechanics on the A2-ADAMTS13 biochemistry, similar to what has been shown for disulfide bond cleavage by DTT and thioredoxin [38].

We here assumed that the stretching force in VWF propagates to A2 primarily along the protein backbone. A full A1-A2-A3 structure is needed to re-examine the unfolding mechanism taking inter-domain interactions into account, as a next important step towards deciphering the molecular details of VWF mechanical response.

Another example for a Rossmann fold in which the termini are locked together by a disulfide bond is the VWF type A domain of human capillary morphogenesis protein 2, interestingly again a collagen-binding adhesion protein [39]. To what extent nature has made use of the Rossmann fold as a module that can be reversibly switched into a force-resistant state remains to be seen.

Addendum

C. Baldauf, R. Schneppenheim and F. Gräter designed the research described in this article. C. Baldauf, W. Stacklies and J. Zhou performed modeling and simulation described in the article. T. Obser, A. Pieconka and S. Schneppenheim performed experiments. C. Baldauf, R. Schneppenheim, W. Stacklies and U. Budde analyzed and interpreted the data. C. Baldauf, R. Schneppenheim, W. Stacklies and F. Gräter wrote the paper.

Acknowledgements

The authors thank M. F. Schneider and A. Alexander-Katz for fruitful discussions and W. Wong and T. Springer for early communication of their findings. CB thanks R. Meier and W. Sippl (Institute of Pharmaceutical Chemistry, Martin-Luther University of Halle-Wittenberg) for the possibility of a research

stay and for help with the creation of the homology model. CB and FG thank Y. Yin for performing structure searches. CB is grateful for a Feodor Lynen Fellowship by the Alexander von Humboldt foundation.

Disclosure of Conflict of Interests

The authors state that they have no conflict of interests.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. (A) Superposition of the average structures under 10 and 100 pN in the folded state used for FDA. Structures are averages over 300 ns, respectively. (B) Superposition of the average structures under 10 and 100 pN of the unfolding intermediate. Structures are averages over 300 ns, respectively.
Figure S2. Multiple sequence alignment used as basis for homology modeling.
Figure S3. (A) Verification of the Homology Model with ProSA 2003: The energy analysis is smoothed with a window size of 30 aa. Characterizing the model with ProSA-Web shows a Z-score for the raw model of -6.99, and of -8.11 for the model after 10 ns MD simulation. The Z-score for the structure model published by Sutherland et al. is -7.84. (B) Backbone rmsd of the wild type A2 domain monitored in three independent 30 ns MD simulations. (C) Backbone rmsd of the mutant A2 (N1493C/C1670S) domain monitored in three independent 30 ns MD simulations.
Figure S4. Superposition of 3GXB (silver) and the homology model (orange), the main chain RMSD is 0.189 nm.
Figure S5. The force profiles for three independent force-probe MD simulations of our VWF A2 domain model.
Dataset S1. Homology model of the A2 domain including VWF residues 1488 to 1676 in PDB-format.
Video S1. Visualization of a VWF A2 Domain Force Probe MD Simulation.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

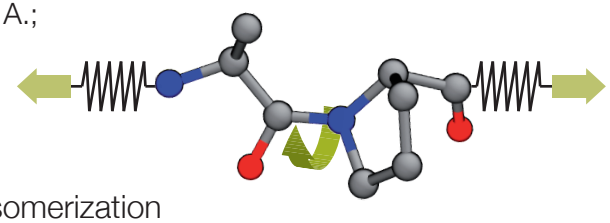
References

1 Sadler JE. New concepts in von Willebrand disease. *Annu Rev Med* 2005; **56**: 173–91.
2 Ruggeri ZM. The role of von Willebrand factor in thrombus formation. *Thromb Res* 2007; **120**: S5–9.
3 Millar CM, Brown SA. Oligosaccharide structures of von Willebrand factor and their potential role in von Willebrand disease. *Blood Rev* 2006; **20**: 83–92.
4 Dent JA, Berkowitz SD, Ware J, Kasper CK, Ruggeri ZM. Identification of a cleavage site directing the immunochemical detection of molecular abnormalities in type IIa von Willebrand-factor. *Proc Natl Acad Sci USA* 1990; **87**: 6306–10.
5 Sadler JE. A new name in thrombosis, ADAMTS13. *Proc Natl Acad Sci USA* 2002; **99**: 11552–4.
6 Alexander-Katz A, Schneider MF, Schneider SW, Wixforth A, Netz RR. Shear-flow-induced unfolding of polymeric globules. *Phys Rev Lett* 2006; **97**: 138101.
7 Schneider SW, Nuschele S, Wixforth A, Gorzelanny C, Alexander-Katz A, Netz RR, Schneider MF. Shear-induced unfolding triggers adhesion of von Willebrand factor fibers. *Proc Natl Acad Sci USA* 2007; **104**: 7899–903.
8 Perutelli P, Amato S, Molinari AC. ADAMTS-13 activity in von Willebrand disease. *Thromb Res* 2006; **117**: 685–8.
9 Shida Y, Nishio K, Sugimoto M, Mizuno T, Hamada M, Kato S, Matsumoto M, Okuchi K, Fujimura Y, Yoshioka A. Functional imaging of shear-dependent activity of ADAMTS13 in regulating mural thrombus growth under whole blood flow conditions. *Blood* 2008; **111**: 1295–8.
10 Sadler JE. Von Willebrand factor, ADAMTS13, and thrombotic thrombocytopenic purpura. *Blood* 2008; **112**: 11–8.
11 Von Willebrand EA. Hereditar pseudohemofili. *Fin Lakaresallsk Handl* 1926; **68**: 7–112.
12 Emsley J, Cruz M, Handin R, Liddington R. Crystal structure of the von Willebrand factor A1 domain and implications for the binding of platelet glycoprotein Ib. *J Biol Chem* 1998; **273**: 10396–401.
13 Huizinga EG, van der Plas RM, Kroon J, Sixma JJ, Gros P. Crystal structure of the A3 domain of human von Willebrand factor: implications for collagen binding. *Structure* 1997; **5**: 1147–56.
14 Sutherland JJ, O'Brien LA, Lillicrap D, Weaver DF. Molecular modeling of the von Willebrand factor A2 domain and the effects of associated type 2A von Willebrand disease mutations. *J Mol Model* 2004; **10**: 259–70.
15 Zhang Q, Zhou Y-F, Zhang C-Z, Zhang X, Lu C, Springer TA. Structural specializations of A2, a force-sensing domain in the ultra-large vascular protein von Willebrand factor. *Proc Natl Acad Sci USA* 2009; **106**: 9226–31.
16 Dong JF, Moake JL, Nolasco L, Bernardo A, Arceneaux W, Shrimpton CN, Schade AJ, McIntire LV, Fujikawa K, Lopez JA. ADAMTS-13 rapidly cleaves newly secreted ultralarge von Willebrand factor multimers on the endothelial surface under flowing conditions. *Blood* 2002; **100**: 4033–9.
17 Zhang XH, Halvorsen K, Zhang CZ, Wong WP, Springer TA. Mechanoenzymatic cleavage of the ultralarge vascular protein von Willebrand factor. *Science* 2009; **324**: 1330–4.
18 Stacklies W, Vega MC, Wilmanns M, Gräter F. Mechanical network in titin immunoglobulin from force distribution analysis. *PLoS Comput Biol* 2009; **5**: e1000306.
19 LindahlE,HessB,VanderSpoelD.GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model* 2001; **7**: 306–17.
20 Jorgensen WL, Ulmschneider JP, Tirado-Rives J. Free energies of hydration from a generalized Born model and an ALL-atom force field. *J Phys Chem B* 2004; **108**: 16264–70.
21 Lawrence CP, Skinner JL. Flexible TIP4P model for molecular dynamics simulation of liquid water. *Chem Phys Lett* 2003; **372**: 842–7.
22 Hess B, Bekker H, Berendsen HJC, Fraaije J. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem* 1997; **18**: 1463–72.
23 Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 2008; **4**: 435–47.
24 Hassenpflug WA, Budde U, Obser T, Angerhaus D, Drewke E, Schneppenheim S, Schneppenheim R. Impact of mutations in the von Willebrand factor A2 domain on ADAMTS 13-dependent proteolysis. *Blood* 2006; **107**: 2339–45.
25 Furlan M, Robles R, Lammle B. Partial purification and characterization of a protease from human plasma cleaving von Willebrand factor to fragments produced by in vivo proteolysis. *Blood* 1996; **87**: 4223–34.

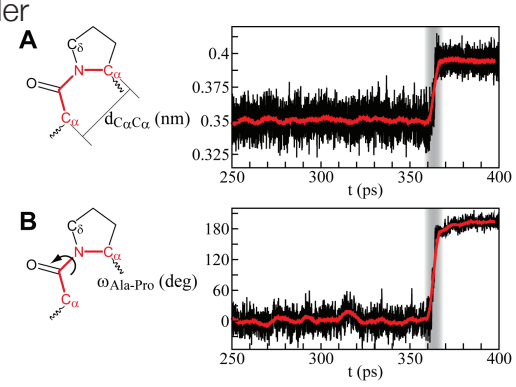
26 Schneppenheim R, Brassard J, Krey S, Budde U, Kunicki TJ, Holmberg L, Ware J, Ruggeri ZM. Defective dimerization of von Willebrand factor subunits due to a Cys- > Arg mutation in type IID von Willebrand disease. *Proc Natl Acad Sci USA* 1996; **93**: 3581–6.
27 Budde U, Schneppenheim R, Eikenboom J, Goodeve A, Will K, Drewke E, Castaman G, Rodeghiero F, Federici AB, Batlle J, Perez A, Meyer D, Mazurier C, Goudemand J, Ingerslev J, Habart D, Vorlova Z, Holmberg L, Lethagen S, Pasi J, *et al*. Detailed von Willebrand factor multimer analysis in patients with von Willebrand disease in the European study, molecular and clinical markers for the diagnosis and management of type 1 von Willebrand disease (MCMDM-1VWD). *J Thromb Haemost* 2008; **6**: 762–71.
28 Ruggeri ZM, Zimmerman TS. The complex multimeric composition of factor-VIII-von Willebrand factor. *Blood* 1981; **57**: 1140–3.
29 Schneppenheim R, Plendl H, Budde U. Luminography – an alternative assay for detection of von Willebrand factor multimers. *Thromb Haemost* 1988; **60**: 133–6.
30 Di Stasio E, Lancellotti S, Peyvandi F, Palla R, Mannucci PM, De Cristofaro R. Mechanistic studies on ADAMTS13 catalysis. *Biophys J* 2008; **95**: 2450–61.
31 Möhle K, Gussmann M, Hofmann H-J. Structural and energetic relations between beta turns. *J Comput Chem* 1997; **18**: 1415–30.
32 Fischer G. Chemical aspects of peptide bond isomerisation. *Chem Soc Rev* 2000; **29**: 119–27.
33 Valiaev A, Lim DW, Oas TG, Chilkoti A, Zauscher S. Force-induced prolyl cis-trans isomerization in elastin-like polypeptides. *J Am Chem Soc* 2007; **129**: 6491–7.
34 Wu JJ, Fujikawa K, McMullen BA, Chung DW. Characterization of a core binding site for ADAMTS-13 in the A2 domain of von Willebrand factor. *Proc Natl Acad Sci USA* 2006; **103**: 18470–4.
35 Gao WQ, Anderson PJ, Sadler JE. Extensive contacts between ADAMTS13 exosites and von Willebrand factor domain A2 contribute to substrate specificity. *Blood* 2008; **112**: 1713–9.
36 Evans E, Ritchie K. Dynamic strength of molecular adhesion bonds. *Biophys J* 1997; **72**: 1541–55.
37 Martin C, Morales LD, Cruz MA. Purified A2 domain of von Willebrand factor binds to the active conformation of von Willebrand factor and blocks the interaction with platelet glycoprotein Ib alpha. *J Thromb Haemost* 2007; **5**: 1363–70.
38 Wiita AP, Perez-Jimenez R, Walther KA, Grater F, Berne BJ, Holmgren A, Sanchez-Ruiz JM, Fernandez JM. Probing the chemistry of thioredoxin catalysis with force. *Nature* 2007; **450**: 124–7.
39 Lacy DB, Wigelsworth DJ, Scobie HM, Young JAT, Collier RJ. Crystal structure of the von Willebrand factor A domain of human capillary morphogenesis protein 2: an anthrax toxin receptor. *Proc Natl Acad Sci USA* 2004; **101**: 6367–72.

2.2 On the *cis* to *trans* isomerization of prolyl-peptide bonds under tension

Chen, J.; Edwards, S. A.; Gräter, F.; Baldauf, C.



On the *cis* to *trans* isomerization of prolyl-peptide bonds under tension



J. Phys. Chem. B
2012 (31), 9346-9351.

DOI: 10.1021/jp3042846

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

On the Cis to Trans Isomerization of Prolyl−Peptide Bonds under Tension

Jian Chen,[†] Scott A. Edwards,[‡,†] Frauke Gräter,*[,§,†] and Carsten Baldauf*[,∥,†]

[†]CAS-MPG Partner Institute and Key Laboratory for Computational Biology (PICB), 320 Yue Yang Road, Shanghai 200031, China,

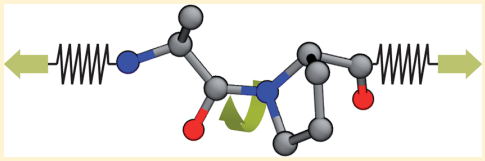
[‡]College of Physics and Technology, Shenzhen University, Shenzhen 518060, Guangdong, China,

[§]Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany, and

[∥]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin-Dahlem, Germany

Supporting Information

ABSTRACT: The cis peptide bond is a characteristic feature of turns in protein structures and can play the role of a hinge in protein folding. Such cis conformations are most commonly found at peptide bonds immediately preceding proline residues, as the cis and trans states for such bonds are close in energy. However, isomerization over the high rotational barrier is slow. In this study, we investigate how mechanical force accelerates the cis to trans isomerization of the prolyl−peptide bond in a stretched backbone. We employ hybrid quantum mechanical/molecular mechanical force-clamp molecular dynamics simulations in order to describe the electronic effects involved. Under tension, the bond order of the prolyl−peptide bond decreases from a partially double toward a single bond, involving a reduction in the electronic conjugation around the peptide bond. The conformational change from cis to extended trans takes place within a few femtoseconds through a nonplanar state of the nitrogen of the peptide moiety in the transition state region, whereupon the partial double-bond character and planarity of the peptide bond in the final trans state is restored. Our findings give insight into how prolyl−peptide bonds might act as force-modulated mechanical timers or switches in the refolding of proteins.



INTRODUCTION

The imino acid proline is unique among the proteinogenic amino acids, as the side chain is linked back to the backbone via a bond to the nitrogen, making it a secondary amine. This feature is the basis for a number of distinctive structural characteristics of proline in the context of peptide sequences: (i) the lack of a polar hydrogen prohibits proline from acting as a H-bond donor; (ii) the backbone torsion angle ϕ (angle $CNC_\alpha C$) is part of the heterocycle and thus conformationally restricted; and (iii) the trans state of the prolyl peptide bond is energetically only slightly preferred over the cis state, as in either case the C_β of the preceding residue is close to a carbon atom of proline (either C_α or C_δ). This is reflected in the analysis of high-resolution X-ray structures from the RCSB protein data bank,[1] where more than 90% of the cis peptide bonds in proteins are Xaa-Pro imide bonds.[2] Still, the barrier between both states is high, and the interconversion from cis to trans is slow in equilibrium.[3−6] The general mechanism of peptide bond isomerization involves a change of the bond order from a partial double bond to a single bond, followed by a return to partial double-bond character. This is caused by the pyrrolidine N changing from a planar sp^2 state to a pyramidal sp^3 state with a lone pair orbital. The resulting lone pair dipole interacts with the C═O dipole and influences the peptide bond rotation during isomerization.[7] Furthermore, the transition state is stabilized by the interaction between the N−H of the following peptide bond with the lone pair of the sp^3 N.[5,8] This isomerization can be catalyzed by peptidyl−prolyl cis−trans isomerases.[9−12]

Proline is found in a number of structural proteins, including tropoelastin. Elastin-like polypeptides (ELP) are models of tropoelastin and consist of multiple repeats of the sequence Val-Pro-Gly-Xaa-Gly. The high Pro content results in a high share of cis-prolyl peptide bonds. Single-molecule force-spectroscopy experiments by Zauscher and co-workers on ELP and poly-Pro peptides revealed a temperature-independent extensional transition upon application of a stretching force. This increase of the contour length was interpreted as force-induced cis to trans isomerization of multiple prolyl−peptide bonds.[13] A scheme of the cis to trans isomerization is shown in Figure 1A.

Another place where cis peptide bonds and Pro residues are frequently found in proteins is at type VI β turns, with a cis prolyl−peptide bond between residues 2 and 3 of the turn.[14,15] Such turns act as hinges during protein folding and arrange helices and strands in their native three-dimensional fold. The structural difference between the cis and the trans conformation of the peptide bond is significant, and an isomerization might substantially alter protein (re)folding pathways and timings.[6] An example where mechanical force meets the regulation of

Received: May 3, 2012
Revised: July 4, 2012
Published: July 6, 2012

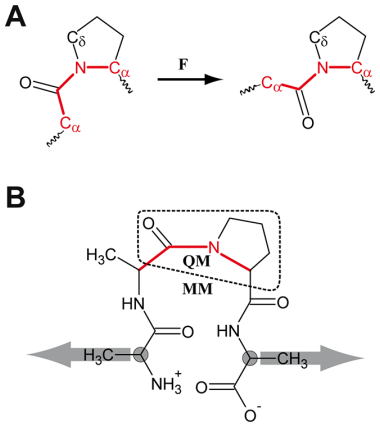


Figure 1. (A) Schematic representation of the cis to trans isomerization. Bonds and atoms defining the peptide bond torsion are highlighted in red. (B) Schematic representation of the AAPA peptide as modeled in our simulations. QM region is separated from the MM region by a dashed line. In the force-clamp simulations, the C_α atoms of Ala1 and Ala4 (gray spheres) were subjected to a constant force in opposite directions.

physiological function is the giant blood protein von Willebrand factor (VWF). The multimers of VWF are sensitive to the shear forces present in flowing blood and translate shear flow to an extensional force along its length axis.[16,17] As a result, individual domains of VWF partially unfold and eventually βVIa turns with cis prolyl–peptide bonds are under direct extensional force. Possible refolding is hindered if the prolyl–peptide bond of the βVIa turn isomerizes to the trans form.[18] Indeed, in a subset of the optical tweezers experiments reported by Springer and co-workers refolding of the tethered A2 was delayed.[19] This can be interpreted as force-induced cis to trans isomerization of the prolyl–peptide bond in the A2 domain that hampers refolding.

The question arises how mechanical force facilitates the prolyl cis to trans interconversion over the rotational barrier of 60–80 kJ/mol.[5,8] Because of a general interest in mechanochemistry and with regard to the potential physiological importance of this process, we present here a study of the forced cis to trans isomerization of the prolyl–peptide bond by classical molecular mechanics and hybrid quantum mechanics/molecular mechanics (QM/MM) force-clamp molecular dynamics (FCMD) simulations. Previous theoretical efforts to reveal the mechanism of isomerization have employed optimizations along the reaction pathway from trans to cis, based on a QM or QM/MM description.[5,7,8,20] We obtained dynamic trajectories and kinetic information as a function of the applied external force while also taking full solvation into account. We find mechanical stretching of the peptide to weaken the peptide bond, making the distorted nonplanar transition state region with reduced electronic delocalization accessible within shorter time scales.

RESULTS AND DISCUSSION

To trigger the isomerization transition from cis to trans prolyl–peptide bond, we applied an extensional force to the backbone of a short model peptide, AAPA (Figure 1B), dissolved in water during FCMD simulations. The C_α of residues Ala1 and Ala4 were subjected to a constant pulling force (Figure 1B). In response to the applied force, AAPA adopted an extended configuration, with the prolyl–peptide bond maintaining its cis conformation.

Starting from this mechanically stretched peptide conformation, we next investigated the isomerization reaction of the prolyl peptide bond in AAPA. The cis to trans isomerization reaction under force can be expected to require a change of bond character and hybridization of the involved atoms. These electronic structure effects are not accurately described by classical force fields, which are parametrized on the basis of equilibrium states without considering transition states. We therefore relied on hybrid quantum mechanical and molecular mechanical simulations (QM/MM) to consider electronic effects relating to the prolyl–peptide bond (Figure 1B). Constant forces ranging from 1.1 to 5 nN were applied; the lowest force for which isomerization was observed within a time of 500 ps was 3 nN. The experiments of Valiev et al. on ELP were performed with a constant pulling velocity; the forces they measured prior to the isomerization event were lower than ours by a factor of 10, in the range of 200–260 pN.[13] Higher forces induced isomerization on shorter time scales, with a subpicosecond transition in the 4–5 nN range (Figure 2). The force

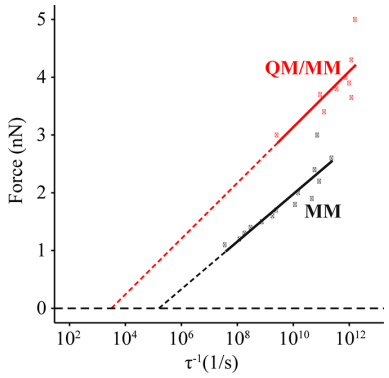


Figure 2. Isomerization lifetimes τ at different forces as obtained from FCMD simulations. Lifetimes from QM/MM simulations are shown in red and those from pure MM simulations in black. Fits of the linear Bell model[21] to the data are shown as lines.

dependence of the transition time was fitted with the linear Bell model,[21,22] yielding a distance Δx of 0.02 ± 0.007 nm between the reactant (cis) state and the transition state. The model predicts a lifetime at zero force (and similarly at the relatively low forces in experiments[13]) on the order of $\tau_0 \approx 10^{-4}$ s. However, the limited range of time scales accessible in the QM/MM simulations entails a large uncertainty of these values and does not allow us to distinguish between linear (like the Bell model) and nonlinear models like the Dudko–Hummer model.[23] Nevertheless, our simulations can be semiquantitatively validated by comparison to experiment. Using the Eyring equation, the estimated lifetime can be converted into an activation free energy of ~60 kJ/mol, which is in line with the experimental zero force barrier of ~60–80 kJ/mol.[13,24−26] The barrier appears indeed slightly underestimated, as we predict life times on the order of 0.1 ms instead of the lifetimes in atomic force microscopy experiments, which are larger than milliseconds.[13] However, the overall agreement is satisfying given the difference in force application (here constant force, in

experiments constant velocity pulling) and the limited time scales of our simulations necessitating a defective extrapolation.

At a first glance, performing the same FCMD simulations of the AAPA peptide in a pure MM description appears to reproduce the main findings of the QM/MM simulations (Figure 2). Indeed, a similar linear dependency of the logarithm of the lifetime of the cis isomer to the applied force was observed, now spanning a larger force and time range. The Bell model fit gives $\Delta x = 0.023 \pm 0.004$ nm, comparable to the value obtained from our QM/MM simulations. However, for a given lifetime, smaller forces are required in the MM simulations as compared to QM/MM, as reflected by a shorter zero-force lifetime of $\tau_0 \approx 10^{-5}$ s. Even with an error of at least 1 order of magnitude up or down, we can conclude that the MM description underestimates the transition free energy barrier for isomerization. As we will show further below, this is due to the lack of changes in hybridization and bond order during the process.

How does the cis to trans isomerization proceed? We next analyze the mechanism of the isomerization via geometrical properties of the model system. In the following we will focus on the QM/MM trajectory at 3 nN, the lowest force for which isomerization was observed on accessible time scales. The macroscopic order parameter observed in pulling experiments with optical tweezers or atomic force microscopy is an increase of the contour length, which here corresponds to the distance $d_{C\alpha C\alpha}$ of the C_α atoms adjacent to the prolyl−peptide bond. During the simulation a sudden jump of $d_{C\alpha C\alpha}$ from about 0.35 to roughly 0.4 nm can be observed between 360 and 370 ps of simulation time (Figure 3A), which indicates a two-step process. The resulting difference between the stretched cis and trans states is 0.05 nm; the difference between the equilibrium cis state and the stretched trans state is about 0.1 nm, well in line with an investigation by Reimer and Fischer, who measured C_α distances around Pro residues of selected X-ray structures from the Protein Data Bank. For residues directly adjacent to the Pro residue, they reported an increase in $d_{C\alpha C\alpha}$ from cis to trans of about 0.08 nm.[27] The change in prolyl peptide bond geometry between the two isomers translates into larger shifts in the adjacent backbone segments. This is why Valiev et al. consider an increase of the contour length by 0.2 nm per cis to trans isomerized prolyl−peptide bond as an effect on the overall peptide conformation.[13] Simultaneously with the increase in contour length, the dihedral angle $\omega_{\text{Ala-Pro}}$ (defined by the backbone atoms C_α−C−N−C_α) rotates from 0° (cis) to 180° (trans) (Figure 3B). This behavior confirms that experimental observations of jumps in contour length during the stretching of ELP[13,28] and VWF[29,19] can be interpreted as prolyl−peptide bond isomerization into the trans state.

The C_α distance corresponds to the experimental observables, yet it offers only limited insight into the mechanism of isomerization. Thus, we focus on further order parameters describing the transition mechanism. One such parameter is the bond length d_{CN} of the peptide bond, which is plotted in Figure 3C. In equilibrium, the C−N peptide bond is shorter than a typical single C−N bond, as the p orbitals of N and carboxyl C are conjugated. The single peak of d_{CN} from 0.14 to 0.145 nm and back coincides with the jumps in contour length and peptide bond torsion angle. This illustrates that the change of bond order from partially double to single and back is a consequence of the elimination of orbital conjugation and its reformation between C and N. The connected change in hybridization state of the peptide bond N can be tracked by the

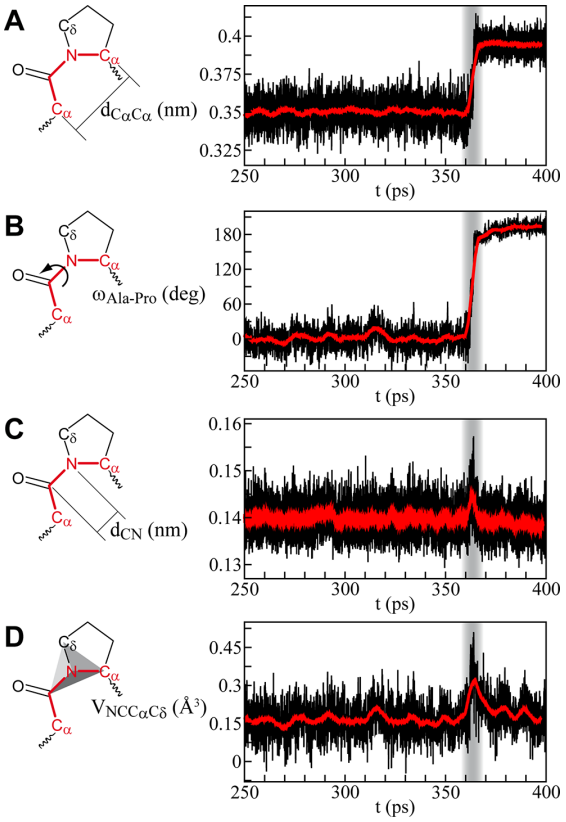


Figure 3. Mechanism of a cis to trans isomerization event (QM/MM, F = 3 nN). Different order parameters are shown with respect to the simulation time: (A) contour length given by the distance between the C_α atoms of Ala2 and Pro3 ($d_{C\alpha C\alpha}$); (B) torsion angle of the peptide bond between Ala2 and Pro3 ($\omega_{\text{Ala-Pro}}$); (C) length of the peptide bond between Ala2 and Pro3 (d_{CN}); (D) volume $V_{NCC\alpha C\delta}$ of the tetrahedron defined by the atoms C (of Ala2), N, C_α, and C_δ (of Pro3) as a measure for N hybridization. Black lines show the measured values from the FCMD simulation; red lines are running averages with a window size of 100 data points. Time range of isomerization is highlighted by gray rectangles.

volume of the tetrahedron consisting of the nitrogen and the atoms bound to it (C, C_α, C_δ) as shown in Figure 3D. In the stretched cis state, the pyramid N, C, C_α, C_δ has a volume of 0.15 Å^3. During isomerization, this volume peaks at 0.3 Å^3 and relaxes back to about 0.15 Å^3 (Figure 3D) following the transition of the N hybridization from sp^2 to sp^3 and back to sp^2.

The force-induced cis to trans isomerization of the prolyl−peptide bond occurs through a rotation of the peptide bond torsion angle by roughly 100° (Figure 4A). According to an overview of the possible transition states by Fischer,[30] this transition state can be characterized as syn/exo. Only the Ala2-Pro3 peptide bond rotates, while the other torsions have been already stretched out by the pulling force and remain at extended values around 180° (except the ϕ_{Pro3} which is fixed in the heterocycle). The transition state can be more readily analyzed by relating the major order parameter, the prolyl dihedral angle ω, to the other degrees of freedom affected by the isomerization, namely, $d_{C\alpha C\alpha}$, d_{CN}, and $V_{NCC\alpha C\delta}$ (Figure

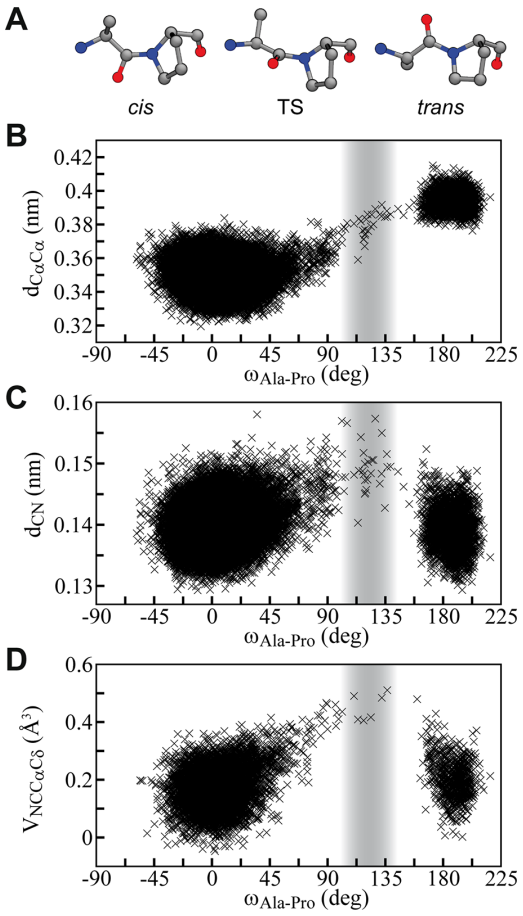


Figure 4. (A) Representative snapshots from the QM/MM trajectory at F = 3 nN, exemplifying cis, transition (TS), and trans state. (B) Distance of C_α atoms of Ala2 and Pro3 ($d_{C\alpha C\alpha}$) plotted versus the torsion angle of the peptide bond between Ala2 and Pro3 ($\omega_{Ala-Pro}$). (C) Length of the peptide bond between Ala2 and Pro3 (d_{CN}) plotted versus $\omega_{Ala-Pro}$. (D) Volume $V_{NCC\alpha C\delta}$ of the tetrahedron defined by the atoms C (of Ala2), N, C_α, and C_δ (of Pro3) plotted versus $\omega_{Ala-Pro}$.

4B–D). The change in peptide bond torsion angle $\omega_{Ala-Pro}$ coincides with the change in contour length measured by $d_{C\alpha C\alpha}$ in Figure 4B. Even though our force-clamp MD simulations do not allow us to draw quantitative conclusions on the transition state, Figure 4B clearly features an area with a low population of states (gray shade), within which the transition state is likely to be located. Notably, this transition state region features an angle $\omega \geq 100°$, slightly beyond the halfway rotation, and a contour length $d_{C\alpha C\alpha}$ between 0.36 and 0.39 nm. This corresponds to an elongation with respect to the product state by 0.01–0.04 nm. Thus, the change in contour length from the cis to the ~100° rotated state serves as a direct structural interpretation of the Δx of 0.02 ± 0.007 nm obtained from the Bell model fit (see above) and suggests this as the transition state. Both the peptide bond length (d_{CN}) and $V_{NCC\alpha C\delta}$ only show increased values within the sparsely populated region of conformational space that includes the transition state and then return to equilibrium values (Figure 4C and 4D). Again, this highlights the change in bond character (double to single and back to double) and hybridization (sp^2 to sp^3 and back to sp^2) during isomerization. In all our simulations, employing a range of different forces, isomerization occurs by rotation from 0° via 90° to 180°, with the lone pair of the sp^3 passing an ecliptic conformation with the preceding C_α. Rotation via −90°, with an ecliptic orientation of the lone pair and the carboxyl O, was never observed. In a previous theoretical study of the isomerization in the absence of mechanical strain both rotation directions were observed.[8] Another important feature of the mechanism shown in previous studies is that the pyramidal state of the peptide bond N is stabilized by a N···H interaction with a downstream backbone NH.[5,8] Here, instead, activation of the prolyl–peptide bond is not promoted by an attacking nucleophile but by mechanical force.

How can the described force-induced isomerization mechanism (Figure 4) be related to the increase in rate (Figure 2)? Forces of a few nN extend the C–N bond to a length that is closer to the single bond (0.147 nm) than to the double bond (0.132 nm). This change in length of the C–N bond in the cis and trans state as well as for the transition state region is shown in Supporting Information Figure 2 for the sampled range of forces. The transition state region features bond lengths typical of a single bond, while the product trans state relaxes back to partial double-bond character. Stress relief after isomerization allows bond lengths even smaller than that of the cis state at the same force. Thus, application of mechanical force destabilizes the reactant cis state, thereby moving it closer to the transition state. It may be that a stabilization of the transition state by force is another factor for lowering the activation barrier, but the limited sampling of transitions does not allow the inference of a relationship between the transition state energy of bond character and force (Supporting Information Figure 2).

From a technical point of view, it is interesting to compare the mechanism of isomerization as observed in the QM/MM description with the pure OPLS-AA force field treatment. In both setups, sudden and concurrent changes of the contour length and the peptide bond torsion angle, the two main order parameters representing the peptide bond isomerization, can be observed (Figure 3 and Supporting Information Figure 3). A clear difference becomes obvious when monitoring d_{CN} and $V_{NCC\alpha C\delta}$. The isomerization mechanism involves transient femtosecond scale changes in both parameters with the QM/MM treatment (Figure 3), whereas no such changes occur in the classical force field model (Supporting Information Figure 3). In the pure MM treatment, the force constant of the prolyl peptide bond potential remains unmodified and cannot reflect the bond order changes of the peptide bond undergoing isomerization. Furthermore, the change of the hybridization state of the nitrogen, measured via the tetrahedron volume $V_{NCC\alpha C\delta}$, is prevented by the improper dihedral potentials exerted on the peptide bond in the force field description of the system. Thus, as expected, only the QM/MM simulation is able to reflect the nature of the isomerization of the prolyl–peptide bond, most importantly including the transient sp^3 hybridization of N associated with a loss of π-electron conjugation of the prolyl–peptide bond.

CONCLUSION

In this work, we were able to elucidate the mechanism of force-induced cis to trans isomerization of the prolyl–peptide bond with QM/MM FCMD simulations. Tensile force releases the partial double bond and converts it to a single bond, rather like

a clutch. Isomerization (rotation around the peptide bond) can then occur, and afterward the 'clutch' closes and the partial double bond reforms. In quantitative agreement with experiments, force increases the isomerization rate. One important factor for the force-induced acceleration is a lengthening of the reactive C−N bond by force toward the single-bond character of the transition state.

A previously discussed mechanism for the cis to trans isomerization in equilibrium by Karplus and co-workers is based on the interaction of a downstream amide hydrogen with the free electron pair of the transition state nitrogen.[5,8] Such a mechanism is impossible in the stretched state of the peptide under tensile force. Instead, force appears to take up the role of the activating stimulus and stabilizes the transition state (the change of bond order and hybridization) which allows for isomerization. Another interesting observation of our study is that the direction of the isomerization always proceeds from 0° via 90° to 180°. Apparently, the chirality of the peptide in combination with the strain along its length axis results in a preferred rotation direction. The semiquantitative agreement of our MM and QM/MM simulations on the lifetime−force relation is of methodological interest. However, the limitations of molecular mechanics become obvious when studying the actual isomerization mechanism whose main features, the changes of the bond order and the hybridization state, can only be accounted for by electronic structure theory.

Our study demonstrates that cis to trans isomerization can be triggered by mechanical force. It contributes to an interpretation of experimental findings on the behavior of ELP under tensile force.[13] Our findings are of special relevance to the purported regulatory role played by force-triggered cis to trans isomerization of the prolyl−peptide bond, in which it acts as a folding timer: in force-responsive proteins, cis prolyl−peptide bonds can isomerize to the trans state and hinder refolding until spontaneous isomerization returns it to the original state. We note that our results suggest only a minor acceleration of the cis to trans isomerization (by a factor of less than 10) at physiological forces of less than a few 100 pN. Nevertheless, if isomerization is a rate-limiting step for protein folding, tensile forces can tune the competition between proline isomerization and folding, thereby potentially altering folding pathways. We speculate that such a mechanism may be functionally important for the blood protein VWF.

METHODS

The tetrameric peptide Ala1-Ala2-Pro3-Ala4 (AAPA) was modeled in a type VIa β-turn conformation and immersed within a cubic TIP4P[31] water box. Classical MM MD simulations were performed with Gromacs 3.3.1[32] and the OPLS-AA force field,[33] with an MD step size of 2 fs. Cut offs were applied to van der Waals interactions, and the particle-mesh Ewald method was used for long-range electrostatics.[34] The simulations were carried out in an NPT ensemble coupling to a Nosé−Hoover thermostat[35,36] of 300 K and to a Parinello−Rahman barostat of 1 atm.[37] The initial system was prepared by performing a free MD simulation in equilibrium for 50 ns. Later, external stretching force was applied to the C_α atoms of residues Ala1 and Ala4 (highlighted in Figure 1B). A series of FCMD simulations[38] was performed with constant forces ranging from 0.1 to 3 nN; a list with all forces for which cis to trans isomerization occurred can be found in Supporting Information Table S1.

The combined QM/MM simulations under tension were performed with Gromacs-3.3.1[32] and Gaussian03.[39] As shown in Figure 1B, the tetrapeptide AAPA was divided into a QM region and a MM region by cutting the carbon−carbon bonds, as the dashed line shows. The QM region with 16 atoms was simulated with B3LYP/6-31G* hybrid density functional theory[40,41] as implemented in Gaussian03. The carbon−carbon bonds connecting the QM and MM part were capped with hydrogens for QM calculations.[42,43] The QM part of the system was modeled under a Coulomb field of all MM atoms. The QM/MM FCMD simulations were carried out with constant forces from 2 to 5 nN with an integration step size of 1 fs, starting from a structure sampled from MM simulations at the same force.

ASSOCIATED CONTENT

Supporting Information

Details of the simulation setups and results of the equilibrium MD and pure MM FCMD simulations; brief analysis of the pure MM equilibration MD simulation of AAPA; tables with the lifetimes of cis states in pure MM and QM/MM FCMD simulations; plot with d_{CN} bond distances of the prolyl peptide bond in the cis, trans, and transition state over a range of applied forces; order parameters $d_{C\alpha C\alpha}$, $\omega_{Ala-Pro}$, d_{CN}, and $V_{NCC\alpha C\delta}$ of the prolyl peptide bond from a pure MM FCMD simulation; exemplary structure files of the initial conformation, extended cis conformation, and extended trans conformation of Ala-Ala-Pro-Ala. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: frauke.graeter@h-its.org; baldauf@fhi-berlin.mpg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Support by Gerrit Groenhof (Göttingen) for realization of the QM/MM scheme is gratefully acknowledged. C.B. thanks Hans-Jörg Hofmann (Leipzig) for his comments on the manuscript. The authors thank the Klaus Tschira foundation for financial support. J.C. acknowledges support by the Postdoctoral Research Program of the Shanghai Institutes for the Biological Sciences, Chinese Academy of Sciences (2011KIP514). F.G. and C.B. acknowledge funding from Deutsche Forschungsgemeinschaft (Forschergruppe 1543: Shear flow regulation in hemostasis−Bridging the gap between nanomechanics and clinical presentation), and F.G. acknowledges funding from the DAAD (Sino-German Junior Research Group for Biotechnology). S.A.E. acknowledges support from the Max Planck Society, a CAS Young International Scientist Fellowship (O91GC11401), and an NSFC Research Fellowship for International Young Scientists (O93DC11401).

REFERENCES

(1) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(2) Weiss, M.; Jabs, A.; Hilgenfeld, R. *Nat. Struct. Mol. Biol.* **1998**, *5*, 676−676.

(3) Salahuddin, A. *J. Biosci.* **1984**, *6*, 349−355.

(4) Stewart, D.; Sarkar, A.; Wampler, J. *J. Mol. Biol.* **1990**, *214*, 253−260.

(5) Fischer, S.; Dunbrack, R. L.; Karplus, M. *J. Am. Chem. Soc.* **1994**, *116*, 11931−11937.
(6) Wedemeyer, W.; Welker, E.; Scheraga, H. *Biochemistry* **2002**, *41*, 14637−14644.
(7) Wiberg, K. B.; Laidig, K. E. *J. Am. Chem. Soc.* **1987**, *109*, 5935−5943.
(8) Yonezawa, Y.; Nakata, K.; Sakakura, K.; Takada, T.; Nakamura, H. *J. Am. Chem. Soc.* **2009**, *131*, 4535−4540.
(9) Gothel, S.; Marahiel, M. *Cell. Mol. Life Sci.* **1999**, *55*, 423−436.
(10) Leuzzi, R.; Serino, L.; Scarselli, M.; Savino, S.; Fontana, M.; Monaci, E.; Taddei, A.; Fischer, G.; Rappuoli, R.; Pizza, M. *Mol. Microbiol.* **2005**, *58*, 669−681.
(11) Siegrist, R.; Zürcher, M.; Baumgartner, C.; Seiler, P.; Diederich, F.; Daum, S.; Fischer, G.; Klein, C.; Dangl, M.; Schwaiger, M. *Helv. Chim. Acta* **2007**, *90*, 217−259.
(12) Braun, M.; Hessamian-Alinejad, A.; De Lacroix, B.; Alvarez, B.; Fischer, G. *Molecules* **2008**, *13*, 995−1003.
(13) Valiaev, A.; Lim, D.; Oas, T.; Chilkoti, A.; Zauscher, S. *J. Am. Chem. Soc.* **2007**, *129*, 6491−6497.
(14) Hutchinson, E. G.; Thornton, J. M. *Protein Sci.* **1994**, *3*, 2207−2216.
(15) Möhle, K.; Gußmann, M.; Hofmann, H. *J. Comput. Chem.* **1997**, *18*, 1415−1430.
(16) Alexander-Katz, A.; Schneider, M. F.; Schneider, S. W.; Wixforth, A.; Netz, R. R. *Phys. Rev. Lett.* **2006**, *97*, 138101.
(17) Schneider, S. W.; Nuschele, S.; Wixforth, A.; Gorzelanny, C.; Alexander-Katz, A.; Netz, R. R.; Schneider, M. F. *Proc. Natl. Acad. Sci. U.S.A* **2007**, *104*, 7899−7903.
(18) Baldauf, C.; Schneppenheim, R.; Stacklies, W.; Obser, T.; Pieconka, A.; Schneppenheim, S.; Budde, U.; Zhou, J.; Gräter, F. *J. Thromb. Haemost.* **2009**, *7*, 2096−2105.
(19) Zhang, X.; Halvorsen, K.; Zhang, C.; Wong, W.; Springer, T. *Science* **2009**, *324*, 1330.
(20) Fischer, S.; Michnick, S.; Karplus, M. *Biochemistry* **1993**, *32*, 13830−13837.
(21) Bell, G. I. *Science* **1978**, *200*, 618−627.
(22) Evans, E.; Ritchie, K. *Biophys. J.* **1997**, *72*, 1541−1555.
(23) Dudko, O.; Hummer, G.; Szabo, A. *Phys. Rev. Lett.* **2006**, *96*, 108101.
(24) Fischer, G.; Schmid, F. X. *Biochemistry* **1990**, *29*, 2205−2212.
(25) Dugave, C.; Demange, L. *Chem. Rev.* **2003**, *103*, 2475−2532.
(26) Aliev, A.; Bhandal, S.; Murias, D. C. *J. Phys. Chem. A* **2009**, *113*, 10858−10865.
(27) Reimer, U.; Fischer, G. *Biophys. Chem.* **2002**, *96*, 203−212.
(28) Valiaev, A.; Lim, D.; Schmidler, S.; Clark, R.; Chilkoti, A.; Zauscher, S. *J. Am. Chem. Soc.* **2008**, *130*, 10939−10946.
(29) Springer, T. A. *J. Thromb. Haemost.* **2011**, *9*, 130−143.
(30) Fischer, G. *Chem. Soc. Rev.* **2000**, *29*, 119−127.
(31) Jorgensen, W.; Jenson, C. *J. Comput. Chem.* **1998**, *19*, 1179−1186.
(32) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306−317.
(33) Jorgensen, W.; Ulmschneider, J.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264−16270.
(34) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(35) Hoover, W. *Phys. Rev. A* **1985**, *31*, 1695.
(36) Nose, S. *Mol. Phys.* **1984**, *52*, 255−268.
(37) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(38) Grubmüller, H.; Heymann, B.; Tavan, P. *Science* **1996**, *271*, 997.
(39) Frisch, M. J.; et al. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
(40) Becke, A. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(41) Lee, C.; Yang, W.; Parr, R. *Phys. Rev. B* **1988**, *37*, 785.
(42) Field, M.; Bash, P.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700−733.
(43) Groenhof, G.; Bouxin-Cademartory, M.; Hess, B.; de Visser, S. P.; Berendsen, H. J. C.; Olivucci, M.; Mark, A. E.; Robb, M. A. *J. Am. Chem. Soc.* **2004**, *126*, 4228−4233.

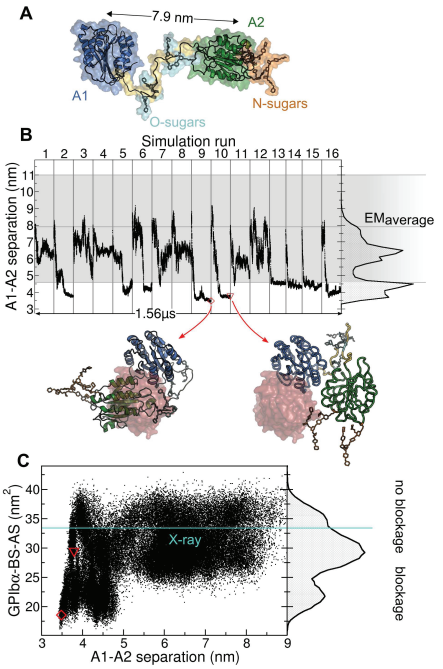
2.3 Force-sensitive autoinhibition of the von Willebrand factor mediated by interdomain interactions

Aponte-Santamaría, C.; Huck, V.; Posch, S.; Bronowska, A.K.; Grässle, S.; Brehm, M.A.; Obser, T.; Schneppenheim, R.; Hinterdorfer, P.; Schneider, S.W.; Baldauf, C.; Gräter, F.

Force-sensitive autoinhibition of the von Willebrand factor mediated by interdomain interactions

Biophys. J.
2015 (108), 2312-2321.

DOI: 10.1016/j.bpj.2015.03.041



Article

Force-Sensitive Autoinhibition of the von Willebrand Factor Is Mediated by Interdomain Interactions

Camilo Aponte-Santamaría,[1] Volker Huck,[2] Sandra Posch,[3] Agnieszka K. Bronowska,[1] Sandra Grässle,[2] Maria A. Brehm,[4] Tobias Obser,[4] Reinhard Schneppenheim,[4] Peter Hinterdorfer,[3] Stefan W. Schneider,[2] Carsten Baldauf,[5,*] and Frauke Gräter[1,*]

[1]Molecular Biomechanics Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany; [2]Experimental Dermatology, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany; [3]Department of Applied Experimental Biophysics, Institute of Biophysics, Johannes Kepler University, Linz, Austria; [4]Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; and [5]Theory Department, Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany

ABSTRACT Von Willebrand factor (VWF) plays a central role in hemostasis. Triggered by shear-stress, it adheres to platelets at sites of vascular injury. Inactivation of VWF has been associated to the shielding of its adhesion sites and proteolytic cleavage. However, the molecular nature of this shielding and its coupling to cleavage under shear-forces in flowing blood remain unknown. In this study, we describe, to our knowledge, a new force-sensory mechanism for VWF-platelet binding, which addresses these questions, based on a combination of molecular dynamics (MD) simulations, atomic force microscopy (AFM), and microfluidic experiments. Our MD simulations demonstrate that the VWF A2 domain targets a specific region at the VWF A1 domain, corresponding to the binding site of the platelet glycoprotein Ibα (GPIbα) receptor, thereby causing its blockage. This implies autoinhibition of the VWF for the binding of platelets mediated by the A1-A2 protein-protein interaction. During force-probe MD simulations, a stretching force dissociated the A1A2 complex, thereby unblocking the GPIbα binding site. Dissociation was found to be coupled to the unfolding of the A2 domain, with dissociation predominantly occurring before exposure of the cleavage site in A2, an observation that is supported by our AFM experiments. This suggests that the A2 domain prevents platelet binding in a force-dependent manner, ensuring that VWF initiates hemostasis before inactivation by proteolytic cleavage. Microfluidic experiments with an A2-deletion VWF mutant resulted in increased platelet binding, corroborating the key autoinhibitory role of the A2 domain within VWF multimers. Overall, autoinhibition of VWF mediated by force-dependent interdomain interactions offers the molecular basis for the shear-sensitive growth of VWF-platelet aggregates, and might be similarly involved in shear-induced VWF self-aggregation and other force-sensing functions in hemostasis.

INTRODUCTION

Von Willebrand Factor (VWF) is a giant extracellular protein playing a key adhesive role in blood clotting. Activated by shear-stress, this protein cross-links the extracellular matrix of the endothelium with blood platelets, at sites of vascular injury (1,2). It efficiently participates in the shear-induced reversible formation of biopolymer-colloid aggregates (3), and its malfunction leads to pathological bleeding and thromboembolic disorders (1).

Functional VWF is a linear multimer of tens of covalently linked monomers (4), extending up to 15 μm (5). Each monomer, with a length of 60 to 80 nm (2,6), comprises 2050 amino acids in domains of few nm in size (7). The large size in the μm range enables VWF multimers to sense changes in the shear flow of blood and to translate them into a mechanical stretching force along the protein chain (5,8,9). Shear-forces, by inducing a tumbling motion alternating between globular and extended states, facilitate the adhesion of VWF to the extracellular matrix (5,10) and to flowing platelets (3).

The VWF A1 and A2 domains are critical for the activation of VWF to bind platelets and for its deactivation by size control. These two domains are adjacent to each other and connected by a linker of $\sim$ 30 amino acids (Fig. 1 *A*). X-ray crystallography revealed that both domains adopt a stable Rossmann α/β-fold (11,12), stabilized by calcium in the case of A2 (13,14). Platelets bind through the glycoprotein Ibα (GPIbα) to a region of the A1 domain (15,16), in a shear-dependent manner (17–21). For size control, the A2 domain is cleaved by the metalloprotease ADAMTS13 (22), after exposure of the Y1605-M1606 (YM) cleavage site, because of shear-induced domain unfolding (23–27).

Under equilibrium or under low shear-stress conditions, VWF is incapable of binding platelets. This inactivation has been associated with a shielding of the GPIbα binding site of A1. Recent experiments revealed that, in addition to the D'D3 domains (28) and the linker connecting them to the A1 domain (29), isolated A2 domains modulate glycoprotein Ib (and thereby platelet) binding (30,31). However, electron microscopy (EM) images established the separation between these two domains, within the same VWF molecule, from 4.4 to 11 nm (6), challenging the inhibitory

Submitted October 10, 2014, and accepted for publication March 18, 2015.

*Correspondence: frauke.graeter@h-its.org or baldauf@fhi-berlin.mpg.de

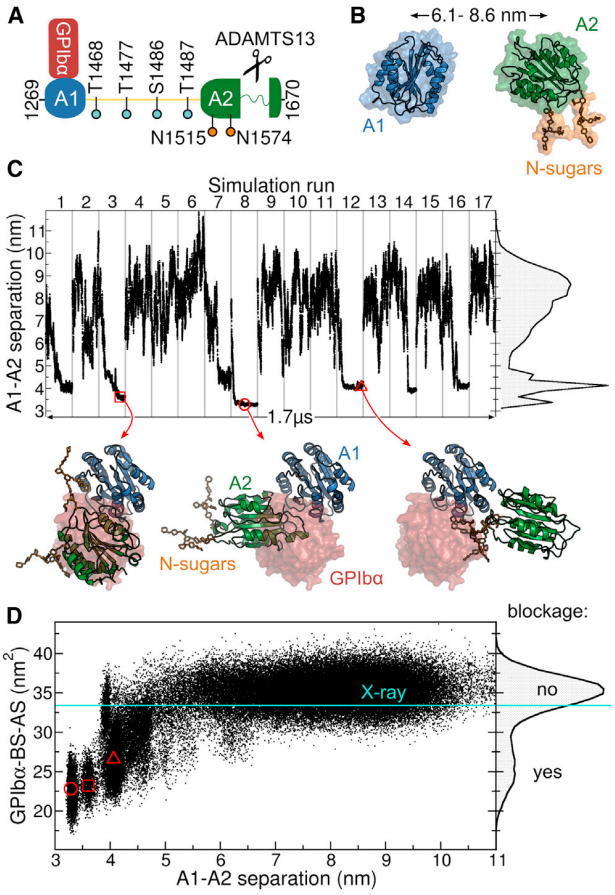


FIGURE 1 Blockage of the GPIbα binding site in the VWF revealed by MD simulations of the VWF A1 and A2 domains. (*A*) Scheme illustrating the human VWF-A1A2 fragment (residues 1269 to 1670). The A1 and A2 domains are connected by a 30 residue linker (*yellow*). GPIbα anchors platelets to VWF by binding to the A1 domain. VWF size is controlled by cleavage of the unfolded A2 domain by ADAMTS13. O-linked (*cyan*) and N-linked sugars (*N-sugars, orange*) are found within the fragment. (*B*) One of the multiple starting conformation used in the MD simulations (*protein as cartoon and surface and sugars as sticks*). The domain-domain center of mass (A1-A2) separation is indicated with the black arrow. (*C*) A1-A2 separation along the concatenated MD simulation time. Gray lines separate individual MD runs. The right plot shows the normalized histogram of the A1-A2 separation. Conformations at the bottom show examples with the two domains in contact (*cartoon*) contrasted to the region occupied by GPIbα when it binds to A1 (*red surface*), taken at the instants marked with the red symbols. (*D*) GPIbα binding site accessible surface (GPIbα-BS-AS) as a function of the A1-A2 separation (*main panel*) and its normalized histogram (*right plot*), recovered from MD simulations. Reduced GPIbα-BS-AS values indicate blockage of the GPIbα binding site. The GPIbα-BS-AS derived from the VWF A1-GPIbα complex x-ray structure (16) is depicted by the cyan line. The red symbols correspond to the conformations shown in (*C*). To see this figure in color, go online.

role of A2 on A1. Hence, little is known on how these two domains interact with each other, causing inhibition, and how sensitive this interaction is to shear-forces in flowing blood. It also remains unclear how VWF activation, through the release of the GPIbα binding site, and VWF deactivation, through unfolding of the A2 domain, are mechanically regulated to balance the propagation and attenuation of hemostasis. We addressed these questions by performing molecular dynamics (MD) simulations of the VWF A1 and A2 domains, under equilibrium and force-probe conditions, together with molecular docking calculations, atomic force microscopy (AFM) binding measurements, and microfluidic experiments. To our knowledge, our results suggest a novel mechanism for shear-dependent primary hemostasis, involving a force-sensitive autoinhibition state, in which platelets are incapable to bind to VWF because of direct (intra- or intermolecular) A1-A2 interactions precluding the A1-GPIbα interaction.

MATERIALS AND METHODS

Equilibrium MD simulations

In the first simulation system, the A1 and A2 domains of the VWF were not covalently connected by their interdomain linker. They were either initially separated by distances from 6.1 to 8.6 nm to monitor association or already bound in conformations blocking the GPIbα binding site (obtained by docking, see below) for refinement. The second simulation system corresponded to the VWF-A1A2 fragment consisting of the A1 and A2 domains connected by a 30 amino acid linker, with an initial interdomain separation of 7.9 nm based on EM estimates (6). The most predominantly found sugars in the VWF glycome (32,33) were attached to the protein (Fig. 1 *A* and Fig. S1 in the Supporting Material). Simulations were carried out with the GROMACS package (4.5 version) (34–36). Sixteen or 17 runs, considering multiple interdomain initial orientations, were performed for each condition ($\geq$ 82 ns per run) yielding a concatenated simulated time of 4.86 μs. The GPIbα binding site accessible surface (GPIbα-BS-AS) was computed by monitoring the amount of exposed surface of the GPIbα binding site in the A1 domain. A principal component analysis (PCA), consisting in the calculation and diagonalization of the covariance matrix of the atomic coordinates (37), was employed to monitor the interdomain orientations (Fig. 2). The solvent accessible hydrophobic surface (SAHS) reduction was estimated as [SAHS(A1A2) $-$ SAHS(A1) $-$ SAHS(A2)]/[SAHS(A1) $+$ SAHS(A2), computing separately the surface for the complex (A1A2) and for the domains A1 and A2.

Force-probe MD simulations

The A1 and A2 domains of the VWF were subjected to external harmonic forces on the N-terminus of the A1 domain and on the C terminus of the A2 domain (Fig. 3 *A*). Harmonic springs (with elastic constants of 500 kJmol^{-1}nm^{-2}) were attached to these termini and moved away from each other at a speed of 0.2 m/s. These simulations were started from 17 different starting conformations: one was extracted from an equilibrium MD run showing spontaneous binding (run number eight in Fig. 1 *C*) and the remaining 16 corresponded to representative conformations of the equilibrium simulations of the VWF-A1A2 complex (one conformation taken from each run presented in Fig. 2 *A*). Hence, starting conformations with high but also moderate stability were considered. The two monomers were not connected, first, to resemble dissociation of the A1-A2 complex either within or across VWF monomers (preventing from possible artifacts by the inclusion of the flexible linker for which the structure is unknown), and second, to have a direct comparison with our AFM experiments (also carried out with nonconnected domains, see below). Dissociation was assigned to the moment when the interdomain number of contacts was zero. Detachment of the A2-β5 strand from the core of the A2 domain

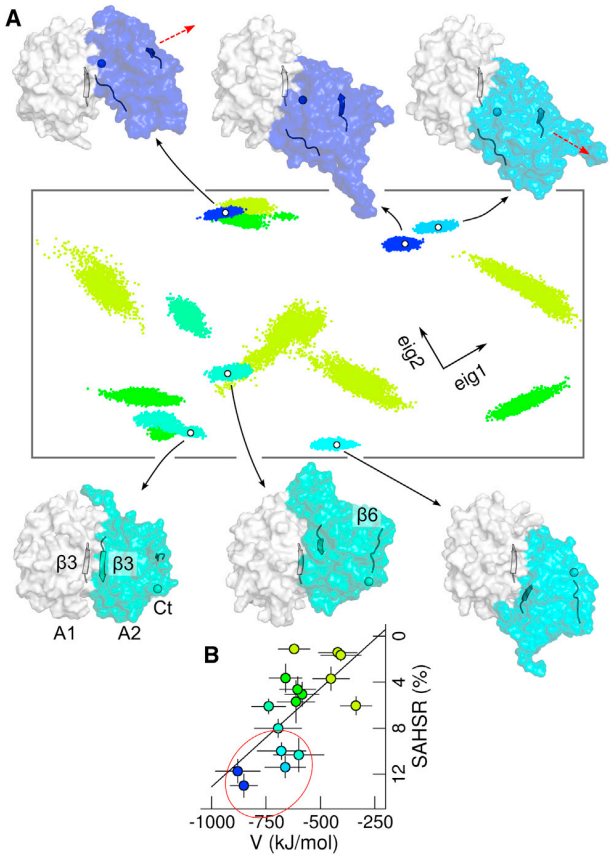


FIGURE 2 Orientational preferences of the VWF-A1A2 complex in the blocked state. (*A*) Principal component analysis (PCA) of the structures of the not-covalently linked VWF-A1A2 complex, with the GPIbα binding site blocked, predicted by molecular docking, yielded two main collective vectors (eig1 and eig2). MD trajectories (the last 50 ns) starting from these structures were projected onto the two-dimensional (2D) space created by these two vectors (projections in arbitrary units). Each dot, representing a simulation snapshot, reflects an adopted interdomain orientation. Each run is colored according to its interdomain potential interaction energy, V, and average solvent accessible hydrophobic surface reduction, SAHSR (see *B*). Representative orientations of runs with both high V and SAHSR (*enclosed by the red circle in B*) are displayed (*A1 domain*, *white*; *A2 domain*, *color*; *β3 strands*, *cartoon*; *A2 β6 strand*, *ribbon*, *and A2 C terminus*, *sphere*). The red arrows illustrate the change in orientation of A2 on horizontal changes in the 2D-PCA space. (*B*) SAHSR as a function of V (time-average $\pm$ standard deviation from the last 50 ns of each run). Colors indicate the projection along a linear fit (*black line*), with both V and SAHSR ranging from small (*light green*) to large (*blue*) values. To see this figure in color, go online.

was monitored by measuring the distance between V1625-P1627 (at β5) and V1604-Y1605 (at β4).

Molecular docking

To augment the MD-generated conformational ensemble of the VWF-A1A2 complex, with a blocked GPIbα binding site, we used molecular docking. Two independent docking approaches, either using Patchdock (38) with further refinement with Firedock (39) or using RosettaDock (40) were considered. Starting conformations of the MD simulations with

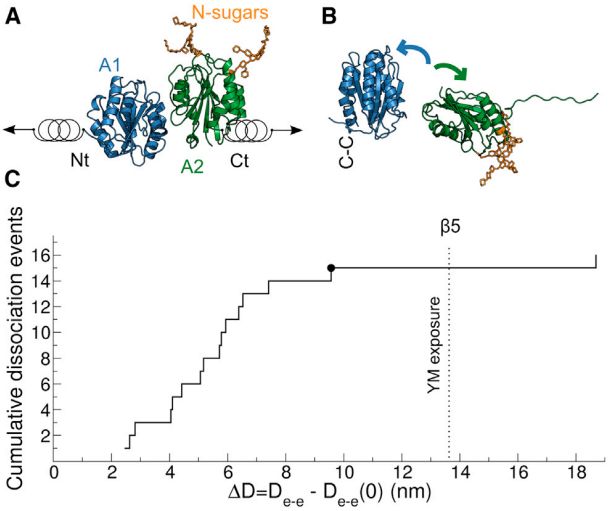


FIGURE 3 Force response of the VWF-A1A2 complex from force-probe MD simulations. (*A*) The N-terminus (Nt) of the A1 domain and the C terminus (Ct) of the A2 domain were pulled away from each other by harmonic springs. The domains were initially in contact but not connected by a linker (*domains in cartoon and N-linked sugars in stick representation*). (*B*) Snapshot illustrating a typical dissociation event of the VWF-A1A2 complex induced by the applied force (same representation as in *A*). Slight unfolding of the C-terminal part of the A2 domain was observed. The disulfide bond Cys1272-Cys1458 (C-C) prevented the A1 domain from unfolding. (*C*) Cumulative dissociation events (from 17 runs) as a function of the distance D_{e-e} between the pulled N- and C termini at the moment of dissociation. Here, $\Delta D = D_{e-e} - D_{e-e}(0)$, subtracting the initial distance $D_{e-e}(0)$, is shown. The Y1605-M1606 (YM) ADAMTS13 cleavage site was exposed after separation of the A2 C-terminal β5 strand from the core of the protein (*event indicated by the dotted line*). The black circle corresponds to the dissociation event illustrated in (*B*). To see this figure in color, go online.

the domains in contact were generated by Patchdock and Firedock (see selection criterion in Fig. S3).

Cloning, expression, and purification of VWF constructs

The cDNAs coding for either the full-length human VWF, or the A1, A2, and A3 domain, the latter three with 6x His-tag, were cloned into the mammalian expression vector pcDNA3 (41). ΔA1-VWF and ΔA2-VWF mutants were obtained by deleting either the A1 or the A2 domain from the full-length cDNA, by site-directed mutagenesis, employing the QuickChange kit (Stratagene, La Jolla, CA). All primers are available on request. Recombinant expression of VWF constructs in HEK293-EBNA cells was performed as described (42) and the His-tagged VWF domain constructs were purified employing the His-Pur Ni-NTA Resin (Thermo Scientific, Waltham, MA).

AFM

Force distance cycles (FDC) were acquired by approaching and retracting the VWF A1 domain (C-terminally linked to the AFM cantilever by maleimide-polyethylene glycol (PEG)-NHS -mPN- molecules) to VWF A2 domains (C-terminally immobilized on a mica surface by mPN linkers). The disulfide bond Cys1272-Cys1458, connecting the N- and C terminus of the A1 domain, ensured a high similarity of the pulling geometry in

the force-probe MD simulations (pulling the N-terminus) and the AFM experiments (pulling the C terminus). Binding events were discerned from nonspecific adhesion by how much they differed in the approach and retraction force signals. To have an unbiased choice of binding events, FDC displaying a characteristic worm-like-chain-type force signal, as well as FDC not showing such behavior, were included for further analysis. To validate specific binding, control experiments were carried out either in the presence of 0.1 mg/ml soluble A2 domains or by replacing either the A1 or the A2 domain by VWF A3 domains. The latter case constitutes a critical control experiment, because A3 is a protein domain that is in the vicinity of A1 and A2 in physiological conditions, and also has the Rossmann topology. For each system, four cantilever tips were utilized. At least 1000 FDC were recorded for each of the tips at a pulling speed of 600 nm/s.

The elongation L corresponded to the extension of the A1 and A2 domains, together with the ones of the mPN linkers and 3-aminopropyltriethoxy silane (APTES) coating molecules. It was measured, during a binding event, as the distance in which the attraction and retraction force-distance curves differed minus the cantilever deflection CD (Fig. 4 *A*). In practice, L + CD was measured by fitting a second-order polynomial to the force curves, followed by the determination of the point in the retraction curve where the force abruptly returned back to zero. The cantilever deflection CD was determined as the applied stretching force F (extracted at the moment of rupture during the FDC) divided by the actual spring constant of the cantilever (30 pNnm^{-1}). The expectation value of L (EV) was estimated as EV = $\sum_i P_i L_i$, with P_i the measured probability to have an elongation of L_i, summing over all the measured L_i values. To account for the size of the A1 domain and the mPN linkers, $l_{A1} + 2l_l$ was subtracted to each measured elongation L. The size of the A1 domain (l_{A1}) was estimated as 2 × its radius of gyration (1.6 nm, derived from MD simulations of the isolated A1 domain (43)). A worm-like-chain model was employed to compute the extension l_l of the mPN linkers as a function of the force F. It reads as follows:

$$\frac{FP}{k_BT} = \frac{1}{4}\left(1 - \frac{l_l}{lc}\right)^{-2} - \frac{1}{4} + \frac{l_l}{lc},$$

where P is the persistence length (0.38 nm (44)), l_c is the mPN linker contour length (8.9 nm, considering 27 PEG units and 0.33 nm per unit), k_B is the Boltzmann constant, and T is the temperature. The cumulative distributions of both the original elongation L and its theoretical reduction (accounting for the size of A1 and linkers) were shown.

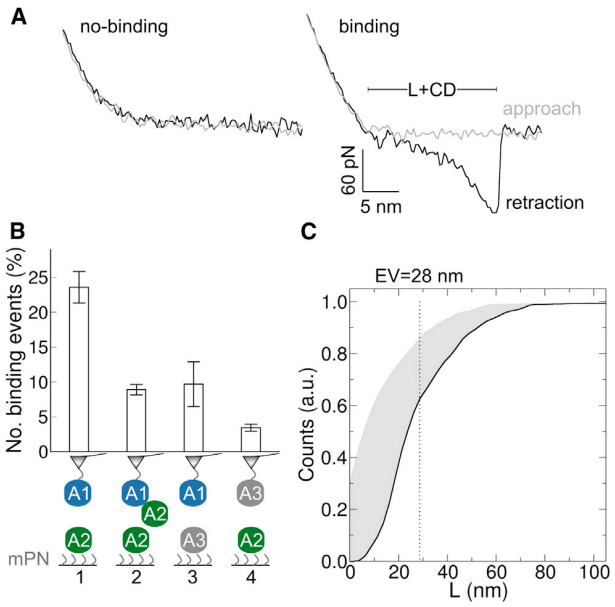


FIGURE 4 Force response of the VWF-A1A2 complex from AFM. (*A*) Typical approach-retraction force-distance profiles associated to no-binding and binding events. The elongation L, of the A1 and A2 domains, together with the mPN linkers and the 3-aminopropyltriethoxy silane coating molecules, summed to the cantilever deflection (CD) was determined by the difference between approach and retraction curves. (*B*) (*1*) Number of binding events between VWF A1 and A2 domains. A1 was connected to the tip of the AFM cantilever (*triangle*) using malemide-PEG-NHS (mPN) linkers. It was approached to and retracted from the surface carrying mPN-linked A2 domains. Force-distance cycles presented in (*A*) correspond to this situation. (*2–4*) Number of binding events measured in control AFM experiments, in which the A1 domain was blocked by soluble A2 domains (*2*), or either the A2 domains on the surface (*3*) or the A1 domain connected to the cantilever (*4*) were replaced by VWF A3 domain. (*C*) Cumulative distribution of L (*black line*) and its correction by subtracting the size of A1 and the mPN linkers (*gray area*). Dotted line indicates the expectation value (EV) of L. To see this figure in color, go online.

Microfluidic experiments

For distinct shear rate application, air-pressure driven microfluidic channels were coated with recombinant wild-type VWF, ΔA2-VWF, or ΔA1-VWF. For the functional characterization, the coated microfluidic channels were mounted onto an inverted fluorescence microscope and perfused, as previously published (45), with wild-type VWF, VWF with the A2 domain deleted, or VWF with the A1 domain deleted. Live cell fluorescence images were taken and analyzed at shear rates in the range of 500 s^{-1} to 4000 s^{-1}. To track the motion of VWF-platelet fibers and aggregates, an image composition of 20 sequential frames (taken at a frequency of two frames/s) was implemented. Increasing number of frames was considered for the composition (from one to all 20 frames), subtracting identical pixels among frames. Dynamical monitoring allowed the exact determination of the critical shear rate for VWF-platelet fiber and aggregate formation.

See further details of the simulations and the experimental procedures in the Supporting Material.

RESULTS

Blockage of the VWF GPIbα binding site in A1 by A2

We first investigated whether the VWF A2 domain spontaneously binds to the A1 domain. To this end, we carried out 17 independent 100 ns equilibrium MD simulations, starting with these two domains separated by distances (between their center of masses) from 6.1 to 8.6 nm and adopting different orientations with respect to each other (Fig. 1 *B*). The linker connecting the two domains was not considered (in the following, this situation will be referred as not connected domains). The two domains spontaneously came into contact and remained stably bound in seven out of 17 simulation runs, as reflected by drops in their separation to values smaller than 5 nm (Fig. 1 *C*).

We next analyzed if the GPIbα binding site in the A1 domain was blocked upon binding of the A2 domain. We quantified the amount of blockage by computing the GPIbα binding site accessible surface (GPIbα-BS-AS) (Fig. 1 *D*). The GPIbα-BS-AS histogram recovered from our simulations revealed a major peak close to the value estimated

from the x-ray structure of the VWF A1-GPIbα complex (16) (33.4 nm^2), indicating no blockage. In addition, the histogram contained a tail extending to values smaller than 20 nm^2, reflecting substantial blockage (of more than 40% of the x-ray GPIbα-BS-AS). Remarkably, blockage was found correlated with the separation between domains, with the GPIbα binding site fully accessible (large GPIbα-BS-AS) only for large interdomain separations, whereas completely blocked (small GPIbα-BS-AS) when the A2 domain approached the A1 domain. Thus, from our simulations, A2 binding to A1 implies blockage of the VWF-GPIbα interaction site.

We also tested the blockage of the GPIbα binding site within a VWF-A1A2 fragment, with the A1 and A2 domains connected by the linker. We simulated the dynamics of such fragment, in 16 independent MD runs of 82 to 100 ns, with initial interdomain separations (∼ 7.9 nm) and linker extensions (∼ 6.0 nm) taken from EM estimates (6) (Fig. S2 *A*). The fragment populated the lower range of separations measured in the EM experiments (6) (Fig. S2 *B*). Again, the MD-generated conformations included several instances of direct A1-A2 interactions (Fig. S2 *B*). The presence and involvement of the O-linked glycosylated linker now alleviated the strong correlation between A1-A2 binding and blockage of the GPIbα binding site as observed for not connected domains (Fig. S2 *C*).

Orientational preferences in the blocked state

Our simulations raised the question on the most-favorable conformation of the two domains with GPIbα binding blocked. We addressed this by performing molecular docking followed by MD refinement. We generated a set of conformations by docking the A2 domain to the A1 domain. From this set, we selected representative conformations with both the GPIbα binding site blocked and high docking score as starting positions of 16 MD simulations of 100 ns each (see Fig. S3 and the Supporting Material for the selection criterion). Similar conformations presenting blockage were predicted by two independent docking approaches (Fig. S4). Furthermore, an enrichment of blocked conformations over random conformations was observed, because of their large interdomain shape complementarity and favorable protein-protein interactions, thus justifying our selection criterion of only blocked and high-docking-score structures (Fig. S5 and the Supporting Material).

During the simulations the domains remained bound causing blockage, while maintaining their internal structure almost intact (backbone root-mean-square deviation to the initial structure below 1.5 Å for A1 and 2.3 Å for A2), but accommodating with respect to each other in multiple orientations. To capture the extent of stable blocking interdomain orientations we carried out a PCA of the conformations predicted by docking (yielding two main collective eigenvectors covering 68% of the possible interdomain orientations), followed by projections of the MD trajectories onto the two-dimensional (2D) space generated by these two vectors (Fig. 2 *A*). Furthermore, we narrowed the orientations to those with high interdomain potential energy, V, and substantial solvent accessible hydrophobic surface reduction (SAHSR) (Fig. 2 *B*). Remarkably, in all orientations with large V and SAHSR contributions, the A2 domain was found directly obstructing the A1-domain β3 strand (the one connecting with GPIbα (15,16)) and displaying only small orientational deviations (small point clouds in the 2D-PCA projections), indicating high structural integrity. Within this preferred set of VWF-A1A2 complexes, the A2 domain oriented in two main modes: either with its C terminus in proximity to the A1 domain or—on ~180° relative rotation—with its β3 strand in proximity, almost forming a stable interdomain β-sheet in the latter case (compare top with bottom projections and snapshots in Fig. 2 *A*). The residues Arg1668 and Asp1587, both in A2, were found to strongly interact with A1: Arg1668, when the C terminus was in proximity to A1, and Asp1587, when the β3-strand was in vicinity. Destabilizing mutations Arg1668Asp and Asp1587Lys are thus potential candidates to detect the most favored conformation of the complex among the two observed orientational modes. In addition, replacement of Val1548 located directly at the β3 strand of A2, for instance by a bulky polar residue such as serine or asparagine, would further distort the orientational mode that features a quasi interdomain β-sheet.

We validated the observed orientational preferences by comparing this with our previous set of simulations (Fig. S6). The docking-MD refined region was also sampled during the MD simulations starting from separated domains, with the A2 domain located directly in front of the β3 strand of the A1 domain. However, the conformational ensemble in the blocked state was further broadened presumably because of the sugars and also the linker between A1 and A2.

VWF-A1A2 complex under force: activation versus cleavage

Induced by shear-forces, the release of the GPIbα binding site in the A1 domain would allow platelet-binding activation, whereas exposure of the YM catalytic site after unfolding of the A2 domain would enable cleavage and degradation. We studied how a stretching force balances these two processes. For this purpose we performed 17 independent force-probe MD simulations, starting from a diverse set of conformations of the two domains, not connected, forming a complex, and with the GPIbα binding site obstructed (Fig. 3 *A*). We pulled the N-terminus of the A1 domain and the C terminus of the A2 domain away from each other, until dissociation of the complex (and thereby unblocking of the GPIbα binding

site) occurred (Fig. 3 *B*). The A2 domain slightly unfolded in its C-terminal part, while the A1 domain remained folded because of its Cys1272-Cys1458 disulfide bond (Fig. 3 *B*).

We quantified the extent of unfolding of the C terminus of the A2 domain by monitoring the increase in the distance between the pulled termini, D_{e-e}, with respect to the initial distance $D_{e-e}(0)$. Exposure of the YM cleavage site, as an initial requirement for ADAMTS13 cleavage, occurred after the detachment of the $\beta 5$ strand ($D_{e-e} - D_{e-e}(0) \simeq 13.6$ nm). In comparison, dissociation of the fragment, as needed for activation, occurred before YM exposure, in 15 of the 17 runs (88% of the cases) (Fig. 3 *C*).

We next probed the physical interaction between A1 and A2 and the coupling between dissociation and unfolding, as suggested by our simulations, at the single-molecule level by using AFM (Fig. 4). FDC were acquired by approaching the A1 domain (linked to the AFM cantilever) to A2 domains (immobilized on a surface) and retracting it again. A retracting force signal differing from the approaching one, with an abrupt drop to zero at dissociation, was used as an indicator for a binding event (Fig. 4 *A*). It was observed in ~23% of the cycles (*1* in Fig. 4 *B*). In contrast, a substantially reduced number of binding events (less than 10%) was observed in the presence of soluble A2 domains, presumably because of the blocking of the A1 domain at the cantilever (*2* in Fig. 4 *B*). As a control, reduction in the number of binding events was also observed when replacing either the A2 domains at the surface (*3* in Fig. 4 *B*) or the A1 domain at the cantilever (*4* in Fig. 4 *B*) by VWF A3 domains. This implies that binding events are exclusively through A1-A2 interactions, thus confirming the observation from our MD simulations and from previous binding assays (30) that the VWF A1 and A2 domains specifically interact.

To further investigate the coupling between dissociation of the VWF-A1A2 complex and unfolding of the A2 domain, we measured the elongation of the complex (together with linkers and coating molecules) before dissociation by AFM (Fig. 4 *C*). The measured expectation value of the elongation (~28 nm) was substantially lower than the extension of a fully stretched unfolded A2 domain (~80 nm (23–27)). In fact, in all FDC, the elongation remained below those levels of extension. Although the noise in the length distribution is expected to be large because of the tip and surface chemistry, our AFM data speak against full unfolding of A2 before dissociation. Instead, it suggests a small extent of unfolding of A2 before dissociation.

Functional characteristics of VWF with the A2 domain deleted in shear-induced fiber formation

We next examined if the A2 domain inhibits VWF-platelet binding in a shear-dependent manner, by performing microfluidic experiments, in wild-type VWF-coated channels, under replacement of the plasmatic wild-type VWF by recombinant VWF with the A2 domain deleted (ΔA2-VWF), and in a wide shear range. In the presence of wild-type VWF in the perfusion medium, above a critical shear rate of 4000 s^{-1}, large aggregates of VWF and platelets were observed to roll along the surface coated with VWF (Fig. 5, *top right*). At lower shear rates, rolling VWF-platelet aggregates were absent. Here, we only observed either rolling of single platelets along the microfluidic channel (at 500 s^{-1}, Fig. 5, *top left*) or reversibly formed platelet-decorated VWF fibers, which stayed attached to the channel surface (at 2500 s^{-1}, Fig. 5, *top middle*). Instead, in the presence of ΔA2-VWF in the perfusion medium, the critical shear rate for rolling aggregate formation was decreased to 2500 s^{-1}, indicating a gain of function for the VWF by deletion of its A2 domain (Fig. 5, *middle*, and Movie S1). Identical results were obtained using ΔA2-VWF instead of wild-type VWF for coating of the microfluidic channels (data not shown). In a multimer analysis, similar VWF size distributions were observed for the mutants and for the wild-type VWF, just slightly shifted down because of the deletions in the mutant proteins (Fig. S7). Changes in the VWF distribution size are thus discarded as the reason for the gain in function of the ΔA2-VWF mutant. As expected, neither fibers nor VWF-platelet aggregates were formed in the presence of VWF with an A1-domain deletion (Fig. 5, *bottom*). Furthermore, coating with ΔA1-VWF led to a complete absence of both single platelet rolling and the formation of rolling

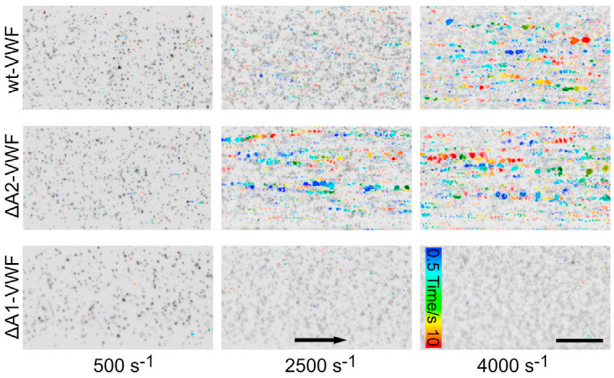


FIGURE 5 Changes in shear-induced fiber and aggregate formation on deletion of the VWF A2 domain. Live-cell fluorescence images of platelet-decorated VWF fibers and platelet-VWF aggregates observed in microfluidic experiments at the indicated shear rates (*different columns*). Microfluidic channels were perfused with plasmatic wild-type VWF (wt-VWF, *top row*), VWF with the A2 domain deleted (ΔA2-VWF, *middle row*), or VWF with the A1 domain deleted (ΔA1-VWF, *bottom row*). A static image is presented as background, displaying platelets, fibers, and aggregates in black. Moving fibers and aggregates are highlighted in color. Their positions were tracked during 10 s after taking the static image. Flow direction is indicated with the arrow and the line corresponds to 100 μm. To see this figure in color, go online.

VWF-platelet aggregates independent of the VWF present in the perfusion medium.

DISCUSSION

Blockage of the GPIbα binding site mediated by A1-A2 interactions implies autoinhibition

Our extensive set of simulations (in the μs time range) demonstrates that the GPIbα binding site of VWF (located in the A1 domain) can be significantly blocked, upon spontaneous binding of the A2 to the A1 domain (Fig. 1). In addition, the binding of these two domains was further confirmed at the single-molecule level by AFM (Fig. 4). The increase in blockage with reducing interdomain separation observed in our simulations suggests that the A2 domain does not recognize a random region in the A1 domain but instead it specifically targets the GPIbα binding site. This observation was further supported by our docking calculations, which showed enrichment toward blocked conformations over random conformations, by enhanced shape complementarity and favorable protein-protein interactions (Fig. S5). With the GPIbα binding site blocked, platelets are prevented to bind and thus the VWF remains inactive. Our results, together with the experimentally observed platelet-binding modulation in the presence of A2 domains (30), thus imply an autoinhibition mechanism for the binding of platelets to the VWF mediated by A1-A2 interactions.

Additional simulations, this time with the two domains connected (also in the μs time range), revealed broad dynamics of the VWF-A1A2 fragment (Fig. S2). Although the A2 domain (bound to the A1 domain) was sometimes observed causing no shielding, presumably stabilized by the connecting linker, it was also found in many other times substantially blocking the GPIbα binding site. This indicates that not only not connected, but also vicinal, covalently linked, A1 and A2 domains can interact with each other causing blockage, further supporting the hypothesis of VWF autoinhibition because of A1-A2 interactions.

Our simulations of connected domains sampled a range from compact to extended conformations, covering the lower region of the interdomain separations measured by EM (6). In fact, compact conformations are expected from a direct A1-A2 interaction, as established in previous assays (30) and confirmed in our AFM experiments. Also, the presence of a third domain (e.g., D'D3 or A3) or deposition on the surface may favor more extended conformations in the EM experiments compared with the ones sampled in our simulations.

Autoinhibition driven by A1-A2 interactions provides a molecular picture of the shielding of the GPIbα (platelet) binding site, crucial to maintain the VWF inactive under equilibrium conditions. This is a complementary scenario to previous shear-dependent models (18,20) for GPIbα binding, but is the only one reconciling previous inactivation experiments (30).

Main orientational modes of the autoinhibited state

From our simulations, the minimum structural requirement to block GPIbα binding is to have A1-A2 binding and this is effectively achieved by the A2 domain specifically targeting the GPIbα binding site in A1. Our docking calculations and further extensive MD refinement narrowed the interdomain conformational variability to two main orientational modes of blockage, stabilized by an attractive interdomain potential energy and a reduction in the amount of solvent accessible hydrophobic surface (Fig. 2). A2 located either with its C terminus or with its β3 strand in proximity to the β3 strand of A1, resulting in a quasi-extended cross-domain β-sheet in the latter case. Notably, as a general feature, the A2 domain obstructs the A1-domain β3 strand (which connects to GPIbα (15,16)), thus suggesting drastic VWF autoinhibition. Direct blockage of the main interaction partner of GPIbα in the A1 domain (the β3 strand) was also observed in our simulations started from unbiased positions, with the domains separated, further supporting our proposed mode of autoinhibition. In addition, the agreement between our force-probe MD simulations and AFM experiments (see below) stresses on the validity of the chosen conformations from docking, followed by MD refinement, and the robustness of the MD simulation results. Our structural predictions are anticipated to motivate future structural studies aiming at determining the structure of the A1-A2 complex, in the nonconnected and connected situations, both of physiological relevance. Mutants Arg1668Asp, Asp1587Lys, and Val1548Ser(Asn) may serve as initial candidates for mutagenesis studies to discern among the two proposed modes of blockage.

In the simulations started from separated domains, additional blocking orientations were observed. Here, the presence of the N-linked sugars or the O-glycosylated linker may also play stabilization roles. An additional stabilization of the blocked (autoinhibited) state of the VWF by the sugars is consistent with recent microfluidic experiments that showed an increase in platelet adhesion when the VWF was N-deglycosylated (46).

Force unblocks the GPIbα-binding site before exposure of the ADAMTS13 cleavage site, ensuring VWF activation before cleavage

In our force-probe simulations, we induced the dissociation of the complex formed by the A1 and A2 domains by applying an external stretching force. In complex, the VWF A2 domain showed only marginal unfolding, which proceeded from the C terminus, in line with the unfolding mechanism previously observed for this domain in isolation (with different force fields) (23,47). Dissociation was found to occur before exposure of the ADAMTS13 cleavage site in the A2 domain with a very high probability (~0.88) (Fig. 3).

This is consistent with our AFM measurements, which yielded in the majority of the binding events small elongations of the VWF-A1A2 complex at rupture (Fig. 4). Our simulations and AFM thus support that a stretching force unblocks the GPIbα-binding site, by detaching the A1 and A2 domains, and that this process is coupled to the exposure of the ADAMTS13 cleavage site in the A2 domain after its unfolding. The stretching force ensures, however, that the VWF is activated for platelet binding predominantly before deactivation through cleavage. In this respect, the interactions between A1 and A2 may also serve to clarify the role of ristocetin, coupling platelet binding and ADAMTS13 cleavage (48).

Deletion of the A2 domain results in a VWF with a gain of function

Our microfluidic experiments showed a reduction of the critical shear-rate for the formation of VWF-platelet fibers and rolling aggregates, when the A2 domain was deleted. This implies a VWF with a gain in function (Fig. 5). Our results in consequence expand the experiments by Martin et al. (30), proving that not only present in solution but also within the VWF molecule, the A2 domain critically influences platelet binding in a shear-dependent manner. In addition, our combined computational and experimental results suggest that the A2 domain stabilizes a VWF inactive state, by direct A1-A2 interactions, either within or across VWF monomers. However, additional inhibitory mechanisms must be at play, because the ΔA2-VWF mutant still requires intermediate shear rates for the formation of rolling aggregates (2500 s^{-1} for ΔA2-VWF instead of 4000 s^{-1} for the wild-type VWF). We speculate that the exposure of the GPIbα binding site requires both a global globule-to-stretch transition, eventually involving other—specific or nonspecific—domain-domain interactions (e.g., between D'D3 and A1 (28)), and VWF-A1A2 dissociation.

CONCLUSIONS

In this study, we examined the inactivation of VWF for platelet binding, induced by a specific domain-domain interaction, and its coupling to VWF cleavage degradation driven by force, by using MD simulations, molecular docking, AFM, and microfluidic experiments. We demonstrate that under equilibrium conditions the VWF A1 and A2 domains bind to each other, with the A2 domain specifically targeting the GPIbα binding site in the A1 domain, thus blocking the binding of GPIbα (and thereby of platelets) to VWF. This implies autoinhibition of the VWF mediated by A1-A2 interactions. We identified two main orientational blocking modes, which have the shielding of the A1 β3 strand, the site critical for GPIbα binding, in common. Detachment of the two domains, induced by a stretching force, unblocked the GPIbα binding site most predominantly before exposure of the cleavage site in the A2 domain. This suggests that A2 blocks GPIbα binding in a force-dependent manner, but guaranteeing that the VWF is ready for activation before cleavage, to mechanically balance the propagation and attenuation of hemostasis. Deletion of the A2 domain enhanced platelet binding, corroborating the key autoinhibition role of this domain. In summary, our results suggest, to our knowledge, a new interdomain-mediated autoinhibition mechanism that explains the inactivation of VWF under equilibrium conditions while allowing shear-sensitive growth of blood coagulates. This mechanism reconciles previous and can be tested by future experiments. It will be highly interesting to investigate if this or other domain-domain interactions are a common regulatory mechanism, not only for the shear-sensitive binding of VWF to its partners, but also potentially for the shear-dependent self-aggregation of VWF.

SUPPORTING MATERIAL

Supporting Materials and Methods, seven figures, and one movie are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(15)00302-1.

AUTHOR CONTRIBUTIONS

C.A.-S. performed the MD simulations and docking calculations with Patchdock and Firedock. C.A.-S., C.B., and F.G. analyzed the computational results. V.H., S.G., and S.W.S. performed the microfluidic experiments. S.P. and P.H. carried out the AFM experiments. A.K.B. performed the docking calculations with RosettaDock. M.A.B., T.O., and R.S. generated the VWF constructs. C.B. and F.G. conceived the project. All authors discussed the results. C.A.-S., V.H., S.P., A.K.B., M.A.B., C.B., and F.G. wrote the manuscript.

ACKNOWLEDGMENTS

This study was supported by research funding from the German Research Foundation, to the Research Group FOR1543: "Shear flow regulation of hemostasis—bridging the gap between nanomechanics and clinical presentation" (C.A.-S., V.H., S.P., S.G., M.A.B., T.O., R.S., P.H., S.W.S., C.B., and F.G.), the Klaus Tschira Stiftung (F.G.), and the BIOMS program of the Heidelberg University (A.K.B.). We thank the Jülich Supercomputing Centre (J.S.C.) (HHD24 and HHD25 projects). We thank Gesa König for technical assistance with protein purification.

SUPPORTING CITATIONS

References (49–74) appear in the Supporting Material.

REFERENCES

1. Schneppenheim, R., and U. Budde. 2011. von Willebrand factor: the complex molecular genetics of a multidomain and multifunctional protein. *J. Thromb. Haemost.* 9 (*Suppl. 1*):209–215.
2. Springer, T. A. 2011. Biology and physics of von Willebrand factor concatamers. *J. Thromb. Haemost.* 9 (*Suppl. 1*):130–143.
3. Chen, H., M. A. Fallah, …, A. Alexander-Katz. 2013. Blood-clotting-inspired reversible polymer—colloid composite assembly in flow. *Nat. Commun.* 4:1333. http://dx.doi.org/10.1038/ncomms2326.

4. Lippok, S., T. Obser, …, J. O. Rädler. 2013. Exponential size distribution of von Willebrand factor. *Biophys. J.* 105:1208–1216.
5. Schneider, S. W., S. Nuschele, …, M. F. Schneider. 2007. Shear-induced unfolding triggers adhesion of von Willebrand factor fibers. *Proc. Natl. Acad. Sci. USA.* 104:7899–7903.
6. Zhou, Y.-F., E. T. Eng, …, T. A. Springer. 2011. A pH-regulated dimeric bouquet in the structure of von Willebrand factor. *EMBO J.* 30:4098–4111.
7. Zhou, Y.-F., E. T. Eng, …, T. A. Springer. 2012. Sequence and structure relationships within von Willebrand factor. *Blood.* 120:449–458.
8. Alexander-Katz, A., M. F. Schneider, …, R. R. Netz. 2006. Shear-flow-induced unfolding of polymeric globules. *Phys. Rev. Lett.* 97:138101.
9. Sing, C. E., and A. Alexander-Katz. 2010. Elongational flow induces the unfolding of von Willebrand factor at physiological flow rates. *Biophys. J.* 98:L35–L37.
10. Sing, C. E., J. G. Selvidge, and A. Alexander-Katz. 2013. Von Willlebrand adhesion to surfaces at high shear rates is controlled by long-lived bonds. *Biophys. J.* 105:1475–1481.
11. Emsley, J., M. Cruz, …, R. Liddington. 1998. Crystal structure of the von Willebrand factor A1 domain and implications for the binding of platelet glycoprotein Ib. *J. Biol. Chem.* 273:10396–10401.
12. Zhang, Q., Y.-F. Zhou, …, T. A. Springer. 2009. Structural specializations of A2, a force-sensing domain in the ultralarge vascular protein von Willebrand factor. *Proc. Natl. Acad. Sci. USA.* 106:9226–9231.
13. Zhou, M., X. Dong, …, J. Ding. 2011. A novel calcium-binding site of von Willebrand factor A2 domain regulates its cleavage by ADAMTS13. *Blood.* 117:4623–4631.
14. Xu, A. J., and T. A. Springer. 2012. Calcium stabilizes the von Willebrand factor A2 domain by promoting refolding. *Proc. Natl. Acad. Sci. USA.* 109:3742–3747.
15. Huizinga, E. G., S. Tsuji, …, P. Gros. 2002. Structures of glycoprotein Ibalpha and its complex with von Willebrand factor A1 domain. *Science.* 297:1176–1179.
16. Dumas, J. J., R. Kumar, …, L. Mosyak. 2004. Crystal structure of the wild-type von Willebrand factor A1-glycoprotein Ibalpha complex reveals conformation differences with a complex bearing von Willebrand disease mutations. *J. Biol. Chem.* 279:23327–23334.
17. Chen, Z., J. Lou, …, K. Schulten. 2008. Flow-induced structural transition in the beta-switch region of glycoprotein Ib. *Biophys. J.* 95:1303–1313.
18. Lou, J., and C. Zhu. 2008. Flow induces loop-to-beta-hairpin transition on the beta-switch of platelet glycoprotein Ib α. *Proc. Natl. Acad. Sci. USA.* 105:13847–13852.
19. Zou, X., Y. Liu, …, K. Schulten. 2010. Flow-induced beta-hairpin folding of the glycoprotein Ibalpha beta-switch. *Biophys. J.* 99:1182–1191.
20. Kim, J., C.-Z. Zhang, …, T. A. Springer. 2010. A mechanically stabilized receptor-ligand flex-bond important in the vasculature. *Nature.* 466:992–995.
21. Blenner, M. A., X. Dong, and T. A. Springer. 2014. Structural basis of regulation of von Willebrand factor binding to glycoprotein Ib. *J. Biol. Chem.* 289:5565–5579.
22. Sadler, J. E. 2002. A new name in thrombosis, ADAMTS13. *Proc. Natl. Acad. Sci. USA.* 99:11552–11554.
23. Baldauf, C., R. Schneppenheim, …, F. Gräter. 2009. Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis. *J. Thromb. Haemost.* 7:2096–2105.
24. Chen, W., J. Lou, and C. Zhu. 2009. Molecular dynamics simulated unfolding of von Willebrand factor A domains by force. *Cell Mol. Bioeng.* 2:75–86.
25. Zhang, X., K. Halvorsen, …, T. A. Springer. 2009. Mechanoenzymatic cleavage of the ultralarge vascular protein von Willebrand factor. *Science.* 324:1330–1334.
26. Wu, T., J. Lin, …, C. Zhu. 2010. Force-induced cleavage of single VWFA1A2A3 tridomains by ADAMTS-13. *Blood.* 115:370–378.
27. Ying, J., Y. Ling, …, J.-Y. Shao. 2010. Unfolding the A2 domain of von Willebrand factor with the optical trap. *Biophys. J.* 98:1685–1693.
28. Ulrichts, H., M. Udvardy, …, H. Deckmyn. 2006. Shielding of the A1 domain by the D'D3 domains of von Willebrand factor modulates its interaction with platelet glycoprotein Ib-IX-V. *J. Biol. Chem.* 281:4699–4707.
29. Auton, M., K. E. Sowa, …, M. A. Cruz. 2012. N-terminal flanking region of A1 domain in von Willebrand factor stabilizes structure of A1A2A3 complex and modulates platelet activation under shear stress. *J. Biol. Chem.* 287:14579–14585.
30. Martin, C., L. D. Morales, and M. A. Cruz. 2007. Purified A2 domain of von Willebrand factor binds to the active conformation of von Willebrand factor and blocks the interaction with platelet glycoprotein Ibalpha. *J. Thromb. Haemost.* 5:1363–1370.
31. Lenting, P. J., and C. V. Denis. 2007. von Willebrand factor A1 domain: stuck in the middle. *J. Thromb. Haemost.* 5:1361–1362.
32. Canis, K., T. A. J. McKinnon, …, A. Dell. 2010. The plasma von Willebrand factor O-glycome comprises a surprising variety of structures including ABH antigens and disialosyl motifs. *J. Thromb. Haemost.* 8:137–145.
33. Matsui, T., K. Titani, and T. Mizuochi. 1992. Structures of the asparagine-linked oligosaccharide chains of human von Willebrand factor. Occurrence of blood group A, B, and H(O) structures. *J. Biol. Chem.* 267:8723–8731.
34. Van Der Spoel, D., E. Lindahl, …, H. J. C. Berendsen. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.
35. Hess, B., C. Kutzner, …, E. Lindahl. 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4:435–447.
36. Pronk, S., S. Páll, …, E. Lindahl. 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics.* 29:845–854.
37. Amadei, A., A. B. M. Linssen, and H. J. C. Berendsen. 1993. Essential dynamics of proteins. *Proteins Struct. Funct. Bioinformatics.* 17: 412–425.
38. Schneidman-Duhovny, D., Y. Inbar, …, H. J. Wolfson. 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33:W363–W367.
39. Mashiach, E., D. Schneidman-Duhovny, …, H. J. Wolfson. 2008. FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res.* 36 (*Suppl. 2*):W229–W332.
40. Lyskov, S., and J. J. Gray. 2008. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.* 36 (*Suppl. 2*):W233–W238.
41. Schneppenheim, R., J. J. Michiels, …, U. Budde. 2010. A cluster of mutations in the D3 domain of von Willebrand factor correlates with a distinct subgroup of von Willebrand disease: type 2A/IIE. *Blood.* 115:4894–4901.
42. Schneppenheim, R., U. Budde, …, J. Oldenburg. 2001. Expression and characterization of von Willebrand factor dimerization defects in different types of von Willebrand disease. *Blood.* 97:2059–2066.
43. Grässle, S., V. Huck, …, S. W. Schneider. 2014. von Willebrand factor directly interacts with DNA from neutrophil extracellular traps. *Arterioscler. Thromb. Vasc. Biol.* 34:1382–1389.
44. Kienberger, F., V. P. Pastushenko, …, P. Hinterdorfer. 2000. Static and dynamical properties of single poly(ethylene glycol) molecules investigated by force spectroscopy. *Single Molecules.* 1:123–128.
45. Brehm, M. A., V. Huck, …, R. Schneppenheim. 2014. von Willebrand disease type 2A phenotypes IIC, IID and IIE: a day in the life of shear-stressed mutant von Willebrand factor. *Thromb. Haemost.* 112:96–108.
46. Fallah, M. A., V. Huck, …, M. F. Schneider. 2013. Circulating but not immobilized N-deglycosylated von Willebrand factor increases platelet adhesion under flow conditions. *Biomicrofluidics.* 7:044124.
47. Interlandi, G., M. Ling, …, W. E. Thomas. 2012. Structural basis of type 2A von Willebrand disease investigated by molecular dynamics simulations and experiments. *PLoS ONE.* 7:e45207.

48. Chen, J., M. Ling, …, D. W. Chung. 2012. Simultaneous exposure of sites in von Willebrand factor for glycoprotein Ib binding and ADAMTS13 cleavage: studies with ristocetin. *Arterioscler. Thromb. Vasc. Biol.* 32:2625–2630.

49. Woods Group. 2005–2013. GLYCAM Web. *In* R. J. Woods, editor. Complex Carbohydrate Research Center, University of Georgia, Athens, GA.

50. Schrödinger, L. L. C. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

51. Hornak, V., R. Abel, …, C. Simmerling. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinformatics.* 65:712–725.

52. Best, R. B., and G. Hummer. 2009. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B.* 113:9004–9015.

53. Lindorff-Larsen, K., S. Piana, …, D. E. Shaw. 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinformatics.* 78:1950–1958.

54. Kirschner, K. N., A. B. Yongye, …, R. J. Woods. 2008. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* 29:622–655.

55. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

56. Joung, I. S., and T. E. Cheatham, 3rd. 2008. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B.* 112:9020–9041.

57. Hess, B., H. Bekker, …, J. G. E. M. Fraaije. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.

58. Feenstra, K. A., B. Hess, and H. J. C. Berendsen. 1999. Improving efficiency of large timescale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* 20:786–798.

59. Miyamoto, S., and P. A. Kollman. 1992. Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* 13:952–962.

60. Hockney, R. W., and J. W. Eastwood. 1988. Computer Simulation Using Particles. Hilger, Bristol, UK.

61. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an N·log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

62. Essmann, U., L. Perera, …, L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577–8593.

63. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101.

64. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

65. Parrinello, M., and A. Rahman. 1981. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* 52:7182–7190.

66. Connolly, M. L. 1983. Analytical molecular surface calculation. *J. Appl. Cryst.* 16:548–558.

67. Budde, U., R. Schneppenheim, …, I. Peake. 2008. Detailed von Willebrand factor multimer analysis in patients with von Willebrand disease in the European study, molecular and clinical markers for the diagnosis and management of type 1 von Willebrand disease (MCMDM-1VWD). *J. Thromb. Haemost.* 6:762–771.

68. Budde, U., R. Schneppenheim, …, T. S. Zimmerman. 1990. Luminographic detection of von Willebrand factor multimers in agarose gels and on nitrocellulose membranes. *Thromb. Haemost.* 63:312–315.

69. Schneppenheim, R., H. Plendl, and U. Budde. 1988. Luminography—an alternative assay for detection of von Willebrand factor multimers. *Thromb. Haemost.* 60:133–136.

70. Johannes Kepler University Linz. Crosslinkers and protocols for AFM tip functionalization. Published online April 8, 2015. http://www.jku.at/biophysics/content/e257042.

71. Hutter, J. L., and J. Bechhoefer. 1993. Calibration of atomic-force microscope tips. *Rev. Sci. Instrum.* 64:1868–1873.

72. Ebner, A., P. Hinterdorfer, and H. J. Gruber. 2007. Comparison of different aminofunctionalization strategies for attachment of single antibodies to AFM cantilevers. *Ultramicroscopy.* 107:922–927.

73. Zhu, R., S. Howorka, …, P. Hinterdorfer. 2010. Nanomechanical recognition measurements of individual DNA molecules reveal epigenetic methylation patterns. *Nat. Nanotechnol.* 5:788–791.

74. Rickham, P. P. 1964. Human experimentation: code of ethics of World Medical Association. Declaration of Helsinki. *BMJ.* 2:177.

3 Peptide foldamers in the gas phase

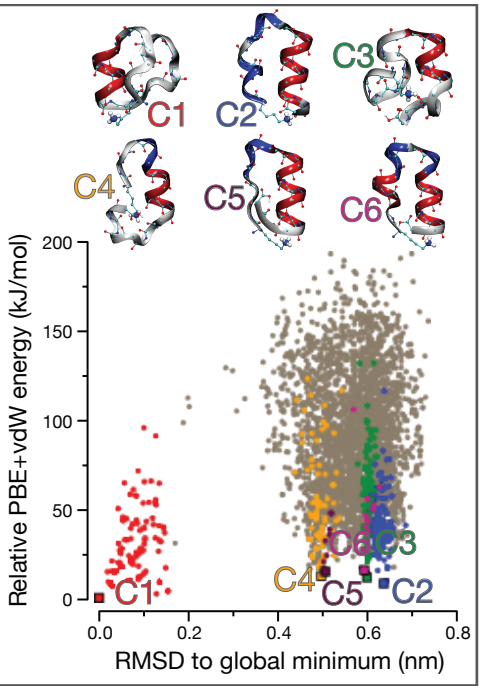
3.1 Going clean: Structure and dynamics of peptides in the gas phase and paths to solvation

Baldauf, C.; Rossi, M.

Going clean: Structure and dynamics of peptides in the gas-phase from first principles and paths to sovation

J. Phys. Cond. Matt. **2015** (49), 493002.

DOI: 10.1088/0953-8984/27/49/493002



OPEN ACCESS
IOP Publishing
Journal of Physics: Condensed Matter
J. Phys.: Condens. Matter **27** (2015) 493002 (27pp)
doi:10.1088/0953-8984/27/49/493002

Topical Review

Going clean: structure and dynamics of peptides in the gas phase and paths to solvation

Carsten Baldauf[1] **and Mariana Rossi**[2]

[1] Fritz Haber Institute, Faradayweg 4–6, 14195 Berlin, Germany
[2] University of Oxford, South Parks Road, OX1 3QZ Oxford, UK

E-mail: baldauf@fhi-berlin.mpg.de and mariana.rossi@chem.ox.ac.uk

Received 8 December 2014, revised 29 September 2015
Accepted for publication 19 October 2015
Published 24 November 2015

Abstract
The gas phase is an artificial environment for biomolecules that has gained much attention both experimentally and theoretically due to its unique characteristic of providing a clean room environment for the comparison between theory and experiment. In this review we give an overview mainly on first-principles simulations of isolated peptides and the initial steps of their interactions with ions and solvent molecules: a bottom up approach to the complexity of biological environments. We focus on the accuracy of different methods to explore the conformational space, the connections between theory and experiment regarding collision cross section evaluations and (anharmonic) vibrational spectra, and the challenges faced in this field.

Keywords: biomolecules, peptides, first-principles electronic structure, vibrational spectroscopy, gas phase

(Some figures may appear in colour only in the online journal)

1. Biomolecules in the gas phase

In spite of bearing little resemblance to biological environments, the experimental and theoretical study of biomolecules in the gas phase has been steadily gaining importance in the past decades, especially among physical scientists. Pioneer experimental studies starting in the late 90s encompassing all main groups of biomolecules [1–9] were able to show that much physical insight on structure formation and dynamics of these molecules can be gained from transfering them to the gas phase. The reason is that the gas phase offers clean conditions, under which theory and experiment can meet on equal footing and can follow a stepwise bottom-up approach towards the full complexity of the real biological environment. The reduced size of the systems allows their treatment with a range of theoretical methods that rely on approaches to solving the quantum mechanical Schrödinger equation, usually referred to as 'first-principles' methods. These methods are typically much more accurate than empirical models—but due to the intrinsic approximations in them, it is also *a priori* unclear how well they are actually able to describe the structure and dynamics of biomolecules in the gas phase. In this synergistic combination, experiments can serve as a benchmark for testing how appropriate the theoretical treatment of these complex systems is, while theory can be employed to give a physical interpretation to experiments.

In this review, we give a brief survey of the current state of the field regarding the study of, in particular, peptides in the gas phase. We focus on the theoretical side of this field, summarizing what is the current state-of-the-art with respect to

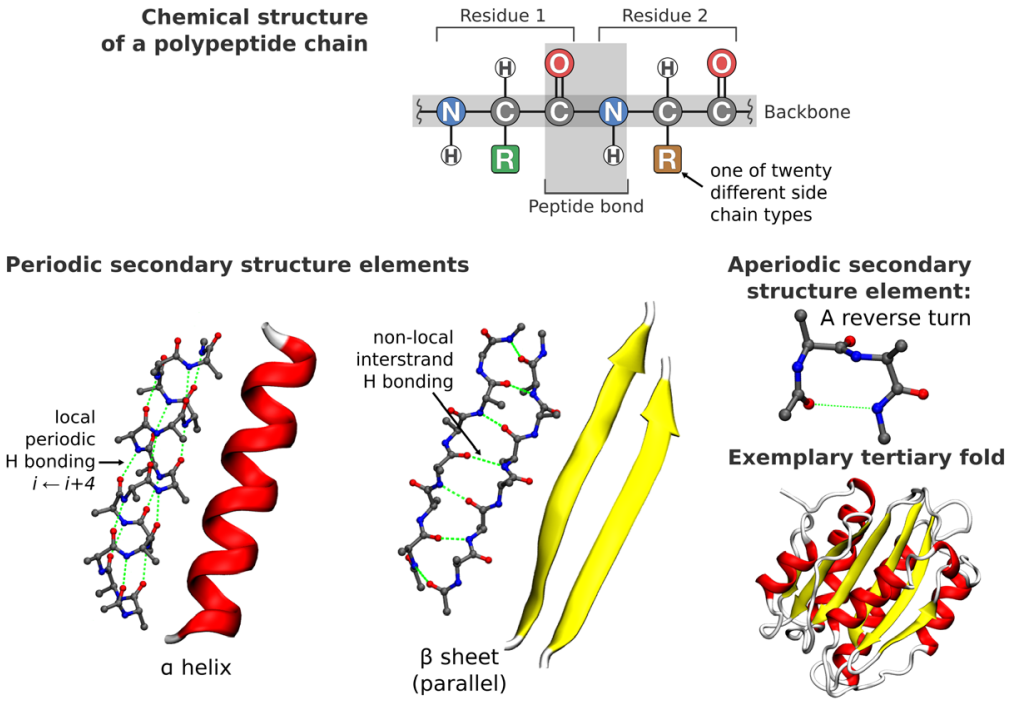


Figure 1. Overview over the structure levels of proteins with the chemical structure of a peptide chain, periodic and aperiodic secondary structure elements, and an example of a tertiary protein fold. The three-dimensional structure examples are taken from PDB-ID 3PPY [10]. Copyright 2011 American Society & Haematology.

accuracy of such calculations, the systems sizes they can treat, what is their predictive power, and where there is room for improvement—basing a good portion of it on works in which the authors were involved. We also give special attention to the dynamical nature of these molecules, and the importance of grasping at least local entropic, anharmonic and temperature effects. There will be less of a focus on long time scale dynamics of these molecules, which involve large conformational rearrangements (e.g. folding). We choose to concentrate on local dynamics because these span time scales currently accessible to first principles potentials. Also they can be connected to most state-of-the-art experiments available in the literature for medium-sized peptides.

1.1. Polymeric biomolecules in the gas phase

There are three main classes of biomolecular oligomers and polymers, namely peptides and proteins (see figure 1), nucleic acids (figure 2(A)) and carbohydrates (figure 2(B)). Below we briefly describe each of them, with a stronger focus on peptides and proteins, which will be the main subject of this review.

Peptides and proteins make up the machinery of life and are involved in virtually all of its manifestations, from comparably small signaling peptides to gigantic protein complexes. A peptide or protein is a linear chain (oligomer) of amino acids (residues) that are linked by so-called peptide bonds (see figure 1, top). Peptide bonds are formed between the amino group and carboxylic acid group of two building blocks. In addition, amino acids carry a side chain '**R**' of differing chemical functionality. The sequence of the different amino acid side chains **R** is called primary structure and defines the structure and dynamics of the peptide or protein. Oligomers beyond a certain length (from about 50 amino acids on) that are able to form distinct structural motifs are called proteins. Structure formation at the level of peptides (secondary structure) is mainly dependent on the conformational properties of the monomers and backbone hydrogen bonding. In larger oligomers, i.e. in proteins, side chain interactions and packing gain importance and govern tertiary structure formation. These larger proteins or even complexes thereof can be studied in isolation as well (see, for example, a recent review by Carol Robinson [11]).

Among other biological functions, nucleic acids are the carriers of genetic information. In an organism, a sequence of nucleotides in deoxyribose nucleic acids (DNA) can be transcribed into ribose nucleic acids (RNA) that then serves as template for the stepwise linkage of the amino acids into peptide or protein chain. They feature a sugar-phosphate backbone with nucleobases connected to the (deoxy)ribose moieties (see figure 2(A) for a pictorial representation of the different groups). Structure formation is mainly triggered via intermolecular hydrogen bonding between specific pairs of bases (base pairing) in case of DNA or intramolecular base pairing in case of RNA. A recent review by Abi-Ghanem and Gabelica [12] may serve as an entry point to the literature

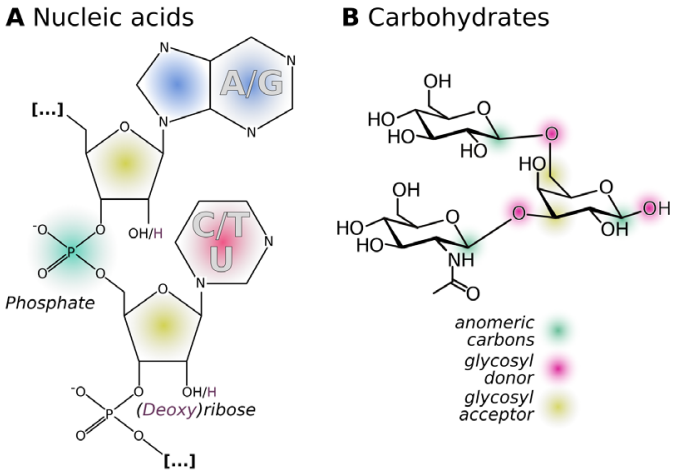


Figure 2. Schematic representations of (A) nucleic acids and (B) carbohydrates.

about nucleic acids in the gas phase. Arcella *et al* [13] investigate DNA in the gas phase by combining ion-mobility mass spectrometry and extensive classical molecular dynamics (MD) and *ab initio* molecular dynamics (AIMD) simulations. They describe rich dynamics of DNA that quickly looses memory of its solution structure in the gas phase and explores a large conformational space. Of special interest is the observation of protons hopping between phosphates of the DNA backbone that was seen in AIMD simulations.

Polymeric carbohydrates serve as nutrition and energy source or as structural scaffolds. They can also be linked to proteins acting as recognition molecules and possibly playing a role in protein folding. Most of the known carbohydrates are composed of around 20 different monosaccharide units connected to each other by what is called the glycosidic bond or linkage (see figure 2(B)). In contrast to the backbone of peptides or the backbone of nucleic acids, carbohydrates are not necessarily composed as a linear chain. The building blocks have one donor (the anomeric C) but multiple acceptors for glycosidic bonds, such that branched structures can be realized. In addition, due to chirality, glycosidic bonds can be formed in two chiral (enantiomeric) forms (α or β). These contributions result in a diversity of possible topologies of carbohydrates that surpasses the number of possible sequences in nucleic acids and peptides by orders of magnitude, even with relatively small numbers of building blocks [14]. The significant conformational degrees of freedom are rotations around the single bonds of the glycosidic linkages and the conformation of the monosaccharide rings.

The main focus in this review will be on secondary structure stabilization and dynamics in peptides containing from a few to some tens of amino acids. These motifs are shown in figure 1. Briefly there are three main elements of secondary structure, namely, helices, pleated-sheets, and turns. The turns are regarded as non-periodic motifs, while helices and sheets are regarded as periodic, in the sense that a repeating unit can be defined, allowing for a characterization based on pairs of torsional angles. The nomenclature given to the helices depend on the hydrogen bonding pattern that arise from their constituting residues (amino acids). The most famous types, the α and the 3_{10} helices are characterized by H-bonds between residue i to $i+4$ and residue i to $i+3$, respectively. Sheets are also stabilized by backbone H bonds and can be characterized as parallel and anti-parallel depending on the relative orientation of their peptide chains. Finally, turns are necessary motifs to reverse the propagation of sheets and helices, so that compact structures can be formed. It is not necessary for a H-bond to form in order for the motif to be characterized as a turn, but many do form through the formation of H-bonds. The most common type is known as the β turn. Turns cause a complete reversal of the direction of structure propogation.

1.2. Experimental techniques probing conformation and dynamics in the gas phase

The study of (bio)molecules in the gas phase has become more popular in the past decades mainly due to the development of experimental techniques in the late eighties, that can gently transfer intact biomolecules to the gas phase, like MALDI (matrix-assisted laser-desorption ionization [15]) and ESI (electro-spray ionization [16]), in combination with high-accuracy mass spectrometers [17, 18] or in molecular beams [19].

When dealing with peptides, it is possible to isolate secondary structure motifs in the gas phase, so that their 'unperturbed' energy landscape and stabilizing intermolecular interactions can be carefully studied. The environmental effects can then be added in a controlled way, for example by the stepwise addition of water molecules to the polypeptide or by adding ions to the complexes. At the same time, these clean experiments in the gas phase allow to benchmark theoretical methods, at system sizes that can be treated in a fully first-principles manner. There is much debate as to how biologically relevant the study of biomolecules in the gas phase actually is [20–22], since it is to be expected that due to the lack of solvent and hydrophobic/hydrophilic interactions one

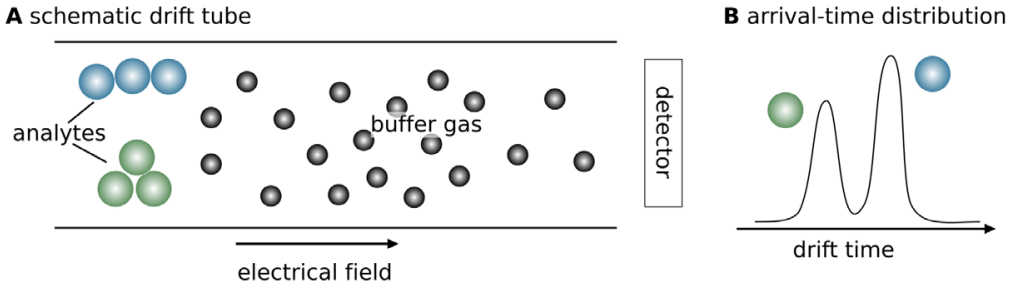


Figure 3. Schematic representation of an ion mobility spectrometry experiment.

can actually stabilize different conformations in the gas phase. From a more physical perspective, it is undeniable though that in these experiments much insight about the fundamental stabilizing interactions can be gained, also encompassing what is the role of the protonation state and the first shells of solvation, or the interaction with only ions, ions and water, etc. This understanding can be certainly transferred to the more complex biological environment.

There are several reviews about studies of biomolecules (peptides, proteins, sugars, etc) in the gas phase in the literature, from which we highlight only a few for the interested reader, namely [1–3, 6, 8, 9, 18, 20–25]. Below we give a brief overview of the main experimental techniques that yield quantities which can be connected to theoretical calculations that we will review in the next sections.

1.2.1. Ion mobility-mass spectrometry.
Mass spectrometry (MS) is a powerful gas-phase experimental technique that separates ionic clusters or molecular ions based on their mass-to-charge ratio (m/z). With ion mobility (IM), or gas chromatography, charged molecules and clusters can be separated according to their different mobility in a buffer gas. Especially the combination of both techniques, ion mobility-mass spectrometry (IM-MS), first accomplished in 1962, [26] can allow for the separation and characterization of mixtures of compounds or conformers that would otherwise not be distinguishable. In the context of this review, we focus especially on the ability of IM-MS to investigate structural and dynamical properties of peptides.

In IM-MS experiments, an electric field drags the ions through a drift tube of a certain length. This drift tube is filled with a buffer gas (often He or N_2) and collisions between buffer gas and ions slow down the ions depending on their shape and size. As a result, an arrival-time distribution (ATD) of m/z selected ions is measured by a detector, as sketched in figure 3. The arrival times can be transformed into collision cross sections (CCSs) by the Mason–Schamp equation [27]

$$\mathrm{CCS} = \frac{3ze}{16N}\left(\frac{2\pi}{\mu k_\mathrm{B} T}\right)^{\frac{1}{2}}\frac{1}{K_0}, \qquad (1)$$

where ze refers to the net charge of the system, μ is the reduced mass of the ion and the buffer-gas particles (usually He atoms), and k_B is Boltzmann's constant. The reduced mobility K_0 is the proportionality constant that relates drift velocity v_d and the electric field E of the apparatus following the relation $v_d = K_0 E$. The resulting CCS is a geometrical property of the molecule and it is ideally independent of the apparatus used.

A few examples of the use of IM-MS experiments to study structure and dynamics of peptides in the gas phase are

- Jarrold and coworkers have developed a high-temperature drift-tube instrument and studied polyalanine helices in the gas phase from room temperature to 725 K [28]. The surprising finding is that helical structures can be observed still at these high temperatures for the peptide Ac-Ala_{15}-Lys(H^+). Tkatchenko *et al* identified van der Waals (vdW) interactions as the crucial stabilizing contribution in DFT-based molecular dynamics simulations [29], being essential to explain the high temperature stability of the helical structure observed in experiment.
- Based on ion-mobility measurements, Shelimov and Jarrold were able to show the unfolding and refolding of Cytochrome C in vacuum [30]. The folded versus unfolded state is linked to different charge states with a folded to unfolded transition between charge states +5 and +7.
- By using a combination of IM-MS and MD simulation, von Helden and co-workers studied different combinations of *cis/trans* isomerization states of prolyl peptide bonds of ubiquitin [31]. CCS measurements and computations are sensitive enough to reveal the *cis* or *trans* conformation of a single peptide bond in a biological macromolecule.
- The group of Clemmer has played a leading role in devising drift-tube apparatus using them to investigate different kinetically-trapped conformations of, for example, Bradykinin [32, 33].
- Russel and co-workers have used a cryogenic drift tube at 80 K to investigate the structures of singly-protonated water clusters [34]. They were able to measure small (1–30 water molecules) and large clusters (31 up to about 120 water molecules) and to assign changes of H bonding upon loss of single water molecules from the clusters.

1.2.2. Vibrational spectroscopy. The low concentration of molecules in the gas phase renders it difficult to obtain vibrational spectra through absorption spectroscopy, the technique commonly used in the condensed phase. Instead, in what is called action spectroscopy, an intense tunable laser that acts on a comparably small number of molecules. When a resonance

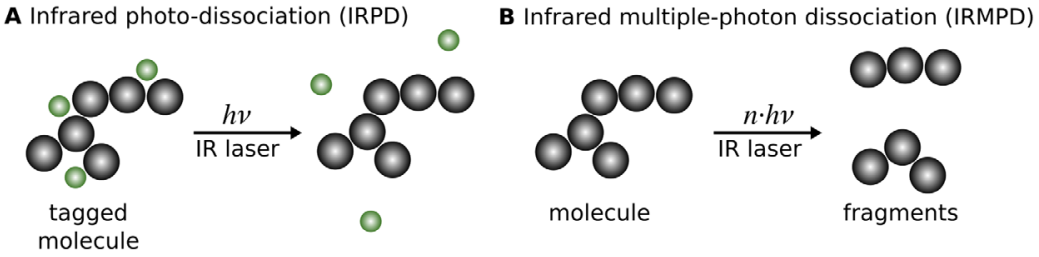


Figure 4. Vibrational action-spectroscopy techniques.

is encountered, due to the absorption of single or multiple photons, the sample dissociates or fragments and detection via mass spectrometry is possible. One can detect either the fragments or the depletion of the molecular beam. At this point, two types of action spectroscopy can be performed. One technique, commonly called infrared photo-dissociation IRPD (sketched in figure 4(A)), is usually performed at lower temperatures and uses inert tags (e.g. H_2, Ar, Ne, etc) on the target molecule that are released after the absorption of one or very few photons, due to the low binding energy of the tag. Another technique, called infrared multiple-photon dissociation IRMPD (sketched in figure 4(B)) does not use any tag and simply measures the fragmentation of whole molecules due to the sequential absorption of at least a few tens of photons. For detailed reviews of the experimental techniques, we point the reader to [18, 23, 35, 36].

In both action spectroscopy techniques mentioned above, non-linear effects can arise due to the absorption of more than one photon. Therefore, different from absorption spectroscopy where the spectra can be safely approximated by a linear response theory, here it is not *a priori* clear that the vibrational spectra measured in this manner will allow a linear response modelling. Especially for IRMPD, where indeed many photons are absorbed sequentially, causing induced and spontaneous emission as well as energy redistribution among vibrational modes, it is clear that a linear response approximation may fail. It has been shown that while the line shape and intensity of the peaks can be strongly influenced by these non-linear absorption effects, the peak positions usually follow the ones calculated by linear response [37, 38] with slight red-shifts due to anharmonicities. In certain systems, it is found in particular, for lower frequencies below $1200\,\mathrm{cm}^{-1}$, some peaks are transparent to IRMPD (but not to IRPD), as was shown by the group of Asmis for microhydrated nitrate-nitric acid clusters [39] and bisulfate/sulfuric acid/water clusters [40]. The reason they propose is that the absorption of photons disrupts the hydrogen bond network of these systems and causes the modes to go out of resonance with the frequency of the laser. More specific comparisons regarding theoretical modeling and experimental IR spectra will be given in section 4.2.

In the experimental studies, several different parts of the vibrational spectra can be probed, which are sensitive to different conformational properties: (i) The amide A/B regions, comprising localized CH and NH stretch vibrations above $\approx 2500\,\mathrm{cm}^{-1}$, sensitive to the H-bonding pattern; (ii) the amide I (mainly collective CO stretch vibrations), amide II (mainly collective NH bend vibrations), and amide III (collective and localised CH and CN bend vibrations) regions between 2000 and $800\,\mathrm{cm}^{-1}$, sensitive to backbone conformation; and (iii) the 'far-infrared' region, below $800\,\mathrm{cm}^{-1}$, which contains collective vibrations and is also sensitive to backbone conformation. While much focus has been given to the amide A/B and amide I and II regions in most studies, attention has been called to the amide III region in mid-sized polypeptides [41, 42] and to the far-infrared region in small polypeptides [43] as regions that can be used to differentiate conformations, if anharmonicities of the potential-energy surface are taken into account. As an illustration, we show these regions and the harmonic normal modes of vibrations calculated with the PBE exchange correlation functional for the formamide molecule in figure 5.

Gas phase investigations can be used to study distinct aspects of protein secondary structure, peptide bond properties, and aspects of microsolvation. In the following we discuss a few outstanding examples:

- Tanabe *et al* have used UV/IR pump-probe experiments on clusters of acetanilide and water to investigate the motion of a single water molecule from the hydrogen bond acceptor (CO group) to the hydrogen bond donor (NH group) of a **peptide bond** [44].
- Gerhards and co-workers studied dimers of the short peptide Ac-Val-Tyr(Me)-NHMe in molecular beam experiments [45]. The combination of IR/UV double-resonance spectroscopy and simulated vibrational spectra (harmonic, B3LYP/cc-pVDZ) identifies the formation of an anti-parallel β **sheet-like structure**. The study shows that sheet-formation can be regarded as an intrinsic trend of peptides that is not necessarily linked to aqueous solution.
- The group of Rizzo has a long-standing experience in UV/IR experiments on **helical peptides** Ac-Phe-Ala_n-LysH^+ with a C terminal protonated Lys and a Phe residue as UV chromophore [46]. The helical pattern has been elucidated with a ^{15}N labeling technique. The C terminal capping motif that is present in the longer helices with $n \geqslant 5$ has recently been shown to be present also in short peptides with $n = 1$ [47]. These results confirm predictions about the helix onset made by Rossi *et al* for very related systems [48]. These systems with the aromatic

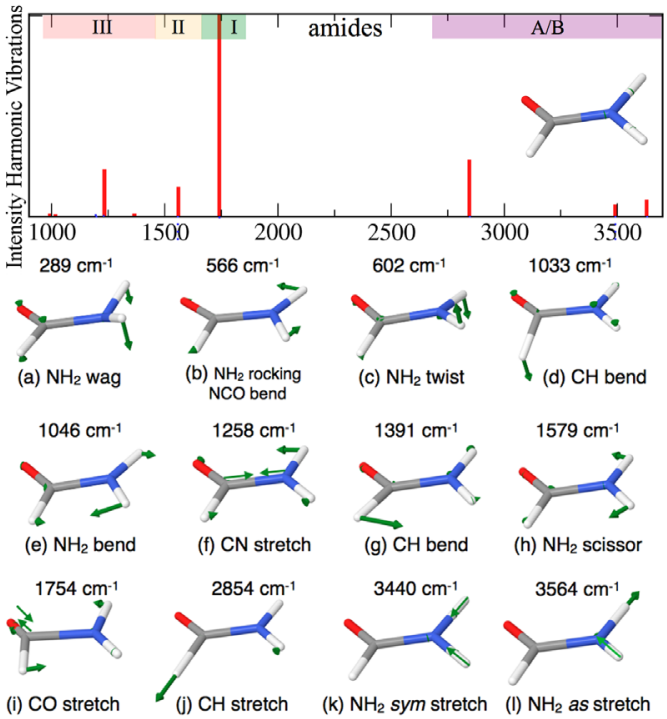


Figure 5. Gas phase spectrum and normal modes of vibrations calculated with DFT-PBE functional for the formamide molecule. Amides I, II, III, and A/B regions are marked on top.

Phe side chain are a challenge to theory and will be discussed further on in this review.

- Besides strands/sheets and helices, **turns** are the third main secondary structure motif in proteins. The group of Mons studied peptides Ac-Phe-NH_2, Ac-Phe-Pro-NH_2, and Ac-Pro-Phe-NH_2 from supersonic molecular beams wit UV/IR double resonance spectroscopy [49]. The authors assign various turn types and indicate the dependence of Phe conformations on the neighboring residues.
- Johnson and his group have most elegantly shown how gas-phase infrared spectroscopy of cold ionic **complexes** can be used to elucidate not only molecular structure, but also the way two molecules interact with each other [50]. They use site-specifically placed ^{13}C labels as conformational reporters. Difference spectra between the distinctly labeled systems allow for structural investigations of a single peptide ion and also complex formation through binding to sodium cations or with other molecules.
- In order to directly estimate the energy barriers between different conformers, Zwier and co-workers developed a double resonance conformer selective pump and dump technique that excites molecules to a higher electronic level and then relaxes them back into a specific vibrational ground state [51]. With this approach the authors were able to **reconstruct the potential-energy surface** of tryptamine.
- Compagnon and coworkers have carried out seminal work on the FELIX free electron laser on peptides in the gas-phase, for example looking at backbone preferences [52], internal proton transfer that can stabilize zwitterionic structures in the gas-phase [53], and microsolvation of amino acids [54]. More recently they have been looking at sugars in the gas phase [55], focusing especially on the issue of 'anharmonicities in vibrational modes'.
- The group of Lisy has a body of work based on IRPD regarding the influence of charge due to the interaction with 'metal ions and temperature' on the conformational preferences of small biomolecules [56, 57].
- Vaden, Snoek, and coworkers have measured IRMPD spectra of a variety of peptides in the gas phase, also performing extensive structural searches involving density functional theory. They have, for example, studied the Ala_nH^+, $n = 3, 4, 5, 7$ series of peptides [58] in the amide A/B region concluding that these peptides form mostly globular structures at larger sizes, despite the high propensity of the Ala amino acid to for helices. They have also looked at peptide sequences relevant to amyloid formation, showing that even if the isolated structure of Ac-VQIVYK-NHMe is folded, the simple interaction with another monomer in the gas phase seems be energetically favorable enough to trigger a conformational change and 'β-sheet aggregation' [59].

Vibrational spectroscopy techniques can also be combined with ion mobility-mass spectrometry. A first example of, in that case, electronic spectroscopy of mobility selected peptides was published by Rizzo and coworkers [60]. They

use a field asymmetric waveform ion mobility spectrometry (FAIMS) setup combined with UV photofragment spectroscopy in order to decompose the electronic spectrum of doubly-protonated bradykinin in a conformer-specific manner. Also Voronina and Rizzo demonstrate how to use a combination of ion-mobility selection and cold-ion spectroscopy to study kinetically trapped conformers of triply-protonated bradykinin [61]. von Helden, Pagel, and co-workers have used ion mobility in order to separate conformers of protonated benzocaine and to record vibrational spectra [62].

The spectral resolution can be improved by measuring vibrational spectra of cold species. This can, for example, be realized in cold traps that are utilized in IR/UV double resonance experiments where changes of UV fragmentation yield are recorded as a function of IR excitations [63, 64]. Alternatively, ions can be embedded in liquid He nanodroplets [65] and therewith cooled to an equilibrium temperature of about 0.4 K. Employing such a setup, von Helden and coworkers have measured vibrational spectra of the short peptide leucine-enkephalin [66].

2. Potential-energy surfaces for peptides

2.1. Accuracy of the potential-energy surface

The potential-energy surface (PES) of a system is often defined as an energy function of the coordinates that tells how energy changes with respect to a change in any atomic position. This definition assumes an adiabatic separation of the electronic and nuclear degrees of freedom (known as the Born-Oppenheimer approximation). Moreover, it is usually (but not necessarily) also connected to the assumption that nuclei are classical particles. Even if both of these assumptions can break down in many situations (some of them discussed in the next sections), they are, in most cases, a good approximation or at least a good starting point to map the energy profile of a system.

When dealing with biomolecules in general, the challenging aspect is that the PES are often far from simple: due to the existence of several soft and anharmonic degrees of freedom these PES tend to have several different local minima—all of which will contribute to the partition function and thus define the thermodynamical properties of the system. If this PES is not rigorously described also all thermodynamic properties and structural preferences of the system will not be reliable. Especially the amount of anharmonic degrees of freedom make most harmonic approximations fail for these systems.

Perhaps the most popular way of evaluating PES are the so-called force fields. Force fields are parametrized empirical energy functions that represent the energy of a given system in terms of the sum of qualitatively different interactions. In the case of molecules (and especially peptides) the different contributions are separated into bonded interactions (e.g. potentials for bond lengths, bond angles, and torsions) and non-bonded interactions (e.g. van der Waals and electrostatics). For all of these terms, the functional form is physically motivated but arbitrary, and the parameters are fitted to either experimental data or quantum chemistry methods. The advantages of such an approach is that energy evaluations are computationally cheap. Therefore, these methods (if used in combination with smart sampling techniques [67]) typically allow enough statistical sampling to enable the evaluation of thermodynamical properties and to treat system sizes that can bear more connection to biological size- and time-scales with respect to more accurate methods that are too computationally expensive. If used with caution, these potentials can yield good physical insight on the structure and dynamics of biomolecules. However, it is becoming more clear that their performance in many situations is far from optimal. Especially regarding polypeptides, recent literature has shown that force fields have several limitations when compared and benchmarked against higher level quantum chemistry methods. Relative energies between different peptide conformations are not well reproduced [68–71] and differ quite drastically between different force fields. Regarding the interaction of peptides with ions, force fields have been shown to yield even poorer energetics with respect to high level theoretical benchmark data [72, 73], even when especially tailored parameters and polarizable potentials are used. More recently, a study has shown that kinetic models derived from converged simulations based on different non-polarizable force fields largely disagree [74].

The desired solution would be to describe the potential-energy surface (PES) at least as accurately as possible for the electronic degrees of freedom—which would mean to use methods like full configuration interaction (full CI), coupled cluster with a high enough order of excitations (e.g. with single, double, and perturbative triple excitations CCSD(T)), or quantum Monte Carlo (QMC). These methods are considered the gold standard of quantum chemistry, and do indeed provide a very accurate description of potential energy surfaces, but of course, are very costly to compute. Even if they can be used for benchmarking purposes it is not computationally feasible to routinely use them for PES exploration and the simulation of other physical properties.

A good compromise can be found among the wave function based methods, for example with Møller-Plesset perturbation theory (MP2) or coupled-cluster methods with lower-order excitations, e.g. singles and doubles (CCSD). A promising route is to use approximations like the domain based local pair natural orbital coupled cluster method with single-, double-, and perturbative triple excitations (DLPNO–CCSD(T)) [75]. The method is described as efficient enough to perform rather accurate coupled cluster calculations even for relatively large molecules with hundreds of atoms. However, some of the approximations must be carefully balanced [76]. It is typically computationally cheaper to use electronic density based methods like density-functional theory (DFT). DFT, with its approximate exchange correlation functionals, is arguably the best compromise between cost and accuracy in the market of electronic structure theory methods. Its advantage is that it allows one to treat molecules of sizes up to a few thousand atoms and reach time scales of hundreds of picoseconds in its most optimized implementations (Big-DFT [77], ONETEP [78], FHI-aims [79], CASTEP [80], CP2K [81], etc).

Table 1. Mean absolute error and maximum error for the energy hierarchies of 16 conformers of Gly-Phe-Ala (GFA), 15 conformers of Gly-Gly-Phe (GGF), 15 conformers of Phe-Gly-Gly (FGG), and 27 conformers of Ac-Ala$_3$-NMe, compared to CCSD(T) reference data from [70, 82].

	PBE	PBE + vdW^{TS}	PBE + MBD	PBE0	PBE0 + vdW^{TS}	PBE0 + MBD
FGG						
MAE	43(1.0)	37(0.8)	36(0.8)	35(0.8)	23(0.5)	23(0.5)
Max.	160(3.7)	59(1.4)	88(2.0)	132(3.0)	38(0.9)	59(1.4)
GFA						
MAE	53(1.2)	32(0.7)	44(1.0)	40(0.9)	17(0.4)	25(0.6)
Max.	108(2.5)	88(2.0)	76(1.7)	89(2.0)	72(1.7)	61(1.4)
GGF						
MAE	48(1.1)	36(0.8)	40(0.9)	38(0.9)	26(0.6)	28(0.6)
Max.	143(3.3)	99(2.3)	84(1.9)	119(2.7)	78(1.8)	66(1.5)
Ac-Ala$_3$-NMe						
MAE	55(1.3)	21(0.5)	22(0.5)	54(1.2)	18(0.4)	20(0.5)
Max.	131(3.0)	72(1.7)	66(1.5)	132(3.0)	47(1.1)	54(1.2)
	OPLS-aa	Amber99sb	Charmm22	AmoebaPro04		
Ac-Ala$_3$-NMe						
MAE	108(2.5)	42(1.0)	91(2.1)	53(1.2)		
Max.	246(5.7)	86(2.0)	271(6.2)	112(2.6)		
GGF						
MAE				91(2.1)		
Max.				606(14.0)		

Note: Values for the mean-absolute errors (MAE) and maximal errors (Max.) are reported in meV (in parentheses: converted to kcal mol^{-1}).

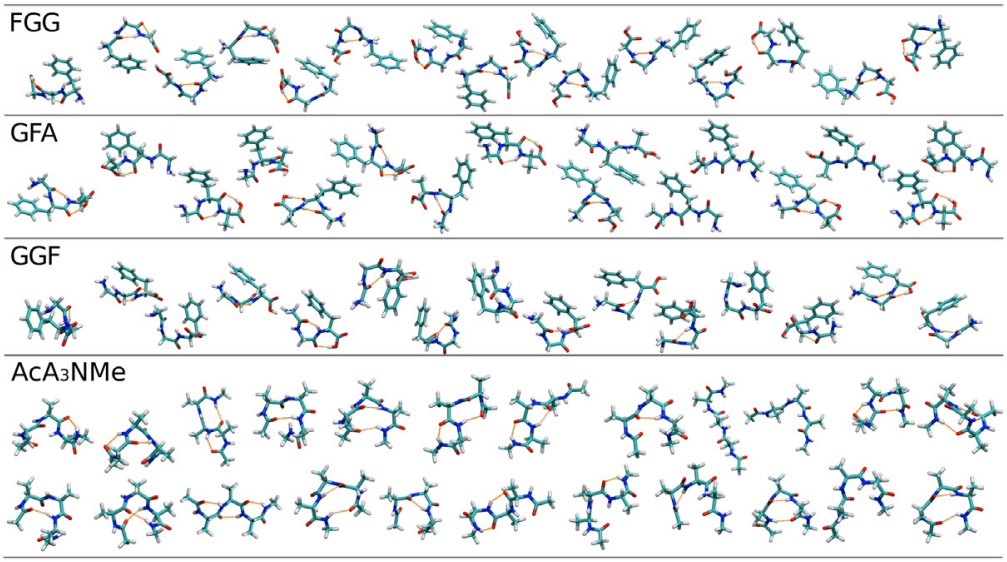


Figure 6. Conformers of the peptides Phe-Gly-Gly (FGG), Gly-Phe-Ala (GFA), Gly-Gly-Phe (GGF), and Ac-Ala$_3$-NMe (AcA$_3$NMe) used for energy benchmark calculations appearing in references [70] and [82].

It is well known that results from DFT can depend on the choice of exchange-correlation functional. However, since the theory itself is based on the first principles of quantum mechanics, it is possible to obtain accurate results as long as one ensures that the chosen functional can describe the relevant physical properties of the system. For example, most standard DFT functionals lack, by construction, long range van der Waals (vdW) dispersion. It is, however, now widely accepted that these interactions have a critical impact on the structure [48, 69, 70, 72, 73, 83, 84] and dynamics [29, 83] of peptides, especially for the larger sizes. It becomes thus almost mandatory to include these interactions in the most accurate manner in DFT calculations of peptides in any type of environment, and several schemes for including these corrections have been proposed in the last decade, which were nicely reviewed in [85]. Also, the inclusion of Hartree–Fock exchange can mitigate the self-interaction/delocalization problem of DFT and substantially change the strength of H

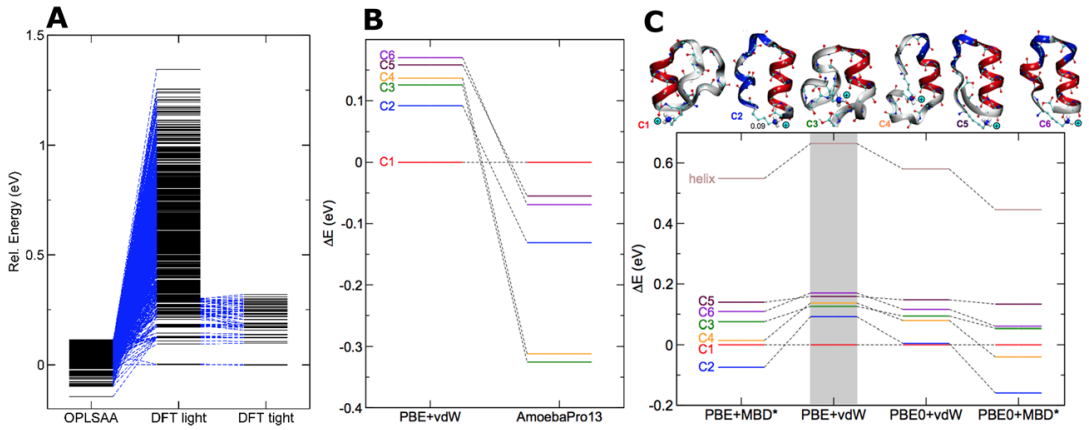


Figure 7. Ac-Lys-Ala$_{19}$-LysH^{+} reproduced and adapted from [98], copyright 2015 Royal Society of Chemistry and [99], with permission from F Schubert.

bonds, the description of polarizability, or barriers for conformational dynamics. What is more prudent to avoid in DFT is to blindly use different types of functionals without any kind of physical reasoning or benchmarks.

As an example of the type of accuracy that can be reached with state-of-the-art DFT methods nowadays, we show in table 1 mean absolute errors and maximum errors on relative energies for three-residue peptides, shown in figure 6 (FGG, GFA, GGF, Ac-Ala$_3$-NMe), of DFT functionals with respect to CCSD(T) reference benchmark data. We test a generalized gradient exchange correlation functional (PBE [86]) and include both a pairwise van der Waals correction with C_6 coefficients that depend on the electronic density [87] (vdW$^{\mathrm{TS}}$), and another that includes both electrostatic screening and many body effects up to infinite order through a coupled fluctuating dipole model [88, 89] (MBD@rsSCS, which we here call MBD). We also test a hybrid exchange correlation functional with these corrections, namely PBE0 [90]. For comparison, we also calculate the same relative energies with popular non-polarizable force fields (OPLS-aa [91], Amber99sb [92], Charmm22 [93, 94]) and the polarizable force field AmoebaPro04 [95, 96]. Augmenting DFT approaches with a correction for long-range van der Waals interactions leads to energy estimates that agree very well with CCSD(T) calculations, which is evident by low mean-absolute errors (MAE) and low maximal errors. For example PBE0 + MBD yields MAEs of only up to 28 meV (0.6 kcal mol^{-1}) and a maximal error of 66 meV (1.5 kcal mol^{-1}). The force fields tested here and the bare functionals alike give higher MAE and also higher maximal errors that severely limit their predictive power.

In order to illustrate how such errors can impact larger polypeptides, the experimental benchmark helix-forming peptide Ac-Phe-Ala$_5$-LysH^{+} is ideal. From very accurate conformer selective UV-IR double resonance experiments in the gas-phase by Stearns and coworkers [46], it was established that four conformers are present in the experimental beam, which have been satisfactorily assigned to helix-forming structures, based on the similarity of their harmonic IR spectra to the measured ones. A subsequent study [97] considered 19 density functionals, plus Hartree–Fock and MP2 methods, finding that the spread of the relative energies of these four conformers could vary by around 0.15 eV for these methods. None of the functionals considered included long-range van der Waals interactions. Further studies on the same system by Rossi and coworkers [69] considered a larger pool of conformers coming from an extensive first-principles scan of the PES of this peptide. Based on the benchmarks shown in figure 6, the authors found that when considering the energy hierarchies at the PBE0 + MBD level and (harmonic) zero point energy contributions on this system, the four conformers observed in experiment are indeed predicted to be the ones with lowest energies. The spread of their energy differences is also consistent with what is estimated from experiment (≈50 meV), and within the estimated error bars, such that the detailed energy hierarchy between them cannot be safely predicted by any DFT method. Interestingly, [69] finds that the relative abundances for different conformers observed in experiment are better explained by a kinetic trapping from higher temperatures.

Finally for even larger peptides, where the experimental data is also not so conclusive, small energy differences can be even more important as the conformational landscape can get more congested. We take as an example the 20-residue peptide Ac-Lys-Ala$_{19}$-H^{+}, studied in [98] by Schubert, the author of this review, and coworkers. We show in figure 7 (data reproduced from [98] and [99]) in panel (A) the comparison between the force field relative energies for thousands of conformers predicted by the OPLS-aa force field, and relative energies of the same conformers when further relaxed with PBE + vdW$^{\mathrm{TS}}$ 'light settings' (smaller basis sets and integration grids in the FHI-aims [79] code) and 'tight settings' (larger basis sets and integration grids). The scatter is huge, spanning up to 1.5 eV in DFT for conformers that were 0.5 eV apart in OPLS-aa. We also show in figures 7(B) and (C), for a set of selected conformers of this molecule the comparison between the energy hierarchies of PBE + vdW$^{\mathrm{TS}}$ and the AmoebaPro13 force field [100], and the comparison between the different functionals and van der Waals corrections discussed above. The energy differences between the

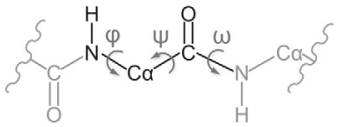


Figure 8. Backbone torsion angles of a prototypical amino acid building block embedded in a peptide chain.

functionals are much smaller than the difference comparing to AmoebaPro13. Many body van der Waals dispersion do indeed have an impact in molecules of this size, which in this case also improves agreement to experimental data, as discussed further in section 4.2.

2.2. Sampling the PES connecting to first-principles methods

The degrees of freedom (DOF) that define a PES are the positions of all atoms of the molecules expressed in, for example, Cartesian space, internal coordinates, etc. For molecules, one can often simplify that (reducing the number of DOF) by assuming a fixed configuration of the molecular system (basically assuming that covalent bonds do not break). As a consequence, an internal coordinate system consisting of bond lengths, bond angles, and torsion angles can be used to describe a molecule's structural (conformational) space. Since bond lengths and bond angles typically vary around a single equilibrium value, torsion angles are often the most descriptive internal coordinates for a molecular system. An exploration of a molecule's potential energy surface must sample the space defined by the combination of all its torsional degrees of freedom. For a typical peptide molecule with three backbone torsion angles per residue and further torsions in the side chain, the problem easily gets too large for a systematic grid-based enumeration of possible points on the PES. A single alanine building block in a peptide chain has three torsional DOF. (See figure 8: the torsions ϕ and ψ represent rotations around single bonds and the peptide bond torsion angle ω adopts *cis* or *trans* conformations. Assuming a grid of 60 degrees for discretization of the single-bond rotations yields $6 \times 6 \times 2 = 72$ conformations to test for a single building block. For a chain of N building blocks this number virtually explodes already for short peptides with 72^N. A variety of strategies has been developed and employed to explore these conformational spaces connecting to first principles methods. Below, we will give a rough definition and some examples of them.

- Systematic searches can be performed by discretization of the involved degrees of freedom with sufficiently fine grids. All combinations of torsion angles are either subject to a single point energy calculation or serve as starting point for local geometry optimizations. Such an approach is well applicable to small molecular systems, e.g. dipeptides. With a more 'target-oriented' objective, also bigger systems can be studied in a systematic way, if only a particular region of the search space is of interest. An example is the search for all possible helical structures in homologous peptides, i.e. peptides which have their backbones extended by methylene units. With the aim of finding such periodic and hydrogen bonded structures, the same combination of backbone torsion angles is applied to all subunits and only geometries that are (i) clash free and (ii) feature a backbone hydrogen-bonding pattern of interest are considered [101–103].
- Systematic searches can easily be performed for monomers. The knowledge gained in this way can then be combined in the creation of starting structures for longer oligomers of the respective building block(s). This approach has been successfully employed for example to β-peptides, which are homologous peptides with an addition of one methylene unit. [103–106].
- Parallel-tempering or replica-exchange molecular dynamics (REMD) can substantially enhance the sampling of conformational space in comparison to standard MD simulations [107–112]. REMD requires only limited human interaction and no definition of collective variable or alike. Robust protocols exist for a wide range of simulation programs. Several copies (a.k.a. replicas) are simulated in parallel by means of MD simulations at different temperatures. At predefined intervals, pairs of replicas with neighboring temperatures are eventually swapped based on a Metropolis criterion. The individual copies traverse a wide temperature range and can overcome barriers.
- Basin hopping [113] reduces the PES to attraction basins centered on local minima. In contrast to REMD, moves on the landscape do not follow realistic pathways. The basic algorithm starts with a structure guess and a local optimization to the next local minimum. A perturbation of coordinates generates a new staring point for a geometry optimization that leads to the next minimum. This sequence of coordinate perturbation and local optimization is repeated until a convergence criterion is met. Frequently used implementations are for example in the programs TINKER [114] or GMIN [113].
- Genetic algorithms (GAs) are frequently used for global structure search and optimization of chemical compounds [115–117]. They use a 'survival of the fittest' concept. Starting from a population of random solutions, genetic operations are applied and energy-optimal solutions are selected. GAs use the accumulated information to explore promising regions of conformational space. Examples are the program foldaway by Damsbo *et al* [118] and the program Fafoom [119, 120] that can employ first-principles techniques.

A complete sweep of the potential-energy surface with any of the above mentioned methods is anything but trivial. All methods require parameter choices that have to be made by the respective user as well as a careful selection of the energy function to be used. While force fields offer low computational costs and therefore allow for a more exhaustive sampling of the PES, the results can suffer from the systematic energy errors that were discussed in the previous section. First-principles methods offer a description of the energetics that is unbiased by empirical parameters, but that may demand far more computational resources. Clever combinations of search

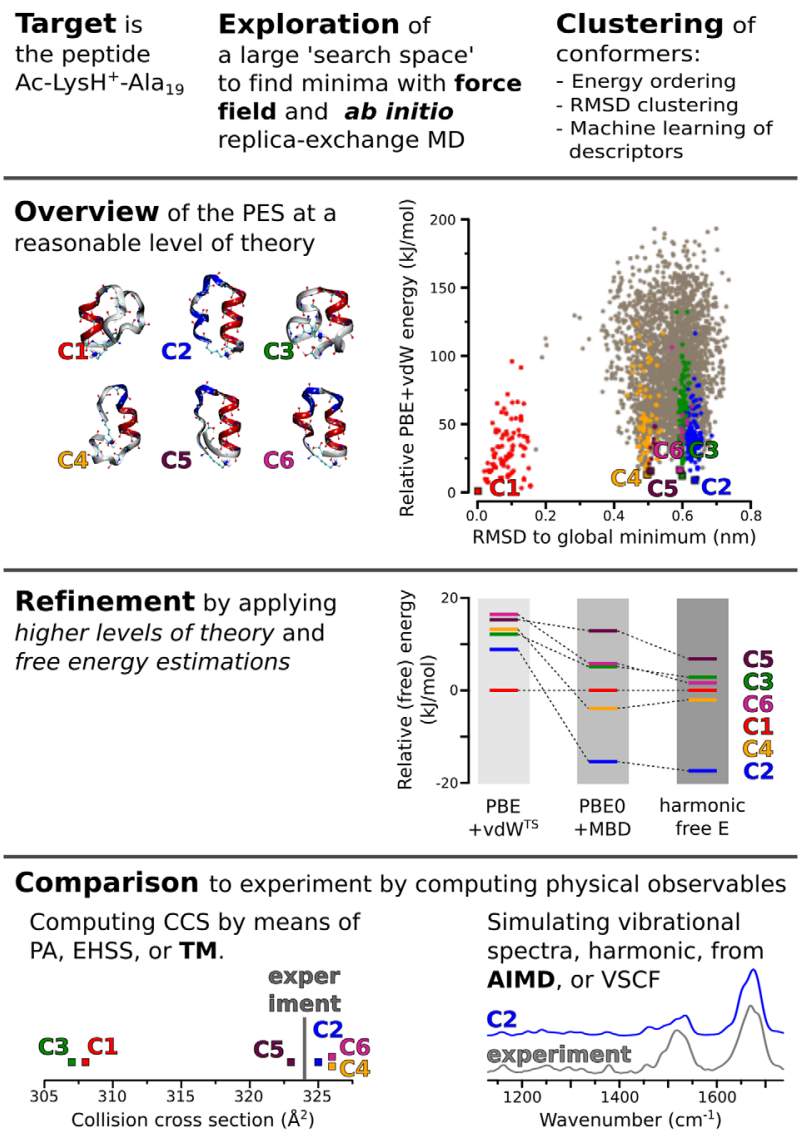


Figure 9. Typical steps followed by theoretical studies regarding structure search and prediction based on first principles methods.

techniques and stepwise increase of accuracy can be a way out that, however, requires experience. In the next section, we will review some of these combination methods.

3. How can theory predict structure and dynamics?

As presented in the last sections, several benchmark works have shown that force fields may not be accurate enough to predict quantitative energy differences between peptide conformations in the gas phase. However, as also mentioned in the previous section, the high dimensionality of the potential-energy surface renders the direct exploration with first-principles potentials an elusive task. Therefore, theoretical studies that aim to explore the PES of larger polypeptides (and biomolecules in general) with first principles methods tend to follow an overall similar work flow [48, 69, 72, 98, 121–125].

The general aim is to balance a broad sampling of conformers and an accurate description of the energetics with the available computer power. We exemplify this work flow in general below, illustrating it by the technique followed in [98], which we believe to be among the largest current computational efforts to study the conformational space of alanine based polypeptides from first principles. The work flow is also schematically represented in figure 9.

The **first step** involves a thorough enumeration of different conformers using a force field. These conformers are commonly local minima in the force field, found by different sampling techniques, like basin hopping, replica exchange, genetic algorithms, or any other sampling method. The idea is to perform a global and thorough exploration of structure space. For example in [98] replica-exchange molecular dynamics (REMD) simulations were performed with the OPLS-aa force field [91, 126] with 16 replicas for a total of 500 ns per replica. From these simulations conformations at each 2 ps were considered to generate an overall set of conformations. The less reliable the PES is at this step, the more conformers will have to be considered in the second step.

The **second step** is choosing which conformers from the force field sampling will be considered for the treatment with higher level methods (e.g. density-functional theory or other quantum chemistry methods). The conformers can be ranked by energy from lowest to highest. As described above, there can be large possible errors related to the force fields. The discrepancy between empirical and first principles descriptions is highlighted, for example, in figure 7(A). Many conformers (hundreds to thousands, depending on the system's characteristics) should be considered, otherwise low-energy conformers may be completely missed. Alternatively, conformers can be sorted by structural criteria in order to generate a pool of candidate structures that is as diverse as possible for investigation. Examples are clustering algorithms based on the root-mean-square deviation (RMSD) of Cartesian coordinates (e.g. in [98]) or sorting of structures according to hydrogen-bonding patterns (e.g. in [42, 101]). Other descriptors for structural similarity can, for example, be found by using machine learning methods similar to the ones presented in [127–129]. The chosen conformers are typically fully optimized with higher-level methods. Especially the local geometry optimization of force field minima with first-principles methods can involve large conformational changes that may lead to new local minima, which are not present in the force field. In [48, 69, 72, 98, 122, 123] we could highlight the importance of considering a large pool of conformers: Considering only a couple of tens of conformers would have led to missing many of the relevant structures discussed in these papers. The discrepancy in relative energies from FF and DFT illustrated in figure 7(A) also raises the question if all relevant local minima can be located from simply re-relaxing the force-field conformers. As a means of ameliorating the situation, it is possible to introduce a **third step** a local first-principles sampling. In [98], for example, 16 × 20 ps *ab initio* REMD simulations were performed and the most stable conformer (C2) of the study was only found in this refinement step.

After that step, one can continue increasing the accuracy for a subset of the conformers from the previous step. The conformers can again be clustered and a new smaller set can be chosen according to the same criteria as in the first step or others. The accuracy can be increased either by increasing numerical settings of the calculations (basis sets, grids, etc) or by going to even higher level theoretical methods. In [98] both were done, by going to a higher numerical accuracy as well as using computationally more expensive (and often more accurate) hybrid DFT functionals, and many-body van der Waals dispersion corrections [88]. Other works have also used MP2 and CCSD(T) methods for smaller systems in this step [70].

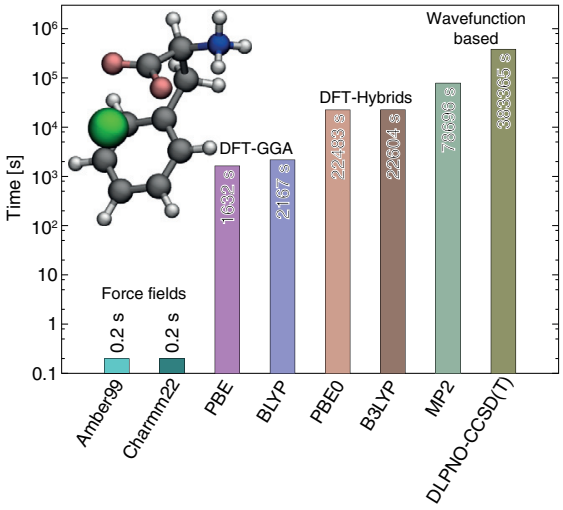


Figure 10. Timings for typical single point calculations of conformers of phenylalanine with Zn^{2+}. Standard protein force fields (Amber 99 and Charmm22) were computed with Tinker [114]. DFT calculations in the generalized gradient approximation (PBE and BLYP) and with hybrid functionals (PBE0 and B3LYP) were done with FHI-aims [79] (including pairwise Tkatchenko-Scheffler van der Waals correction and *really_tight* computational settings). Wavefunction-based calculations (MP2 and DLPNO-CCSD(T) [75]) were performed with the Orca code [130] using Ahlrich's basis sets for a 3–4 extrapolation to the complete-basis-set limit. The timings for the DFT calculations include force evaluations. The timings for the wavefunction calculations include both steps, the triple- and quadruple-ζ calculations. If the calculations were running in parallel (DFT and wavefunction), the real timings were multiplied with the number of cores. Please note, the numbers are meant to give a rough qualitative idea about the range of timings that can be expected with different methods. Different codes, settings, systems, and computer infrastructures will result in quantitatively different timings.

In order to exemplify the range of computational costs of different methods, we present in figure 10 timings that were measured for a comparably small system, namely phenylalanine with a Ca^{2+} cation. Please note that the accuracy level of the DFT (*really_tight* settings mean a very large basis and very fine integration grid) and the wavefunction calculations (with 3–4 extrapolation to the complete basis-set limit) are chosen rather high compared to what one would perform as standard calculation. The specific timings for each method can vary considerably when using different (smaller or larger) basis sets, when using different codes, or when treating larger and denser systems. The nominal scaling with system size N is for DFT N^3, for MP2 N^5, and for CCSD(T) N^7. In all cases however, developments are ongoing to reduce the respective scaling by the use of smart algorithms [75, 78, 131]. Nevertheless, the timings presented in figure 10 are good guidelines for what to expect in computational cost when increasing accuracy.

Having finished with a smaller subset of the most-likely structure candidates, it is desirable to connect to more physical quantities than a simple scan of the potential energy surface. Free energies and thermodynamic properties at realistic/experimental conditions can be explored at this step, either by performing anharmonic free energy evaluations with a method of choice (steered dynamics, metadynamics, umbrella sampling, replica exchange, etc) or at least considering these contributions in the harmonic approximation. If the system is too large, again it becomes unfeasible to calculate more accurate anharmonic quantities with a higher-level electronic-structure method, such that the harmonic approximation remains as the last resort. Its predictive power, though, has to be critically assessed for these soft and flexible systems.

With the low (free) energy conformers at hand, the connection to experiment can be established by computing experimentally accessible observables. In the present work we focus especially on collision cross sections that are experimentally derived from IM-MS (see section 1.2.1) and on vibrational spectra (see section 1.2.2). Other possible quantities of interest are electronic spectra, neutron scattering data, or any other experimental technique that is the most applicable to the environment where the biomolecule is measured in experiment.

Another important application of first-principles based conformational searches are studies that compare properties across chemical space. An example is the search for essentially all conformers of 20 proteinogenic amino acids alone and interacting with either of the cations Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, and Hg^{2+} [124]. As a result, one obtains comparable data for sets of compounds and/or complexes, generated on equal footing with respect to the search technique and the employed energy function. Based on such grounds, physical observables can be computed and compared across chemical compound space. The workflow employed by Ropo and coworkers [124] starts from a force field based structure search (Tinker scan [114] with the OPLS-AA force field [91]) and the relaxation with DFT-PBE + vdW. Again, it is necessary to refine the search results with a local first-principles search step. The bias from the initial treatment with empirical potentials can only be compensated by *ab initio* REMD simulations. The multi-step search procedure yielded an essentially unbiased first-principles data set of more than 45,000 stationary points on the PESs of the different molecular systems. The data can be used as a starting point for, e.g. the parameterization of empirical potentials, comparisons of properties like cation binding strength across chemical space, or as input for spectra calculations. The data is available from the website http://aminoacidb.rz-berlin.mpg.de and from the NoMaD repository [132].

4. Theory-experiment comparison—computation of experimentally accessible observables

A major challenge when performing simulations is to match the experimental conditions in a simulation setup. An effort on both ends is needed. Experimental conditions should be well controlled and the data recorded precise and sharp and the system size and character that is considered in the simulation should be as realistic as computationally feasible. The gas phase is an excellent environment in this respect, where it is possible to simulate physical observables on a very similar footing with experiments.

In the next section we focus on the calculation of collision cross sections and vibrational spectra. In addition, there are several optical spectroscopy techniques that can probe also electronic excitations and dynamics of excited states in the gas phase, connected to UV and visible probes. For example, in the UV-IR pump-probe experiments mentioned above, the UV laser induces electronic excitations that can be used to select different conformers. Reviews and perspectives of such optical spectroscopies in the gas phase, applied to peptides and other biomolecules can be found in [8, 9, 7, 133]. Antoine and Dugourd report the possibility of recording electron photo-detachment following electronic excitation in negatively charged peptides to obtain gas-phase optical spectra for large systems (even proteins), since this process does not suffer from limitations brought by energy redistribution into vibrational modes with system size and is less congested than a vibrational spectrum for large systems [133]. Theoretical modelling of electronic excited states and the resulting processes and dynamics is a major challenge, since it requires the use of time-dependent or explicitly correlated electronic structure techniques [134–136] that can treat excited states. These are very computationally expensive if compared to ground state techniques and have many further limitations included in the approximations, such that their application to large biomolecular systems is still limited, but growing fast.

4.1. Collision cross sections

From the Cartesian coordinates of conformers that result from a structure search for a particular molecular ion, it is possible to compute CCS values. The underlying collisions of the ion with the buffer-gas atoms (e.g. He) or molecules (e.g. N_2) can be modeled including different levels of detail. We will review here the three most-commonly used methods, the projection approximation [137], the exact hard-sphere scattering [138], and the trajectory method [139].

The projection approximation, or in short PA [137], takes the shape of the molecule into account, modelling the interaction between ion and buffer-gas particles by means of Lennard-Jones and charge-dipole interactions. The averaged collision cross section in the PA (CCS_{PA}) is calculated by using the collision parameters θ, ϕ, and γ as well as the minimal impact parameter b_{min} as follows:

$$CCS_{PA} = \frac{1}{4\pi^2}\int_0^{2\pi} d\theta \int_0^{\pi} d\phi \, \sin\phi \int_0^{2\pi} d\gamma \, \pi b_{min}^2 \qquad (2)$$

In practice, b_{min} is tabulated as atom-wise impact parameters, and in a simplified view they are stored as up-scaled atomic radii. The CCS value for a given molecular conformation is computed numerically by: (i) projecting the atoms of the molecule onto a randomly chosen plane, (ii) drawing the collision radii around positions of the nuclei, and (iii)

repeatedly selecting random points within an area A enclosing the projected molecule. Out of step (iii), a CCS value for a planar orientation N is computed following the formula $CCS_N = (h/t) * A$, where h is the number of hits within the projected outline of the molecule and t is the number of overall tries. Steps (i) to (iii) are repeated for different planes and an average CCS value out of CCS_N values is computed until convergence to a given threshold is reached. PA is shown to work well especially for largely convex molecules.

PA neglects scattering events as well as multiple collisions between buffer-gas particles and the ion. However, such effects are especially pronounced for concave molecular surfaces where certain surface areas can be shielded by parts of the molecule, while in others multiple collisions may occur. The projection-superposition approximation (PSA) aims to compensate for this with a shape factor that accounts for the concavity of a molecule [140]. Alternatively, scattering and multiple-collision effects can be considered by regarding ion and buffer-gas particles as hard-spheres. The exact hard-sphere scattering (EHSS) approach [138] explicitly follows the trajectory of a He atom that is shot at the molecule or cluster through all possible collisions until it leaves the molecule/cluster for good. Here, the scattering angle χ (the angle between the trajectories before and after a collision event between the molecular ion and a buffer-gas particle) is computed as a function of the collision parameters θ, ϕ, and γ and the impact parameter b for multiple collision geometries and thus an average CCS_{EHSS} can be obtained:

$$CCS_{EHSS} = \frac{1}{4\pi^2}\int_0^{2\pi} d\theta \int_0^{\pi} d\phi \, \sin\phi \int_0^{2\pi} d\gamma \times \int_0^{\infty} db \, 2b(1 - \cos\,[\chi(\theta,\phi,\gamma,b)]\,) \quad (3)$$

The trajectory method (TM) models one extra bit of the physics defining the drift of an ion through a buffer gas, namely long-range interactions between the drifting ion and the buffer gas. The importance of this contribution depends on the polarizability of the buffer gas, which is for example stronger in N_2 than in He, and on the charge distribution in the (molecular) ion. The charge(s) of the drifting ion induces dipoles in the buffer gas atoms altering its drift velocity without 'physical contact' [139].

$$CCS_{TM} = \frac{1}{4\pi^2}\int_0^{2\pi} d\theta \int_0^{\pi} d\phi \, \sin\phi \int_0^{2\pi} d\gamma \times \int_0^{\infty} db \, 2b(1 - \cos\,[\chi(\theta,\phi,\gamma,b)]\,) \times \left(\frac{\mu}{k_B T}\right)^3 \int_0^{\infty} dg e^{-\mu g^2/2k_B T} g^5 \quad (4)$$

In addition to the symbols explained above, the reduced mass μ and the relative velocity g are being used. The interaction between the ion and the buffer-gas particles is modeled by two terms: a Lennard–Jones 12–6 potential and a term that accounts for the interaction between the charge (distribution) of the ion and the charge-induced dipole of the buffer-gas particle. This treatment can consider differences in polarizability between buffer gases, for example between He and N_2.

Table 2. CCS values computed with PA or PSA and TM for different conformers/protomers of three molecules compared to the respective experiment-derived CCS.

Structure	$CCS_{PA/PSA}$ in Å^2	CCS_{TM} in Å^2	CCS_{Exp} in Å^2
Ac-Ala_6-Lys(H^+) from [123]			
α helix	180	181	180
Compact	172	171	
Ac-β^2hAla$_6$-Lys(H^+) from [123]			
H12	203	204	190
H16	191	193	
H20	182	182	
Compact	183	182	
Benzocaine from [62]			
O-prot./trans	131	133	135
O-prot./gauche	132	133	
N-prot./trans	133	144	155
N-prot./gauche	130	144	

We note, though, that in principle all methods are designed to work with He as the buffer gas. When comparing to measurements made with, for example, N_2, parameters going into the calculations have to be adapted. An overview about specific contributions to the collision cross section can be found in a paper by Wyttenbach *et al* [141] where for a wide range of systems experimental and PSA-simulated CCS are compared. There are several programs described in the literature, which can be more or less straightforward to obtain. We list here only some of the more popular ones:

MOBCAL is developed in the group of Jarrold and incorporates PA, EHSS, and TM. It can be downloaded at www.indiana.edu/nano/software.html.

sigma is developed in the group of Bowers and it computes CCSs according to the PA and EHSS method. It is available under this URL: bowers.chem.ucsb.edu/theory_analysis/cross-sections/sigma.shtml.

FHIsigma is a spin-off of `sigma` by Wesemann and von Helden and comes with a graphical user interface. The program is available at: sigma.fhi-berlin.mpg.de.

IMPACT is intended for structural proteomics applications and claims to compute extremely fast PA-CCSs [142]. The software is available at: benesch.chem.ox.ac.uk/resources.html.

The choice of method, for example between PA, PSA, EHSS, and TM, can be critical for the predictive power of the CCS calculation. Some examples are collected in table 2. Depending on the nature of the ionic cluster/complex or molecular ion under investigation, the alternative methods can agree, like in the case of two peptides from reference [123], where PA amd TM give virtually the same results. But there are also examples where the methods give qualitatively different results. Different protonation states (protomers) of benzocaine exist that result in either the distribution of the positive charge over the molecule or in its localization at a protonated amino function [62]. In the experiment, both forms can be separated with a polarizable buffer gas (N_2). In simulations, the CCSs computed with the PA are indistinguishable, while

TM predicts distinct values for the protomers and allows an interpretation of the experiment.

The interpretation of an experimental arrival-time distribution or of the derived CCS distribution is not unambiguous. The theoretical CCS of a single conformer represents a projection of the conformational degrees of freedom onto a single coordinate. As a consequence similar CCSs may still result from different structures. Also, in the experimental CCS, even a single sharp peak represents not only a projection of spatial coordinates, but also of the dynamics of the molecular or cluster ion over the drift time. Consequently, measuring a single sharp peak can mean that either (i) there is only a single conformational family present in the ion cloud, (ii) there are multiple (more than one) conformational families present in the ion cloud that have the same CCS, or even (iii) the time average over multiple interconverting conformers for a single molecule is converged during the drift time and the measured CCS basically represents a converged average over the CCSs of the different structures. An example was shown in [123], where IMS data of a β peptide is interpreted to represent the interconversion between related helix types. In a sense, ion-mobility experiments, especially in conjunction with molecular simulations, can be used to deduce not only the structure of molecules, but also their dynamics.

4.2. Vibrational spectra

As mentioned in section 1.2.2, several experiments probe the vibrational spectra of biomolecules in the gas phase. These spectra contain more detailed structural information than CCS experiments and simulations. However, especially for larger and more anharmonic systems, a comparison to theoretical simulations is necessary in order to interpret the experimental signal. Good reviews on several types of theoretical spectroscopy methods that can be used in connection to first principles potential-energy surfaces for biomolecules can be found, e.g. in [143, 144].

Theoretically, the 'zeroth-order' way to model the vibrational properties of any system is the harmonic (or double harmonic) approximation. In this approximation a Taylor expansion of the Born-Oppenheimer potential with respect to displacements of nuclear coordinates is truncated on the second (quadratic) order and harmonic frequencies of vibrations are calculated for the problem of coupled harmonic oscillators with force constants corresponding to the second derivative of the potential [145]. From Fermi's golden rule, it is known that the IR intensities are proportional to the square of the matrix elements of dipole-allowed transitions. One can thus Taylor expand the dipole moment with respect to nuclear displacements, solve the quantum mechanical Hamiltonian in the harmonic approximation, and find the allowed transitions. By truncating the expansion of the dipole moment at first order, one arrives at the expressions for the so called 'double harmonic' approximation. Not only this approximation does not contain any anharmonicities, it also does not allow any other transition beyond the fundamental ones. For Raman spectra, similar expressions can be calculated for the harmonic approximation relying on the estimation of matrix elements of allowed transitions from the polarizability tensor [145]. This type of approximation is frequently used for a first comparison of structural properties in connection with scaling factors that compensate for the complete lack of anharmonicities (both of the classical PES and connected to the quantum nature of the nuclei).

A fundamental problem with the harmonic approximation for the study of biomolecules is that these molecules can have very anharmonic potential-energy surfaces. A well known way to calculate IR transitions including anharmonicities is to relate Fermi's Golden Rule to time correlation functions—a derivation found in many textbooks (e.g. [146]). One finds that the IR absorption spectrum can be written as the product of the frequency-dependent refractive index $n(\omega)$ and the Beer-Lambert absorption coefficient $\alpha(\omega)$ as

$$n(\omega)\alpha(\omega) = \frac{\pi\omega(1-\mathrm{e}^{-\beta\hbar\omega})}{3cV\hbar\epsilon_0} I_{\mu\mu}(\omega), \tag{5}$$

where β is the inverse temperature, V the volume, ϵ_0 the dielectric permittivity of vacuum, c the speed of light and $I_{\mu\mu}(\omega)$ is the Fourier transform of the dipole auto-correlation function, here defined in the canonical ensemble $C_{\mu\mu}(t) = \mathrm{Tr}[\mathrm{e}^{-\beta H}\boldsymbol{\mu}(0)\boldsymbol{\mu}(t)]/Z$, where the partition function $Z = \mathrm{Tr}[\mathrm{e}^{-\beta H}]$ and $\boldsymbol{\mu}(t) = \mathrm{e}^{\mathrm{i}Ht/\hbar}\boldsymbol{\mu}\mathrm{e}^{\mathrm{i}Ht/\hbar}$. Since the correlation functions are usually approximated by classical (nuclei) or semi-classical dynamics, the correlation function that is in fact better approximated is the Kubo-transformed one, defined as $\tilde{C}_{\mu\mu}(t)$:

$$\tilde{C}_{\mu\mu}(t) = \frac{1}{\beta}\int_0^\beta C^\lambda_{\mu\mu}(t)\mathrm{d}\lambda, \tag{6}$$

$$C^\lambda_{\mu\mu}(t) = \mathrm{Tr}\left[\mathrm{e}^{-(\beta-\lambda)H}\boldsymbol{\mu}\mathrm{e}^{-\lambda H}\boldsymbol{\mu}(t)\right]/Z. \tag{7}$$

The Kubo transformed correlation has the same symmetries as a classical correlation function [147] and arises naturally in several approximate quantum dynamics schemes [147, 148]. The Fourier transform of the Kubo transformed time correlation $\tilde{I}_{\mu\mu}(\omega)$ and the one of the canonical time correlation $I_{\mu\mu}(\omega)$ are related by

$$I_{\mu\mu}(\omega) = \frac{\beta\hbar\omega}{1-\mathrm{e}^{-\beta\hbar\omega}}\tilde{I}_{\mu\mu}(\omega). \tag{8}$$

Thus, the commonly coined 'quantum correction factor' [37, 149] arises naturally from the relationship of these two correlations. The expression that one usually calculates for IR absorption is

$$n(\omega)\alpha(\omega) = \frac{\pi\beta\omega^2}{3cV\epsilon_0}\tilde{I}_{\mu\mu}(\omega) = \frac{\pi\beta\omega^2}{3cV\epsilon_0}\int \mathrm{d}t\mathrm{e}^{-\mathrm{i}\omega t}\langle\boldsymbol{\mu}(0)\boldsymbol{\mu}(t)\rangle \tag{9}$$

where the brackets denote a time average, and $\boldsymbol{\mu}(t)$ is generated by classical or approximate quantum dynamics for the nuclei. Similar expressions for Raman spectra can be found with respect to the autocorrelation functions of the polarizability tensor [150]. When classical dynamics (e.g. Born-Oppenheimer *ab initio* molecular dynamics) is employed to approximate these autocorrelation functions only the anharmonicities of the underlying

(classical) potential-energy surface are taken into account. The remaining discrepancies when comparing to benchmark experiments can be due to the lack of considering the quantum nature of the nuclei (which introduces what is sometimes referred to as quantum anharmonicities), the use of an approximate potential-energy surface, or sampling of the wrong (ensemble of) conformers—all of which can cause the spectra to change considerably, as discussed in more detail below.

Other techniques to obtain anharmonic vibrational spectra are, e.g. vibrational self consistent field (VSCF) and second order vibrational perturbation theory (VPT2). These methods and their applications to biomolecules have been reviewd by Roy and Gerber [151], and Barone and coworkers [152] recently. In both of them the quantum nuclear Hamiltonian is approximately solved either in a mean field approximation or a perturbation theory one, thus including quantum anharmonicities. However, the inclusion of temperature and explicit dynamics (where many conformations may be sampled) is not straightforward [153, 154], and the methods are expensive to treat very large molecules. An impressive recent work from a computational point of view was the application of VSCF-PT2 with the B3LYP functional to the spectra of two conformers of Gramicidin S, comparing to cold gas-phase IR-UV double resonant spectra, obtaining satisfactory agreement [155].

Even though the evaluation of IR and other vibrational spectra from autocorrelation functions has been popular for decades especially for condensed phase systems and empirical potentials, Gaigeot and coworkers have pioneered its use in connection to first-principles (DFT) potential energy surfaces and applying it to isolated and solvated small polypeptides [24, 156–158]. It is remarkable how well the simulated spectra based on a linear absorption regime (see equation (9)) agree with those measured with the IRMPD technique. Great examples are spectra for Ala_2H^+, Ala_3H^+ that were derived from *ab initio* molecular dynamics simulations employing the BLYP functional [159, 160]. The authors observe that at room temperature the peptides interconvert between a few different structures and that these dynamics are important for the comparison with the IRMPD spectra. This type of studies serves also as an indirect probe of the dynamics. They also reported sensitivity to different conformations in the amide III regions for polyalanine peptides [24], and good structure selectivity and comparison to IR-UV IRMPD spectra in the far-infrared region for Ac-Phe-Gly-NH_2 and Ac-Phe-Ala-NH_2 [43]. This is very interesting, since vibrations in this lower wavenumber region are more classical in nature and can be more accurately represented by classical (*ab initio*) molecular dynamics, not requiring simulation techniques that incorporate quantum effects of the nuclei.

As an illustration of their work about the importance of anharmonicities in comparison to experiments, we highlight a larger peptide, Ala_7H^+, for which IRMPD spectra were measured by Vaden and coworkers [58]. In that study, Vaden and coworkers also performed extensive structural searches starting with a force field, then passing through a cascade of more accurate (standard) DFT functionals (until B3LYP), identifying conformational families, and finally performing single point calculations with MP2 for the energetically most favored conformers and calculating harmonic vibrations at the B3LYP level. The most likely globular structures, and the comparison of their harmonic IR spectra at the B3LYP level with the measured room temperature IRMPD spectrum is shown in figures 11(A) and (B) (reproduced from [58]). Gaigeot and coworkers then took these structures and calculated IR spectra from *ab initio* molecular dynamics at the BLYP and level and $T = 350$ K in [161]. The comparison between this anharmonic spectrum and the same experiment is shown in figure 11(C), reproduced from [161]. It is immediately apparent that even if the agreement is not perfect, anharmonicities in this NH and CH stretch regions are necessary to reproduce the experimentally observed intensities below $\approx 3100\,cm^{-1}$. The authors conclude that these structures adopt more globular conformations with the NH_3^+ group self solvated within CO groups of the molecule. As will be shown below, the exact placement of the position of the simulated peaks with respect to experiment in the anharmonic case may be a fortuitous cancellation of errors, since the inclusion of van der Waals interactions can change considerably the dynamics of the molecule and inclusion of nuclear quantum effects cause large red shifts in this spectral region.

It is worth noting that intensities are typically not to be trusted when comparing theory and IRMPD experiments due to the strong non-linear effects expected in the multiple-photon abosption process. Attempts have been made by Calvo and coworkers to model specifically IRMPD [162] with all relevant dynamical effects, which can yield good results for small molecules albeit relying on some empirical modelling. Comparisons to IRPD would be interesting, since it is less prone to to non-linearity in the lineshape and peak positions. However, the tag which is often used can also disturb the spectrum (as observed in Kr tagged gold clusters [163] and Ar tagged protonated water clusters [164]), and one is usually restrained to low temperatures due to the low binding energy of the tag. In most of the work present in the literature so far, it must be said, though, that the modelling of the IR spectra within linear response theory (including anharmonicity) has been able to provide important interpretations to vibrational signatures obtained from IRPD or IRMPD.

Blum and coworkers (including the authors of this review) have focused on the study of larger polypeptides, especially in the fundamental characterization of interactions governing structure formation and dynamics. For the benchmark series of helix-forming alanine based polypeptides Ac-Ala_n-$LysH^+$ the authors have studied many different aspects related to secondary structure formation using DFT and *ab initio* molecular dynamics. Regarding the smaller members of this polypeptide series, $n = 4$–8, the authors have reported that beyond the formation of stable H-bond chains with increasing n, an important contribution to helix stabilization comes from the vibrational entropy of very soft modes that are present in the helices but not in more compact structures [48]. Helices are predicted to be the most stable isolated structures in the gas phase starting at $n = 8$, in agreement with experimental evidence from IMMS measurements [165].

For a more direct structural characterization, Rossi and coworkers have also calculated the (classical-nuclei) anharmonic IR spectra of $n = 5$, 10, and 15, and compared to

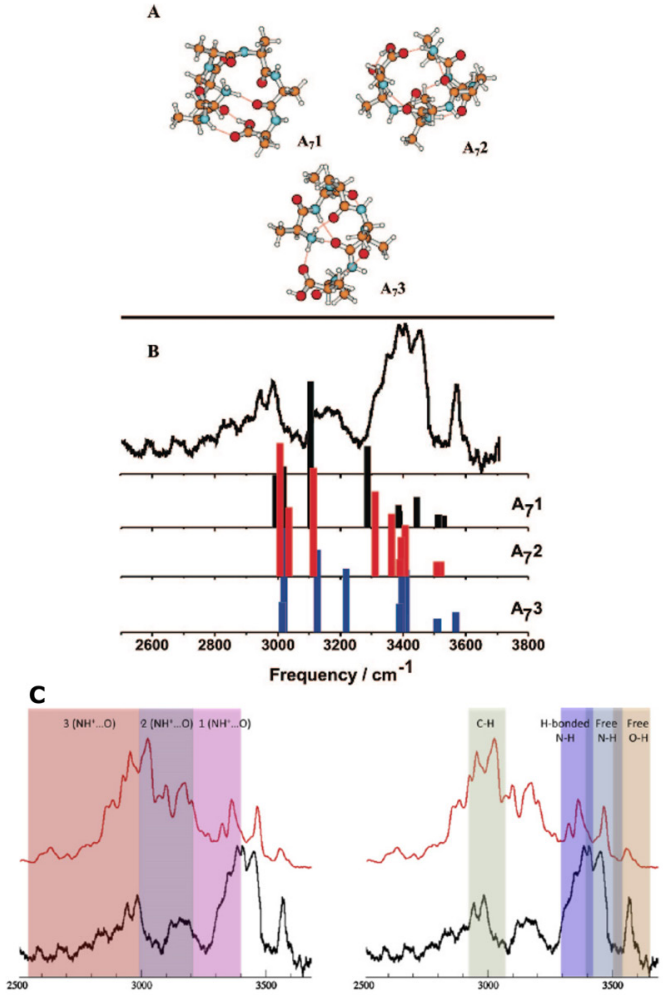


Figure 11. (A) and (B): Structures of Ala_7H^+ and their corresponding harmonic IR spectra with the B3LYP functional, compared to the measured IRMPD, reproduced from [58], copyright 2008 American Chemical Society. Anharmonic IR spectrum (classical nuclei) with the BLYP functional (red) for the same molecule, compared to the experimental IRMPD spectrum (black), reproduced from [161], copyright 2011 Elsevier.

experimental IRMPD measurements at room temperature [42]. In general, the structural characterization of gas-phase peptides based on vibrational spectra requires an objective metric of agreement between simulation and experiment. To that end, Rossi and coworkers have employed the Pendry reliability factor R_P [166] in an implementation that was distributed with [167]. Since, as it was already discussed, the IRMPD spectra could have peak intensities that are distorted due to the absorption of many photons, a simple overall least squares fit for the intensities would not suffice for a comparison between theory and experiment. The Pendry R-factor, originally used in low energy electron diffraction experiments [166], addresses the need to match mainly peak positions, rather than the intensities. Given two continuous curves with intensities $I_{exp}(\omega)$ and $I_{th}(\omega)$, this R-factor compares the renormalized logarithmic derivatives of the intensities, given by:

$$Y(\omega) = L^{-1}(\omega)/[L^{-2}(\omega) + W^2] \tag{10}$$

with $L(\omega) = I'(\omega)/I(\omega)$, and W approximately the half width of peaks in the spectra. The advantage is that the L functions have a sign inversion exactly where the maximum of the peak is, and if peaks are far enough apart, relative intensities are completely ignored, while if they are close together, $L(\omega)$ is moderately sensitive. However, the L functions would be too sensitive to zeroes in the intensity, since the logarithmic derivatives would have singularities in this case. The Y function is a simple transformation of L, which avoids such singularities, by giving similar weights to maxima and zeroes in the intensities. The Pendry R-factor (R_P) is then defined as:

$$R_P = \int d\omega (Y_{th} - Y_{exp})^2/(Y_{th}^2 + Y_{exp}^2), \tag{11}$$

which leads in practice to values of $R_P = 0$ for perfect agreement, $R_P = 1$ for uncorrelated spectra, and $R_P = 2$ for

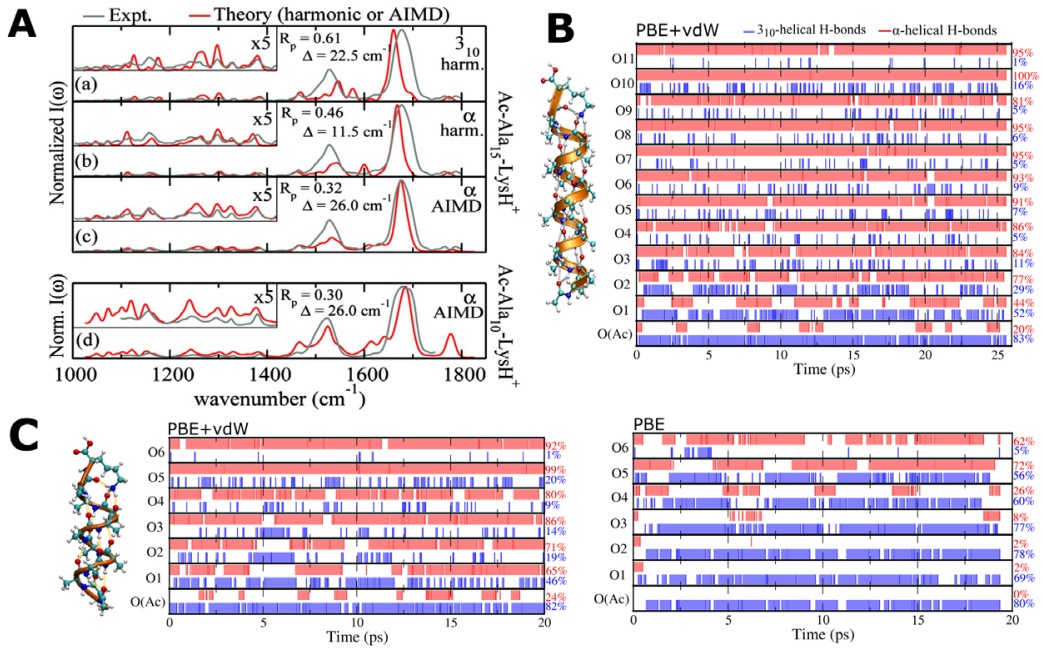


Figure 12. (A): Reproduced from [42], copyright 2010 American Chemical Society. Comparison between experimental (gray lines) and theoretical (red lines) (PBE + vdW functional) vibrational spectra, all normalized to 1 for the highest peak. ((a), (b)) Ac-Ala$_{15}$-LysH$^+$: calculated spectra based on the harmonic approximation, for a 3_{10}-helical (a) and an α-helical (b) local minimum of the potential-energy surface. (c) Ac-Ala15-LysH$^+$: calculated spectrum from AIMD (including anharmonic effects), starting from an α-helix and α-helical in character throughout the simulation. (d) Same as panel (c), for Ac-Ala10-LysH$^+$. Pendry R-factors and rigid shifts Δ between measured and calculated spectra are included in each graph (calculated spectra are shifted by Δ for visual comparison). (B): Illustration of the hydrogen bond network evolution of Ac-Ala$_{15}$-LysH$^+$ during a PBE + vdW microcanonical simulation. On the right side of the plot, the ratios of α-helical and 3_{10}-helical bonds observed during the simulation for each oxygen, labeled from N to C-terminus is shown. (C): Illustration of the hydrogen bond network evolution of Ac-Ala$_{10}$-LysH$^+$ during a PBE + vdW^{TS} and a PBE microcanonical simulation (labels are the same as in (B)).

complete anti-correlation. R_P is always defined with respect to a rigid shift Δ between the two curves considered. A python script for the calculation of this and other reliability factors is available from Github[3].

We reproduce in figure 12(A) the theoretical IR spectra obtained with DFT-PBE adding pairwise van der Waals corrections (PBE + vdW [87]) for helical structures of Ac-Ala$_{10}$-LysH$^+$ and Ac-Ala$_{15}$-LysH$^+$ compared to experiment. For $n = 15$ the comparison of the harmonic spectra of a helix containing mostly 3_{10} helical H-bonds, another containing α-helical H-bonds, and the anharmonic spectra obtained from equation (9) from PBE + vdW molecular dynamics shows (quantitatively) how the agreement to experiment increases in the anharmonic case. A Pendry reliability factor R_P of 0.32, obtained with respect to a rigid shift Δ of the whole spectrum by 26 cm^{-1} is an indication that the structure of this molecule is indeed the α-helical one shown in figure 12(B), where the lysine residue is completely self-solvated in the backbone carbonyl groups. Also in panel B, we show the H-bond dynamics of the molecule in the trajectory generating that spectrum, highlighting 3_{10}- and α-helical H bonds. Although fluctuations are observed, the molecule maintains a mostly α-helical structure throughout. For Ac-Ala$_{10}$-LysH$^+$ we also find a good agreement between the theoretical (anharmonic) and experimental IR spectrum for the α-helix. Examining the dynamics of this molecule when switching off the vdW interactions, we can show in panel (C) that the structure becomes more extended, stabilizing a 3_{10} helical motif, and worsening the agreement with experiment (shown only in [83]). This observation is also in line with a study of interplay between H-bond cooperativity and vdW contributions in polyalanine helices: H-bonds get systematically strengthened by vdW interactions, and the high temperature stability of Ac-Ala$_{15}$-LysH$^+$ is increased, while at lower temperatures the lack of vdW interactions also stabilize a more extended 3_{10}-helical structure [29].

The effect of the location of the charge and the peptide sequence was also studied for even larger alanine-based polypeptides, namely Ac-Ala$_{19}$-LysH$^+$ and Ac-Lys-Ala$_{19}$-H$^+$ [98]. Ac-Ala$_{19}$-LysH$^+$ was seen to form helices, consistent with measured ion mobility cross sections. Ac-Lys-Ala$_{19}$-H$^+$ presented cross sections consistent with more compact, globular conformers (as expected due to the unfavorable interaction of the charge with the possible helix macrodipole), but its IR spectrum was very similar to helical structures. Theoretical calculations could solve this puzzle: even if of a compact/globular nature, energetically favored conformers of Ac-Lys-Ala$_{19}$-H$^+$ still retained a large helical content.

[3] https://github.com/mahrossi/r-factors

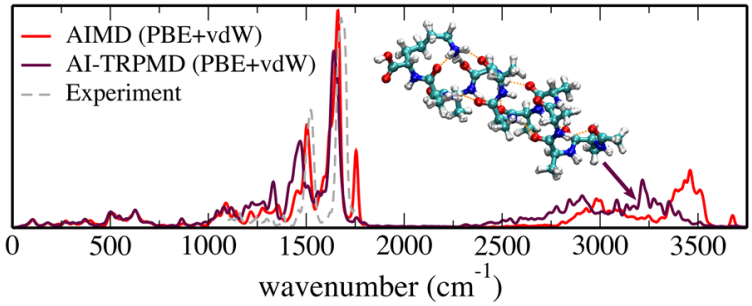


Figure 13. Infrared absorption spectra of Ac-Ala$_{10}$-LysH$^+$ calculated with *ab initio* molecular dynamics (AIMD-PBE + vdW) at 300 K, with *ab initio* thermostatted ring polymer molecular dynamics [168] (TRPMD-PBE + vdW) at 300 K, and the experimental IRMPD room-temperature spectrum from [42].

Here we take the opportunity to address a commonly adopted approximation in these simulations, namely that of performing dynamics considering classical nuclei. Hydrogen atoms, ubiquitous in these molecules, are quite quantum entities even at temperatures as high as room temperature. These effects are known to affect the structure and dynamics of condensed phase systems (especially water) [169, 170] and hydrogen bonds [171, 172]. A simulation technique that has been progressively gaining more attention to include nuclear quantum effects (NQE) beyond the harmonic approximation at least in non time-dependent observables is path integral molecular dynamics (PIMD). This technique exploits an exact isomorphism between the statistical properties of a quantum system and that of a classical ring polymer, where each bead is a repetition of the original system, connected to each other by harmonic springs. A detailed explanation of this technique is beyond the scope of this manuscript, but good descriptions can be found in [173, 174]. This technique is especially suited to massively parallel architectures, since the replicas of the system can be run in parallel given that there are enough CPUs available. For time-dependent observables, e.g. time correlation functions, the situation is much trickier, due to the difficulties of modelling true quantum dynamics. Also within path integral molecular dynamics there are a few approximations to time correlation functions that have been proposed, namely centroid molecular dynamics [175], ring polymer molecular dynamics [147], and thermostatted ring polymer molecular dynamics (TRPMD) [168]. Albeit approximate these methods can give reliable results especially for larger systems and/or extended systems [176], and are the only methods so far that can be applied on a more routine basis to realistic multidimensional systems. At room temperature, even for the most efficient of these methods, one must use several tens of replicas of the system, making these simulations still substantially more costly than their classical-nuclei counterparts.

We used TRPMD to calculate the IR spectrum of Ac-Ala$_{10}$LysH$^+$, shown in figure 13. We used the FHI-aims program package [79] in connection to the i-PI program [177] in order to perform the dynamics. We simulated 20 ps of TRPMD dynamics, starting from the thermalized α-helical structure, a time step of 0.5 fs for the integration, 16 replicas of the system (beads), and light settings in FHI-aims for the PBE + vdW force evaluation. In figure 13 we compare the IR spectrum thus obtained with the AIMD-PBE + vdW spectrum (tight settings, without any shifts applied) and the IRMPD room temperature experimental spectrum already published in [42]. We observe that while for very low frequency modes the classical and quantum nuclei simulations agree pretty well, above 1000 cm^{-1} most of the modes are softened (red-shifted) in the quantum case, something that becomes progressively more pronounced for all modes above 2500 cm^{-1}. This observation is in line with the fact that higher frequency modes are more quantum in nature. Even if TRPMD is known to over-broaden the line-shapes [168], the red-shifts should be reliable, modulo the limitations of the DFT functional itself (lower barriers, softer H-bonds). As also shown in figure 13, this effect goes in the opposite direction of the experimental data, which is already slightly blue shifted from the classical nuclei simulation. This is an indication that the PBE + vdW functional itself is here at fault. In these systems, when calculating harmonic frequencies of vibration with, e.g. the PBE0 + vdW functional, they are all blue shifted with respect to PBE + vdW. The over-softening of the modes is one more manifestation of the self-interaction problem. It seems, thus, that in order to get better agreement of peak positions with experiment in a fully anharmonic picture, one should perform a simulation with van der Waals corrected hybrid functionals (which are, unfortunately, considerably more expensive than standard generalized gradient ones) and include nuclear quantum effects.

So far, only studies of polypeptides in isolation have been discussed. As mentioned in the introduction, the gas phase is ideal not only due to its 'clean room' conditions, but also to the fact that it is straightforward to control the gradual inclusion of 'external agents', as for example ions, metal cations and small metallic clusters, and solvent molecules, for example water. We dedicate a following section to the discussion of microsolvation. Here we briefly review the interaction with ions. Since the early 2000s, IMS experiments have pointed to the role of cations stabilizing helical structures in polyalanine peptides [178], and more recently evidence for helix stabilization has been established based on the measurement of gas phase IRMPD spectra in the Amide A/B range of sodiated polyalanine peptides of various sizes [179]. Through

measurement of IR spectra, also the role of metal cations to stabilize the zwitterionic form of some amino acids in the gas phase has been studied [18].

We had a detailed look at the effect of small cations (Li^+ and Na^+) on the structure of prototypical turn-forming peptides Ac-Ala-Ala-Pro-Ala-NMe and Ac-Ala-Asp-Pro-Ala-NMe [72]. The different systems were investigated by means of theoretical and experimental vibrational spectroscopy. First of all it was evident that in the gas phase, the interaction of the peptide carbonyl groups with the strong positive charge of the cations enforces conformations on the backbone that would not be possible for the peptide alone. Furthermore, the preferred conformations differ depending on the cation. The comparison between experimental and simulated spectra revealed that multiple conformers co-exist and probably interconvert in the gas phase. Consequently, the computed spectra for individual conformers have to be mixed in order to match the spectra recorded in the experiments, but a good agreement is reached. One can raise the question of how relevant are these results in solution. Hints come from short *ab initio* MD simulations that were performed on energetically stable conformations of peptide-cation systems with a few dozens of waters. Within the time scales accessible, the interactions between the cation and the peptide backbone remained preferred over direct solvation of the cation by the water molecules.

4.3. Towards first-principles free energies

Even if the PES is really the basis for all thermodynamic quantities, the sole knowledge of the PES does not allow a direct connection with real-world physics. For equilibrium properties, what is really needed is a good estimate of the partition function from statistical mechanics and all thermodynamic quantities that can be derived from it, most importantly, free energies.

Unfortunately, estimating free energy values for biomolecules is not an easy task. The harmonic approximation for the free energy (discussed in many textbooks, e.g. [146]), is the most common approximation. The reason is that it is the only one feasible with more costly (e.g. first principles) potentials and for larger molecular sizes. Due to the anharmonic nature of these molecules, it is not guaranteed though that this approximation will be plausible even at relatively low temperatures.

In order to get vibrational contributions to the free energy it is possible to use, for example, the VSCF and VPT2 methods, already discussed in the last section. For small molecules, Basire and coworkers have developed a technique which relies on the estimation of microcanonical densities of states and partition functions, that gives access to temperature effects and relative populations connected to a second order vibrational perturbation theory [153, 154] approach. However for higher dimensional and flexible systems this technique becomes very challenging. Quasi-harmonic analysis, in which dynamics can be decomposed into principal components and entropies calculated from this decomposition can be used as an approximation, provided there is enough sampling, but again, they rely on a quasi-harmonic picture that is likely to fail in many situations.

We have shown in the previous sections that it is possible to extract, for example, vibrational spectra from first-principles molecular dynamics (MD) simulations. However, the estimation of (relative) free energies requires a sampling of the conformational space that can currently only be realized for rather small molecular systems with few well defined degrees of freedom [180]. For larger systems, with hundreds of atoms, it is a much larger (and close to impossible) effort to gather the required statistical sampling of conformational space in order to estimate these free energies. It is worth noting though that with smart algorithms and optimized codes these quantities are becoming accessible [181]. There are two main points do be addressed [182]: (i) The simulation has to be long enough to ensure that the time-average of the simulations resembles the ensemble average of the system and (ii) free energies from MD simulations require the definition of collective variables, that are not trivial to define. In the field of biomolecular simulations a variety of MD-based simulation techniques are being used to solve point (i), we only summarize some frequently used types here:

- A straightforward approach is the computation of long (μs to ms time-scale) trajectories. This idea brought to the extreme is the construction of dedicated hardware like the molecular-dynamics supercomputer Anton [183] that provides access to the kinetics and thermodynamics of, for example, protein folding [184, 185].
- Alternatively, many short MD trajectories can be combined by using Markov-chain models [186–190]. This approach is striking because it is inherently parallel and allows the use of distributed computational resources [191, 192].
- The necessity of either very long or large numbers of independent shorter MD simulations comes from the nature of the transitions between the different meta-stable states on the free-energy landscape of a given system. These transition are often rare events and in order to obtain converged values, these events have to be observed sufficiently often. In order to enhance sampling and therewith shorten the required simulation times, multiple methods are available: replica-exchange MD, umbrella sampling [193–196], metadynamics [197, 198], etc.

One or several collective variables are needed as degrees of freedom (DOF) that define the free-energy surface. In case of, e.g. umbrella sampling or meta-dynamics, these collective variables have to be known *a priori*, while they can be defined *a posteriori* in non-biased MD simulations. Overall, it would be interesting to pursue methods that can be even more efficient in sampling, or methods that can reach convergence with a small amount of statistics.

5. Challenges towards solvation

A biomolecule immersed in a solvent presents three different qualitative types of interactions that need to be described. These are the intramolecular interactions, the biomolecule-solvent

interactions, and the intra- and inter-molecular interactions of the solvent. The interactions between the biomolecule and the solvent and the influence of the collective interactions of the solvent on the biomolecule are the ones that will ultimately define the solvated state. It is important to note that the solvent is often not a simple homogeneous environment, but includes ions and other inhomogeneities that also need to be accurately captured. Studying biomolecules directly in solution has the drawback that the resulting measurements are quite congested by the amount of different interactions that play a role. It is thus desirable to build up the solvated state step by step, so that theory and experiment can work in synergy towards a consistent and reliable description of these molecules in solution.

Experimentally, regarding the solute–solvent and solvent–solvent interactions, perhaps the most detailed characterizations of physical properties are connected to mass spectrometry (MS), where it is possible perform thermochemical equilibrium measurements [199] and, if connected to spectroscopy techniques, to measure also more detailed structural information. In these experiments, solvation with water molecules or ions (or both) can be investigated in a stepwise manner, such that the physical properties of the very first stages of solvation can be identified. For example, it is possible to measure equilibrium constants, binding enthalpies, and vibrational spectra that can be directly connected to calculations.

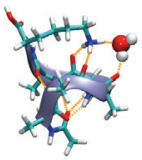
Using only IM-MS, thermochemical equilibrium properties and overall geometric information have been gathered for a range of biomolecules and the first stages of their interaction with the solvent (microsolvation) [3, 178, 200–203]. A review in this area can be found in [3]. More recently, also the measurement of vibrational spectra of mass selected species in the gas phase were able to probe more detailed conformational properties of clusters of solvent molecules [164, 204–207, 208] or the first stages solvation especially of peptides [209], and sugars [6, 210, 211]. We highlight here two recent experimental works dealing with peptides to illustrate the state of the field. Impressive results have been reported by Nagornova and coworkers [209] on the microsolvation of Gramicidin S cooled to 12 K. By performing conformer selective double resonance IR-UV spectroscopy they were able to connect IR features to structural changes caused by the absorption of 1–15 water molecules. Another work by Warnke and coworkers [200] instead used ion mobility-mass spectrometry to show how crown-ethers can micro-solvate charged Lys side chains in cytochrome-C and other proteins. The authors were able to decompose the effects responsible for the unfolding of highly-charged states in the gas phase into Coulomb repulsion and side chain to backbone interactions that interrupt backbone hydrogen bonding.

Experiments nowadays are able to provide more and more accurate data on thermochemical and structural properties of (micro)solvated biomolecules, but without the support of theoretical calculations, the understanding of the results is limited. It is not straightforward to obtain quantitative data for these systems from simulations, though. The difficulties are at least two fold: (i) One still has the high conformational freedom of the biomolecule itself, but now further complicated by the presence of ions and solvent which introduce an extra range of qualitatively different interactions to be modeled; (ii) It is known to be difficult to simulate even the solvent in isolation, with most quantum chemical methods failing to correctly describe overall structural properties like radial distribution functions, or diffusion coefficients [212–218], or the correct relative energies of hydrogen bonded structures [219, 220].

The main challenge is to correctly and thoroughly explore the potential energy surface (PES) *and* the entropic contributions to the free energy—even more important when related to the solvent. These simulations must involve an accurate evaluation of the potential energy and span a long time scale (or a huge volume of phase space). Unfortunately nowadays one can have either one or the other: an accurate evaluation of points in the PES can be achieved by the highest-level quantum chemistry methods but these are too computationally expensive to allow a thorough sampling of the PES, while empirical potentials allow a thorough sampling of the PES but do not provide quantitative estimates. It is also important to note that only describing the electronic structure of these systems is not enough—especially in connection with the solvent, the inclusion of nuclear quantum effects beyond the harmonic approximation is necessary [170, 221–225].

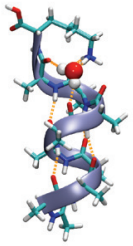
Nevertheless some successes from theory have been achieved for the microsolvation of model peptides, for example the Ac-Ala$_n$-LysH$^+$ series already mentioned in this review. The groups of Bowers [203] as well as of Jarrold [165] performed IM-MS experiments for the monohydrated structures of a few conformers (different sizes) of this peptide series. In these experiments, they had access to equilibrium constants of the monohydration reaction, derived from the ratio between the intensity of the peaks corresponding to the bare and the monohydrated structures. Based on previous observations that more globular/compact structures had a lower propensity to adsorb one water molecule than helical ones, they concluded that the shortest helical member of this series would happen at $n = 8$—without thorough theoretical support, it is difficult though to understand what is the atomistic mechanism for this difference in water adsorption propensity. In [122], Chutia and coworkers have performed extensive first principles conformational scans of $n = 5$ and $n = 8$ microsolvated by up to 5 water molecules. One conclusion is that the intramolecular hydrogen bonds of the self-solvated ammonium group, in both cases, are the most stable hydration sites. For one water molecule the most stable conformers are shown in figure 14, together with the calculated standard (Gibbs free) energy of formation ΔG_0 of the reaction. The agreement with experimental values is pretty good (also at other temperatures, shown in [122]). From the theoretical work, the authors concluded that the decrease in water adsorption propensity is not due to a radically different binding site, but instead only to modified internal free energy contributions (harmonic vibrational free energy) in the specific H_2O adsorption site at the LysH$^+$ termination, in an example of how theory can help to gather a deeper understanding of experimental data. However, it is still a challenge for theory to be able to give even more reliable results for larger peptides surrounded by more solvent molecules. In this respect, theoretical advances as proposed by Gaigeot *et al* [226] that allow a separation of solute and solvent vibrational spectra in simulations are of great importance.

Ac-Ala$_5$-LysH$^+$ • H$_2$O



Monohydrated peptide	Method	ΔG_0(eV), T=0K	ΔG_0(eV), T=223K
Ac-Ala$_5$-LysH$^+$	Theory/PBE+vdW	-0.53	-0.24
Ac-Ala$_8$-LysH$^+$	Theory/PBE+vdW	-0.51	-0.20
Ac-Ala$_5$-LysH$^+$	Expt.[a]	-	-0.20 ± 0.02
Ac-Ala$_8$-LysH$^+$	Expt.[a]	-	-0.15

Ac-Ala$_8$-LysH$^+$ • H$_2$O



[a] *Kohtani and Jarrold, JACS 126, 8454 (2004), values converted from K_1 equilibrium constants that were read from Figure 2. Error bars from our estimate.*

Figure 14. Calculated $\Delta G_0(T)$ (in eV, and corresponding to a reference pressure of $p_0 = 1.01325 \times 10^5$ Pa = 760 Torr) for monohydration of Ac-Ala$_5$-LysH$^+$ and Ac-Ala$_8$-LysH$^+$ compared to literature data. Also shown, the most stable conformations of monohydrated Ac-Ala$_5$-LysH$^+$ and Ac-Ala$_8$-LysH$^+$ from theory (PBE + vdW). Values and structures from [122].

It is thus pressing to build a tighter relationship between the quantum and the empirical world. While for water there is an appreciable effort to build better and more accurate potentials based on quantum mechanical calculations [221, 227–229], for the solvent-biomolecule (or ion-biomolecule) interaction these efforts are much less pronounced. An improvement in this area can be achieved precisely by performing these theory-experiment benchmarks of the stepwise build-up of solvation, and modifying empirical potentials according to this data.

6. Conclusions

The aim of this review was to give an overview on the interplay of experiment and simulation regarding the structure and dynamics of biomolecules in the gas phase. Given the scientific fields of the authors, the focus was clearly on first-principles calculations on peptides towards the computation of physical observables like vibrational spectra and collision cross sections. For flexible molecular systems, for which biomolecules are a prime example, a thorough search of the accessible conformational space is crucial before any attempt to compare simulated properties with their experimental counterpart.

A typical work flow is outlined in the following (and in figure 9):

(i) The exact chemical structure (connectivity of the atoms) of the molecular system has to be known. This includes knowledge about possible alternative protonation states (protomers). In cases where, for example, cations like H$^+$ or Na$^+$ are involved, their presence and location relative to the molecule has to be considered as well.

(ii) An initial enumeration of structural candidates can be performed by the sampling of a computationally-cheap potential-energy surface (PES), for example of an empirical force field.

(iii) As we have outlined in this review, the limited accuracy of force-field methods requires a refinement at the level of electronic-structure theory. This can be facilitated by using density-functional theory (DFT) methods or quantum-chemistry methods like Møller–Plesset perturbation theory (MP2). Higher-level methods, like coupled cluster, quantum Monte Carlo, or full configuration interaction, are computationally very demanding and thus normally limited to small systems and benchmark-type calculations.

(iv) In order to remove a possible bias from the initial sampling of the force-field based PES, further exploration of the first-principles PES in the proximity of already located low-energy structures is advisable. This can be facilitated by, for example, (replica-exchange) *ab initio* molecular dynamics simulations.

(v) Free-energy estimations in the harmonic approximation should be considered, not the least because they also offer a first glimpse at the vibrational spectrum of the molecular system. Further MD-based sampling can potentially be used to obtain more accurate thermodynamical observables (free energies, enthalpies, etc). However, the size of structure space and the computational cost of the required converged simulations again restrict such approaches to either rather small or rigid molecular systems.

(vi) The comparison to experiment serves as (i) validation of the method (search strategy and energy function) and (ii) as a way to add structural resolution to the experiment. Both can be achieved by the computation of physical observables, e.g. collision cross sections, vibrational spectra, optical spectra, etc.

Each simulation represents an approximation to reality and inherently produces errors. The gas phase is a clean-room environment and gas-phase experiments can produce accurate and sharp data that represents a challenge to theory and simulation. We would dare to say that the higher signal-to-noise ratio that is present in condensed-phase experiments might actually cover some of the involved systematic errors in the theoretical description. This highlights the importance of the gas phase as an ideal environment for validating energy functions and simulation techniques.

An important point that we can conclude is that it is not sufficient to focus on a single or a few structures, given the complex dynamics observed in the gas phase (and even more

so in solution). Most of the larger sources of uncertainties in the theoretical treatment have to do with an insufficient or still inaccurate treatment of dynamics. If an accurate free-energy surface could be accessed and sampled, most of the remaining problems would be solved. This would allow, for example, the correct prediction of the conformational ensembles observed in ion-mobility measurements (CCS/ATD) or in vibrational spectroscopy. In addition, it would give access to reliable barriers and a natural inclusion of anharmonic effects in vibrational spectra. In order to reach this goal, we need to compute potential energies and forces including the correct physics, which then need to be sampled faster and for long time scales. We note that the correct physics may go even beyond just grasping the physics of the electronic structure but also the quantum nature of the nuclei, which can cause much stronger anharmonicities (as shown in this review) and change considerably effective barrier heights. Going even further, for these highly anharmonic and high-dimensional systems, in many situations the dynamics of nuclei and electrons are coupled. These non-adiabatic effects are truly difficult to treat from a theoretical point of view in these structures.

The efficient exploration of conformational space for high-dimensional flexible systems in an accurate manner thus poses one of the most pressing issues in this field. For it to be solved, either the accuracy of force fields must be improved, or the computational limitations of first-principles methods, when it comes to larger length scales and longer time scales, needs to be lifted. Possible routes that can be followed in methodological developments involve, for example, better parametrization of force fields based on the increasing number of first-principles data present in the literature, development of smarter free energy evaluation methods that can deal with fewer statistical sampling, and/or even better scaling of first-principles codes in massively parallel architectures. As these issues are already recognized by the community, several efforts in all fronts are paving the way to treat larger systems with state-of-the-art accuracy (e.g. [75, 98, 183, 230–234] and many others).

Nevertheless, as it has been shown in this review, both the time and length scale currently accessible to first-principles methods already allow an accurate treatment of systems with hundreds of atoms in simulations. On the experimental side, it is routinely possible to transfer large biomolecules, e.g. large proteins and even complexes, to the gas phase by electrospray ionization and to study them by mass spectrometry and ion mobility-mass spectrometry (IM-MS) [11]. However, with the size of the molecular systems, vibrational spectroscopy investigations get hindered by more and more congested spectra. A promising route that is currently being followed to circumvent this problem is to measure conformer selective spectra by either using (i) UV/IR double-resonance techniques and (ii) pre-selecting conformers by using IM-MS. A way to get sharper spectra is to measure them at low temperatures for example by using either cold-ion traps [63, 64] or helium droplets [65, 66]. Conformational selection and cold-ion spectroscopy can also be combined.

The investigation of biomolecules in the gas phase is a dynamically growing field and a constant challenge to experimentalists and theorists alike. The constant developments and improvements of experimental techniques trigger the use of more and more sophisticated simulations and *vice versa*. As such this line of research pushes our understanding of the very basics of biomolecular structure formation and dynamics. For the development of simulation methods, the precise data that can be obtained from gas-phase experiments is ideal to develop and test new methodologies that will also have an impact in condensed-phase simulation.

Acknowledgments

We are grateful to our current and previous colleagues in the 'Bio Group': S Chutia, M Ropo, J Wieferink, F Schubert, M Schneider, M Marianski, A Supady and the previous group leader V Blum. We acknowledge the continuous support of our work by M Scheffler and the inspiring discussions on the experimental side, especially G von Helden, K Pagel, L Voronina, T Rizzo, and K Asmis. MR acknowledges funding from the German Research Foundation (DFG) under project RO 4637/1-1 and office space in the group of Prof D Manolopoulos.

References

[1] Jarrold M 2000 *Annu. Rev. Phys. Chem.* **51** 179–207
[2] Jarrold M F 2007 *Phys. Chem. Chem. Phys.* **9** 1659–71
[3] Wyttenbach T and Bowers M T 2009 *Chem. Phys. Lett.* **480** 1–16
[4] Simons J 2003 *C. R. Chim.* **6** 17–31
[5] Simons J, Jockusch R, ÇarÇabal P, Hünig I, Kroemer R, Macleod N and Snoek L 2005 *Int. Rev. Phys. Chem.* **24** 489–531
[6] Simons J 2009 *Mol. Phys.* **107** 2435–58
[7] De Vries M S and Hobza P 2007 *Annu. Rev. Phys. Chem.* **58** 585–612
[8] Schermann J P 2008 *Spectroscopy and Modeling of Biomolecular Building Blocks* (Amsterdam: Elsevier)
[9] Rijs A M and Oomens J 2015 Ir spectroscopic techniques to study isolated biomolecules *Gas-Phase IR Spectroscopy and Structure of Biological Molecules* (*Topics in Current Chemistry* vol 364) ed A M Rijs and J Oomens (Cham: Springer) pp 1–42
[10] Zhou M, Dong X, Baldauf C, Chen H, Zhou Y, Springer T A, Luo X, Zhong C, Gräter F and Ding J 2011 *Blood* **117** 4623–31
[11] Marcoux J and Robinson C 2013 *Structure* **21** 1541–50
[12] Abi-Ghanem J and Gabelica V 2014 *Phys. Chem. Chem. Phys.* **16** 21204–18
[13] Arcella A, Dreyer J, Ippoliti E, Ivani I, Portella G, Gabelica V, Carloni P and Orozco M 2015 *Angew. Chem. Int. Ed.* **1** 477–81
[14] Werz D B, Ranzinger R, Herget S, Adibekian A, von der Lieth C W and Seeberger P H 2007 *ACS Chem. Biol.* **19** 685–91
[15] Karas M, Bachmann D, Bahr U and Hillenkamp F 1987 *Int. J. Mass Spectrom.* **78** 53–68
[16] Fenn J, Mann M, Meng C, Wong S and Whitehouse C 1989 *Science* **246** 64–71
[17] Aebersold R and Mann M 2003 *Nature* **422** 198–207
[18] Polfer N C and Oomens J 2009 *Mass Spectrom. Rev.* **28** 468–94
[19] Smalley R E, Wharton L and Levy D H 1977 *Acc. Chem. Res.* **10** 139–45

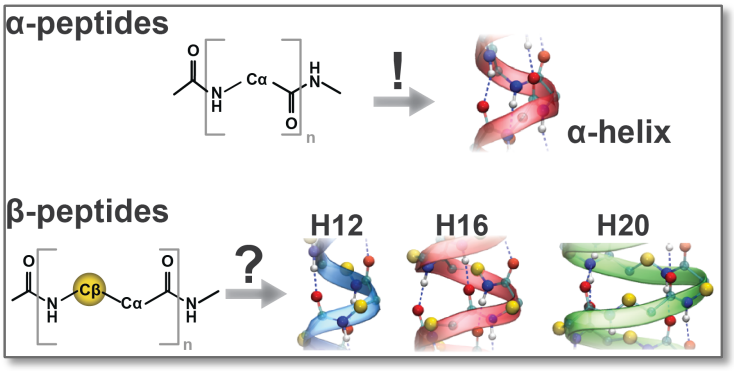
[20] Meyer T, Gabelica V, Grubmüller H and Orozco M 2013 *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **3** 408–25
[21] Barran P E *et al* 2005 *Int. J. Mass Spectrom.* **240** 273–84
[22] Benesch J L P and Robinson C V 2009 *Nature* **462** 576–7
[23] Rizzo T R, Stearns J A and Boyarkin O V 2009 *Int. Rev. Phys. Chem.* **28** 481–515
[24] Gaigeot M P 2010 *Phys. Chem. Chem. Phys.* **12** 3336–59
[25] Simons J P 2004 *Phys. Chem. Chem. Phys.* **6** E7
[26] McDaniel E W, Martin D W and Barnes W S 1962 *Rev. Sci. Instrum.* **33** 2–7
[27] Mason E A and Schamp Jr H W 1958 *Ann. Phys.* **4** 233–70
[28] Kohtani M, Jones T C, Schneider J E and Jarrold M F 2004 *J. Am. Chem. Soc.* **126** 7420–1
[29] Tkatchenko A, Rossi M, Blum V, Ireta J and Scheffler M 2011 *Phys. Rev. Lett.* **106** 118102
[30] Shelimov K B and Jarrold M F 1996 *J. Am. Chem. Soc.* **118** 10313–4
[31] Warnke S, Baldauf C, Bowers M T, Pagel K and von Helden G 2014 *J. Am. Chem. Soc.* **136** 10308–14
[32] Pierson N A, Valentine S J and Clemmer D E 2010 *J. Phys. Chem.* B **114** 7777–83
[33] Pierson N A, Chen L, Valentine S J, Russell D H and Clemmer D E 2011 *J. Am. Chem. Soc.* **133** 13810–3
[34] Servage K A, Silveira J A, Fort K L and Russell D H 2014 *J. Phys. Chem. Lett.* **5** 1825–30
[35] Stedwell C N, Galindo J F, Roitberg A E and Polfer N C 2013 *Annu. Rev. Anal. Chem.* **6** 267–85
[36] Heine N and Asmis K R 2015 *Int. Rev. Phys. Chem.* **34** 1–34
[37] Borysow J, Moraldi M and Frommhold L 1985 *Mol. Phys.* **56** 913–22
[38] Oomens J, van Roij A J A, Meijer G and von Helden G 2000 *Astrophys. J.* **542** 404
[39] Heine N, Yacovitch T I, Schubert F, Brieger C, Neumark D M and Asmis K R 2014 *J. Phys. Chem.* A **118** 7613–22
[40] Yacovitch T I, Heine N, Brieger C, Wende T, Hock C, Neumark D M and Asmis K R 2013 *J. Phys. Chem.* A **117** 7081–90
[41] Weymuth T, Jacob C R and Reiher M 2010 *J. Phys. Chem.* B **114** 10649–60
[42] Rossi M, Blum V, Kupser P, von Helden G, Bierau F, Pagel K, Meijer G and Scheffler M 2010 *J. Phys. Chem. Lett.* **1** 3465–70
[43] Jaeqx S, Oomens J, Cimas A, Gaigeot M P and Rijs A M 2014 *Angew. Chem. Int. Ed.* **53** 3663–6
[44] Tanabe K, Miyazaki M, Schmies M, Patzer A, Schütz M, Sekiya H, Sakai M, Dopfer O and Fujii M 2012 *Angew. Chem. Int. Ed.* **51** 6604–7
[45] Fricke H, Funk A, Schrader T and Gerhards M 2008 *J. Am. Chem. Soc.* **130** 4692–8
[46] Stearns J A, Seaiby C, Boyarkin O V and Rizzo T R 2009 *Phys. Chem. Chem. Phys.* **11** 125–32
[47] Zabuga A V and Rizzo T R 2015 *J. Phys. Chem. Lett.* **6** 1504–8
[48] Rossi M, Scheffler M and Blum V 2013 *J. Phys. Chem.* B **117** 5574–84
[49] Chin W, Mons M, Dognon J P, Piuzzi F, Tardivel B and Dimicoli I 2004 *Phys. Chem. Chem. Phys.* **6** 2700–9
[50] Garand E, Kamrath M Z, Jordan P A, Wolk A B, Leavitt C M, McCoy A B, Miller S J and Johnson M A 2012 *Science* **335** 694–8
[51] Dian B C, Clarkson J R and Zwier T S 2004 *Science* **303** 1169–73
[52] Compagnon I, Oomens J, Meijer G and von Helden G 2006 *J. Am. Chem. Soc.* **128** 3592–7
[53] Rijs A M, Ohanessian G, Oomens J, Meijer G, von Helden G and Compagnon I 2010 *Angew. Chem. Int. Ed.* **49** 2332–5
[54] Blom M N, Compagnon I, Polfer N C, von Helden G, Meijer G, Suhai S, Paizs B and Oomens J 2007 *J. Phys. Chem.* A **111** 7309–16
[55] Barnes L, Schindler B, Allouche A R, Simon D, Chambert S, Oomens J and Compagnon I 2015 *Phys. Chem. Chem. Phys.* **17** 25705–13
[56] Nicely A L and Lisy J M 2011 *J. Phys. Chem.* A **115** 2669–78
[57] Nicely A L, Miller D J and Lisy J M 2009 *J. Am. Chem. Soc.* **131** 6314–5
[58] Vaden T D, de Boer T S J A, Simons J, Snoek L C, Suhai S and Paizs B 2008 *J. Phys. Chem.* A **112** 4608–16
[59] Vaden T D, Gowers S A N and Snoek L C 2009 *J. Am. Chem. Soc.* **131** 2472–4
[60] Papadopoulos G, Svendsen A, Boyarkin O V and Rizzo T R 2011 *Faraday Discuss.* **150** 243–55
[61] Voronina L and Rizzo T R 2015 *Phys. Chem. Chem. Phys.* **17** 25825–36
[62] Warnke S *et al* 2015 *J. Am. Chem. Soc.* **137** 4236–42
[63] Burke N L, Redwine J G, Dean J C, McLuckey S A and Zwier T S 2015 *Int. J. Mass Spectrom.* **378** 196–205
[64] Wassermann T N, Boyarkin O V, Paizs B and Rizzo T R 2012 *J. Am. Soc. Mass Spectrom.* **23** 1029–45
[65] Toennies J P and Vilesov A F 2004 *Angew. Chem. Int. Ed.* **43** 2622–48
[66] Flórez A I G, Ahn D S, Gewinner S, Schöllkopf W and von Helden G 2015 *Phys. Chem. Chem. Phys.* **17** 21902–11
[67] Ballard A J, Martiniani S, Stevenson J D, Somani S and Wales D J 2015 *WIREs Comput. Mol. Sci.* **5** 273–89
[68] Penev E, Ireta J and Shea J E 2008 *J. Phys. Chem.* B **112** 6872–7
[69] Rossi M, Chutia S, Scheffler M and Blum V 2014 *J. Phys. Chem.* A **118** 7349–59
[70] Valdes H, Pluhackova K, Pitonak M, Rezac J and Hobza P 2008 *Phys. Chem. Chem. Phys.* **10** 2747–57
[71] Tzanov A T, Cuendet M A and Tuckerman M E 2014 *J. Phys. Chem.* B **118** 6539–52
[72] Baldauf C, Pagel K, Warnke S, von Helden G, Koksch B, Blum V and Scheffler M 2013 *Chem. Eur. J.* **19** 11224–34
[73] Rossi M, Tkatchenko A, Rempe S B and Varma S 2013 *Proc. Natl Acad. Sci. USA* **110** 12978–83
[74] Vitalini F, Mey A S J S, Noé F and Keller B G 2015 *J. Chem. Phys.* **142** 084101
[75] Riplinger C and Neese F 2013 *J. Chem. Phys.* **138** 034106
[76] Liakos D G, Sparta M, Kesharwani M K, Martin J M L and Neese F 2015 *J. Chem. Theory Comput.* **11** 1525–39
[77] Genovese L *et al* 2008 *J. Chem. Phys.* **129** 014109
[78] Haynes P D, Mostof A A, Skylaris C K and Payne M C 2006 *J. Phys. Conf. Ser.* **26** 143
[79] Blum V, Gehrke R, Hanke F, Havu P, Ren X, Reuter K and Scheffler M 2009 *Comput. Phys. Commun.* **180** 2175–96
[80] Clark S J, Segall M D, Pickard C J, Hasnip P J, Probert M J, Refson K and Payne M C 2005 *Z. Kristallogr.* **220** 567–70
[81] Hutter J, Iannuzzi M, Schiffmann F and VandeVondele J 2014 *WIREs Comput. Mol. Sci.* **4** 15–25
[82] DiStasio R A, Steele R P, Rhee Y M, Shao Y and Head-Gordon M 2007 *J. Comput. Chem.* **28** 839–56
[83] Rossi M 2011 *Ab initio* study of alanine-based polypeptide secondary-structure motifs in the gas phase *PhD Thesis* Technical University Berlin and Fritz Haber Institute http://opus.kobv.de/tuberlin/volltexte/2012/3429/
[84] Schwabe T and Grimme S 2007 *Phys. Chem. Chem. Phys.* **9** 3397–406
[85] Klimeš J and Michaelides A 2012 *J. Chem. Phys.* **137** 120901
[86] Perdew J, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865–8
[87] Tkatchenko A and Scheffler M 2009 *Phys. Rev. Lett.* **102** 073005
[88] Tkatchenko A, DiStasio R A, Car R and Scheffler M 2012 *Phys. Rev. Lett.* **108** 236402
[89] DiStasio R A Jr, Gobre V V and Tkatchenko A 2014 *J. Phys.: Condens. Matter* **26** 213202
[90] Adamo C and Barone V 1999 *J. Chem. Phys.* **110** 6158–70

[91] Jorgensen W L, Maxwell D S and Tirado-Rives J 1996 *J. Am. Chem. Soc.* **118** 11225–36
[92] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A and Simmerling C 2006 *Proteins* **65** 712–25
[93] Mackerell A D, Feig M and Brooks C L 2004 *J. Comput. Chem.* **25** 1400–15
[94] MacKerell A D *et al* 1998 *J. Phys. Chem.* B **102** 3586–616
[95] Ren P and Ponder J W 2003 *J. Phys. Chem.* B **107** 5933–47
[96] Ponder J W and Case D A 2003 *Protein Simulations* (*Advances in Protein Chemistry* vol 66) ed V Daggett (London: Academic) pp 27–85
[97] Xie Y, Schaefer H F, Silaghi-Dumitrescu R, Peng B, Li Q, Stearns J A and Rizzo T R 2012 *Chem. Eur. J.* **18** 12941–4
[98] Schubert F *et al* 2015 *Phys. Chem. Chem. Phys.* **17** 7373–85
[99] Schubert F 2014 Conformational equilibria and spectroscopy of gas-phase homologous peptides from first principles *PhD Thesis* Free University Berlin and Fritz Haber Institute www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000097733?lang=en
[100] Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder J W and Ren P 2013 *J. Chem. Theory Comput.* **9** 4046–63
[101] Baldauf C, Günther R and Hofmann H J 2003 *Helv. Chim. Acta* **86** 2573–88
[102] Baldauf C, Günther R and Hofmann H J 2006 *J. Org. Chem.* **71** 1200–8
[103] Baldauf C and Hofmann H J 2012 *Helv. Chim. Acta* **95** 2348–83
[104] Wu Y D and Wang D P 1998 *J. Am. Chem. Soc.* **120** 13485–93
[105] Wu Y D and Wang D P 1999 *J. Am. Chem. Soc.* **121** 9352–62
[106] Möhle K, Günther R, Thormann M, Sewald N and Hofmann H J 1999 *Biopolymers* **50** 167–84
[107] Swendsen R H and Wang J S 1986 *Phys. Rev. Lett.* **57** 2607–9
[108] Hukushima K and Nemoto K 1996 *J. Phys. Soc. Japan* **65** 1604–8
[109] Okabe T, Kawata M, Okamoto Y and Mikami M 2001 *Chem. Phys. Lett.* **335** 435–9
[110] Seibert M M, Patriksson A, Hess B and van der Spoel D 2005 *J. Mol. Biol.* **354** 173–83
[111] van der Spoel D and Seibert M M 2006 *Phys. Rev. Lett.* **96** 238102
[112] Lei H and Duan Y 2007 *Curr. Opin. Struct. Biol.* **17** 187–91
[113] Wales D J and Doye J P K 1997 *J. Phys. Chem.* A **101** 5111–6
[114] Pappu R V, Hart R K and Ponder J W 1998 *J. Phys. Chem.* B **102** 9725–42 http://dasher.wustl.edu/tinker
[115] Hartke B 1995 *Chem. Phys. Lett.* **240** 560–5
[116] Clark D E and Westhead D R 1996 *J. Computer-Aided Mol. Des.* **10** 337–58
[117] Nair N and Goodman J M 1998 *J. Chem. Inf. Comput. Sci.* **38** 317–20
[118] Damsbo M, Kinnear B S, Hartings M R, Ruhoff P T, Jarrold M F and Ratner M A 2004 *Proc. Natl Acad. Sci. USA* **101** 7215–22
[119] Supady A, Blum V and Baldauf C 2015 *J. Chem. Inf. Model* (doi: 10.1021/acs.jcim.5b00243)
[120] Fafoom—Flexible algorithm for optimization of molecules https://github.com/adrianasupady/fafoom
[121] Hartke B 1998 *Theor. Chem. Acc.* **99** 241–7
[122] Chutia S, Rossi M and Blum V 2012 *J. Phys. Chem.* B **116** 14788–804
[123] Schubert F, Pagel K, Rossi M, Warnke S, Salwiczek M, Koksch B, von Helden G, Blum V, Baldauf C and Scheffler M 2015 *Phys. Chem. Chem. Phys.* **17** 5376–85
[124] Ropo M, Baldauf C and Blum V 2015 arXiv:1504.03708 [physics, q-bio] http://arxiv.org/abs/1504.03708
[125] Bhattacharya S, Levchenko S V, Ghiringhelli L M and Scheffler M 2014 *New J. Phys.* **16** 123016
[126] Price M, Dennis O and William J 2001 *J. Comput. Chem.* **22** 1340–52
[127] Ceriotti M, Tribello G A and Parrinello M 2011 *Proc. Natl Acad. Sci. USA* **108** 13023–8
[128] Ceriotti M, Tribello G A and Parrinello M 2013 *J. Chem. Theory Comput.* **9** 1521–32
[129] Gasparotto P and Ceriotti M 2014 *J. Chem. Phys.* **141** 174110
[130] Neese F 2012 *WIREs Comput. Mol. Sci.* **2** 73–8
[131] Auckenthaler T, Blum V, Bungartz H J, Huckle T, Johanni R, Krämer L, Lang B, Lederer H and Willems P 2011 *Parallel Comput.* **37** 783–94
[132] NOMAD repository http://nomad-repository.eu doi: 10.17172/NOMAD/20150526220502
[133] Antoine R and Dugourd P 2011 *Phys. Chem. Chem. Phys.* **13** 16494–509
[134] Hutter J 2006 Excited-state dynamics in finite systems and biomolecules *Time-Dependent Density Functional Theory* (*Lecture Notes in Physics* vol 706) ed M Marques *et al* (Berlin: Springer) pp 217–26
[135] Mercier S R, Boyarkin O V, Kamariotis A, Guglielmi M, Tavernelli I, Cascella M, Rothlisberger U and Rizzo T R 2006 *J. Am. Chem. Soc.* **128** 16938–43
[136] Guglielmi M, Doemer M, Tavernelli I and Röthlisberger U 2013 *Faraday Discuss.* **163** 189–203
[137] Wyttenbach T, von Helden G, Batka J J, Carlat D and Bowers M T 1997 *J. Am. Soc. Mass. Spectrom.* **8** 275–82
[138] Shvartsburg A A and Jarrold M F 1996 *Chem. Phys. Lett.* **261** 86–91
[139] Mesleh M F, Hunter J M, Shvartsburg A A, Schatz G C and Jarrold M F 1996 *J. Phys. Chem.* **100** 16082–6
[140] Bleiholder C, Wyttenbach T and Bowers M T 2011 *Int. J. Mass Spectrom.* **308** 1–10
[141] Wyttenbach T, Bleiholder C and Bowers M T 2013 *Anal. Chem.* **85** 2191–9
[142] Marklund E G, Degiacomi M T, Robinson C V, Baldwin A J and Benesch J L P 2015 *Structure* **23** 791–9
[143] Herrmann C and Reiher M 2007 First-principles approach to vibrational spectroscopy of biomolecules *Atomistic Approaches in Modern Biology* (*Topics in Current Chemistry* vol 268) ed M Reiher (Berlin: Springer) pp 85–132
[144] Herrmann C, Neugebauer J and Reiher M 2007 *New J. Chem.* **31** 818–31
[145] Wilson E, Decius J and Cross P 2003 *Molecular Vibrations: the Theory of Infrared and Raman Vibrational Spectra* new edn (New York: Dover)
[146] McQuarrie D 2000 *Statistical Mechanics* 1st edn (Sausalito, CA: University Science Books)
[147] Craig I R and Manolopoulos D E 2004 *J. Chem. Phys.* **121** 3368–73
[148] Cao J and Voth G A 1994 *J. Chem. Phys.* **100** 5106–17
[149] Ramirez R, Lopez-Ciudad T, Kumar P and Marx D 2004 *J. Chem. Phys.* **121** 3973–83
[150] Berne B J and Pecora R 2000 *Dynamic Light Scattering—with Applications to Chemistry, Biology, and Physics* new edn (New York: Dover)
[151] Roy T K and Gerber R B 2013 *Phys. Chem. Chem. Phys.* **15** 9468–92
[152] Barone V, Biczysko M and Bloino J 2014 *Phys. Chem. Chem. Phys.* **16** 1759–87
[153] Calvo F, Basire M and Parneix P 2011 *J. Phys. Chem.* A **115** 8845–54
[154] Basire M, Parneix P, Calvo F, Pino T and Bréchignac P 2009 *J. Phys. Chem.* A **113** 6947–54
[155] Roy T K, Kopysov V, Nagornova N S, Rizzo T R, Boyarkin O V and Gerber R B 2015 *ChemPhysChem* **16** 1374–8
[156] Gaigeot M P and Sprik M 2003 *J. Phys. Chem.* B **107** 10344–58

[157] Gaigeot M P, Martinez M and Vuilleumier R 2007 *Mol. Phys.* **105** 2857–78
[158] Gaigeot M P and Spezia R 2015 Theoretical methods for vibrational spectroscopy and collision induced dissociation in the gas phase *Gas-Phase IR Spectroscopy and Structure of Biological Molecules* (*Topics in Current Chemistry* vol 364) ed A M Rijs and J Oomens (Cham: Springer International Publishing) pp 99–151
[159] Cimas A, Vaden T D, de Boer T S J A, Snoek L C and Gaigeot M P 2009 *J. Chem. Theory Comput.* **5** 1068–8
[160] Gregoire G, Gaigeot M P, Marinica D C, Lemaire J, Schermann J P and Desfrancois C 2007 *Phys. Chem. Chem. Phys.* **9** 3082–97
[161] Sediki A, Snoek L C and Gaigeot M P 2011 *Int. J. Mass Spectrom.* **308** 281–8
[162] Parneix P, Basire M and Calvo F 2013 *J. Phys. Chem.* A **117** 3954–9
[163] Ghiringhelli L M, Gruene P, Lyon J T, Rayner D M, Meijer G, Fielicke A and Scheffler M 2013 *New J. Phys.* **15** 083003
[164] Hammer N I, Diken E G, Roscioli J R, Johnson M A, Myshakin E M, Jordan K D, McCoy A B, Huang X, Bowman J M and Carter S 2005 *J. Chem. Phys.* **122** 244301
[165] Kohtani M and Jarrold M F 2004 *J. Am. Chem. Soc.* **126** 8454–8
[166] Pendry J 1980 *J. Phys. C: Solid State* **13** 937–44
[167] Blum V and Heinz K 2001 *Comput. Phys. Commun.* **134** 392–425
[168] Rossi M, Ceriotti M and Manolopoulos D E 2014 *J. Chem. Phys.* **140** 234116
[169] Morrone J A and Car R 2008 *Phys. Rev. Lett.* **101** 017801
[170] Marx D, Tuckerman M E, Hutter J and Parrinello M 1999 *Nature* **397** 601–4
[171] Li X Z, Walker B and Michaelides A 2011 *Proc. Natl Acad. Sci. USA* **108** 6369–73
[172] Pérez A, Tuckerman M E, Hjalmarson H P and von Lilienfeld O A 2010 *J. Am. Chem. Soc.* **132** 11510–5
[173] Feynman R P and Hibbs A R 1964 *Quantum Mechanics and Path Integrals* (New York: McGraw-Hill)
[174] Tuckerman M 2010 *Statistical Mechanics: Theory and Molecular Simulation* (Oxford: Oxford University Press)
[175] Cao J and Voth G A 1994 *J. Chem. Phys.* **101** 6168–83
[176] Rossi M, Liu H, Paesani F, Bowman J and Ceriotti M 2014 *J. Chem. Phys.* **141** 181101
[177] Ceriotti M, More J and Manolopoulos D E 2014 *Comput. Phys. Commun.* **185** 1019–26
[178] Kohtani M, Kinnear B and Jarrold M F 2000 *J. Am. Chem. Soc.* **122** 12377–8
[179] Martens J K, Compagnon I, Nicol E, McMahon T B, Clavaguéra C and Ohanessian G 2012 *J. Phys. Chem. Lett.* **3** 3320–4
[180] Fox S, Wallnoefer H G, Fox T, Tautermann C S and Skylaris C K 2011 *J. Chem. Theory Comput.* **7** 1102–8
[181] Fox S J, Pittock C, Tautermann C S, Fox T, Christ C, Malcolm N O J, Essex J W and Skylaris C K 2013 *J. Phys. Chem.* B **117** 9478–85
[182] Freddolino P L, Harrison C B, Liu Y and Schulten K 2010 *Nat. Phys.* **6** 751–8
[183] Shaw D *et al* 2009 Millisecond-scale molecular dynamics simulations on Anton *Proc. of the Conf. on High Performance Computing Networking, Storage and Analysis* pp 1–11
[184] Piana S, Lindorff-Larsen K and Shaw D E 2012 *Proc. Natl Acad. Sci. USA* **109** 17845–50
[185] Piana S, Lindorff-Larsen K and Shaw D E 2013 *Proc. Natl Acad. Sci. USA* **110** 5915–20
[186] Noé F, Horenko I, Schütte C and Smith J C 2007 *J. Chem. Phys.* **126** 155102
[187] Shirts M R, Pitera J W, Swope W C and Pande V S 2003 *J. Chem. Phys.* **119** 5740–61
[188] Lane T J, Shukla D, Beauchamp K A and Pande V S 2013 *Curr. Opin. Struct. Biol.* **23** 58–65
[189] Schwantes C R, McGibbon R T and Pande V S 2014 *J. Chem. Phys.* **141** 090901
[190] Bowman G R, Pande V S and Noé F 2014 Introduction and overview of this book *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (*Advances in Experimental Medicine and Biology* vol 797) ed G R Bowman *et al* (Berlin: Springer) pp 1–6
[191] Larson S M, Snow C D, Shirts M and Pande V S 2009 Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology arXiv: 0901.0866
[192] Kohlhoff K J, Shukla D, Lawrenz M, Bowman G R, Konerding D E, Belov D, Altman R B and Pande V S 2014 *Nat. Chem.* **6** 15–21
[193] Torrie G M and Valleau J P 1977 *J. Comput. Phys.* **23** 187–99
[194] Kong X and Brooks C L III 1996 *J. Chem. Phys.* **105** 2414–23
[195] Bartels C and Karplus M 1997 *J. Comput. Chem.* **18** 1450–62
[196] Young W S and Brooks I I 1996 *J. Mol. Biol.* **259** 560–72
[197] Bonomi M, Branduardi D, Bussi G, Camilloni C, Provasi D, Raiteri P, Donadio D, Marinelli F, Pietrucci F, Broglia R A and Parrinello M 2009 *Comput. Phys. Commun.* **180** 1961–72
[198] Tribello G A, Bonomi M, Branduardi D, Camilloni C and Bussi G 2014 *Comput. Phys. Commun.* **185** 604–13
[199] Kebarle P 2000 *Int. J. Mass Spectrom.* **200** 313–30
[200] Warnke S, von Helden G and Pagel K 2013 *J. Am. Chem. Soc.* **135** 1177–80
[201] Kohtani M, Jarrold M F and Wee S 2004 *J. Phys. Chem.* B **108** 6093–7
[202] Kohtani M and Jarrold M F 2002 *J. Am. Chem. Soc.* **124** 11148–58
[203] Liu D, Wyttenbach T and Bowers M T 2004 *Int. J. Mass Spec.* **236** 81–90
[204] Headrick J M 2005 *Science* **308** 1765–9
[205] Mizuse K and Fujii A 2011 *Phys. Chem. Chem. Phys.* **13** 7129
[206] Asmis K R 2003 *Science* **299** 1375–7
[207] Heine N, Fagiani M R, Rossi M, Wende T, Berden G, Blum V and Asmis K R 2013 *J. Am. Chem. Soc.* **135** 8266–73
[208] Guasco T L, Elliott B M, Johnson M A, Ding J and Jordan K D 2010 *J. Phys. Chem. Lett.* **1** 2396–401
[209] Nagornova N S, Rizzo T R and Boyarkin O V 2012 *Science* **336** 320–3
[210] Cocinero E, Stanca-Kaposta E, Dethlefsen M, Liu B, Gamblin D, Davis B and Simons J 2009 *Chem. Eur. J.* **15** 13427–34
[211] Mayorkas N, Rudić S, Cocinero E J, Davis B G and Simons J P 2011 *Phys. Chem. Chem. Phys.* **13** 18671
[212] DiStasio R A, Santra B, Li Z, Wu X and Car R 2014 *J. Chem. Phys.* **141** 084502
[213] Grossman J C, Schwegler E, Draeger E W, Gygi F and Galli G 2004 *J. Chem. Phys.* **120** 300
[214] Schwegler E, Grossman J C, Gygi F and Galli G 2004 *J. Chem. Phys.* **121** 5400
[215] Fernández-Serra M V and Artacho E 2004 *J. Chem. Phys.* **121** 11136
[216] VandeVondele J, Mohamed F, Krack M, Hutter J, Sprik M and Parrinello M 2005 *J. Chem. Phys.* **122** 014515
[217] Lee H S and Tuckerman M E 2007 *J. Chem. Phys.* **126** 164501
[218] Ben M D, Schönherr M, Hutter J and VandeVondele J 2013 *J. Phys. Chem. Lett.* **4** 3753–9
[219] Santra B, Michaelides A and Scheffler M 2007 *J. Chem. Phys.* **127** 184104

[220] Santra B, Michaelides A, Fuchs M, Tkatchenko A, Filippi C and Scheffler M 2008 *J. Chem. Phys.* **129** 194111
[221] Habershon S, Markland T E and Manolopoulos D E 2009 *J. Chem. Phys.* **131** 24501
[222] Habershon S, Fanourgakis G S and Manolopoulos D E 2008 *J. Chem. Phys.* **129** 74501
[223] Schmitt U W and Voth G A 1999 *J. Chem. Phys.* **111** 9361
[224] Paesani F and Voth G A 2010 *J. Chem. Phys.* **132** 014105
[225] Vendrell O, Gatti F and Meyer H D 2007 *J. Chem. Phys.* **127** 184303
[226] Gaigeot M P, Vuilleumier B, Sprik M and Borgis D 2005 *J. Chem. Theory Comput.* **1** 772–89
[227] Liu H, Wang Y and Bowman J M 2015 *J. Chem. Phys.* **142** 194502
[228] Voora V K, Ding J, Sommerfeld T and Jordan K D 2013 *J. Phys. Chem.* B **117** 4365–70
[229] Medders G R, Babin V and Paesani F 2014 *J. Chem. Theory Comput.* **10** 2906–10
[230] Hartke B and Grimme S 2015 *Phys. Chem. Chem. Phys.* **17** 16715–8
[231] Fox S J, Dziedzic J, Fox T, Tautermann C S and Skylaris C K 2014 *Proteins* **82** 3335–46
[232] Pinski P, Riplinger C, Valeev E F and Neese F 2015 *J. Chem. Phys.* **143** 034108
[233] Schiffmann F and VandeVondele J 2015 *J. Chem. Phys.* **142** 244117
[234] Hasnip P J, Refson K, Probert M I J, Yates J R, Clark S J and Pickard C J 2014 *Phil. Trans. R. Soc.* A **372** 20130270

3.2 Native like helices in a specially designed β peptide in the gas phase



Schubert, F.; Pagel, K.; Rossi, M.; Warnke, S.; Salwiczek, M.; Koksch, B.; von Helden, G.; Blum, V.; Baldauf, C.; Scheffler, M.

Native like helices in a specially designed β peptide in the gas phase

Phys. Chem. Chem. Phys. **2015** (17), 5376-5385.
DOI: 10.1039/C4CP05216A

Cite this: *Phys. Chem. Chem. Phys.*, 2015, **17**, 5376

Native like helices in a specially designed β peptide in the gas phase†

Franziska Schubert,*[a] Kevin Pagel,*[ab] Mariana Rossi,[ac] Stephan Warnke,[a] Mario Salwiczek,‡[b] Beate Koksch,[b] Gert von Helden,[a] Volker Blum,*[ad] Carsten Baldauf*[a] and Matthias Scheffler[a]

Received 10th November 2014,
Accepted 7th January 2015

DOI: 10.1039/c4cp05216a

www.rsc.org/pccp

In the natural peptides, helices are stabilized by hydrogen bonds that point backward along the sequence direction. Until now, there is only little evidence for the existence of analogous structures in oligomers of conformationally unrestricted β amino acids. We specifically designed the β peptide Ac-$(\beta^2hAla)_6$-LysH$^+$ to form native like helical structures in the gas phase. The design follows the known properties of the peptide Ac-Ala_6-LysH$^+$ that forms a α helix in isolation. We perform ion-mobility mass-spectrometry and vibrational spectroscopy in the gas phase, combined with state-of-the-art density-functional theory simulations of these molecular systems in order to characterize their structure. We can show that the straightforward exchange of alanine residues for the homologous β amino acids generates a system that is generally capable of adopting native like helices with backward oriented H-bonds. By pushing the limits of theory and experiments, we show that one cannot assign a single preferred structure type due to the densely populated energy landscape and present an interpretation of the data that suggests an equilibrium of three helical structures.

1 Introduction

Proteins – the polymers of α amino acids – play an essential role in virtually all biochemical processes. Their often highly specific function is directly correlated to their distinctive ability to fold into a well-defined, three-dimensional structure, in which functional groups are spatially arranged to form reaction centers, binding sites, *etc.* Utilizing the toolbox of organic synthesis, chemists have long sought to mimic these folding characteristics using polymers that contain non-natural amino acids – so-called "peptide foldamers".[1] The advantage here is that peptide bonds involving non-natural building blocks are less prone to proteolytic cleavage and, as such, of enormous interest for drug development.[2–4]

The first step toward successful foldamer design is the identification of polymeric backbones which fold into a well-defined structure that is ideally native-like. In this context, much effort has been spent to design peptide foldamers that imitate the characteristics of the most prominent secondary structure element – the α helix.[5–15] A promising route to achieve this goal is backbone homologation, *i.e.* the extension of the amino acid's backbone by methylene units.[5] The first homologs of natural α amino acids are β amino acids (Fig. 1a), followed by γ amino acids, δ amino acids, *etc.* In particular, β peptides were found to form secondary structures, which are similar in shape to α helices, and some of them have been used to design modulators for native protein–protein interactions.[3,16–18] Surprisingly, none of these structures directly resembles the periodically repeating backbone H-bonding pattern of α helices. The characteristic α helical $i \leftarrow (i + 4)$ H-bonding pattern[19] is depicted in Fig. 1b: H-bonds form between the NH of residue $(i + 4)$ and the backbone carbonyl group of residue i. As a result pseudocycles of 13 atoms are formed. The alternative H-bonding patterns in Fig. 1b are either tighter wound $(i \leftarrow (i + 3))$ and characterize the 3_{10} helix with 10-membered pseudocycles or feature the wider 16-membered H-bonded pseudocycles $(i \leftarrow (i + 5))$ of the π helix. The interconversion between these helices is possible by tightening or widening the helix, that is by changing the H-bonding pattern from $i \leftarrow (i+3)$ to $i \leftarrow (i+4)$ to $i \leftarrow (i+5)$ and back. By this mechanism, transitions will always happen from or to (*via*) the α helix.[20,21] In experimental and theoretical structural

[a] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany. E-mail: schubert@fhi-berlin.mpg.de, baldauf@fhi-berlin.mpg.de*
[b] *Institut für Chemie und Biochemie, Freie Universität Berlin, Takustr. 3, D-14195 Berlin, Germany. E-mail: kevin.pagel@fu-berlin.de*
[c] *Physical and Theoretical Chemistry Laboratory, University of Oxford, OX13QZ Oxford, UK*
[d] *Mechanical Engineering and Material Science Department and Center for Materials Genomics, Duke University, Durham, NC 27708, USA. E-mail: volker.blum@duke.edu*

† Electronic supplementary information (ESI) available: Cartesian coordinates for all structures displayed in the manuscript, unprocessed experimental spectra, and simulated spectra ranging from 0 to 3500 cm^{-1}. See DOI: 10.1039/c4cp05216a

‡ Present address: CSIRO Materials Science and Engineering, Bayview Avenue, Clayton, Victoria 3168, Australia.

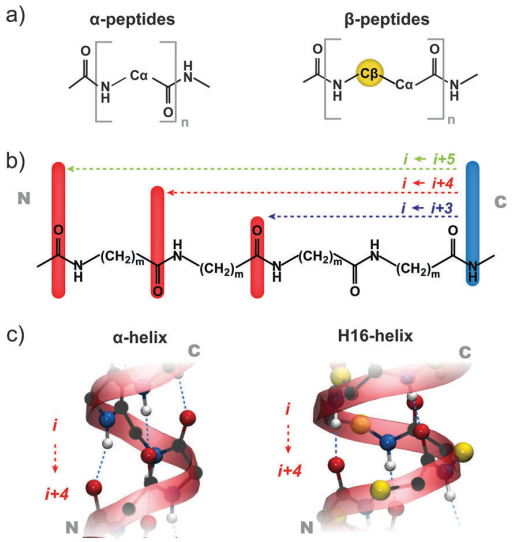


Fig. 1 Structure of α and β amino acids and their oligomers. (a) α Amino acids and β amino acids are homologs that differ by a single backbone CH_2 group. (b) In α peptides, $-(CH_2)_1-$, and β peptides, $-(CH_2)_2-$, different backbone H-bonding patterns may lead to helical structures with H-bonds pointing in backward direction along the sequence, for example, the α helix and the H16 helix (see c).

studies, mainly the α helix is found. This is not only due to enthalpy, *e.g.* H-bond cooperativity, but also due to a significant vibrational entropic stabilization that sets helices apart from competing compact conformers at room temperature.[22]

It is well established that polyalanine sequences form α helices in the gas-phase, especially in the presence of a protonated lysine residue at the C terminus.[22–29] These prototypical peptides follow the sequence Ac-Ala_n-$LysH^+$; members of this series have been extensively studied by ion mobility-mass spectrometry (IM-MS),[23–25] gas-phase vibrational spectroscopy,[26,28] and density-functional theory (DFT).[22,26–29] The placement of a positive charge at the C-terminus stabilizes the helix *via* coordination of dangling backbone carbonyls and favorable interaction with the helix macro-dipole. As an example for these polyalanine systems, we study here the peptide Ac-Ala_6-$LysH^+$, for which the formation of an α helical structure at room temperature has been predicted.[22]

β Peptides have been demonstrated to form various helices with H bonds pointing in forward (from N to C terminus), in backward (from C to N terminus), or in alternating direction ("mixed" helices) along the sequence.[5–10,30–32] We are here however specifically interested in helix types that resemble the α helix, *i.e.* with H bonds that point backward relative to the sequences direction (from C to N terminus), as indicated in the H bonding scheme in Fig. 1b. The resulting helices are characterized by H bonds that form pseudo cycles with 12, 16, or 20 atoms and are therefore consistently named H12, H16, and H20, respectively. An illustrative example for the helix H16 is shown in Fig. 1b along with its α peptide equivalent, the α helix. Both feature H bonds with the same $i \leftarrow (i+4)$ pattern as depicted in Fig. 1b. According to the H-bonding patterns, H12, H16, and H20 are related to the 3_{10}, α, and π helix motifs of the α peptides.[10,33] The H12 helix was first described by Gellman and co-workers. Its formation, however, required cyclic β amino acids that are sterically restricted.[34–36] The α helix equivalent H16 helix has been proposed theoretically by Hartree–Fock calculations,[10] but to date there has been only limited experimental evidence for its existence, most of it stemming from diffraction patterns of Nylon-3 polymers,[37,38] which have the same backbone structure as their oligomeric β peptide relatives.

In order to study the formation of helices with native like H bonds in backward direction along the sequence (see Fig. 1b and c), we employed the above-described design principle of Ac-Ala_6-$LysH^+$.[22–29] To obtain a β peptide, we replaced the alanine residues by (*R*)-β-aminoisobutyric acid (β^2hAla) – an alanine derivative with an extended backbone (Fig. 1a). The resulting foldamer Ac-$(\beta^2hAla)_6$-$LysH^+$ was investigated by gas-phase experiments and simulations.

2 Methods

2.1 Ion mobility-mass spectrometry

IM-MS experiments to determine collision cross sections (CCSs) were performed using an in-house built drift-tube instrument following a design described previously.[39] Briefly, ions are formed in a nano-electrospray ionization source (nESI) and transferred into the vacuum. An electrodynamic ion funnel collects and pulses ions into the drift region where they move through a buffer gas (He) under the influence of a weak electric field. At the end of the drift-tube, a second electrodynamic ion funnel guides the ions into a quadrupole mass spectrometer, which separates the ions according to their mass-to-charge ratio (m/z). By measuring the time-dependent ion current of m/z selected ions, characteristic arrival time distributions (ATDs) can be obtained. From these ATDs, absolute CCSs of a particular ion species can be determined.[40]

2.2 Gas-phase vibrational spectroscopy

The experiments were performed at the free-electron laser facility FELIX[41] (Nieuwegein, the Netherlands) using a Fourier-transform ion cyclotron (FT-ICR) mass spectrometer.[42] For ionization a nESI source (MS Vision, Almere, NL) and capillaries prepared in-house were used. Ions were accumulated in a hexapole ion trap and transferred into a home-built FT-ICR mass spectrometer that is optically accessible *via* a KRS-5 window at the back end. The ions were irradiated by IR photons of the free electron laser FELIX. Resonance of the IR light with an IR active vibrational mode in the molecule results in the absorption of multiple photons, which causes the dissociation of the ions. Monitoring the depletion of the individual parent ion signals as a function of IR wavelength leads to the IR spectra.

2.3 Simulation details

The conformational search for the peptide Ac-Ala_6-$LysH^+$ was described previously by Rossi *et al.*[22] For the β-peptide

Ac-(β^2hAla)$_6$-LysH$^+$ an extensive sampling of the potential energy surface (PES) of the OPLS-AA force field[43] has been performed independently by two approaches. We employed the basin hopping algorithm that is implemented in Tinker.[44,45] Furthermore, we employed replica-exchange molecular dynamics (REMD) simulations with the Gromacs program.[46] The simulations yielded an overall sampling time of 8 μs distributed over 16 replicas, finally, snapshots in 2 ps intervals were extracted from the 300 K trajectory and clustered.[47]

Altogether, basin hopping and REMD simulations yielded 13 119 structures that were then relaxed by density-functional theory (DFT) calculations employing the PBE functional[48] that was corrected for long-range dispersion interactions[49] (PBE + vdW). Electronic structure theory calculations, including geometry optimizations, harmonic vibrational frequencies from finite differences, AIMD simulations, and replica-exchange AIMD simulations, were performed with the FHI-aims program package which employs numeric atom-centered orbitals as basis sets.[50] In order to reduce the bias of the empirical force field and following our focus on helical structures, we further sampled the local conformational space by means of replica-exchange AIMD simulations starting from representative structures of the H12, H16, and H20 helices that were obtained in the OPLS structure search (schemes for $i \leftarrow (i+3)$, $i \leftarrow (i+4)$, and $i \leftarrow (i+5)$ in Fig. 1b). The total sampling times were 486 ps, 576 ps, and 558 ps, respectively, each of them distributed over 18 replicas in a temperature range between 300 K and 687 K. We used a time step of 1 fs and swaps between replicas were attempted every 100 fs. Structure snapshots of all replicas were taken after each ps and post-relaxed with PBE + vdW. In summary, 14 739 PBE + vdW relaxations of candidate structures of the β-peptide Ac-(β^2hAla)$_6$-LysH$^+$ were performed. A free-energy correction that includes vibrational free energies in the harmonic approximation and rotational contributions in the rigid-rotor approximation, both computed with PBE + vdW at T = 300 K, was applied. Additionally, we tested modifications of the theory towards a higher-level functional, PBE0,[51] and with the improved many-body description of the long-range dispersion,[52] similar to a recent study of the validity of exchange–correlation functionals and dispersion corrections for the prediction of peptide secondary structures.[53]

The infrared spectra were calculated from the Fourier transform of the dipole time derivative autocorrelation function[27,28] obtained from micro-canonical AIMD simulations of 25 ps length (after at least 5 ps equilibration at 300 K). Some anharmonicity effects will be missing in the spectra, because averages from classical trajectories were used for the dipole–dipole time correlation instead of exact quantum mechanical averages. However, this approach is currently at the limit of what is computationally feasible. To account for experimental broadening, the simulated spectra were convoluted with a Gaussian function with a variable width of 0.5% of the wavenumber. For a quantitative comparison we employed the Pendry reliability factor,[54] which has been successfully used in the context of IR spectroscopy before.[28,55] Perfect agreement yields R_P = 0 while no correlation between the spectra yields R_P = 1. An optimal fit between two spectra (based on R_P) is achieved by rigid shifts along x and y axes.

3 Results

3.1 Ion mobility-mass spectrometry

In an IM-MS experiment a package of ions is injected into a cell filled with an inert neutral buffer gas (in this work: helium). Aided by a weak electric field, the ions traverse the cell where they undergo many low energy collisions with buffer-gas molecules. Compact ions undergo fewer collisions and therefore traverse the cell faster than ions with a more extended conformation, which allows the separation of species with identical mass and charge but different size and shape. Moreover, the recorded drift or arrival times can be converted into collision cross sections, which are universally comparable values that can be calculated theoretically on the basis of molecular models. The typical arrival time distribution (ATD) depends on the shape of the ions and can be converted into a collision cross section (CCS) *via* the Mason-Shamp equation.[40] CCS values are independent of a specific experimental setup (machine, experimental conditions). An ATD of Ac-Ala$_6$-LysH$^+$ is shown in the upper plot of Fig. 2a. The α-peptide ion cloud arrives at a drift time of 12 ms with a full-width half maximum (FWHM$_{exp.}$) of 0.38 ms. Fig. 2b shows the ATD for the β peptide Ac-(β^2hAla)$_6$-LysH$^+$. The backbone extension from α to β amino acid building blocks results in a longer drift time of about 13 ms and a peak width of FWHM$_{exp.}$ = 0.38 ms.

For each of the two systems, the α-peptide and the β-peptide, single and narrow peaks are observed in the ATD. If one assumes only a single type of conformation to be present in the drifting ion cloud, the peak width depends entirely on the initial pulse width and the broadening due to diffusion. This flux-based broadening of the ATD can be calculated by:[40]

$$\Phi(t) = \int \mathrm{d}t' \left\{ \frac{C}{\sqrt{D(t-t')}} \left(v_\mathrm{d} + \frac{L}{(t-t')} \right) \exp\left[\frac{-(L - v_\mathrm{D}(t-t'))^2}{4D(t-t')} \right] P(t') \right\}, \quad (1)$$

where $P(t')$ is a function describing the shape of the ion cloud as it enters the drift region, for which we assume a rectangle pulse of 100 μs length in this case. C is a constant and D is the diffusion coefficient given by the Einstein relation $D = \frac{v_\mathrm{D} k_\mathrm{B} T}{Eze}$, where ze is the charge of the ion, E is the applied electric field, and v_D is the average drift velocity. L denotes the length of the drift tube. The resulting theoretical flux-based broadening of the experimental ATD peak is plotted as dashed lines in Fig. 2. The experimentally observed FWHM$_{exp.}$ are only slightly broader than the theoretical FWHM$_{flux}$ values (see Fig. 2).

However, a narrow peak is not necessarily linked to a single conformer. The ion cloud traverses the drift tube in a time of 12 ms or 13 ms, respectively; within such a time scale, a molecular system may adopt numerous different conformational states if the barriers that separate them on the free energy surface are not too high. Assuming such a scenario, the width of the peak now provides information whether the interconversion between multiple minima is fast enough to

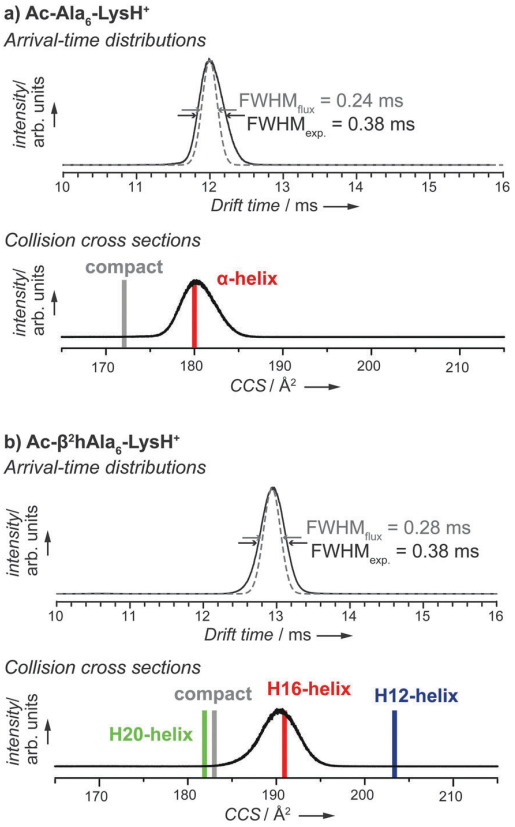


Fig. 2 Ion-mobility mass-spectrometry (IM-MS) of peptides Ac-Ala_6-$LysH^+$ (a) and Ac-$(\beta^2hAla)_6$-$LysH^+$ (b). The experimental arrival-time distributions (ATDs, black lines in the two plots on top) were converted into collision cross sections (CCSs, black lines in the two plots at the bottom). In the plots of the ATDs, a flux-based estimate of the peak width is given as dashed line, the full-width at half-maximum (FWHM) peak width for experiment and flux-based model are given. Vertical bars in the CCS plots indicate CCSs calculated for predicted conformers shown in Fig. 3.

average out over the drift time. In other words, a narrow peak may also indicate that each individual ion in the cloud has reached conformational equilibrium, namely the time average over all accessible conformers. The conformer distribution in the ensemble equals the conformer distribution in the time average of the individual ion due to the relatively long drift time. The relatively narrow peaks we observe by comparing $FWHM_{exp.}$ and $FWHM_{flux}$ indicate that all ions drift with the same average velocity and thus: (i) belong to a single conformational family, or (ii) belong to multiple conformational families with the same drift time, or (iii) interconvert between multiple conformers and reach equilibrium within the drift time of 12 or 13 ms, respectively.

3.2 Tackling the conformational problem by simulation

We narrow down the conformational problem for the α-peptide and the β-peptide applying a two-step procedure. First, the conformational space defined by an empirical force field is sampled in order to generate input for the subsequent first-principles relaxations. In a second step, a local refinement is performed that employs density-functional theory at the PBE + vdW level[48,49] Finally, free energies at 300 K (ΔF_{300K}) were estimated by including harmonic vibrations and rotational contributions in the rigid rotor approximation.

Fig. 3a shows the free energy hierarchy at 300 K (in the harmonic oscillator and rigid rotor approximation) and the two lowest free-energy structures of Ac-Ala_6-$LysH^+$ that were identified by a recent first-principles (PBE + vdW) based conformational search by Rossi *et al.*[22] Vibrational free energy contributions particularly stabilize helical structures with respect to more compact structures.[22] This can be seen in the qualitative changes from the potential energy hierarchy to the free energy hierarchy as displayed in Fig. 3a for the α peptide. Consistently, the α helix is the preferred conformation for Ac-Ala_6-$LysH^+$ confirmed by harmonic free energies at T = 300 K with the PBE + vdW approach. From the Cartesian coordinates of the conformers, theoretical CCSs can be calculated and compared to their experimental counterparts. For this, we employ the projection approximation (PA) method,[56] which is known to yield reliable values for ions with less than 200 atoms.[57] The theoretical CCS of the α-helical conformer of Ac-Ala_6-$LysH^+$ agrees best with the experimental peak of the distribution of CCSs derived from experiment (Fig. 2a).

The β peptide Ac-$(\beta^2hAla)_6$-$LysH^+$ is expected to be structurally more flexible than the α peptide due to the additional methylene group per residue. In order to sample the larger structure space of the β peptide system, a far more extensive first-principles guided conformational search had to be performed. The multi-step search protocol that is described in the methods section yielded approximately 14 000 optimized geometries at the PBE + vdW level within a relative energy window of 156 kJ mol^{-1}. Re-relaxations of all minima within a relative energy window of 38.6 kJ mol^{-1} with tight computational settings and harmonic free-energy calculations were performed. Harmonic free-energy contributions favor helical structures over more compact structures in Ac-$(\beta^2hAla)_6$-$LysH^+$ – an effect observed before for the Ac-Ala_n-$LysH^+$ systems.[22] The high density of structures of Ac-$(\beta^2hAla)_6$-$LysH^+$ with low harmonic free energies is remarkable (see Fig. 3b). However, the comparison to the hierarchy of the α peptide[22] might be misleading. The conformational search strategies differ, especially in the local refinement step of the search results for the β peptide by means of replica exchange AIMD simulations. The three lowest free-energy conformers of Ac-$(\beta^2hAla)_6$-$LysH^+$ at 300 K are the helix H12, a *compact* structure, and the helix H20 (Fig. 3b), all within a free-energy window of about 3 kJ mol^{-1}. The α helix equivalent H16 helix is about 10 kJ mol^{-1} above the global minimum in free energy, among a total of 16 conformers that are present within this narrow energy window. Helix types with forward oriented H bond patterns along the sequence[10] were not found. This is due to the, by design of the peptides, selective stabilization of backward oriented H bonded structures *via* favorable charge dipole interactions.

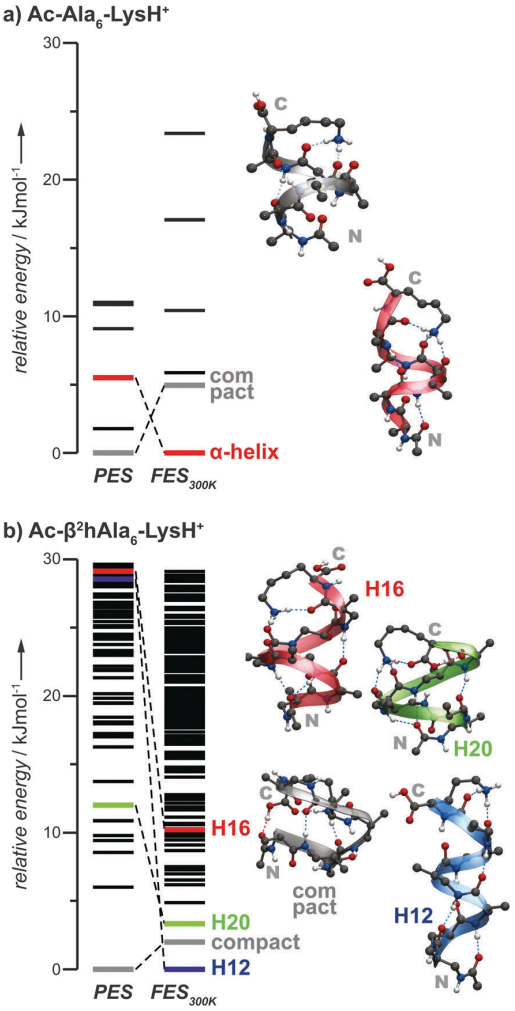


Fig. 3 Free energy hierarchy and examples of conformation of peptides Ac-Ala_6-$LysH^+$ (a) and Ac-$(\beta^2hAla)_6$-$LysH^+$ (b). (a) A first-principles-based conformational search by Rossi *et al.*[22] yielded a compact (grey) and an α-helical (red) structure as the two most likely conformations of α-peptide Ac-Ala_6-$LysH^+$ at room temperature. (b) A compact conformation (grey) as well as the helices H12 (blue), H16 (red), and H20 (green) are displayed along with a plot of the free energy hierarchy. The displayed structures are highlighted in the hierarchy with their assigned color.

For all these β peptide conformers depicted in Fig. 3b theoretical CCSs based on the PA method were computed.[56] Simulation and experiment are compared in Fig. 2b. We get a perfect match between the calculated CCS of H16 and the experimental peak position with a negligible deviation of about 1.5%. The computed CCS values for H12 and H20 as well as for the compact conformer clearly deviate from the CCS value of the experimental peak.

3.3 Gas-phase vibrational spectroscopy

The experimental spectra for the α-peptide Ac-Ala_6-$LysH^+$ and the β-peptide Ac-$(\beta^2hAla)_6$-$LysH^+$ were measured at room temperature and are shown in the upper plots in Fig. 4a and b. At a first glance, the experimental vibrational spectrum of the β peptide (Fig. 4b) shares many features with that of the helical α-peptide (Fig. 4b), with the amide-I (C═O stretch mode, approximately 1679 cm^{-1}) and amide-II (N–H bending mode, approximately 1510 cm^{-1}) resonances being the most prominent peaks. A comparison of both spectra, however, shows characteristic differences in band position, width, and intensity, especially in the region between 1000 and 1400 cm^{-1}. This region is sensitive to the main chemical difference between both peptides, the additional methylene units in the backbone of the β-amino acid building blocks. At the other end of the spectrum (around 1760 cm^{-1}), both experimental spectra feature a vibrational mode of low intensity that hints to a free terminal C═O group. The experimental spectra were each averaged over four individual recordings and the background level was determined multiple times throughout each wavelength scan; the peak is real and not noise. The un-smoothed spectra are shown, together with error bars, in Fig. S1 of the ESI.†

For the α peptide, constant-energy AIMD simulations (with $\langle T \rangle$ = 300 K) were performed for the two lowest-free energy conformers, the α helix and the compact conformer. From this data, theoretical vibrational spectra were derived and compared to the experimental spectrum. A quantitative comparison is crucial here and can be achieved by employing the Pendry reliability factor[54] that was previously introduced to the field of peptide vibrational spectroscopy.[28] Simulated spectra are rigidly shifted in x and y direction in order to yield the optimal R_P with respect to the experiment, values Δ_x and Δ_y are given in Fig. 4. The shift along the x-axis accounts for a mode softening (redshift) in the simulated spectra that probably results from the approximations made. This can for instance be due to the use of the exchange–correlation functional approximation (PBE) to DFT or the classical propagation of the AIMD trajectories that neglects quantum nuclear effects.[27,28] The intensity shift (along the y-axis) accounts for offsets in the experiment. The theoretical spectrum of the α helical conformer fits better to the experimental spectrum (R_P = 0.31) than the compact structure with R_P = 0.46 (Fig. 4a). Theoretical and experimental vibrational spectroscopy strongly support the interpretation of only the α helix being present in the gas phase and at room temperature for the α-peptide Ac-Ala_6-$LysH^+$.

For the β peptide Ac-$(\beta^2hAla)_6$-$LysH^+$ we follow the same approach and select the low free energy conformers H12, compact, and H20 (see Fig. 3) as starting points for AIMD simulations. Even though the H16 conformer is higher in free energy, we still consider it here, as it is the direct analog of the α helix. The computational cost of such simulations is substantial and can only be performed for selected conformers. The individual simulated spectra are again compared to the experimental spectrum. The compact structure as well as the H12 helical structure agree only poorly based on the R_P criterion that rationalizes mismatches in the peak positions. The theoretical spectra of H16 and H20 have a slightly better agreement with experiment based on the R_P criterion, but still much worse than the R_P of 0.31 that we saw with the assignment above for the α peptide. Another possible criterion is the diagnostic peak that

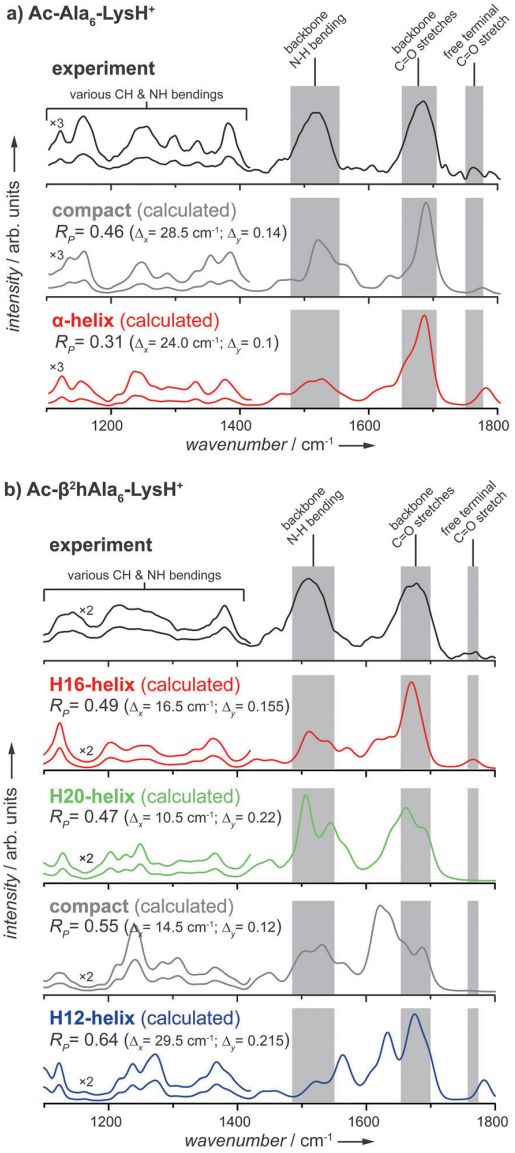


Fig. 4 Gas-phase vibrational spectroscopy of (a) the α-peptide Ac-Ala_6-$LysH^+$ and (b) the β-peptide Ac-$(\beta^2hAla)_6$-$LysH^+$ at room temperature. The plots show the experimental spectra (black lines) and the show simulated spectra (colored lines) from AIMD calculations. Experimental IR spectra were smoothed, see ESI† for the raw data. Vibrational spectra were simulated for the conformers shown in Fig. 3a. A magnification is shown for the wavenumber region from 1000 to 1400 cm^{-1}. Theoretical vibrational spectra were uniformly shifted, not scaled, by Δ_x and Δ_y along the wavenumber and intensity axes to best fit the experiment.[28]

was found in the high wavenumber region (around 1760 cm^{-1}). This diagnostic feature results from the C-terminal carboxyl group not being involved in H bonds and is consequently only reproduced in the simulated spectra of the H12 and H16 helices. However, it is evident that we do not reach a clear conclusion from gas-phase vibrational spectroscopy of the β peptide Ac-$(\beta^2hAla)_6$-$LysH^+$, but there might be slight hints that point towards the H16 helix as possible dominant conformer for the β peptide in the gas phase.

4 Discussion

The data for the α peptide Ac-Ala_6-$LysH^+$ points to one clear and obvious solution: the expected dominance of the α-helix in the gas phase. Contrarily, the data from gas-phase experiments and first-principles simulations for the β-peptide Ac-$(\beta^2hAla)_6$-$LysH^+$ is less clear, even contradictory. The simulation results, specifically the harmonic free energy hierarchy at T = 300 K, point towards the H12 helix as being most stable in the gas phase, next in line are a compact conformer and the helical structure H20, all within a ΔF_{300K} range of about 3 kJ mol^{-1}. The α-helix like conformer H16 is about 10 kJ mol^{-1} higher in this free energy scale. However, the IM-MS measurements find a narrow drift peak with a CCS distribution that agrees very well with the shape of this H16 helix. The vibrational spectroscopy experiments reveal no particularly reliable agreement with any of the theoretically predicted spectra, but might weakly hint towards the H16 helix. In the following, we will discuss in detail two possible interpretations that could help explain the situation.

We then also assess the applicability and accuracy of the applied method by comparing two different density functionals in combination with two different corrections for long-range dispersion.

4.1 A step back

In order to critically assess a possible assignment of the H16 helix as most-likely conformer to be present in the gas phase, we take a step back and evaluate the full pool of structures for which we calculated the harmonic free energy. For each of the 163 low free-energy conformers (up to ΔF_{300K} = 38.5 kJ mol^{-1}) we have a data point that envelopes three values:

- the free energy at 300 K in the harmonic oscillator and rigid rotor approximation,
- the agreement between the experimental and the predicted vibrational spectrum measured by R_P,[54] and
- the agreement between the calculated (PA) and measured CCS expressed by the difference ΔCCS.

The vibrational spectra derived from AIMD simulations are computationally too costly to be routinely computed for a large number of conformers, consequently we can only use *harmonic* vibrational spectra for this number of conformers. Fig. 5 shows again the conformational free-energy hierarchy in the harmonic approximation at 300 K. When considering the full conformational pool up to 38.5 kJ mol^{-1} for plotting R_P and ΔCCS of each conformer (see Fig. 5a), it is hard to draw a conclusion. However, it is obvious that there are conformations for which a good agreement with the experimental observables is predicted (low R_P and ΔCCS close to 0). Fortunately, there is a third dimension to be considered, the computed free energy. Stepwise lowering the cut-off ΔF_{300K} for plotting (see Fig. 5b and c), the conformer H16 appears more and more isolated in the plots

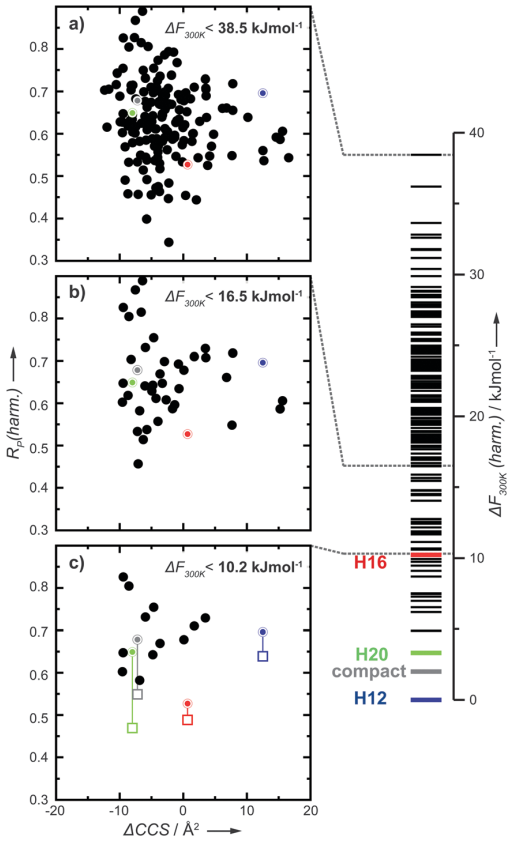


Fig. 5 The conformational free-energy hierarchy of the β peptide Ac-(β^2hAla)$_6$-LysH$^+$ (right panel) and the R_P (harmonic spectra) *versus* ΔCCS plots (left panel). The the filled circles in the three plots represent all predicted structures up to a relative free-energy threshold of (a) 38.5 kJ mol^{-1}, (b) 16.5 kJ mol^{-1}, and (c) 10.2 kJ mol^{-1}, respectively. The compact conformer and the helices H12, H16, H20 are highlighted in gray, blue, red, and green, respectively. In plot (c), also the R_P values for the vibrational spectra of compact, H12, H16, and H20 derived from AIMD simulation are shown as open squares, connected by a straight line to the respective value for the harmonic spectrum.

and suggests itself as a likely conformer to be present in the experiment, especially due to the perfect agreement of experimentally observed and calculated CCS. However, further lowering the energy cut-off will remove the H16 structure from the candidate list.

Another problem with a structure assignment based on R_P is illustrated in Fig. 5c, where also the R_P values for the four spectra (H12, H16, H20, compact) derived from AIMD simulations are plotted as open squares. The MD derived spectra are in better agreement with the experiment than the respective ones calculated in the harmonic approximation as it is indicated by the lower R_P value. However, the improvement is not uniform; while there is, for instance, only minor improvement for H16, the improvement from the harmonic to the MD treatment for H20 is substantial. This limits the applicability of the harmonic vibrational spectra for structure assignment. Overall, a reliable structure assignment seems impossible with the theory-experiment agreement achieved here.

4.2 Equilibrium

In isolation, structural changes that involve the rearrangement of H bonds can be hindered due to the lack of compensation by transient interactions with water molecules. However, the structural interconversion between the helix types shown in Fig. 1b can happen *via* tightening or loosening the helical twist and the intermediate formation of bifurcated H bonds in some sort of "breathing" motion.[20,21] The path of this interconversion always features the H16 helix as an intermediate as it lies in between its relatives H12 and H20 when considering a meaningful reaction coordinate like helical twist, the diameter *vs.* length ratio, or the here used CCSs of the structures. Again, structural transitions between these helices would always be H12 ⇌ H16 ⇌ H20 if we exclude the possibility of full unfolding and the refolding to an alternative helix type. The same concept in turn also holds for the possible helices of the α peptide, where transitions 3_{10} ⇌ α ⇌ π would occur. Combining this view with the relative free energies ΔF_{300} that were calculated yields the two differing pictures shown in Fig. 6. Please note, the free-energy and CCS values for the helical conformers stem from actual calculations, but the gray lines are only an illustrative representation of a possible free-energy surface (FES). In fact we do not have knowledge about barriers (yet). The interpretation of the illustrative α peptide FES is straightforward, the α helix is the most stable structure and the barriers to the neighboring helical structures must be high, as even the respective minima of the 3_{10} and π are above the energy window used in the representation in Fig. 6. Consequently, the experimental CCS distribution only features one peak that fits best the theoretical CCS of the α helix. Also the experimental CCS distribution of the β peptide fits best to the α helix-like H16 structure. However, here the H16 is least stable of three alternative helical structures. How to bring these seemingly contradictory findings in line? First of all, the drift time of 13 ms has to be considered. All individual ions of the ion cloud should be in structural equilibrium and have visited the possible states on our FES, the H12, H16, and H20 minima, several times. As a consequence, the experimental CCS distribution represents an *average* of the visited states that matches the CCS value predicted for the H16 structure that is located between the two lower free-energy conformers H12 and H20. This interpretation brings at least the free-energy prediction and the IM-MS measurements in line. The reasons for the disagreement between the experimental IR spectrum and the four simulated spectra from AIMD simulations remain to be investigated. Straightforward mixing of the four individual spectra with the target function of reducing the R_P to the experiment does not yield satisfying agreement. The R_P values for the helical conformers H12, H16, and H20 of the β peptide are 0.64, 0.49, and 0.47, respectively (see also Fig. 4). The best combination of the three spectra, to which H16 and H20 contribute equally and H12 does not contribute at all, has an R_P of 0.42.

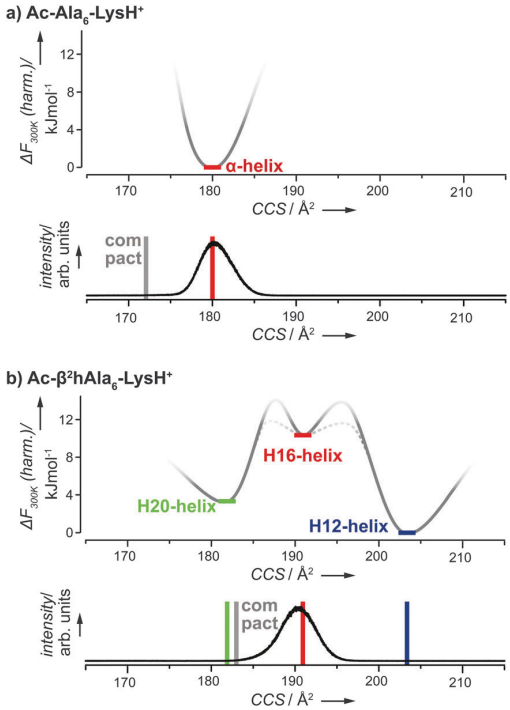


Fig. 6 The predicted CCS can be used as a reduced coordinate together with the computed free energies to draw a free-energy profile that relates the helical structures to each other. Please note, the gray lines are illustrative and do not represent results from simulation. For the α peptide Ac-Ala_6-$LysH^+$ (a), the α helix is the only helical conformer within the considered free energy range. Consequently, we would assume a deep potential well to flank the α helical minimum. For the β peptide Ac-$(\beta^2hAla)_6$-$LysH^+$ (b), three helices are present in the considered energy range of about 12 kJ mol^{-1}. CCS as a conformational coordinate places H16 right between the two alternative helices H12 and H20 like a barrier that has to be overcome whenever the helices interconvert. The CCS plots from Fig. 2 are shown again to illustrate how two very different (hypothetical) energy landscapes can potentially result in a very similar IM-MS signal.

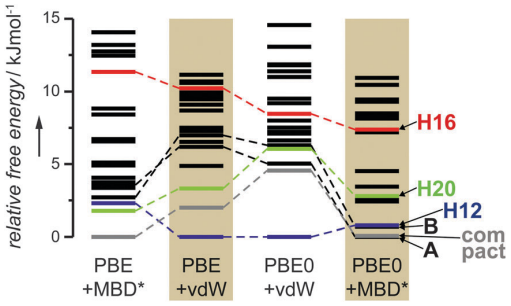


Fig. 7 The potential energy of selected conformers with a relative free energy (PBE + vdW) of 11.2 kJ mol^{-1} was recalculated with PBE + MBD*, PBE0 + vdW, and PBE0 + MBD*. The relative free energies contain the potential energy calculated at the given level and the harmonic free energy contribution computed at the PBE + vdW level. The energy levels of selected conformers are highlighted.

4.3 Exact exchange and many-body dispersion

The conformational free energy hierarchy shown in Fig. 3 is sensitive to the various approximations that we employ. In one of our recent studies[53] we assessed the accuracy of pairwise (vdW)[49] and many-body dispersion corrections (MBD*)[52] in combination with the density functionals PBE (generalized-gradient approximation)[48] and PBE0 (with Hartree–Fock like exchange)[51] for the description of the conformational energy hierarchy of peptides in the gas phase. In the same spirit we have tested how PBE + MBD*, PBE0 + vdW, and PBE0 + MBD*, treat the low energy regime of the β peptide Ac-$(\beta^2hAla)_6$-$LysH^+$ predicted at the PBE + vdW level of theory. The energies of conformers of the β peptide with a relative free energy below 11.2 kJ mol^{-1} were recalculated. The resulting potential energies were then combined with the harmonic vibrational free energy corrections computed with PBE + vdW. The results are summarized in Fig. 7. The change from PBE + vdW to PBE0 + MBD* stabilizes H16, while H12 is slightly destabilized. Furthermore, two additional compact conformers, highlighted as A and B in Fig. 7, are ranked more stable. At the PBE0 + MBD* level, four conformers, namely A, compact, B, and H12, have to be considered within the narrow free energy window of only 1 kJ mol^{-1}. However, for none of them the match between experimental CCS value and computed value is as good as for the H16 conformer (see ESI,† Table S2).

We discuss here a free energy range of about 10 kJ mol^{-1} for the considerably large β-peptide with its 108 atoms. This translates to roughly 0.1 kJ mol^{-1} per atom, in other units: 0.02 kcal mol^{-1} or 1 meV. The comparison of relative energy hierarchies of 27 conformers of Ac-Ala_3-NMe reproduced with PBE + vdW and PBE + MBD*[53] to a CCSD(T) reference hierarchy[58] shows mean absolute errors of only the potential energy description of 0.05 kJ mol^{-1} per atom (in other units: 0.01 kcal mol^{-1} or 0.5 meV). Consequently, the minuscule energy differences that we are discussing here are within the uncertainties of the applied approximations to potential energy (*e.g.* PBE + vdW or PBE0 + MBD*) and free energy (harmonic approximation). Despite these uncertainties, we can identify a group of likely conformers with the error between different functionals being at most 10 kJ mol^{-1}. Within that group making a distinction becomes difficult and we need experimental data to compare with. So it is not only the potential energy description that limits us here, but especially also the conformational and entropic contributions that are of course substantial at 300 K.

5 Conclusion

With this study on the conformational properties of the β-peptide Ac-$(\beta^2hAla)_6$-$LysH^+$, we have clearly pushed the current limits of what is possible in gas-phase experiments and simulation. With respect to the experimental results, the IM-MS experiments give us the simplest, if not even over-simplified answer. The rather narrow peak does, according to our interpretation, not represent a single conformer type but more likely a conformational equilibrium. The gas-phase vibrational spectroscopy on the other hand ideally gives far more structural information that is, however, hard to access, *e.g.*, due to the broadness of the bands in the

experimental spectrum. Furthermore, we here focus recording of spectra to the 1000 to 1800 cm^{-1} region. The flanking wavenumber regions apparently also offer a lot of information as it is evident from the simulated spectra of the β peptide shown in Fig. S2 of the ESI.†

Still we can show that a β peptide that consists of open chain building blocks (not sterically constrained) is generally capable of adopting native like helices with backward oriented H bonds (*cf.* Fig. 1b), similar to the types observed for the natural α peptides. We derive an explanation how these structures can interconvert in isolation that will be investigated in future experiments and simulations.

The density of conformers of the β peptide at low energies is clearly a challenge for the traditional, single-point plus harmonic free energy approach at 300 K, and this is where future work must focus. The interconversion between helices that we discuss here can be described with plausible conformational coordinates and then be studied with techniques like metadynamics[59] in combination with DFT. Low-temperature experiments on the other hand might yield sharper IR bands or, with a cooled drift tube, allow the separation of different conformers by hindering the interconversion between structures.

Acknowledgements

This work was supported by the Center for Supramolecular Interactions of the Freie Universität Berlin. We gratefully acknowledge the "Stichting voor Fundamenteel Onderzoek der Materie" (FOM) for providing the beam time on FELIX as well as support by members of the FELIX staff: Britta Redlich, Lex van der Meer, Rene van Buuren, Jos Oomens, Giel Berden, and Josipa Grzetic.

References

1 S. H. Gellman, *Acc. Chem. Res.*, 1998, **31**, 173.
2 L. K. A. Pilsl and O. Reiser, *Amino Acids*, 2011, **41**, 709.
3 D. Seebach and J. Gardiner, *Acc. Chem. Res.*, 2008, **41**, 1366.
4 J. Frackenpohl, P. I. Arvidsson, J. V. Schreiber and D. Seebach, *ChemBioChem*, 2001, **2**, 445.
5 A. Banerjee and P. Balaram, *Curr. Sci.*, 1997, **73**, 1067.
6 R. P. Cheng, S. H. Gellman and W. F. DeGrado, *Chem. Rev.*, 2001, **101**, 3219.
7 D. Seebach, D. F. Hook and A. Glättli, *Biopolymers*, 2006, **84**, 23.
8 C. M. Goodman, S. Choi, S. Shandler and W. F. DeGrado, *Nat. Chem. Biol.*, 2007, **3**, 252.
9 T. Martinek and F. Fülöp, *Chem. Soc. Rev.*, 2012, **41**, 687.
10 C. Baldauf and H.-J. Hofmann, *Helv. Chim. Acta*, 2012, **95**, 2348.
11 C. Baldauf, R. Günther and H.-J. Hofmann, *J. Org. Chem.*, 2006, **71**, 1200.
12 R. Rezaei Araghi, C. Jäckel, H. Cölfen, M. Salwiczek, A. Völkel, S. C. Wagner, S. Wieczorek, C. Baldauf and B. Koksch, *ChemBioChem*, 2010, **11**, 335–339.
13 R. Rezaei Araghi, C. Baldauf, U. I. M. Gerling, C. D. Cadicamo and B. Koksch, *Amino Acids*, 2011, **41**, 733–742.
14 T. Sawada and S. H. Gellman, *J. Am. Chem. Soc.*, 2011, **133**, 7336.
15 K. Basuroy, B. Dinesh, N. Shamala and P. Balaram, *Angew. Chem.*, 2012, **124**, 8866.
16 J. A. Kritzer, J. D. Lear, M. E. Hodsdon and A. Schepartz, *J. Am. Chem. Soc.*, 2004, **126**, 9468.
17 J. A. Kritzer, N. W. Luedtke, E. A. Harker and A. Schepartz, *J. Am. Chem. Soc.*, 2005, **127**, 14584.
18 Y. Imamura, N. Umezawa, S. Osawa, N. Shimada, T. Higo, S. Yokoshima, T. Fukuyama, T. Iwatsubo, N. Kato, T. Tomita and T. Higuchi, *J. Med. Chem.*, 2013, **56**, 1443.
19 M. Crisma, F. Formaggio, A. Moretto and C. Toniolo, *Pept. Sci.*, 2006, **84**, 3–12.
20 K.-H. Lee, D. R. Benson and K. Kuczera, *Biochemistry*, 2000, **39**, 13737–13747.
21 R. Armen, D. O. Alonso and V. Daggett, *Protein Sci.*, 2003, **12**, 1145–1157.
22 M. Rossi, M. Scheffler and V. Blum, *J. Phys. Chem. B*, 2013, **117**, 5574.
23 R. R. Hudgins, M. A. Ratner and M. F. Jarrold, *J. Am. Chem. Soc.*, 1998, **120**, 12974.
24 R. R. Hudgins and M. F. Jarrold, *J. Am. Chem. Soc.*, 1999, **121**, 3494.
25 M. F. Jarrold, *Phys. Chem. Chem. Phys.*, 2007, **9**, 1659.
26 J. A. Stearns, O. V. Boyarkin and T. R. Rizzo, *J. Am. Chem. Soc.*, 2007, **129**, 13820.
27 M.-P. Gaigeot, *Phys. Chem. Chem. Phys.*, 2010, **12**, 3336.
28 M. Rossi, V. Blum, P. Kupser, G. von Helden, F. Bierau, K. Pagel, G. Meijer and M. Scheffler, *J. Phys. Chem. Lett.*, 2010, **1**, 3465.
29 S. Chutia, M. Rossi and V. Blum, *J. Phys. Chem. B*, 2012, **116**, 14788.
30 D. Seebach, K. Gademann, J. V. Schreiber, J. L. Matthews, T. Hintermann, B. Jaun, L. Oberer, U. Hommel and H. Widmer, *Helv. Chim. Acta*, 1997, **80**, 2033.
31 M. Rueping, J. V. Schreiber, G. Lelais, B. Jaun and D. Seebach, *Helv. Chim. Acta*, 2002, **85**, 2577.
32 C. Baldauf, R. Günther and H.-J. Hofmann, *Angew. Chem., Int. Ed.*, 2004, **43**, 1594.
33 K. Möhle, R. Günther, M. Thormann, N. Sewald and H.-J. Hofmann, *Biopolymers*, 1999, **50**, 167.
34 D. H. Appella, L. A. Christianson, I. L. Karle, D. R. Powell and S. H. Gellman, *J. Am. Chem. Soc.*, 1996, **118**, 13071.
35 D. H. Appella, L. A. Christianson, D. A. Klein, D. R. Powell, X. L. Huang, J. J. Barchi and S. H. Gellman, *Nature*, 1997, **387**, 381.
36 D. H. Appella, L. A. Christianson, I. L. Karle, D. R. Powell and S. H. Gellman, *J. Am. Chem. Soc.*, 1999, **121**, 6206.
37 J. M. Fernández-Santín, S. Muñoz-Guerra, A. Rodríguez-Galán, J. Aymami, J. Lloveras, J. A. Subirana, E. Giraltand and M. Ptak, *Macromolecules*, 1987, **20**, 62.
38 F. López-Carrasquero, C. Alemán, A. M. García-Alvarez, M. de Ilarduya and S. Muñoz-Guerra, *Macromol. Chem. Phys.*, 1995, **196**, 253–268.

39 P. R. Kemper, N. F. Dupuis and M. T. Bowers, *Int. J. Mass Spectrom.*, 2009, **287**, 46–57.

40 E. Mason and W. McDaniel, *Transport properties of ions in gases*, Wiley, 1988.

41 D. Oepts, A. van der Meer and P. van Amersfoort, *Infrared Phys. Technol.*, 1995, **36**, 297–308.

42 J. J. Valle, J. R. Eyler, J. Oomens, D. T. Moore, A. F. G. van der Meer, G. von Helden, G. Meijer, C. L. Hendrickson, A. G. Marshall and G. T. Blakney, *Rev. Sci. Instrum.*, 2005, **76**, 023103.

43 G. A. Kaminski, R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen, *J. Phys. Chem. B*, 2001, **105**, 6474.

44 J. W. Ponder, Tinker – software tools for molecular design, http://dasher.wustl.edu/ffe/, We used version 5.0 of the program and theversions of the force fields distributed with the package.

45 R. V. Pappu, R. K. Hart and J. W. Ponder, *J. Phys. Chem. B*, 1998, **102**, 9725–9742.

46 B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435.

47 X. Daura, K. Gademann, H. Schäfer, B. Jaun, D. Seebach and W. F. van Gunsteren, *J. Am. Chem. Soc.*, 2001, **123**, 2393.

48 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.

49 A. Tkatchenko and M. Scheffler, *Phys. Rev. Lett.*, 2009, **102**, 073005.

50 V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Comput. Phys. Commun.*, 2009, **180**, 2175.

51 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158.

52 A. Ambrosetti, A. M. Reilly, R. A. DiStasio and A. Tkatchenko, *J. Chem. Phys.*, 2014, **140**, 18A508.

53 M. Rossi, S. Chutia, M. Scheffler and V. Blum, *J. Phys. Chem. A*, 2014, **118**, 7349–7359.

54 J. B. Pendry, *J. Phys. C: Solid State Phys.*, 1980, **13**, 937.

55 C. Baldauf, K. Pagel, S. Warnke, G. von Helden, B. Koksch, V. Blum and M. Scheffler, *Chem. – Eur. J.*, 2013, **19**, 11224–11234.

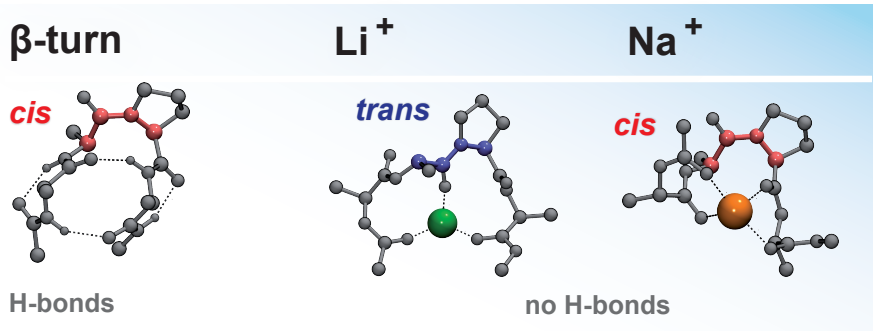
56 G. von Helden, M. T. Hsu, N. Gotts and M. T. Bowers, *J. Phys. Chem.*, 1993, **97**, 8182–8192.

57 The Bowers Group, Theoretical Collision Cross Sections, 2014, http://bowers.chem.ucsb.edu/theory_analysis/cross-sections/index.shtml.

58 R. A. DiStasio, R. P. Steele, Y. M. Rhee, Y. Shao and M. Head-Gordon, *J. Comput. Chem.*, 2007, **28**, 839–856.

59 A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12562–12566.

3.3 How cations change peptide structure



Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Koksch, B.; Blum, V.; Scheffler, M.

How Cations Change Peptide Structure

Chem. Eur. J. **2013** (19), 11224-11234.

DOI: 10.1002/chem.201204554

DOI: 10.1002/chem.201204554

How Cations Change Peptide Structure

Carsten Baldauf,*[a] Kevin Pagel,*[a] Stephan Warnke,[a] Gert von Helden,[a] Beate Koksch,[b] Volker Blum,*[a] and Matthias Scheffler[a]

Abstract: Specific interactions between cations and proteins have a strong impact on peptide and protein structure. Herein, we shed light on the nature of the underlying interactions, especially regarding effects on the polyamide backbone structure. This was done by comparing the conformational ensembles of model peptides in isolation and in the presence of either Li^+ or Na^+ by using state-of-the-art density-functional theory (including van der Waals effects) and gas-phase infrared spectroscopy. These monovalent cations have a drastic effect on the local backbone conformation of turn-forming peptides, by disruption of the hydrogen-bonding networks, thus resulting in severe distortion of the backbone conformations. In fact, Li^+ and Na^+ can even have different conformational effects on the same peptide. We also assess the predictive power of current approximate density functionals for peptide–cation systems and compare to results with those of established protein force fields as well as high-level quantum chemistry calculations (CCSD(T)).

Keywords: density-functional calculations · hydrogen bonds · IR spectroscopy · protein folding · protein structures

Introduction

As early as 1912, Paul Pfeiffer systematically studied the crystallization of short Ala- and Gly-containing peptides from aqueous solution in the presence of alkali salts[1] and postulated that Li^+ exhibits a higher affinity ("Additionsfähigkeit") for peptides than Na^+ and K^+.[2] Indeed, calorimetric studies revealed high interaction enthalpies for the interactions of a series of peptides with Li^+,[3] values that were in the range of solvation enthalpies of peptides. These strong interactions are in practice used to increase the proportion of *cis* prolyl peptide bonds from 10% to 70% through the addition of Li salts in biochemical activity assays of peptidyl prolyl *cis–trans* isomerases.[4] NMR studies of cyclic peptide cyclosporine A (CysA) in organic solvents revealed that Li^+ inhibits the formation of hydrogen bonds and induces unusual backbone conformations.[5,6] One hundred years after Pfeiffer's work, Garand et al.[7] studied the noncovalent interactions of a non-natural peptide-based catalyst by means of gas-phase infrared (IR) spectroscopy. When protonated, the polyamide backbone of the molecule forms intramolecular hydrogen bonds; however, when it is sodiated, there is an apparent complete absence of hydrogen bonds owing to the presence of interactions between the carbonyl groups and Na^+. Both CysA[5,6] and the peptide-based catalyst[7] form narrow turn-like backbone loops, which are well suited to accommodate a cation. Such turns are normally at the outside of globular proteins, where they are exposed to the surrounding medium. Herein, we investigate the atomistic and electronic basis of cation–peptide interactions in turn-forming peptides. The focus of the study is on proline-containing peptides, in which such interactions are expected to have pronounced conformational effects owing to the possible *cis* and *trans* states of the prolyl-peptide bond.[8] This study is based on accurate conformational predictions by using first principles (density-functional theory) in a synergistic combination with gas-phase IR spectroscopy to validate the results.

Structure formation and dynamics in proteins can be primarily attributed to the rotation of the N–C_α and C_α–C bonds, represented by the backbone torsion angles ϕ and ψ, respectively (Figure 1A). This conformational ϕ/ψ space is well described by a Ramachandran diagram,[9] which is used, for example, in the statistical evaluation of high-resolution X-ray data (Figure 1B).[10] The shaded areas are referred to as allowed conformational regions and can be associated with characteristic secondary structure types (Figure 1B).

[a] Dr. C. Baldauf, Dr. K. Pagel, Dipl.-Phys. S. Warnke, Dr. G. von Helden, Dr. V. Blum, Prof. Dr. M. Scheffler
Fritz-Haber-Institut der Max-Planck-Gesellschaft
Faradayweg 4-6, 14195 Berlin-Dahlem (Germany)
E-mail: baldauf@fhi-berlin.mpg.de
pagel@fhi-berlin.mpg.de
blum@fhi-berlin.mpg.de

[b] Prof. Dr. B. Koksch
Institut für Chemie und Biochemie - Organische Chemie
Freie Universität Berlin
Takustr. 3, 14195 Berlin-Dahlem (Germany)

Supporting information for this article contains details of the simulation setup, experimental procedures, energy hierarchies for AAPA+Li^+ geometries in the presence and absence of the respective cation at different levels of theory, backbone torsion angles of low-energy conformers of AAPA and ADPA in isolation and in the presence of Li^+ and Na^+, and Cartesian coordinates of the structures; this information is available on the WWW under http://dx.doi.org/10.1002/chem.201204554.

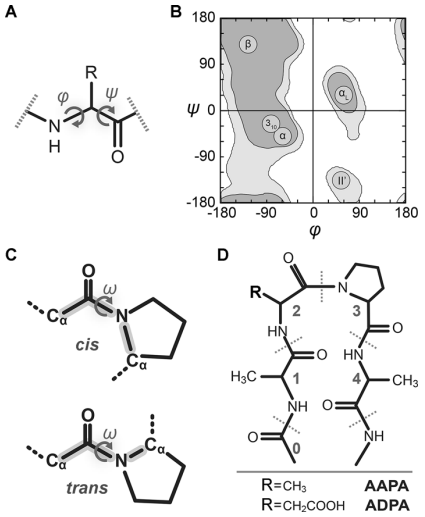


Figure 1. (A) The backbone torsion angles, φ and ψ, of the residues of a polypeptide chain. (B) Backbone torsion angles illustrated by a Ramachandran plot, based on data from ref. [10]; labels highlight characteristic secondary-structure types: the β region in the 2nd quadrant, the 3_{10} and the α-helical region in the 3rd quadrant, and the left-handed α and the βII' region in the 1st and 4th quadrants, respectively. (C) The *cis* and *trans* state of the prolyl-peptide bond. (D) The model peptides, AAPA and ADPA, shown in a schematic βVI-turn conformation.

The double-bond character of the peptide bond hinders free rotation and allows for two distinct conformations. In general, the *trans* conformation is almost exclusively observed with an apparent high barrier for its conversion into the *cis* form.[11,12] A significant fraction of *cis* conformation is only observed for the prolyl peptide bond.[13] In proline, the *cis* and *trans* forms (Figure 1 C) are close in energy because the C_β of the preceding residue encounters a carbon atom of proline (C_α or C_δ) in both states. A *cis* peptide bond, usually preceding a proline residue, is a feature of so-called type β-VI turns (Figure 1 D).[14,15] This notation dates back to work of Venkatachalam, according to which β turns share the feature of a hydrogen bond between residues $i+3$ and i and are further classified by the backbone torsion angles ϕ and ψ of the residues $i+1$ and $i+2$.[16] The β turns of the protein backbone allow for a 180° reversal of the direction of structure propagation within four consecutive residues of a polypeptide chain. Similarly, Hutchinson and Thornton classify β turns according to ranges of values for the backbone torsion angles ϕ and ψ, thus giving eight well-defined classes (I, I′, II, II′, VIa1, VIa2, VIb, and VIII) and a miscellaneous type IV.[17,18] Very prominent are the common (type I) and glycine (type II) turn and their inverse counterparts, I′ and II′. The special β-turn types VIa and VIb have a *cis* peptide bond between central residues $i+1$ and $i+2$; these β-turn types frequently feature proline in position $i+2$.[14,15]

In this study, we make use of the characteristic of proline-containing peptides that allows for the formation of *cis* and *trans* peptide bonds as a potential strong "conformational signal" triggered by the peptide–cation interaction. Indeed, Seebach and co-workers reported ion-induced conformational effects on peptide structure to be especially pronounced in the proximity of proline.[8] Kunz et al. investigated a systematic series of proline-containing peptides using NMR spectroscopy and found that peptides containing an Asp–Pro sequence exhibit *cis*/*trans* ratios that are in opposition to those of all other sequences studied.[19] Therefore, we investigated the sequence, AXPA (Figure 1 D), where P is the single letter code for proline and X is either alanine (A) or aspartate (D), thus allowing us to differentiate the contribution of pure cation–backbone interactions and cation–side-chain interactions to the peptide backbone conformation. Peptides were designed so as to avoid structure-perturbing labels and the peptide termini were protected with acetyl and aminomethyl groups (Figure 1 D) to embed the sequence in a protein-like chain structure, thus avoiding end-group effects, that is, zwitterion formation.

Results and Discussion

We used a combination of exhaustive conformational searches from first principles and both theoretical and experimental gas-phase IR spectroscopy. Such investigations of isolated peptides in the gas phase offer an unbiased view of structure-formation trends intrinsic to the molecule, a strategy that is successful for charged and uncharged amino acids and peptides.[20–30] By the stepwise addition of perturbing contributions, in this case, the presence of cations, we aim to determine the main contributions to protein secondary structure formation in a bottom-up approach. The success of such an approach is critically linked to the quality of the description of the potential-energy surface of the system under investigation. We employ density-functional theory (DFT) in the generalized-gradient approximation with the Perdew–Burke–Ernzerhof (PBE) functional.[31] Van der Waals dispersion interactions are included through a pairwise C_6R^{-6} term for which the C_6 coefficients are derived from the self-consistent electron density, referred to as PBE+vdW.[32] Our use of rather accurate, but computationally efficient approximate DFT is justified by the high-level benchmarks we present in the Computational Methods Section.

Conformational analysis: The theoretical conformational analysis of the short peptide AAPA (Figure 1 D) is challenging. Hypothetically, discretizing the backbone torsion angles with a 30-degree grid and assuming two possible states (*cis* and *trans*) for the peptide bonds would formally result in roughly 35 million conformations for evaluation. To deal with such a large conformational space, we resort to a exhaustive basin-hopping search of the potential-energy surfaces (PES) of conventional protein force fields (either OPLS-AA[33] or AMBER99[34]). We employ the TINKER 5 scan routine[35] in an in-house parallelized version. To achieve a reliable and parameter-free description, we then follow up with a large set (700 to 1800 per peptide–cation system) of

PBE+vdW post-relaxation calculations as a second computational step.

Figure 2 shows our results for AAPA in isolation. The lowest-energy structure of the PES, a βVI turn with a *cis* prolyl peptide bond, also has the lowest free energy in the harmonic approximation. Two alternative βVI turns are 4.5 and 8.3 kJ mol^{-1} higher in $\Delta F_{300\,K}$. The most stable conformer with a *trans* peptide bond is a βII′ turn with $\Delta E =$ 2.8 kJ mol^{-1}. Harmonic free-energy contributions add a further penalty to the structure, yielding $\Delta F_{300\,K} = 8.8$ kJ mol^{-1}. In these cases, the maximum number of four backbone hydrogen bonds is formed. In a DFT study of Ac-Ala-Pro-NMe, Byun et al. also predicted a βVI turn as the most stable conformer in the gas phase.[36] A comparable βII′ turn

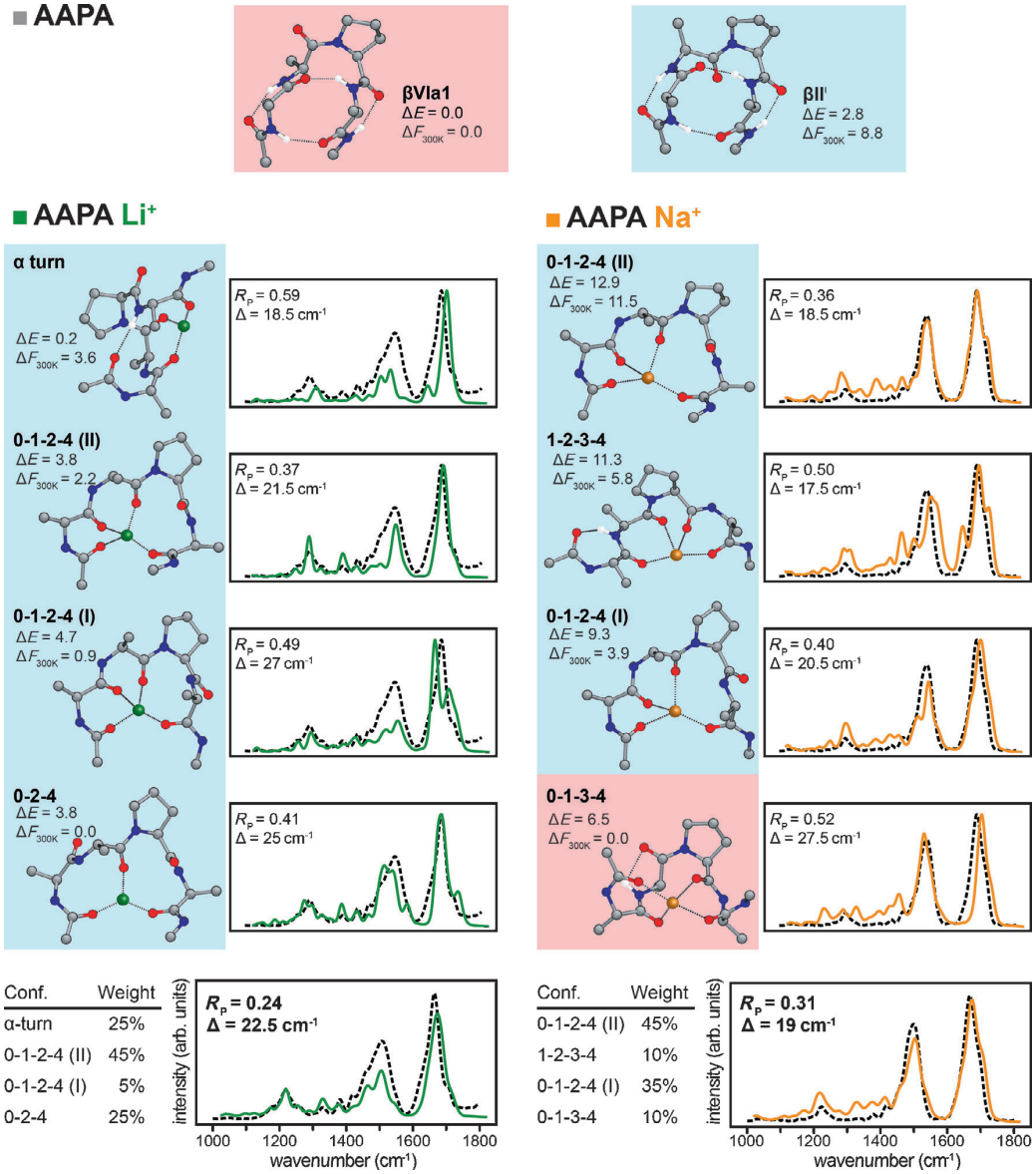


Figure 2. Low free-energy ensembles for AAPA in isolation and in the presence of Li^+ and Na^+, as obtained using exhaustive conformational searches. Potential energies (ΔE, in kJ mol^{-1}) and harmonic free energies ($\Delta F_{300\,K}$, in kJ mol^{-1}) are given. The criterion applied to select the structures shown is lowest free energy, except for conformer 0-1-2-4(II), which was selected because of its relationship with 0-1-2-4(I): both conformers can be interconverted by a backbone crankshaft movement. *Cis* and *trans* conformers are indicated using a red and blue background, respectively. The simulated IR spectra are shown as continuous lines for the individual conformers as well as for the assumed ensemble of conformers (lowest row); the experimental IR spectra are shown as dashed lines. The tables show the relative proportion of each conformer within the respective mixed simulated spectra. Simulated spectra were shifted along the energy axis by a value Δ for an optimal Pendry reliability factor, R_P. The atom colors: C is gray, N is blue, O is red, H is white, Li is green, and Na is orange. Hydrogen atoms are omitted for clarity except where they form part of a hydrogen bond.

was not among the lower-energy conformers of this shorter peptide. The lowest minima of the PES of ADPA (up to 0.7 kJ mol^{-1}) are again βVI turns (Figure 3); the next lowest in energy are two other conformers with relative potential energies of 2 and 4 kJ mol^{-1}. For the conformer that is 4 kJ mol^{-1} higher than the lowest energy conformer, the Asp side chain forms hydrogen bonds with the NH and C=O groups of residue Ala4; this conformer resembles the shape of a β turn, hence we refer to it as SC-β. For the ADPA–cation systems, we confirmed by mass spectrometry (for the ADPA–cation systems) that the Asp side chain is protonated in our experimental setup. Consequently, the Asp side chain is modeled in the protonated neutral state. Harmonic free-energy contributions make SC-β the preferred structure type by approximately 2 kJ mol^{-1}. Notably, the lowest free-energy structure of AAPA features a *cis* prolyl peptide bond, whereas the respective bond in the lowest free-energy structure of ADPA is *trans* configured (Figures 2 and 3).

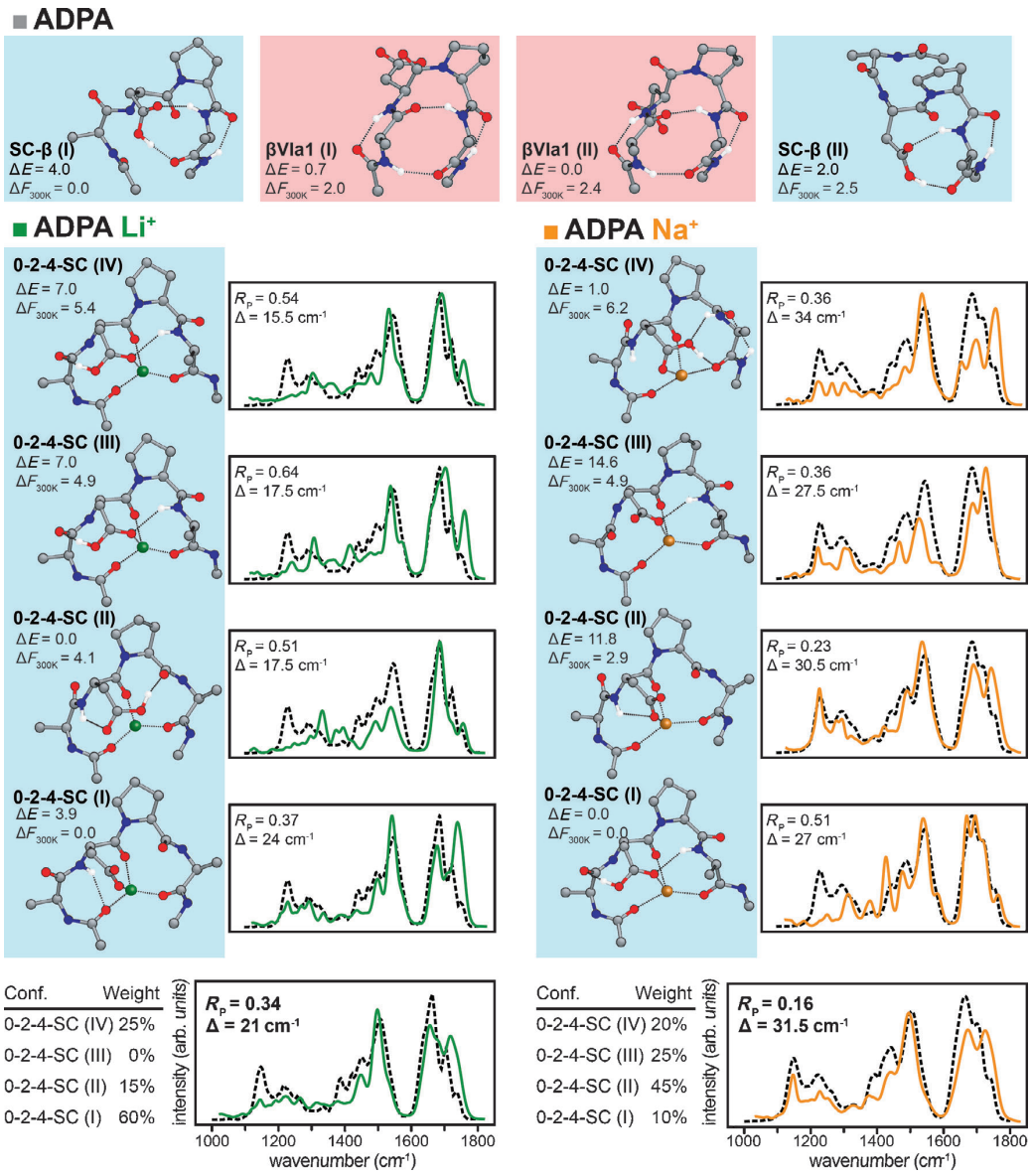


Figure 3. Low free-energy conformers of peptide ADPA in isolation and in the presence of Li^+ and Na^+. Potential energies (ΔE, in kJ mol^{-1}) and harmonic free energies (ΔF_{300K}, in kJ mol^{-1}) are given. *Cis* and *trans* conformers are indicated using a red and blue background, respectively. The experimental and simulated IR spectra are shown as dashed and continuous lines, respectively. The simulated spectra were mixed to account for a conformational ensemble (lowest row). The tables show the relative proportion of the conformers within the mixed spectra. Simulated spectra were shifted by a value Δ along the energy axis for an optimal Pendry reliability factor, R_P. The atom colors: C is gray, N is blue, O is red, H is white, Li is green, and Na is orange. Hydrogen atoms are omitted for clarity except where they form part of a hydrogen bond.

The attraction between backbone carbonyl groups and either Li^+ or Na^+ induces structures that differ substantially from the conformers in the absence of such cations: The hydrogen-bonding networks in the low-energy conformers are disrupted (Figures 2 and 3) and the backbone conformations deviate from those of the isolated peptide. This finding is in line with the above-mentioned results for CysA in apolar Li salt solutions[5,6] and the sodiated peptide-based catalyst in the gas phase.[7] For isolated peptides AAPA and ADPA, the backbone torsion angles ϕ and ψ of the low free-energy conformers ($\Delta F_{300\,K} < 6\ kJ\,mol^{-1}$) are within the allowed regions of the Ramachandran plot (Figure 4). The single outli-

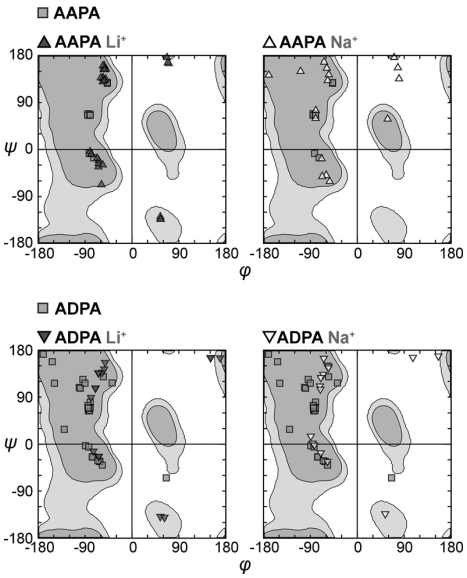


Figure 4. The backbone torsion angles, φ and ψ, of the low free-energy conformers ($\Delta F_{300\,K} < 6\ kJ\,mol^{-1}$) for AAPA and ADPA in isolation (light-gray squares), with Li^+ (dark-gray triangles), and with Na^+ (white triangles) were plotted on top of an empirical contour plot (ref. [10]).

er in the fourth quadrant of the plot for ADPA represents the C-terminal residue, Ala4, of a conformer with $\Delta F_{300\,K} = 4.8\ kJ\,mol^{-1}$. The different possible rotameric states of the Asp side chain prefer different backbone conformations. This leads to more possible backbone conformations (data points) compared to AAPA. The cation–peptide interaction imprints ϕ/ψ combinations (backbone conformations) that differ substantially from those of the unperturbed peptides. Some of them with still low relative free-energy values ($0.9\ kJ\,mol^{-1}$ for AAPA + Li^+ to $2.6\ kJ\,mol^{-1}$ for AAPA + Na^+) are even located outside of the allowed regions of the Ramachandran plot (Figure 4). These outliers do not represent residues at the termini but rather central residues Ala2 or Asp2, which govern the overall structure of the peptides. Interestingly, the cation effects on the two peptides differ. The conformational ensembles of AAPA with Li^+ and Na^+ are different (Figures 2 and 4), whereas those of lithiated and sodiated ADPA are very similar (Figures 3 and 4).

A canonical turn structure, type βII′ (not shown), is the lowest PES minimum of AAPA + Li^+. The second most stable minimum, with $\Delta E = 0.2\ kJ\,mol^{-1}$, is an α turn (Figure 2). Here, the consideration of harmonic free-energy contributions changes the picture dramatically and unusual backbone conformations become dominant. In the lowest free-energy conformer, the Li^+ ion is coordinated by three backbone carbonyl groups of residues 0, 2, and 4 (Figure 2). The conformers resulting form the peptide–cation interactions are named according to the numbers of the interacting oxygen atoms; for example, 0-2-4. In cases of multiple conformations with the same interaction pattern, these are distinguished by roman numerals, which increase in line with the free energy of the conformers. Up to four out of a possible of five binding partners (backbone carbonyl groups) are sterically possible (conformers 0-1-2-4 with $\Delta F_{300\,K} = 0.9$ or $2.2\ kJ\,mol^{-1}$). Although the search for minima does not yield information on the actual barriers connecting different conformers, their high structural similarity suggests dynamic interconversion at a finite temperature. For AAPA + Li^+, the preferred conformation of the prolyl-peptide bond changes from *cis* to *trans*.

Na^+ binding to AAPA results in a similar behavior: canonical structure types (βVI, βII′, α) are lowest in potential energy whereas structures with both unusual backbone conformations and carbonyl groups pointing towards Na^+ are most stable when harmonic free-energy contributions are considered. However, there are substantial differences between the Na^+ and Li^+ adducts: the low free-energy ensemble of the former is more diverse and the central peptide bond of the lowest free-energy conformer 0-1-3-4 is *cis*. In the case of AAPA + Na^+, the second lowest free energy conformer (A1661, $\Delta F_{300\,K} = 2.6\ kJ\,mol^{-1}$) is not shown in Figure 2. This conformer was ruled out because it was proven unstable in the subsequent AIMD simulations for IR spectra (see section below). Please refer to the Supporting Information for all calculated free energies. Instead, we consider 0-1-2-4(II) as fourth conformer, which is much higher in free energy. Interestingly, these two conformers of AAPA + Na^+, 0-1-2-4(I) and 0-1-2-4(II), are almost identical other than the orientation of the peptide bond between Pro3 and Ala4 (Figure 2). This peptide bond is not involved in any interactions and can thus rotate by a concerted motion of adjacent torsion angles ψ and ϕ, a so-called backbone crankshaft rotation.[37,38] During the equilibration AIMD simulations at 300 K, which were carried out in preparation for the simulations to obtain IR spectra, this interconversion between 0-1-2-4(I) and 0-1-2-4(II) was indeed observed within the 10 ps simulation time. The subsequent evaluation of IR spectra also suggests the presence of 0-1-2-4(II) in the experimentally observed conformational ensemble.

In ADPA, the dominant interaction pattern is the complexation of either Li^+ or Na^+ by the backbone oxygen atoms 0, 2, 4 and the Asp side-chain carboxyl group. All conformers in the low-energy range are highly similar and feature no *cis* prolyl peptide bonds (Figure 3). As discussed

above on the basis of the Ramachandran plot (Figure 4), the effects of the cations on AAPA and ADPA differ. Li^+ enforces a *trans* conformation of the prolyl peptide bond of AAPA whereas Na^+ enforces the *cis* conformation (Figure 2). For ADPA, no such selectivity for the cation is observed. With either Li^+ or Na^+ attached, similar structure types with *trans* prolyl peptide bonds are preferred (Figures 3 and 4).

Infrared spectroscopy: To corroborate our structural findings, we obtained gas-phase infrared multi-photon dissociation (IRMPD) spectra, which reflect the same clean-room conditions as used in our simulations. Spectra were recorded from 1000 to 1800 cm^{-1} at the free electron laser facility FELIX[39] using a Fourier-transform ion cyclotron (FT-ICR) mass spectrometer.[40] The experimentally obtained spectra for lithiated and sodiated AAPA and ADPA are shown in Figure 2 and Figure 3. For AAPA, significantly different spectral signatures were obtained for the Li^+ and Na^+ complexed forms, a result that is in line with the results of the conformational analysis described in the previous section. On the other hand, for ADPA, very similar spectra were recorded for both cation complexes.

To allow for a quantitative theory–experiment comparison, IR spectra including anharmonic effects were computed from Born–Oppenheimer ab initio molecular dynamics (AIMD) simulations. The systems were equilibrated using 10 ps of AIMD simulations at 300 K. Subsequently, the microcanonical ensemble was sampled using up to 40 ps long AIMD simulations at constant energy from which IR spectra were derived.[41,27] IR spectra of polyamides feature characteristic bands of high intensity (like the amide I and II regions, 1400–1700 cm^{-1}) but also regions with low intensity (below 1400 cm^{-1}) and fingerprint characteristics. Visual inspection does not allow for a quantitative assessment and is, similar to a simple square of intensity comparison, easily biased by the high-intensity peaks. For a quantitative comparison between the calculated and experimental spectra, we employed the reliability factor R_P, which was introduced by Pendry to the field of low-energy electron diffraction,[42] and an implementation described by Blum and Heinz.[43] For R_P, peak positions are more important than peak intensities, a characteristic that fits the requirements we face herein, especially because we are comparing experimental action spectra and theoretical absorption spectra. Values for R_P range from 0 (perfect agreement) via 1 (no correlation) to 2 (perfect anti correlation). Intensities of the spectra were normalized to 1 and rigidly shifted (not scaled) with a value Δ along the energy axis to account for deviations owing to a systematic mode softening by the density functional we use.[44,27] When comparing the calculated IR spectra of single conformations to the experimental IR spectra we observe only modest agreement (see individual spectra in Figures 2 and 3). Previous studies have shown similar behavior owing to conformational ensembles for peptides in the gas phase at finite temperature.[22–27] Furthermore, the energy differences of the low free-energy conformers lie within the uncertainty of the employed method, as discussed in the section called 'Benchmarks' below. Consequently, an ensemble of conformations is assumed. By mixing the individual theoretical spectra in 5% steps, the R_P for the respective experimental spectrum is optimized. This results in a much better agreement of simulated and experimental spectra of the peptides AAPA and ADPA in complex with single Li^+ or Na^+ ions (Figures 2 and 3). The agreement between predicted and experimental spectra of AAPA+Li^+ and ADPA+Li^+ and between those of the corresponding complexes of Na^+, especially regarding the fine structure below 1400 cm^{-1}, is gratifying. We note for completeness that the spectra for the protonated peptides (not shown) are rather different in appearance, suggesting very different structural effects compared to those induced by the presence of heavier cations.

In a naive way, a correlation between the free-energy estimates in the harmonic approximation and the abundances of the individual spectra in the resulting mixed spectrum could be expected. However, this would be too much to expect for several reasons:

(1) The PBE+vdW method we use is rather accurate as illustrated by the benchmark calculation presented below; however, the systems under investigation here are also large (56 to 60 atoms). The lowest free-energy minima discussed herein are still within the range of uncertainty in the values of the relative (free) energies.
(2) The experimental data base which we are comparing to multiple theoretical spectra is relatively small; fitting many parameters to a small data set has well known limitations.[42] The use of multiple spectra, therefore, is strictly only a consistency check. The spectra of just a single conformer are not sufficient to explain the observed IR spectra. In contrast, the use of spectra of multiple conformers yield a much more consistent description of the spectra, in line with several conformers of similar free energy. This is the primary qualitative statement that we can derive from the experiment–theory comparison.
(3) The here employed free-energy model neglects anharmonicity as well as the entropic effects of a possibly greater accessible conformational space (dynamic interconversion in the case of low barriers) of specific conformers. That the latter can be of special importance is illustrated by the crankshaft rotation discussed above for the two conformers, 0-1-2-4(I) and 0-1-2-4(II) of AAPA+Na^+. That both conformers can interconvert shows the extent to which the anharmonic nature of the potential-energy surface can play a role. In fact, in the combination of the four individual spectra that shows the best agreement with the experimental spectrum (Figure 2), 0-1-2-4(II) is the predominant conformer with 45% of the total population of the species. Again, conformer 0-1-2-4(II) is structurally and dynamically closely related to the 0-1-2-4(I) conformer with the lowest harmonic free energy. This result illustrates the

limits of the harmonic free-energy assignment to potential-energy minima at room temperature, which neglects such conformational and dynamical effects. The result furthermore illustrates the limitations of interpreting IR spectra by using a combination of individual and isolated conformers.

Overall, on the one hand, the accuracy of the harmonic approximation to the free energy is limited by the dynamic character of such molecular systems at finite temperature; on the other hand, the IRMPD spectroscopy setup we use here is limited in its resolution, especially regarding the separation of individual conformers. However, we can unambiguously predict minima by first-principles theory and validate the results by room-temperature IR spectroscopy (keeping the differences of static harmonic free-energy minima and actual room-temperature molecules in mind). The observed cation–peptide effects were certainly qualitatively corroborated by both approaches.

Microsolvation of a peptide–cation complex: In the introduction, we mentioned the presence of turn sequences, which are mainly located at the surface of proteins and thus exposed to the aqueous environment. In this section, a qualitative picture of how the interaction between the peptide backbone and the cation can compete with solvation of the cation is given. AIMD simulations were performed for AAPA+Li^+ alone and with a few water molecules. For the setup of the latter system, 18 water molecules were accommodated within a sphere of radius 4.5 Å around the Li^+ ion. For comparison, Li^+ embedded within both 4 and 10 water molecules was also studied. We characterized the interaction between the Li^+ ion and either the respective oxygen atoms of the peptide backbone or of first-solvation-shell water molecules by the Li^+–O distance and by the O–Li^+–O angle (Figure 5). Previous ab initio studies predict a coordination number of 4 for Li^+ in water.[23,45,46] Consistent with these studies, the cation is complexed by 4 backbone carbonyl groups in, for example, conformer 0-1-2-4(I). During a 100 ps AIMD trajectory at 330 K (Nosé–Hoover thermostat), the Li^+–O distance fluctuates around 1.9 Å, the O-Li^+-O angle distribution is broad, thus indicating the nonideal tetrahedron formed by the interacting carbonyl oxygen atoms. The microsolvation of AAPA+Li^+ within 18 water molecules results in a slight change of the binding site within a few picoseconds: a water oxygen atom substitutes for the backbone C=O group of Ala1. The cation interacts with the three backbone carbonyl oxygen atoms of AAPA and the same water molecule (Figure 5) for the whole 90 ps of remaining AIMD simulation time. As a result, a virtually ideal binding site is formed, characterized by an almost symmetric distribution of the O-Li^+-O angle around the ideal tetrahedral angle of 109.5°. For Li^+ immersed within a small water cluster (either 4 or 10 water molecules), the Li^+–O distance distribution peaks around 2.0 Å. Remarkably, the distribution of the tetrahedron angles O-Li^+-O is multimodal again, accounting for alternative (and less populated) geometries of the Li^+ complex involving 3 or, in the case of the 10 water molecules with Li^+ cluster, even 5 water molecules in the first solvation shell. For now, we can at least qualitatively say that AAPA is able to form an ideal interaction shell that seems to be able to compete with water solvation. A fully correct answer could be given on the basis of free-energy differences from simulations with fully solvated systems. Such simulations are standard for force-field approaches, yet they are computationally very demanding at the level of theory we employ here. A rigorous assessment is thus beyond the scope of this article.

Conclusion

Starting from the isolated peptides that adopt either canonical turn structures (AAPA) or turn-like conformations with hydrogen bonds between the side chains and the backbones

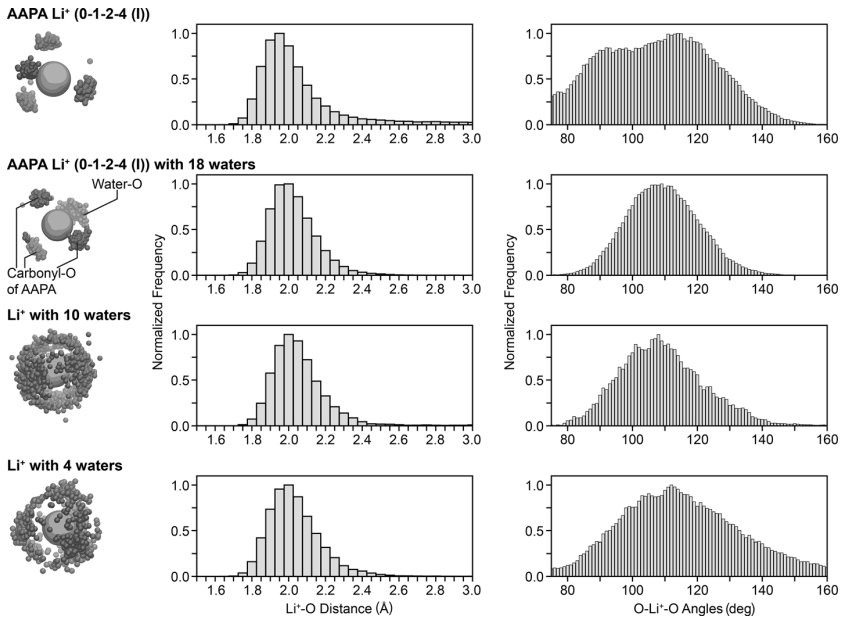


Figure 5. The first solvation shell around a Li^+ ion. Interactions are formed with the oxygen atoms of water molecules or those of backbone carbonyl groups. The histograms were derived from AIMD simulation of different length (20 to 100 ps) and the counts were normalized to 1.

(ADPA), we show the drastic effect of cations on the local secondary structure of peptides: the cation interacts with most of the backbone carbonyl groups and, as a result, completely breaks the local hydrogen-bonding network. This leads to distortions of the peptide backbone and results in conformations with backbone torsion angles ϕ and ψ that are, in part, outside of the allowed regions of the Ramachandran plot (Figure 4). Consequently the question of the range of such ion-induced disruptions arises. Ohanessian and co-workers studied[47,48] polyglycines with a chain length of 2 to 8 residues in complex with Na^+ by simulation and gas-phase IR spectroscopy: for sequences of up to 7 glycine residues, the contact number between the cation and backbone C=O groups is maximized and no hydrogen bonding was observed. With the Gly_8 peptide, backbone hydrogen bonding appeared again in the form of γ- and β turns. Glycine, owing to the lack of a side chain, is a very special case among the canonical amino acids. As a contrast, the helical secondary structure of sodiated polyalanine (8–12 residues) is not broken in the gas phase. Here, the Na^+ ion is attached to the C terminus.[49,50] The importance of considering the effect of side-chain functionalities is highlighted by the sequence dependence of the cation effects we observe. The conformational preferences of AAPA with either Li^+ or Na^+ differ drastically in the *trans*/*cis* state of the central prolyl peptide bond. Noskov and Roux investigated the selectivity of the ion-coupled transporter LeuT. Two Na^+ binding sites (NA1 and NA2) show differences in the Li^+/Na^+ selectivity: NA1 appears to be rather flexible and exhibits no selectivity for one cation over the other as it adapts to the different ionic radii; for NA2, a limited selectivity is apparently induced by a "snug-fit" mechanism (the rigid NA2 interaction site is unable to adapt to different ionic radii).[51] Similarly, lowest free-energy structure 0-1-3-4 of AAPA + Na^+ may be too rigid to adapt to the Li^+ cation, because only backbone carbonyl groups can be involved in the interaction. With ADPA, the Asp side chain prevents such conformation selectivity.

Our findings might even help to understand a basic biochemical principle: in 1888, Hofmeister published an article[52] that laid the basis for a sorting of cations and anions according to their effect on the solubility of biomolecules, colloids, and functional polymers. Although it was believed that the underlying effects can be explained solely by bulk properties stemming from the solvent–ion interactions,[53] evidence was found that most effects of ions on water structure are limited to the first solvation shell[54] and that specific ion–solute interactions can be expected to contribute substantially.[55] These effects are especially clear at high salt concentrations as shown by Dzubiella and co-workers, who employed classical MD simulations,[56–58] and later, experimental approaches.[59] They demonstrated that the perturbing effect of ions on peptide structure results from the breaking of secondary-structure-specific hydrogen bonds in the backbone. Our own findings point to a similar direction, as we have shown here how cations can substantially change the backbone structure of a (bio)polymer. These interactions are not necessarily stable over a very long time range, but our exploratory AIMD simulations suggest time ranges at least in the tens to hundreds of picoseconds. Dzubiella described long-lived loop conformations that are stable over 10 to 20 ns in classical MD trajectories.[56] Similar to the specific interactions between anions and the amide-bond-containing polymers of *N*-isopropylacrylamide described by Cremer and co-workers,[60] we show here the possible interactions between small monovalent cations and peptides and highlight their significant effect on local peptide structure. These effects could be one of the drivers behind the Hofmeister salt effects on proteins.

Computational methods

Scans of the PES were performed with an exhaustive basin-hopping search and conventional protein force fields (either OPLS-AA[33] or AMBER99[34]). We employ the TINKER 5 scan routine[35] in an in-house parallelized version. The required methods to perform DFT-based simulations, including geometry optimization, computation of harmonic vibrations, and ab initio Born–Oppenheimer molecular dynamics (AIMD), are incorporated in the FHI-aims code, which provides an efficient and accurate all-electron description based on numeric atom-centered orbitals.[61] We discuss fully relaxed conformations at the PBE + vdW level and their relative potential energies (ΔE) and relative harmonic free energies at 300 K ($\Delta F_{300\,K}$), all computed with tight convergence settings and an accurate tier-2 basis set.[61] High-level quantum-chemical benchmark calculations, that is, relaxations at the MP2 level of theory and coupled-cluster calculations with singles, doubles, and perturbative triples (CCSD(T)), were performed with the ORCA quantum-chemistry program;[62] CCSD(T) energies extrapolated to the complete basis-set limit (CBS) were obtained by a method described by Truhlar,[63] employing the Dunning basis sets cc-pVDZ and cc-pVTZ.[64]

Benchmarks: We assessed the predictive power of the DFT approximations applied here by benchmarks in two directions with respect to the approximation level: we compare these approximations to high-level quantum-chemistry calculations at the CCSD(T) level of theory extrapolated to the complete basis-set limit. On the other hand, we assess the quality of the force-field description of cation–peptide interactions in comparison to approximate DFT at the PBE + vdW and PBE0 + vdW levels.

Comparing electronic-structure theory methods: There have been several assessments of the accuracy of the PBE + vdW level of theory applied to a variety of systems, such as peptides,[65] weakly bound metal–phtalocyanine systems,[66] and ionic and semiconductor solids.[67] A previous assessment of the accuracy of PBE + vdW level of theory for peptide systems, for the conformational-energy hierarchy of Ace-Ala-NMe and Ace-Ala_3-NMe, shows mean absolute errors (MAE) below 2 kJ mol^{-1} in comparison to CCSD(T) energies.[65] Herein, we investigate cation–peptide systems and thus reassess the accuracy of our DFT-based predictions. We employ high-level quantum chemical theory on the conformational-energy hierarchy of Ac-Ala-NMe + Li^+. A conformational analysis identified five local minima (Figure 6) within a potential-energy range of 35 kJ mol^{-1} at the MP2/cc-pVTZ level of theory.[64,68] The cation closes a 7-membered pseudocycle through interaction with the oxygen atoms of the backbone carbonyl groups. The orientation of the methyl groups relative to the pseudocycle plane defines them as either equatorial (Figure 6; **1**, **3**, **5**) or axial (Figure 6; **2**, **4**). In addition, an important characteristic of our actual systems of interest (AAPA and ADPA) is present here as well: in lower-energy conformers **1** and **2**, the C-terminal peptide bond is *trans* configured, which is in contrast to the other conformers with a C-terminal *cis* peptide bond.

Relative energies at the CCSD(T), PBE + vdW, and PBE0 + vdW levels of theory were compared by estimating the mean absolute error

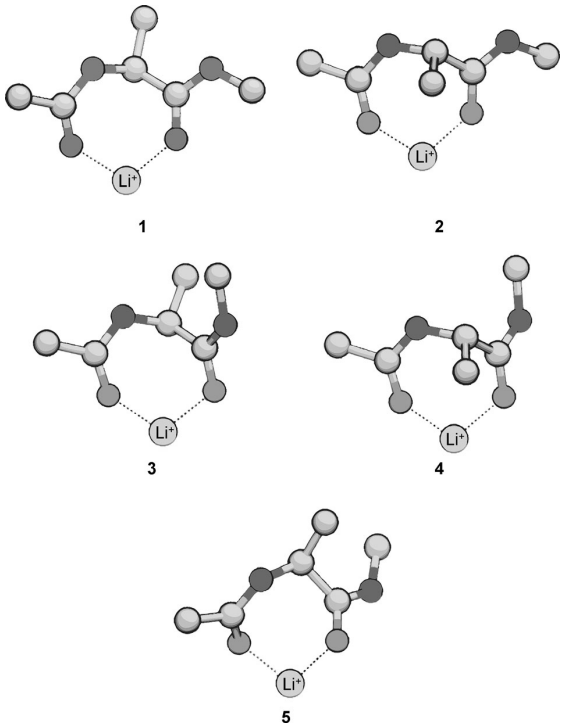


Figure 6. Lowest-energy conformers of Ace-Ala-NHMe+Li^+, fully relaxed at the MP2/cc-pVTZ level of ab initio theory. Hydrogen atoms were omitted for clarity; dashed black lines show the oxygen–lithium interactions.

(MAE $= \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i|$). As an additional method, we also include the frequently used B3LYP method (no vdW correction). For much larger systems, the missing description of dispersion effects represents a deficiency in the description of conformational-energy hierarchies. The DFT methods give low uncertainties (see Table 1), well within the often stressed "chemical accuracy" of 1 kcal mol^{-1} (4.2 kJ mol^{-1}). The conformational changes upon relaxation with DFT are negligible, as indicated by the low maximal RMSD value of 0.38 Å (Table 1). We note that only small contributions of the van der Waals correction can be expected for molecular systems of this size. Of the tested DFT approaches, PBE0+vdW gives the best agreement with the benchmark calculations, as it is obvious from the MAE and the average RMSD (see Table 1). However, for a large-scale conformational screening and the extensive molecular dynamics simulations we undertake in this study, PBE+vdW strikes a perfect balance between computational cost and accuracy.

Standard protein force fields versus electronic structure theory: When comparing the results of different standard protein force fields and DFT, we observe dramatic discrepancies in the conformational hierarchies. Such force fields were parameterized for the solvated state, whereas our assessment was based in the gas phase. Nonetheless, these force fields are also frequently used for conformational investigations irrespective of the environment. Consequently, their performance in vacuo is of interest. As a reference, we employ the conformational-energy hierarchy of AAPA+Li^+ at the PBE+vdW level. In line with the results of the above comparison, PBE[31] and the hybrid density functional, PBE0,[69] (both vdW corrected)[32] give very similar results, illustrated by the low mean absolute (MAE) and maximal errors (E_{max}) listed in Table 2. Both approaches without the vdW correction give higher values for MAE and E_{max}. The widely used protein force fields, Amber99,[34] Charmm22,[70] and OPLS-AA,[33] give MAE values that are at least approximately 15 times larger and very large E_{max} values (see Table 2). The relative energies can be found in the Supporting Information. The main characteristic of the system is apparently the cation–peptide interaction. The effect on the partial charges appears to be better described by the polarizable FF Amoeba,[71] illustrated by a MAE value of about 10 kJ mol^{-1}. The removal of the cation leads to reduced MAE values for the FF methods (see Table 2); also, the energy hierarchies themselves appear more consistent among the different methods. This is apparent either when comparing the two plots (in the presence and absence of Li^+) in Figure 7 or when studying the maximal error values, as given in Table 2. The calculations (single point) were repeated for the same AAPA conformers (fixed geometries) but without the cation. The MAE values obtained using the force-field approaches are consistently much larger than those obtained using the DFT techniques, with significant errors in the energetic hierarchy of the conformers. Apparently, the large errors of the force fields can mainly be attributed to the ill-described cation–peptide interaction. In short, DFT-based approaches for cation–peptide systems appear to be superior to standard force field based approaches tested here.

Table 1. Relative energies and RMSD values of the conformers depicted in Figure 6.[a]

	E (MP2 geometries)				RMSD to MP2		
Conf.	CCSD(T)	PBE+vdW	PBE0+vdW	B3LYP	PBE+vdW	PBE0+vdW	B3LYP
1	0.0	0.0	0.0	0.0	0.03	0.03	0.03
2	6.7	4.4	5.0	5.7	0.03	0.03	0.03
3	20.2	17.1	18.2	22.3	0.38	0.16	0.33
4	23.1	22.0	23.1	27.1	0.12	0.08	0.18
5	33.8	34.2	34.8	38.0	0.04	0.02	0.04
MAE/RMSD		1.4	0.9	2.3	0.12	0.06	0.12

[a] Left columns: CCSD(T), PBE+vdW, PBE0+vdW, and B3LYP relative energies were calculated for MP2/cc-pVTZ geometries. The MAE of the DFT relative-energy hierarchies to CCSD(T) is also given. Right columns: the conformers were also relaxed with the respective DFT methods. With respect to the MP2 geometries, RMSD values for the individual conformers and average RMSD values are given. Relative energies and the mean absolute error (MAE) values are given in kJ mol^{-1}; RMSD values are given in Å.

Experimental methods

Synthesis: Peptides were synthesized by solid-phase assembly using a Multi-Syntech Syro XP peptide synthesizer (MultisynTech GmbH, Witten, Germany) and an Fmoc strategy on Fmoc-Ala-OWang resin (0.5 mmol g^{-1}). The peptides were cleaved from the resin by reaction with 2 mL of a solution containing 10% (*w*/*v*) triisopropylsilane, 1% (*w*/*v*) water, and 89% (*w*/*v*) trifluoroacetic acid (TFA). The crude peptides were purified by reversed-phase HPLC on a Knauer smartline manager 5000 system (Knauer, Berlin, Germany) equipped with a C8 (10 µm) LUNATM Phenomenex column (Phenomenex, Torrance, CA, USA). Peptides were eluted with a linear gradient of acetonitrile/water/0.1% TFA and identified on an Agilent 6210 ESI-TOF mass spectrometer. Peptide purity was determined by analytical HPLC on a Merck LaChrom system (Merck KGaA, Darmstadt, Germany) equipped with a C8 (10 µm) LUNATM Phenomenex column (Phenomenex, Torrance, CA, USA). The gradient used was similar to those used for the preparative HPLC.

Infrared spectroscopy: The gas-phase IR experiments were performed at the free electron laser facility FELIX[39] (Nieuwegein, The Netherlands) using the Fourier-transform ion cyclotron (FT-ICR) mass spectrometer[40] which was temporarily equipped with a nano electrospray ionization (nESI) source

Table 2. MAE and E_{max} values.[a]

	PBE0+vdW	PBE	PBE0	Amoeba	Amber99	Charmm22	OPLS-AA
	Fixed AAPA+Li^+ geometries						
MAE	1.0	5.7	6.1	9.7	15.7	20.5	26.0
E_{max}	2.0	10.9	11.1	22.3	30.1	53.6	69.1
	Same geometries, fixed without Li^+						
MAE	1.1	6.6	6.6	8.1	14.0	15.6	11.6
E_{max}	2.8	14.3	14.8	20.3	37.5	36.8	29.7

[a] MAE and E_{max} values with respect to the PBE+vdW hierarchy, of the energy hierarchies computed with PBE0+vdW, PBE, PBE0, Amoeba (ref. [71]), Amber99 (ref. [34]), Charmm22 (ref. [70]), and OPLS (ref. [33]) for the fixed geometries of AAPA with and without the Li^+ ion (energies in kJ mol^{-1}). Cartesian coordinates and relative energies of the conformations are provided in the Supporting Information.

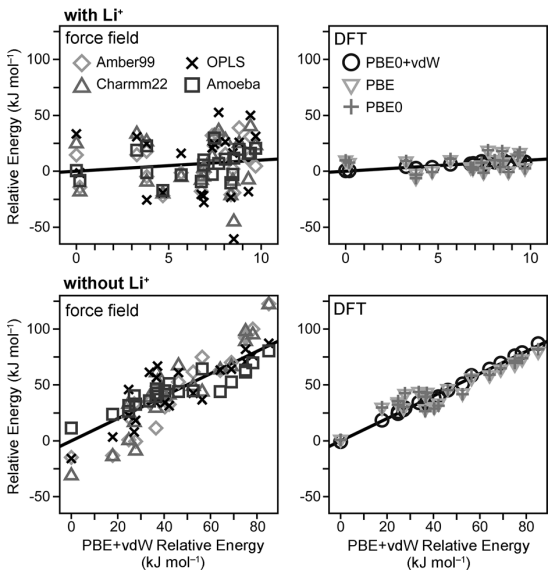


Figure 7. The relative energies of 21 AAPA+Li^+ conformers with a relative potential energy below 10 kJ mol^{-1} at the PBE+vdW level were recalculated with PBE0+vdW, PBE, PBE0, Amoeba (ref. [71]), Amber99 (ref. [34]), Charmm22 (ref. [70]), and OPLS (ref. [33]) for the fixed geometries with and without the Li^+ ion. The conformational energy hierarchies with different FF and DFT methods are plotted against the PBE+vdW values (*x* axis). The relative energy values were shifted according to the overall offset of the individual energy hierarchy. Please note the different scales of the *x* axes in the plots with and without Li^+ cation. The hypothetical perfect correlation is indicated by the straight lines.

(MS Vision, Almere, The Netherlands). Typically, 5 μL of a solution containing 1 mM peptide, 50% water, 50% methanol and, where needed, 10 mM LiCl or NaCl, were placed in gold-coated off-line emitters prepared in-house. To obtain a stable spray, a small backing pressure of approximately 0.5 bar and a relatively low capillary voltage of approximately 850 V was applied to the needle. The nESI-generated ions were accumulated in a hexapole ion trap and subsequently transferred into the FT-ICR mass spectrometer that is optically accessible through a KRS-5 window at the back end. After trapping and SWIFT mass isolation inside the ICR cell, the ions were irradiated by IR photons of the free electron laser FELIX.[72] The light provided by FELIX consists of macropulses of about 5 μs length at a repetition rate of 10 Hz, which contain 0.3 to 5 ps long micropulses with a micropulse spacing of 1 ns. The wavelength is continuously tunable over a range of 40 to 2000 cm^{-1}. Here, typically wavelengths from 500 to 1850 cm^{-1} were scanned. When the IR light is resonant with an IR-active vibrational mode in the molecule, this results in the absorption of many photons, thus causing dissociation of the ion (IRMPD). Monitoring of the fragmentation yield as a function of IR wavelength leads to the IR spectra.

Acknowledgements

The authors acknowledge continuous interest and support from Gerard Meijer (Radboud University Nijmegen). We gratefully acknowledge the "Stichting voor Fundamenteel Onderzoek der Materie" (FOM) for providing the beam time on FELIX as well as the support of the FELIX staff: Britta Redlich, Lex van der Meer, Rene van Buuren, Jos Oomens, Giel Berden, and Josipa Grzetic. Franziska Schubert, Sucismita Chutia, and Mariana Rossi (FHI Berlin) are acknowledged for discussion and technical help. C. B. is grateful to Hans-Jörg Hofmann (Universität Leipzig) for inspiring discussions.

[1] P. Pfeiffer, J. von Modelski, *Hoppe-Seylers Z. Physiol. Chem.* **1912**, *81*, 329–354.
[2] P. Pfeiffer, *Hoppe-Seylers Z. Physiol. Chem.* **1924**, *133*, 22–61.
[3] D. Seebach, H. Bossler, R. Flowers, E. Arnett, *Helv. Chim. Acta* **1994**, *77*, 291–305.
[4] J. Kofron, P. Kuzmič, V. Kishore, E. Colón-Bonilla, D. Rich, *Biochemistry* **1991**, *30*, 6127–6134.
[5] H. Kessler, M. Gehrke, J. Lautz, M. Kock, D. Seebach, A. Thaler, *Biochem Pharmacol.* **1990**, *40*, 169–173.
[6] M. Koeck, H. Kessler, D. Seebach, A. Thaler, *J. Am. Chem. Soc.* **1992**, *114*, 2676–2686.
[7] E. Garand, M. Z. Kamrath, P. A. Jordan, A. B. Wolk, C. M. Leavitt, A. B. McCoy, S. J. Miller, M. A. Johnson, *Science* **2012**, *335*, 694–698.
[8] D. Seebach, A. K. Beck, A. Studer in *Modern Synthetic Methods* (Eds.: B. Ernst, C Leumann), Wiley-VCH, **1995**; pp. 1–178.
[9] G. Ramachandran, C. Ramakrishnan, V. Sasisekharan, *J. Mol. Biol.* **1963**, *7*, 95–99.
[10] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, D. C. Richardson, *Proteins: Struct. Funct. Bioinf.* **2003**, *50*, 437–450.
[11] G. Fischer, *Chem. Soc. Rev.* **2000**, *29*, 119–127.
[12] C. Dugave, L. Demange, *Chem. Rev.* **2003**, *103*, 2475–2532.
[13] M. S. Weiss, A. Jabs, R. Hilgenfeld, *Nat. Struct. Biol.* **1998**, *5*, 676–676.
[14] J. S. Richardson, *Adv. Protein Chem.* **1981**, *34*, 167–339.
[15] For idealized backbone torsion angles for these turn types, see: K. Möhle, M. Gußmann, H.-J. Hofmann, *J. Comput. Chem.* **1997**, *18*, 1415–1430.
[16] C. M. Venkatachalam, *Biopolymers* **1968**, *6*, 1425–1436.
[17] B. L. Sibanda, J. M. Thornton, *Nature* **1985**, *316*, 170–174.
[18] E. G. Hutchinson, J. M. Thornton, *Protein Sci.* **1994**, *3*, 2207–2216.
[19] C. Kunz, G. Jahreis, R. Günther, S. Berger, G. Fischer, H.-J. Hofmann, *J. Pept. Sci.* **2012**, *18*, 400–404.
[20] A. Abo-Riziq, J. E. Bushnell, B. Crews, M. Callahan, L. Grace, M. S. de Vries, *Chem. Phys. Lett.* **2006**, *431*, 227–230.
[21] J. Bakker, L. Aleese, G. Meijer, G. von Helden, *Phys. Rev. Lett.* **2003**, *91*, 203003.
[22] I. Compagnon, J. Oomens, G. Meijer, G. von Helden, *J. Am. Chem. Soc.* **2006**, *128*, 3592–3597.
[23] A. Kamariotis, O. V. Boyarkin, S. R. Mercier, R. D. Beck, M. F. Bush, E. R. Williams, T. R. Rizzo, *J. Am. Chem. Soc.* **2006**, *128*, 905–916.
[24] A. Cimas, T. D. Vaden, T. S. J. A. de Boer, L. C. Snoek, M.-P. Gaigeot, *J. Chem. Theory Comput.* **2009**, *5*, 1068–1078.

[25] W. H. James III, C. W. Müller, E. G. Buchanan, M. G. D. Nix, L. Guo, L. Roskop, M. S. Gordon, L. V. Slipchenko, S. H. Gellman, T. S. Zwier, *J. Am. Chem. Soc.* **2009**, *131*, 14243–14245.
[26] W. H. James III, E. E. Baquero, S. H. Choi, S. H. Gellman, T. S. Zwier, *J. Phys. Chem. A* **2010**, *114*, 1581–1591.
[27] M. Rossi, V. Blum, P. Kupser, G. von Helden, F. Bierau, K. Pagel, G. Meijer, M. Scheffler, *J. Phys. Chem. Lett.* **2010**, *1*, 3465–3470.
[28] R. J. Plowright, E. Gloaguen, M. Mons, *ChemPhysChem* **2011**, *12*, 1889–1899.
[29] S. Chutia, M. Rossi, V. Blum, *J. Phys. Chem. B* **2012**, *116*, 14788–14804.
[30] M. Rossi, M. Scheffler, V. Blum, *J. Phys. Chem. B* **2013**, *117*, 5574–5584.
[31] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
[32] A. Tkatchenko, M. Scheffler, *Phys. Rev. Lett.* **2009**, *102*, 073005.
[33] W. Jorgensen, J. Ulmschneider, J. Tirado-Rives, *J. Phys. Chem. B* **2004**, *108*, 16264–16270.
[34] J. Wang, P. Cieplak, P. Kollman, *J. Comput. Chem.* **2000**, *21*, 1049–1074.
[35] R. Pappu, R. Hart, J. Ponder, *J. Phys. Chem. B* **1998**, *102*, 9725–9742.
[36] B. Byun, I. Song, Y. Chung, K. Ryu, Y. Kang, *J. Phys. Chem. B* **2010**, *114*, 14077–14086.
[37] Z. Wasserman, F. Salemme, *Biopolymers* **1990**, *29*, 1613–1631.
[38] A. R. Fadel, D. Q. Jin, G. T. Montelione, R. M. Levy, *J. Biomol. NMR* **1995**, *6*, 221–226.
[39] D. Oepts, A. van der Meer, P. van Amersfoort, *Infrared Phys. Technol.* **1995**, *36*, 297–308.
[40] J. Valle, J. Eyler, J. Oomens, D. Moore, A. van der Meer, G. von Helden, G. Meijer, C. Hendrickson, A. Marshall, G. Blakney, *Rev. Sci. Instrum.* **2005**, *76*, 023103.
[41] M.-P. Gaigeot, *Phys. Chem. Chem. Phys.* **2010**, *12*, 3336–3359.
[42] J. Pendry, *J. Phys. C: Solid State Phys* **1980**, *13*, 937–944.
[43] V. Blum, K. Heinz, *Comput. Phys. Commun.* **2001**, *134*, 392–425.
[44] G. Grégoire, M. P. Gaigeot, D. C. Marinica, J. Lemaire, J. P. Schermann, C. Desfrancois, *Phys. Chem. Chem. Phys.* **2007**, *9*, 3082–3097.
[45] S. Varma, S. B. Rempe, *Biophys. Chem.* **2006**, *124*, 192–199.
[46] T. Ikeda, M. Boero, K. Terakura, *J. Chem. Phys.* **2007**, *126*, 034501.
[47] D. Semrouni, O. P. Balaj, F. Calvo, C. F. Correia, C. Clavagura, G. Ohanessian, *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 728–738.
[48] O. P. Balaj, D. Semrouni, V. Steinmetz, E. Nicol, C. Clavaguéra, G. Ohanessian, *Chem. Eur. J.* **2012**, *18*, 4583–4592.
[49] M. Kohtani, B. Kinnear, M. Jarrold, *J. Am. Chem. Soc.* **2000**, *122*, 12377–12378.
[50] J. K. Martens, I. Compagnon, E. Nicol, T. B. McMahon, C. Clavaguéra, G. Ohanessian, *J. Phys. Chem. Lett.* **2012**, *3*, 3320–3324.
[51] S. Y. Noskov, B. Roux, *J. Mol. Biol.* **2008**, *377*, 804–818.
[52] F. N.-S. Hofmeister, *Arch. Pharmakol.* **1888**, *25*, 1–30.
[53] P. H. von Hippel, K.-Y. Wong, *J. Biol. Chem.* **1965**, *240*, 3909–3923.
[54] A. W. Omta, M. F. Kropman, S. Woutersen, H. J. Bakker, *Science* **2003**, *301*, 347–349.
[55] W. Kunz, *Curr Opin. Colloid Interface Sci.* **2010**, *15*, 34–39.
[56] J. Dzubiella, *J. Am. Chem. Soc.* **2008**, *130*, 14000–14007.
[57] J. Dzubiella, *J. Phys. Chem. B* **2009**, *113*, 16689–16694.
[58] Y. von Hansen, I. Kalcher, J. Dzubiella, *J. Phys. Chem. B* **2010**, *114*, 13815–13822.
[59] A. Crevenna, N. Naredi-Rainer, D. Lamb, R. Wedlich-Söldner, J. Dzubiella, *Biophys. J.* **2012**, *102*, 907–915.
[60] Y. Zhang, S. Furyk, D. E. Bergbreiter, P. S. Cremer, *J. Am. Chem. Soc.* **2005**, *127*, 14505–14510.
[61] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, M. Scheffler, *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.
[62] F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73–78.
[63] D. G. Truhlar, Chem. Phys. Lett. **1998**, *294*, 45–48.
[64] T. H. Dunning Jr. , *J. Chem. Phys.* **1989**, *90*, 1007–1023.
[65] A. Tkatchenko, M. Rossi, V. Blum, J. Ireta, M. Scheffler, *Phys. Rev. Lett.* **2011**, *106*, 118102.
[66] N. Marom, A. Tkatchenko, M. Scheffler, L. Kronik, *J. Chem. Theory Comput.* **2010**, *6*, 81–90.
[67] G.-X. Zhang, A. Tkatchenko, J. Paier, H. Appel, M. Scheffler, *Phys. Rev. Lett.* **2011**, *107*, 245501.
[68] J. A. Pople, J. S. Binkley, R. Seeger, *Int. J. Quantum Chem.* **1976**, *10*, 1–19.
[69] C. Adamo, V. Barone, J. Chem. Phys. **1999**, *110*, 6158–6170.
[70] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr. , J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus, *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
[71] M. Schnieders, J. Ponder, *J. Chem. Theory Comput.* **2007**, *3*, 2083–2097.
[72] J. Oomens, B. Sartakov, G. Meijer, G. von Helden, *Int. J. Mass Spectrom.* **2006**, *254*, 1–19.

Received: December 21, 2012
Published online: July 12, 2013

4 Sampling biomolecular potential-energy landscapes

4.1 PARADOCKS - A framework for molecular docking with population-based metaheuristics

Meier, R.; Pippel, M.; Brandt, F.; Sippl, W.; Baldauf, C.

PARADOCKS - A framework for molecular docking with population-based metaheuristics

J. Chem. Inf. Model. **2010** (50), 879-889.

DOI: 10.1021/ci900467x

PARADOCKS: A Framework for Molecular Docking with Population-Based Metaheuristics

René Meier,*[,†,‡,⊥] Martin Pippel,[†,⊥] Frank Brandt,[¶] Wolfgang Sippl,[†] and Carsten Baldauf*[,§,∥,¶]

Department of Pharmaceutical Chemistry, Martin-Luther Universität Halle-Wittenberg, Wolfgang-Langenbeck-Strasse 4, 06120 Halle/Saale, Germany, Research Center Pharmaceutical Engineering GmbH, Inffeldgasse 21a/II, 8010 Graz, Austria, Biotechnologisches Zentrum der TU Dresden, Tatzberg 47/49, 01307 Dresden, Germany, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, 200031 Shanghai, P. R. China, and Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

Received December 1, 2009

Molecular docking is a simulation technique that aims to predict the binding pose between a ligand and a receptor. The resulting multidimensional continuous optimization problem is practically unsolvable in an exact way. One possible approach is the combination of an optimization algorithm and an objective function that describes the interaction. The software PARADOCKS is designed to hold different optimization algorithms and objective functions. At the current stage, an adapted particle-swarm optimizer (PSO) is implemented. Available objective functions are (i) the empirical objective function p-Score and (ii) an adapted version of the knowledge-based potential PMF04. We tested the docking accuracy in terms of reproducing known crystal structures from the PDBbind *core set*. For 73% of the test instances the native binding mode was found with an rmsd below 2 Å. The virtual screening efficiency was tested with a subset of 13 targets and the respective ligands and decoys from the directory of useful decoys (DUD). PARADOCKS with PMF04 shows a superior early enrichment. The here presented approach can be employed for molecular docking experiments and virtual screenings of large compound libraries in academia as well as in industrial research and development. The performance in terms of accuracy and enrichment is close to the results of commercial software solutions.

INTRODUCTION

Molecular interactions define all manifestations of life. Accordingly, knowledge of such processes is of paramount importance to current life science research and development in fields as diverse as medicine, biotechnology, and crop science. Upcoming challenges like fast-evolving infectious diseases, personalization of medicine, development of tailor-made enzymes or substrates, as well as the development of crop protective agents mark the need for rapid approaches that feature computational techniques.

Already since the late 1990s, computational approaches have gained considerable attention in the area of drug design.[1,2] In silico techniques from computational chemistry, bioinformatics, and systems biology apparently offer the chance to tackle these problems and to respond faster and in a more resource-conserving way than classical, pure wet-lab approaches. Molecular docking plays a key-role among the variety of approaches and techniques because it offers the chance to gain knowledge on the actual binding pose, the situation at atomic level that defines binding and function.

The three-dimensional structure of the complex formed by protein and ligand is key to the prediction of activities based on physicochemical models that describe the spatial and energetical properties of binding. Still, the high dimensionality and the complicated nature of the problem result in complex energy landscapes with many local minima. These features prohibit an analytic approach to molecular docking and thus, search strategies are employed to find the native pose of ligand and receptor. Most docking approaches generate a large number of complexes and evaluate their quality in terms of binding. Molecular docking thus means the generation and evaluation of molecular complexes to predict binding poses of protein ligand complexes. A way to categorize docking approaches follows the treatment of this high dimensionality:

- The ligand can be subdivided into rigid fragments. These are subsequently reassembled within the binding pocket. Such fragment-based techniques are used by FLEXX,[3] SURFLEX,[4] and eHiTS.[5]
- The docking of ensembles of rigid ligand conformations results in high speed, but has its drawback in the fact that the biologically active conformation of a compound has to be part of the precalculated conformational ensemble. Examples are FRED[6] and early versions of DOCK.[7,8]
- Heuristics-based techniques aim for the global minimum of an objective function, assuming this optimum is the effective complex. The search space of the algorithm is

* To whom correspondence should be addressed: E-mail: rene@paradocks.org (R.M.); caba@paradocks.org (C.B.).
† MLU Halle-Wittenberg.
‡ RCPE Graz.
¶ BIOTEC TU Dresden.
§ CAS-MPG PICB, Shanghai.
∥ HITS, Heidelberg.
⊥ These authors contributed equally to this work.

10.1021/ci900467x
Published on Web 04/26/2010

defined by the degrees of freedom of ligand and protein. Population-based metaheuristics, mainly genetic algorithms (GA), are used by programs like GOLD[9] and AUTODOCK.[10]

Following the line of heuristic-based approaches, alternative search strategies have been proposed. The recently introduced docking program PLANTS[11,12] uses ant-colony optimization (ACO). Based on the AUTODOCK software, a number of particle-swarm optimization (PSO) approaches were presented: AUTODOCK with ClustMPSO,[13] SODOCK,[14] and PSO@AUTODOCK.[15] PSO is inspired by social behavior of animals, for example, bird flocking or fish schooling and was first suggested by Eberhart and Kennedy.[16] Intuitively, the PSO appears perfectly suited to tackle the continuous search space of protein ligand interaction within the molecular docking problem. This assumption is well supported by the performance and success of the published docking methodologies employing PSO variants. Of special interest is the easy adaptability of PSO, and other population-based metaheuristics, for parallel approaches, especially with the current rise of multicore CPU architectures.

The interaction between ligand and protein is described by a mathematical model, the objective or energy function. Important terms are the solvation energies of the protein, the ligand, and their complex ΔG_{sol}^{prot}, ΔG_{sol}^{lig}, and $\Delta G_{sol}^{complex}$, the change in entropy ΔS between bound and unbound state, the interaction energy ΔG_{int}, and the energy change in ligand and protein while the interaction is formed $\Delta\lambda$. All these terms contribute to the binding free energy according to eq 1[17]

$$\Delta G_{bind} = \Delta G_{sol}^{complex} - \Delta G_{sol}^{prot} - \Delta G_{sol}^{lig} + \Delta G_{int} - T\Delta S + \Delta\lambda \quad (1)$$

Practical considerations prohibit the correct estimation of ΔG_{bind}: (i) the large numbers of the individual contributions have to be balanced to avoid errors in the small values of the binding energy, especially with some contributions being only roughly estimated like entropy, and (ii) exact calculation demands a complete sampling of the conformational space for the ligand in the binding pocket, a very time-consuming task that is not feasible for high-throughput molecular docking of compound libraries.[17,18] Thus, a variety of approaches has been introduced that try to correctly rank of protein ligand poses toward the global optimum, the native state. In a test case, this means the reproduction of the X-ray structure. The available approaches can be categorized as follows:

- Force field-derived objective functions are based on the description of nonbonded interactions of established force fields. The terms used are based on physical laws and are accurate representations of the enthalpic contributions. DOCK[7,8] describes the nonbonded interactions partially with terms from the AMBER[19] force field. Within GOLD, the contributions of van der Waals-type interactions (vdW) are estimated by soft 8−4-Lennard-Jones potentials.[9]
- Empirically derived objective functions consist of a number of physics-inspired terms that describe, for example, hydrogen-bonds, ionic interactions, hydrophobic effects, entropy, π-stacking, or π-cation-interactions. These functions are trained to reproduce representative test sets. An advantage of empirical objective functions is their usually fast computational evaluation. GoldScore is in parts an empirically derived scoring function,[9] further examples are SCORE1[20] and X-SCORE.[21]
- Knowledge-based potentials stem from statistical evaluations of large data sets, for example, Protein Database. In contrast to the above-mentioned approaches, there is no limitation to the specifically described interactions because knowledge-based approaches try not to model individual interaction types. Rather, potentials intrinsically include all effects that can be extracted from experimentally derived structures. Well-accepted examples are BLEEP,[22,23] PMF[24] and PMF04,[25] and DRUGSCORE.[26]

Obviously, there is a multitude of energy functions and optimization algorithms available and many new developments can be expected in the future. To us, this clearly renders the need for a platform that allows the convenient incorporation of existing and new approaches either to describe ligand−receptor interaction or to search for the native pose. Even though a wide variety of programs to solve the molecular docking problem exists, there are disadvantages:

- Closed source distributions cover the approaches used for computations for the interaction, as well as for sampling and for energy estimation. This makes results and approaches not comparable and limits progress.
- Restricted licensing policies hinder the redistribution of self-developed code.
- Monolithic code and outdated programing standards limit the extension and further development of several existing approaches.

Our newly developed docking software has the chance to avoid these issues and to satisfy the needs of users and developers from industry and academia. The development of the Parallel Docking *S*uite (PARADOCKS) software follows these rules:

- PARADOCKS will be distributed as open source code under a nonrestrictive license (GPL).
- Design and implementation should result in an as far as possible platform and operating system independent software.
- If actively maintained programs or libraries are available for certain problems, they will be used.
- Parallel computer systems, compute clusters, and multicore workstations become more and more widespread, thus parallel data processing is a major goal of PARADOCKS.
- The program should be usable with automated pipelines for virtual screening and drug design.

Within this article we will describe the PARADOCKS framework for molecular docking. The Materials and Methods section will introduce basic design principles and their implementation and will cover specifics regarding the implementation of optimization algorithms and objective functions. The latter two will be illustrated by example implementations of a PSO, as well as the p-Score and PMF04[25] objective function. The Results section deals with the assessment of the docking accuracy as well as testing the applicability for virtual screening of PARADOCKS.

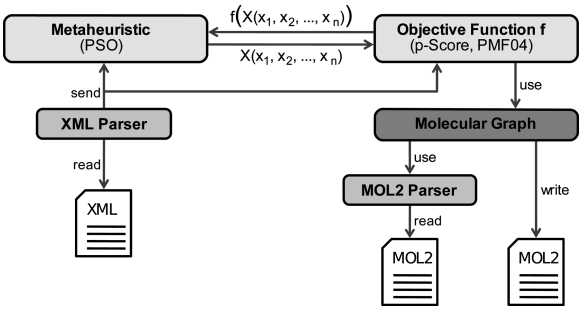


Figure 1. ParaDockS design scheme. Boxes represent classes and arrows represent the interfaces.

MATERIALS AND METHODS

Problem Description. We approximate the ligand as a flexible molecule and the receptor as rigid. The interaction between ligand and receptor is described by an objective function that depends on three types of degrees of freedom: (i) The position of the ligand molecule is described by three values x, y, z in Cartesian coordinate space. (ii) The rotational degrees of freedom are modeled as a quaternion $H = x_1 + x_2i + x_3j + x_4k$. This representation overcomes the gimbal lock problem of Euler angles which, under special conditions, results in the loss of one degree of freedom. Unit quaternions are a non singular representation of rotations and widely used in the field of three-dimensional computer graphics.[27] (iii) The flexibility of the ligand is accounted for by free rotation of torsion angles (single bonds) of the ligand. This results in a variable number of degrees of freedom that depends on the size and topology of a molecule, meaning conformers of a molecular conjugation. The resulting dimensionality of the continuous search space is therefore $7 + T$ (with T being the torsion number). The goal of a molecular docking simulation is the prediction of the native bound structure of a ligand in the binding site of its receptor, which is assumed to be the global optimum of the search space. It is an accepted approach to solve this molecular docking problem, namely the finding of the native pose, using an optimization algorithm.

Framework Design. The ParaDockS software is written in C/C++ and consists of modular functional units. Communication is realized via interfaces (cf. Figure 1). Parallel data processing is implemented via the Message Passing Interface (MPI). The input files are in XML format for simulation setup and in MOL2 format[28] for ligand and receptor coordinates. Subsequently, a molecular graph is created for the ligand. Information on position, orientation, and conformation of the ligand is stored in a $7 + T$-dimensional vector. The information is passed to the objective function for energy evaluation. The energy value (fitness of the solution) is passed back to the metaheuristic and a new iteration starts with the generation of new solutions. ParaDockS can hold different objective functions and optimization algorithms. In addition, basic paradigms can be changed; this includes an increase of the number of degrees of freedom (e.g., by receptor flexibility), the linking to external programs for energy evaluation, or even the employment of multiobjective optimization.

Our aim is to present well working and robust software for molecular docking. Beyond that, we want to invite other scientists to participate in a joint development effort to improve the program and to expand its functionalities. To enable that we publish the source code of ParaDockS under the GNU General Public License (GPLv2)[29] to ensure its free use, the freedom to modify the underlying code, and the redistribution. All parts of the program are as generic as possible and should at least be fit for all metaheuristics-based docking approaches. Implementation of new approaches, namely, energy functions or search strategies is therewith limited to the respective core functionalities, generic components need only little to no modification. All public classes and functions as well as the application programming interface (API) are documented by the documentation system Doxygen.[30] We supply an advanced algorithm for the atom type deduction based on topology subgraph matching similar to the SMARTS[31] system. The parsing of MOL2 files is performed by a robust algorithm based on an Extended Backus-Naur Form[32] grammar. To allow easy testing of self-implemented approaches, we provide an rmsd calculation program for small molecules which takes molecular symmetry into account.

Optimization Algorithm. By design, ParaDockS is able to be used with different optimization algorithms. Based on promising results by others,[13–15] we decided to use particle swarm optimization as an exemplary optimizer implementation. The algorithm implemented here follows PSO as introduced by Eberhart and Kennedy.[16] Optimization starts with a population of random solutions; the search for optima is facilitated by updating generations, making the swarm virtually fly through the search space. The best position in search space so far (best solution achieved) is tracked for the individual particle as well as for the whole swarm. With the change of generations of the swarm, the particles are accelerated toward these best solutions. These accelerations are weighted by random terms. The algorithm is shown in Algorithm 1 and features two modifications to the original algorithm: (i) an inertia weight c_0 decreasing linearly over time[33] and (ii) the reinitialization with random position and velocity of particles leaving the area of interest (the proximity of the binding pocket). Initialization distributes the particles equally in the search space. After evaluation of the objective function, positions of each particle get adjusted toward the best configuration in the particle's history as well as toward the configuration of the current best particle of the swarm. The linearly decreasing inertia weight c_0 in our implementation is intended to force exploration of the search space and convergence to the global minimum (exploitation).

Algorithm 1: Particle Swarm Optimization
for every particle P **do**
  $P_X \Leftarrow$ random_ position ()
  $P_V \Leftarrow$ random_velocity ()
  $P_{BX} \Leftarrow P_X$
  $P_{BF} \Leftarrow f(P_X)$
end for
$i \Leftarrow 0$
while $i <$ maximum iterations **do**
  for every particle P **do**
    if molecule not in binding pocket **then**
      $P_X \Leftarrow$ random_position ()
    end if
    $N \Leftarrow$ neighborhood_best (P)

$$P_V \Leftarrow \left(c_0 - \frac{i}{\text{maximum iterations}}\right)P_V + c_1 r_1 (P_{BX} - P_X) + c_2 r_2 (N_{BX} - P_X)$$

$P_X \Leftarrow P_X + P_V$
$F^* \Leftarrow f(P_X)$
if F^* better P_{BF} **then**
$P_{BF} \Leftarrow F^*$
$P_{BX} \Leftarrow P_X$
end if
end for
end while

LEGEND: P_X = particle position; P_V = particle velocity; P_{BX} = best position of the particle; P_{BF} = best fitness of the particle; $f(x)$ = fitness function; c_0 = inertia weight; c_1 = cognitive weight; c_2 = social weight.

Objective Functions. *p-Score.* The p-Score objective function is an empirically derived energy function. The docking energy E_{dock} is dissected into

$$E_{dock} = E_{vdW} + E_{estate} + E_{internal} \tag{2}$$

The van der Waals (vdW)-type interactions are modeled by a Lennard-Jones potential calculated for pairs of the ligand $\mathscr{L}$ and protein $\mathscr{P}$ atoms

$$E_{vdw} = \sum_{i \in \mathscr{P}} \sum_{j \in \mathscr{L}} \left[\left(\frac{d_{0ij}}{d_{ij}}\right)^8 - 2\left(\frac{d_{0ij}}{d_{ij}}\right)^4 \right] \tag{3}$$

The optimal vdW distance d_{0ij} between atoms i and j is the sum of the vdW radii of atom i and atom j. d_{ij} is the actual distance between atoms i and j. The 8−4 form of the potential is "softer".[34] The resulting reduced penalty for close contacts accounts for a limited flexibility of the receptor without explicitly modeling receptor flexibility.[35] $E_{internal}$ is defined as an 8−4 potential of the same form as E_{vdW}. But with the difference that only destabilizing positive values contribute to E_{dock}. $E_{internal}$ acts solely as penalty for internal vdW clashes.

The second contribution to the p-Score docking energy describes electrostatic interactions. This type of interaction is crucial for a correct description of specificity and affinity and hence crucial for molecular docking. The strength of the interaction depends on orientation and distance and thus E_{estat} is calculated by an angle- and distance-dependent potential

$$E_{estat} = \sum_{i \in \mathscr{P}} \sum_{j \in \mathscr{L}} f(d_{ij}) f(\theta_{1ij}) f(\theta_{2ij}) \tag{4}$$

In all cases, the energy contribution depends on the distance d_{ij} of the atom pairs i and j. The function terms $f(\theta_{1ij})$ and $f(\theta_{2ij})$ are not needed (set to 1) for ionic interactions as there is no angle dependency for this type (cf., Figure 2a), whereas lone-pair or hydrogen bond interactions demand modeling of the angle dependency. The description distinguishes between potentials for ionic interactions and hydrogen bonding with freely rotatable or frozen donor and acceptor atoms. A donor or acceptor atom is considered to be frozen if it is within a chain of heavy atoms, otherwise, if it is the terminal of a chain of heavy atoms, its lone pair or hydrogen can rotate freely. In the case of a freely rotatable hydrogen or lone pair θ_{1ij} and θ_{2ij} are calculated between the heavy atoms of the hydrogen bond as shown for atom j in Figure 2b and atom i in Figure 2c. For frozen donor and acceptor atoms the angles θ_{1ij} and θ_{2ij} correspond to the angle between the hydrogen or lone pair, respectively, and the two heavy atoms of the hydrogen bond as shown in Figure 2d. The linear potentials follow the formulas

$$f(d_{ij}) = \begin{cases} 1 & d_{ij} \le (d_{0ij} - k_1) \\ (1/k_1) \cdot (d_{0ij} - d_{ij}) & (d_{0ij} - k_1) < d_{ij} \le d_{0ij} \\ 0 & d_{ij} > d_{0ij} \end{cases} \tag{5}$$

$$f(\theta_1) = \begin{cases} (1/k_2) \cdot (k_2 - |\theta_1 - k_i|) & 0 \le |\theta_1 - k_i| \le k_2 \\ 0 & |\theta_1 - k_i| > k_2 \end{cases} \tag{6}$$

$$f(\theta_2) = \begin{cases} (1/k_3) \cdot (k_3 - |\theta_2 - k_j|) & 0 \le |\theta_2 - k_j| \le k_3 \\ 0 & |\theta_2 - k_j| > k_3 \end{cases} \tag{7}$$

The last term of eq 2, $E_{internal}$, evaluates the ligand conformation for vdW-clashes by using an 8−4-Lennard-Jones potential for all ligand atoms i and j which have at least 4 bonds distance

$$E_{internal} = \sum_{i \in \mathscr{L}} \sum_{j \in \mathscr{L}} \left[\left(\frac{d_{0ij}}{d_{ij}}\right)^8 - 2\left(\frac{d_{0ij}}{d_{ij}}\right)^4 \right] \tag{8}$$

PMF04. PMF04 is a knowledge-based objective function. This allows the exploitation of the vast amount of experimentally determined protein−ligand structures as a basis for molecular docking. Muegge et al. have shown the capability of statistical potentials for molecular docking by implementing PMF scoring[24] into the DOCK4 program.[36] We implemented the statistical potential PMF04[25] for molecular docking with PARADOCKS. PMF04 is derived from 6611 protein ligand complexes and describes the interactions of 17 protein atom types with 34 ligand atom types in form of pairwise potentials

$$W_{ij}(d_{ij}) = -\ln \frac{g_{ij}(d_{ij})}{g_{ref}} \tag{9}$$

with $g_{ij}(d_{ij})$ the density of the atom pair ij in distance d_{ij} and g_{ref} the average density of atom pair ij. For a detailed description, we point to the original publication.[25] We will continue with the necessary adaptations to use PMF04 as an objective function for molecular docking with PARADOCKS. The original close distance penalty of 3 kcal/mol is far too low for use with molecular docking, since its use results in overlapping of the ligand with receptor atoms after optimization. To circumvent this, the repulsion part of an 8−4 Lennard-Jones-potential has been added as the close distance penalty of PMF04. The Lennard-Jones-potential $E_{internal}$ (cf., eq 8) describes the conformation of the ligand. The docking energy E_{dock} is calculated as follows:

$$E_{dock} = E_{PMF04} + a \cdot E_{internal} \tag{10}$$

with

$$E_{PMF04} = \sum_{i \in \mathscr{P}} \sum_{j \in \mathscr{L}} W_{ij}(d_{ij}) \tag{11}$$

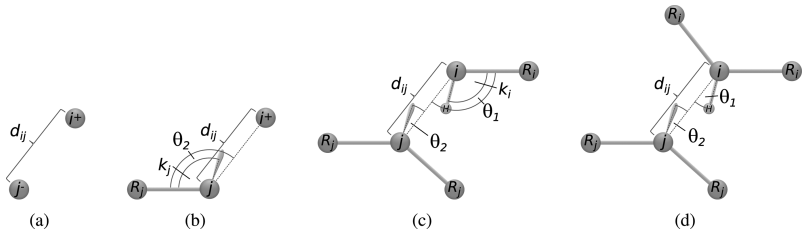


Figure 2. p-Score differentiates between multiple possibilities for electrostatic interactions, for example, (a) ionic interactions, (b) cation−lone pair interactions, (c) frozen acceptor and rotatable donor, and (d) frozen acceptor and donor.

RESULTS AND DISCUSSION

Parameter selection for the PSO and the objective functions, as well as the benchmarking for docking accuracy and virtual screening performance of PARADOCKS, were performed under the following paradigm: a useful molecular docking setup has to distinguish the quality of different poses of a receptor−ligand pair, as well as the quality of different potential ligands with respect to a receptor. The actual performance was compared to GOLD;[9] for further comparison, we point the reader to recent articles that feature extensive performance analysis of docking algorithms.[15,37–39] Parameter selection was performed on the (*Astex Diverse Set*[40]). For the evaluation of the docking accuracy, the PDBbind *core set*[41] was used; for assessing the virtual screening performance the directory of useful decoys (DUD)[42]) was employed. For all docking setups, identical initial coordinates of the ligand and the receptor were used. Where necessary, hydrogens were added to the crystal structures with the MOE program.[43] The initial conformations and orientations of all ligands were randomized.

Parameter Selection. *Particle Swarm Optimizer.* We found a limit of 150 000 iterations and a number of 20 particles to be sufficient for a good sampling and robust results. The search efficency is best with a cognitive weight $c_1 = 1.0$, a social weight $c_2 = 3.4$, and a constricting inertia weight c_0 ranging from 1.0 to 0.2. These parameters were selected in systematic tests of parameter combinations. The complex of the HIV-1 reverse transcriptase and its inhibitor TNK-651 (PDB 1JLA)[44] served as a typical example with 7 rotatable bonds and therefore average dimensionality. Because of its nondeterministic nature, every molecular docking experiment was repeated 400 times to generate comparable average results (this computation takes about four hours on a single 2.53 GHz Intel Xeon CPU). The average of the optimized score was compared and the parameter combination with the best average score is listed above.

p-Score. The parameters for p-Score were derived based on the assumption that an energy function for molecular docking has to evaluate the X-ray structures of a training set always better than alternative structures. The p-Score parameters to be optimized were the optimal vdW distances d_{0ij} as used in eqs 3 and 8 and k_1, k_2, and k_3 as in eqs 5−7.

The Astex Diverse Set,[40] a collection of high resolution (<2.5 Å) crystal structures of proteins and their drug-like ligands, was used as training set. For each protein ligand pair of the test set, 50 ligand conformations (decoys) with an rmsd relative to the X-ray structure above 2 Å and at least 22 decoys with an rmsd < 2 Å were generated. All decoys differed with an rmsd > 2 Å from each other. In the following, each of the up to 80 decoys per protein ligand pair was evaluated with an parameter set for the p-Score objective function. To indicate the quality of a parameter set, the ratio between decoys evaluated better or worse than the crystal structure was estimated

$$\mathrm{QP} = \frac{\text{number of decoys scored better than crystal structure}}{\text{number of decoys scored worse than crystal structure}}$$

An initial set of parameters was taken from X-SCORE[21] and improved by means of a randomized local search minimizing QP until no substantial changes of QP were observed anymore. The parameter optimization result was QP = 0.051, meaning that in more than 95% of all cases the crystal structure scores better than the decoys. The quality of the resulting vdW parameters can be seen in the fact that for 89% of all ligands in the PDBbind core set[41] we find at least one generated conformation which has an rmsd of less than 2 Å to the X-ray structure. The resulting vdW distances d_{0ij} and the electrostatic parameters k_1, k_2, and k_3 for the p-Score function can be found in the Supporting Information.

PMF04. Factor $a = 0.25$ of eq 10 was found by an exhaustive search with the objective of accumulating docking poses from the Astex diverse set that have an *rms*d below 2 Å to the X-ray structure.

Docking Accuracy. The PDBbind core set[41] contains 210 protein ligand pairs in 70 groups. Each group consists of proteins whose sequences are highly similar but that are complexed with ligands of low, medium, or high affinity, respectively. PARADOCKS runs were repeated 50 times per complex, and default parameters for the PSO were used. GOLD was used with automatic parameter settings with a selected search efficiency of 100%. The results of the docking simulations were clustered with a 2 Å rmsd cutoff and compared to the respective X-ray structures. A histogram plot of the results is shown in Figure 3, and numerical values are given in Table 1. 58% of the PMF04 dockings and 63% of the p-Score dockings found the native pose (GOLD 69%) within the three highest-ranking clusters.

We observe a significant decrease of the docking accuracy with the increase of the number of freely rotatable bonds; this effect is also observed for GOLD, but to a lesser extent than for PARADOCKS (cf., Figure 4a). There are two possible reasons for this effect: (*i*) The simple description of the ligand's conformation in p-Score and in our implementation of PMF04 might lead to the observed decrease in docking accuracy for ligands with more than 10 rotatable bonds. (*ii*) The same increase of torsional degrees of freedom leads as well to a substantially larger search space to sample. The simplified description of the ligand by avoiding van der Waals clashes is sufficient to predict meaningful ligand conformations. After fitting of the ligand conformations to

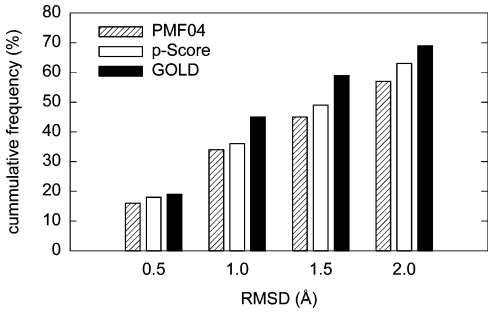


Figure 3. Comparison of the docking accuracy of ParaDockS with p-Score and PMF with GOLD on the PDBbind core set. The data is plotted as an additive histogram for the highest ranked three clusters.

Table 1. Docking Quality of ParaDockS with the Scoring Functions p-Score and PMF04 in Comparison to GOLD[a]

	native pose in clusters		
docking approach	1	1 and 2	1, 2, and 3
ParaDockS/PMF04	47%	52%	58%
ParaDockS/p-Score	52%	61%	63%
GOLD	62%	69%	69%

[a] The threshold for the native pose is a rmsd of 2 Å.

the X-ray structure, for 89% of all ligands in the PDBbind core set,[41] we find at least one conformer with an rmsd below 2 Å. The suggested standard settings for the optimization work well on average-sized problems. For ligands with more than ten torsional degrees of freedom, adapted docking settings should be used.

However, 75% of the substances listed in the world drug index (WDI)[45] have less than ten freely rotatable bonds (cf. Figure 4(b)). The general characteristics of drug molecules, as summarized, among others, by Lipinsky et al.[46] or Veber et al.[47] point toward smaller molecules with less than ten rotatable bonds as well.

Virtual Screening Performance. In virtual screening experiments, molecular docking is employed to find potent lead structures from large compound libraries. Thus it is of paramount importance to avoid false positive solutions. To thoroughly analyze the virtual screening performance of ParaDockS we selected a subset of 13 targets from the directory of useful decoys (DUD)[42] as described by Cheeseright et al.,[48] with at least 15 clusters of active compounds for each target. The 13 targets are: angiotensin-converting enzyme (ace), acetylcholinesterase (ache), cyclin-dependent kinase 2 (cdk2), epidermal growth factor receptor (egfr), factor Xa (fxa), HIV reverse transcriptase (hivrt), enoyl-acyl carrier protein reductase (inha), P38 mitogen-activated protein (p38), phosphodiesterase 5 (pde5), platelet-derived growth factor receptor kinase (pdgfrb), src tyrosine kinase (src) and vascular endothelial growth factor receptor (vegfr2). The data sets were downloaded from DUD in mol2 file format.[49] For ParaDockS we used the default PSO settings with 30 repeats per instance with PMF04 and p-Score, in addition, the results of the p-Score dockings were rescored with PMF04. For GOLD the genetic algorithm with ten repeats was used with each of the three available energy functions GoldScore, ChemScore, and the Astex Statistical Potential (ASP). The virtual screening perfomance is now assessed by the ability to distinguish known-active compounds (P) from the selected decoys (N). For each compound in the sorted row, the true positive rate (TPR) and the false positive rate (FPR) were calculated. Solutions that score better or equal than that particular compound are defined as positive solutions. Active compounds within the range of positive solutions are true positives (TP) and decoys within the range of defined positive solutions are false positives (FP). TPR and FPR are calculated according to

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} \tag{12}$$

and

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} \tag{13}$$

The receiver operator characteristic (ROC) diagrams resulting from plotting the TPR and FPR values are shown in Figure 5. Ideally, ROC curves show a steep early ascent, almost parallel to the y -axis and then, close to the maximal value for y, continue parallel to the x -axis. Such a behavior can be exemplary seen for ParaDockS with p-Score/PMF04 on the hivrt data set and for GOLD with GoldScore on the cox2 data set. However, most of the curves exhibit an sigmoidal shape. A good metric to assess the overall quality of a screening approach is the area under the ROC curve (AUC). The AUC gives the probability that a randomly chosen active is ranked higher than a randomly chosen inactive by the respective method. In Table 2 the AUC values are given, the methods exhibit similar perfomance. GOLD with ChemScore[50] and the Astex statistical potential (ASP)[51]

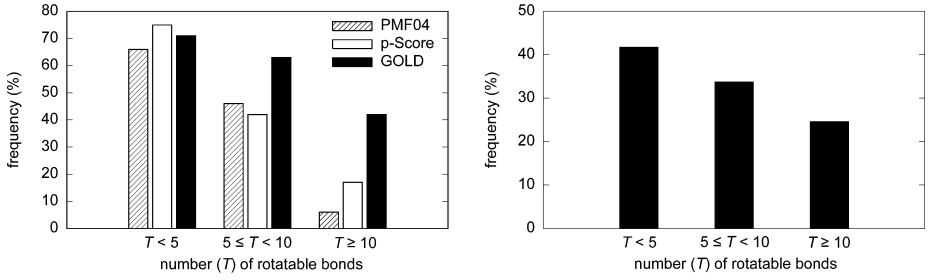


Figure 4. (a) The fraction of successful dockings (rmsd of 2 Å or better) of the PDBbind core set for ParaDockS with p-Score and PMF, respectively, and with GOLD as a function of the number of rotatable bonds of the ligand. (b) Distribution of compounds in the WDI[45] with respect to the number of rotatable bonds.

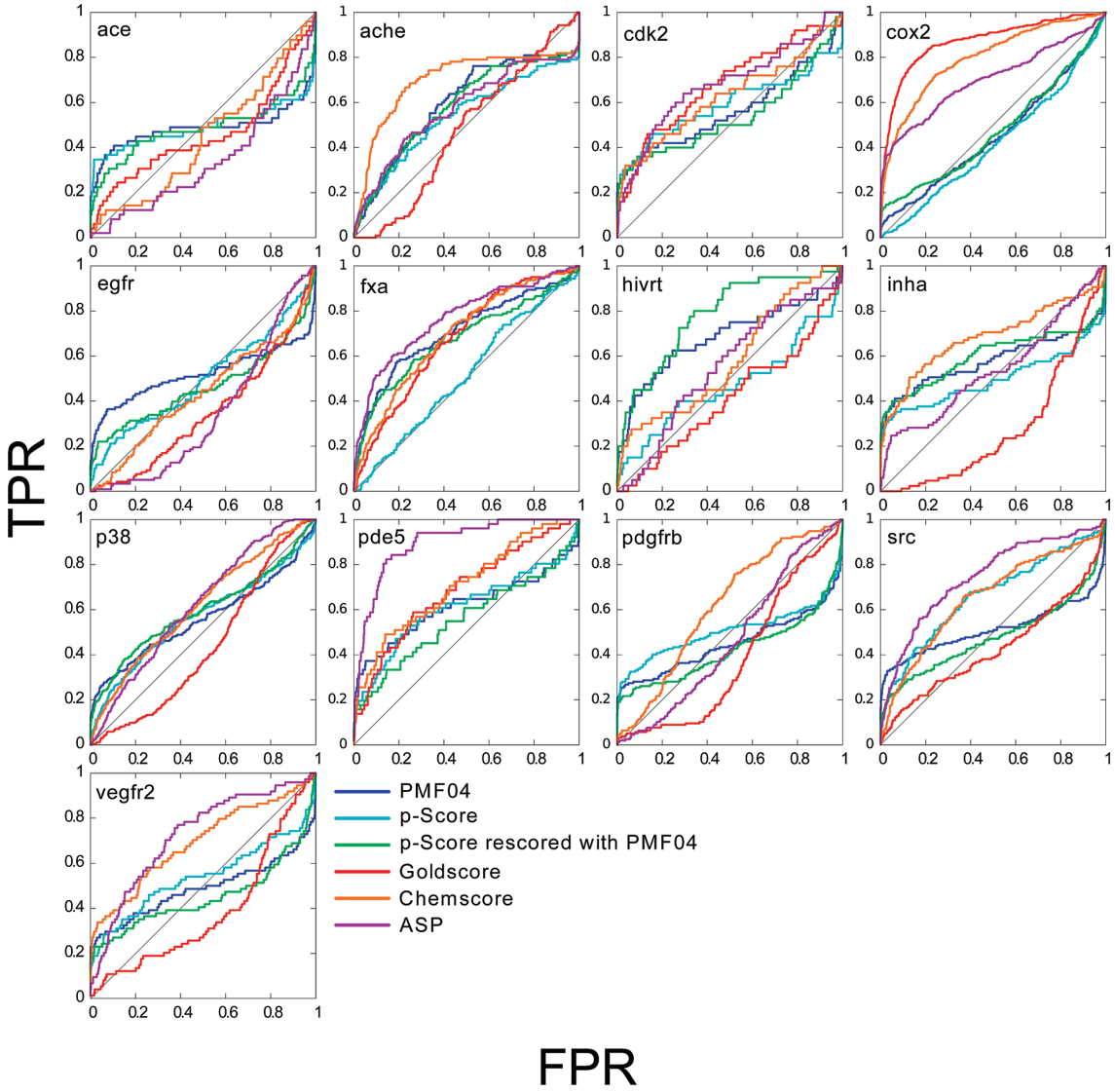


Figure 5. ROC curves to compare the performance of the different VS methods in ParaDockS and GOLD. The lines are colored as follows: ParaDockS with PMF04 in blue, ParaDockS with p-Score in cyan, ParaDockS docked with p-Score and rescored with PMF04 in green, GOLD with GoldScore in red, GOLD with ChemScore in orange, GOLD with ASP in purple.

Table 2. AUC Values for the ROC Curves[a]

target	PMF04	p-Score	p-Score/PMF04	GoldScore	ChemScore	ASP	DOCK
ace	0.49	0.49	0.49	0.46	0.44	0.34	**0.68**
ache	0.60	0.54	0.58	0.47	**0.69**	0.57	0.68
cdk2	0.56	0.59	0.54	**0.68**	0.63	**0.68**	0.57
cox2	0.46	0.42	0.48	**0.87**	0.80	0.71	0.82
egfr	0.52	0.50	0.47	0.36	0.46	0.37	**0.57**
fxa	0.71	0.51	0.68	0.69	0.72	**0.78**	0.73
hivrt	0.68	0.47	**0.78**	0.41	0.59	0.55	0.68
inha	0.58	0.50	0.60	0.29	**0.70**	0.56	0.27
p38	0.56	0.57	0.60	0.45	0.63	**0.64**	0.42
pde5	0.61	0.61	0.56	0.69	0.73	**0.90**	0.56
pdgfrb	0.45	0.51	0.42	0.39	**0.63**	0.49	0.36
src	0.51	0.66	0.48	0.44	0.67	**0.76**	0.48
vegfr2	0.49	0.54	0.45	0.39	0.70	**0.73**	0.38
Average	0.56	0.53	0.55	0.51	**0.65**	0.62	0.55

[a] The highest AUC value for each test set is highlighted in bold numbers. The screening method is abbreviated by the scoring method in use. Results for DOCK were taken from ref 48.

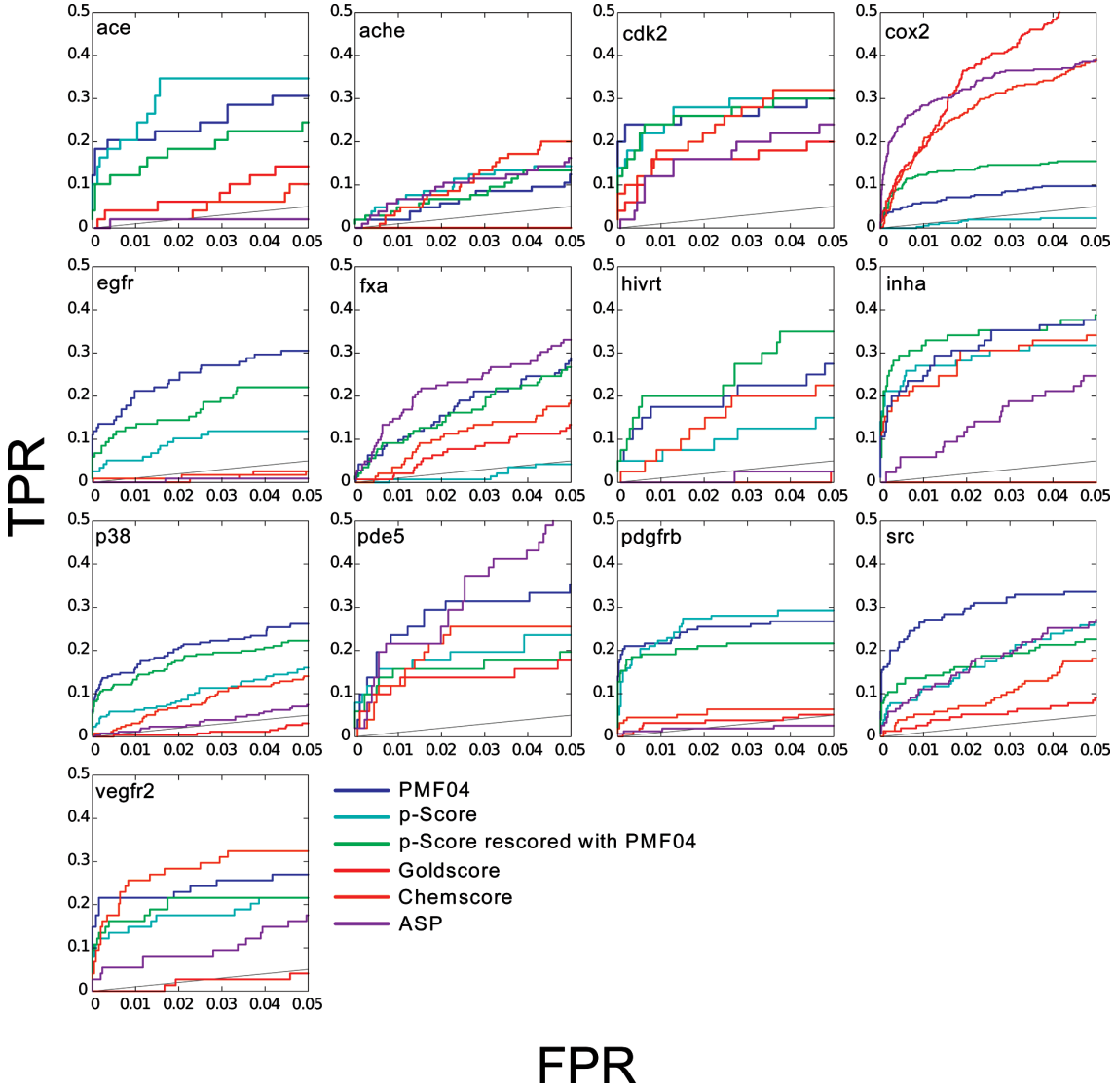


Figure 6. The first 5% of the ROC curves enlarged to compare the early enrichment of the different VS methods in ParaDockS and GOLD. The lines are colored as follows: ParaDockS with PMF04 in blue, ParaDockS with p-Score in cyan, ParaDockS docked with p-Score and rescored with PMF04 in green, GOLD with GoldScore in red, GOLD with ChemScore in orange, GOLD with ASP in purple.

show average values above 0.6, Dock averages at 0.55, and ParaDockS reaches 0.56 with PMF04, 0.53 with p-Score, and 0.55 with PMF04-rescored p-Score results. It is worth mentioning that ParaDockS is hardly producing strong outliers with AUC values significantly below 0.5.

For practical reasons, the enrichment within the few top-ranking solutions is of great interest; economic demands allow only the processing of a limited number of compounds. ROC enrichment[52] is defined as the ratio of TPR to FPR for a given range of decoys and gives a good measure for the "early" enrichment in a virtual screening experiment. The advantage of ROC enrichment values is their independence from the composition of the test set. Most of the ROC curves in Figure 5 are sigmoidal, where it is striking that ParaDockS with PMF04 produces mainly very steep ascending ROC curves. The magnification in Figure 6 puts the focus on the top-ranking 5% of solutions and highlights the high early enrichment. Especially for the very early enrichment (upper 1%) in Table 3, the superiority of ParaDockS is clear, especially in combination with the PMF04 objective function. At 5% ROC enrichment (cf., Table 4) the advantage of ParaDockS is still significant.

In Figure 7, exemplary docking results of two active compounds to the HIV reverse transcriptase are shown. The binding pocket of HIV-RT is narrow, and the shown docking results in Figure 7 are correct. Although the deviations of the predicted structures from the crystal structure are small, the ranking is not necessarily good. The ligand emivirine in Figure 7a is ranked sixth of 1437 by PMF04-rescoring of p-Score results but on position 1310 by GoldScore. The ligand nevirapine in Figure 7b is ranked on position 98 by PMF04, on position 286 by GoldScore, and on position 1012 by p-Score.

Timings and Parallel Efficiency. The computing time is of innate importance for the application of molecular docking techniques especially when performing virtual screenings

Table 3. ROC Enrichment Values at 1% for ParaDockS, GOLD, and DOCK (from ref 48) across the 13 Selected DUD Targets[a]

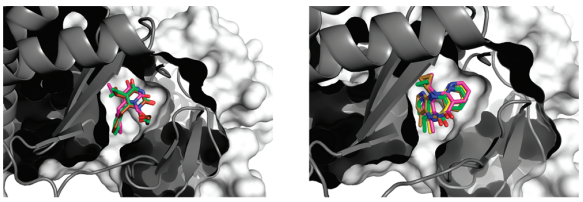
target	PMF04	p-Score	p-Score/PMF04	GoldScore	ChemScore	ASP	DOCK
ace	**44.1**	35.2	17.6	4.4	2.1	2.1	8.7
ache	1.9	**6.4**	4.0	0.0	4.0	**6.4**	0.0
cdk2	**61.0**	32.0	39.5	16.4	16.4	16.4	10.6
cox2	6.6	0.9	16.1	23.4	21.9	**58.5**	16.9
egfr	**24.7**	5.6	16.6	0.0	0.8	0.0	4.1
fxa	11.7	0.7	11.7	2.2	3.7	**20.1**	9.5
hivrt	24.0	5.5	**31.4**	0.0	5.5	0.0	6.2
inha	45.9	45.9	**91.2**	0.0	40.5	6.6	0.0
p38	**24.2**	6.8	18.1	0.4	3.3	1.2	0.0
pde5	**39.4**	25.8	20.7	12.7	16.4	31.9	7.7
pdgfrb	**47.2**	35.8	35.8	3.4	4.9	1.3	0.0
src	**48.9**	10.3	20.1	2.7	5.8	11.4	1.0
vegfr2	**55.2**	23.1	30.9	1.3	35.7	7.8	2.1

[a] The screening method is abbreviated by the scoring method in use. The highest ROC enrichment value for each test set is highlighted in bold numbers.

Table 4. ROC Enrichment Values at 5% for ParaDockS, GOLD, and DOCK (from 48) across the 13 Selected DUD Targets[a]

target	PMF04	p-Score	p-Score/PMF04	GoldScore	ChemScore	ASP	DOCK
ace	7.1	**8.3**	5.0	3.0	2.1	0.4	3.9
ache	2.2	3.0	2.8	0.0	**4.4**	3.0	1.6
cdk2	6.4	6.9	6.9	4.3	**7.5**	4.8	3.0
cox2	2.0	0.5	3.3	**12.3**	8.6	9.0	10.0
egfr	**6.9**	2.5	4.9	0.5	0.5	0.2	3.5
fxa	5.5	0.8	5.5	2.5	3.8	**7.0**	5.1
hivrt	5.6	3.2	**8.4**	0.5	4.4	0.5	3.1
inha	**8.9**	7.5	**8.9**	0.0	7.8	5.0	0.0
p38	**5.8**	3.2	4.9	0.6	2.8	1.4	0.4
pde5	7.9	5.3	3.8	3.3	5.8	**10.4**	6.2
pdgfrb	6.1	**6.8**	4.8	1.0	1.3	0.5	0.2
src	**7.7**	5.6	4.8	1.7	3.7	5.6	0.4
vegfr2	6.5	5.1	5.1	1.1	**8.0**	3.8	0.8

[a] The screening method is abbreviated by the scoring method in use. The highest ROC enrichment value for each test set is highlighted in bold numbers.



(a) (b)

Figure 7. Visualization of example docking results from the virtual screening experiments: (a) the ligand binding domain of the HIV reverse transcriptase (PDB 1RT1) in complex with Emivirine (green) and the docking results with p-Score (magenta), and GoldScore (brown) and (b) the ligand binding domain of the HIV reverse transcriptase (PDB 1VRT) in complex with Nevirapine (green) and the docking results with PMF04 (yellow), p-Score (magenta), and GoldScore (brown).

with large libraries of compounds. To compare the timings, a docking simulation with 50 consecutive runs of a test instance (PDB 1JLA, HIV-1 reverse transcriptase and TNK-652)[44] was performed on an HP server (2.53 GHz Intel Xeon CPUs). ParaDockS with p-Score finishes after about one hour. With PMF04 the timing is almost the same, the advantage of the simpler energy function is reduced due to the all-atom description. GOLD solves the given problem in slightly less than half an hour.

Parallel computing offers a chance for speed-up and is of special interest as there is a clear trend toward multicore CPUs in servers and workstations. In initial studies on parallel efficiency with artificial molecular docking setups we observed almost linearly scaling parallel efficiency with up to 512 CPU cores.[53] However, in the current version the amount of computing time needed to evaluate the interaction was greatly reduced compared to these initial tests. The current parallel implementation suffers from communication overheads; neither the optimization algorithm nor the objective functions are, at the current stage, optimized toward parallel processing. ParaDockS with 4 CPU cores reaches the speed of GOLD with 1 CPU core. While the amount of simultaneous processes for many commercial solutions is limited by the number of licenses purchased, ParaDockS is free software and not limited in the number of processes. This allows to overcome the current speed limitations by data parallel computation. Further details on timings and parallel performance can be found in the Supporting Information.

CONCLUSIONS

In the previous sections, we have introduced the molecular docking software ParaDockS. The main feature, the open and easy extendable design, offers the possibility to implement one's own approaches. In addition, the sofware is equipped with a robust particle swarm optimizer and the two objective functions PMF04 and p-Score. ParaDockS does not need extensive preprocessing of the input data. Input and output of structures as well as parameters and results is

organized in a way that makes the inclusion into existing virtual screening pipelines easy. The code is well structured and is documented by an automatic documentation system to allow easy familiarization. Furthermore, developers of optimization algorithms and scoring approaches will find the open and modular design of the PARADOCKS framework tailored for easy implementation and testing.

The performance was evaluated for three issues critical for molecular docking and virtual screening: accuracy, screening performance, and speed. In all three disciplines PARADOCKS is reaching very promising results. The docking accuracy, tested on reproducing the PDBbind core set, reaches up to 73% with p-Score. To assess the virtual screening performance, extensive testing with 13 targets of the DUD was performed. The early enrichment performance of PARADOCKS with the PMF04 objective function is superior to all other tested approaches. Summarizing, p-Score appears well suited for more accurate evaluations, while PMF04 is apparently well suited for rapid evaluations and high enrichment in virtual high throughput screenings. The particle swarm optimizer performs well and is robust. It offers a straightforward way for parallelization, but with current objective functions the communication overhead is high.

Although PARADOCKS is ready for production use, the software is under constant development. This first status report would be incomplete without an outlook on upcoming improvements and future development directions:

- Improvements of the description of the ligand conformation for the p-Score and PMF04 objective function are planned.
- The receptor flexibility will be accounted for by an explicit modeling of side chain flexibility.
- Further optimization techniques, for example, differential evolution, will be implemented and their performance analyzed.
- Improved load balancing and reduced communication will increase the parallel efficiency. The use of computationally more demanding objective functions will increase the parallel efficiency as well.
- The output of resulting structures will be changed to a trajectory-like format.

PARADOCKS is free software and published under the GNU General Public License (GPLv2).[29]

DOWNLOAD

Please refer to http://www.paradocks.org to download the PARADOCKS source code and to find additional information.

ACKNOWLEDGMENT

The authors, especially F.B. and C.B., gratefully acknowledge the support of this work by M. Teresa Pisabarro (BIOTEC, TU Dresden). The authors thank Daniel Merkle (SDU Odense) and Robert Günther (Universität Leipzig) for inspiring and fruitful discussions and Scott Edwards (PICB Shanghai) for proof reading of the manuscript. C.B. is grateful for a Feodor Lynen fellowship by the Alexander von Humboldt foundation. F.B. and C.B. are grateful for support by Klaus Tschira Foundation during their stay at BIOTEC, TU Dresden. Computational resources at the Centre for Information Services and High Performance Computing of TU Dresden and at the Computing Centre of Martin-Luther-Universität Halle-Wittenberg are gratefully acknowledged. The trademark "Paradocks" is owned by ParaDocks Omnimedia GbR (Freiburg, Germany); we are grateful for the permission to use this name for the molecular docking software.

Supporting Information Available: Data on ligand conformations, timings and parallel efficiency, the active compounds of the two described virtual screening test sets, and p-Score parameters. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bajorath, F. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(2) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580–594.
(3) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.
(4) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similaritybased search engine. *J. Med. Chem.* **2003**, *46*, 499–511.
(5) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, B. S.; Johnson, A. P. eHiTS: an innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.* **2006**, *7*, 421–435.
(6) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76–90.
(7) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(8) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*, 723–732.
(9) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
(10) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
(11) Korb, O.; Stützle, T.; Exner, T. E. An ant colony optimization approach to flexible proteinligand docking. *Swarm Intelligence* **2007**, *1*, 115–134.
(12) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem Inf. Model.* **2009**, *49*, 84–96.
(13) Janson, S.; Merkle, D.; Middendorf, M. Molecular docking with multi-objective particle swarm optimization. *Appl. Soft. Comput.* **2008**, *8*, 666–675.
(14) Chen, H. M.; Liu, B. F.; Huang, H. L.; Hwang, S. F.; Ho, S. Y. SODOCK: Swarm optimization for highly flexible protein-ligand docking. *J. Comput. Chem.* **2007**, *28*, 612–623.
(15) Namasivayam, V.; Günther, R. PSO@AUTODOCK: A fast flexible molecular docking program based on swarm intelligence. *Chem. Biol. Drug Des.* **2007**, *70*, 475–484.
(16) Kennedy, J.; Eberhart, R. Particle swarm optimization. *Proc. IEEE Int. Conf. Neural Networks* **1995**, *4*, 1942–1948.
(17) Muegge, I.; Rarey, M. Small Molecule Docking and Scoring. *Rev. Comput. Chem.* **2001**, *17*, 1–60.
(18) Böhm, H. J.; Stahl, M. The Use of Scoring Functions in Drug Discovery Applications. *Rev. Comput. Chem.* **2002**, *18*, 41–87.
(19) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.
(20) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.
(21) Wang, R. X.; Lai, L. H.; Wang, S. M. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11–26.
(22) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP - potential of mean force describing protein-ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165–1176.

(23) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. BLEEP - potential of mean force describing protein-ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.* **1999**, *20*, 1177–1185.
(24) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791–804.
(25) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902.
(26) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict proteinligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
(27) Shoemake, K. Animating rotation with quaternion curves. *Proc. 12th Annu. Conf. Comput. Graphics Interactive Tech.* **1985**, 245–254.
(28) Tripos L. P. *Tripos Mol2 File Format*; Tripos: St. Louis, MO, 2007; http://www.tripos.com/tripos_resources/fileroot/mol2_format_Dec07.pdf (accessed November 30, 2009).
(29) *GNU General Public License*, version 2; Free Software Foundation, Inc.: Boston, MA, 1991; http://www.gnu.org/licenses/gpl-2.0.txt (accessed November 6, 2009).
(30) Doxygen—Source code documentation generator tool. http://www.stack.nl/~dimitri/doxygen (accessed October 3, 2009).
(31) SMARTS—A language for describing molecular patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed October 3, 2009).
(32) ISO/IEC 14977 Information technology—Syntactic metalanguage—Extended BNF. http://standards.iso.org/ittf/PubliclyAvailableStandards/s026153_ISO_IEC_14977_1996(E).zip (accessed October 3, 2009).
(33) Shi, Y.; Eberhart, R. Parameter selection in particle swarm optimization. LNCS. *Proc. 7th Annu. Conf. Evol. Program., San Diego, CA* **1998**, *1447*, 591–600.
(34) Bytheway, I.; Kepert, D. L. The mathematical modelling of cluster geometry. *J. Math. Chem.* **1992**, *9*, 161–181.
(35) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47*, 5076–5084.
(36) Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* **1999**, *42*, 2498–503.
(37) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
(38) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225–242.
(39) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2007**, *153*, 7–26.
(40) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.
(41) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.
(42) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(43) *MOE 2007.09*; Chemical Computing Group Inc.: Montreal, PQ, Canada, 2009.
(44) Ren, J.; Nichols, C.; Bird, L.; Chamberlain, P.; Weaver, K.; Short, S.; Stuart, D. I.; Stammers, D. K. Structural mechanisms of drug resistance for mutations at codons 181 and 188 in HIV-1 reverse transcriptase and the improved resilience of second generation non-nucleoside inhibitors. *J. Mol. Biol.* **2001**, *312*, 795–805.
(45) *World Drug Index*; Thomson Scientific: Philadelphia, PA, 2001.
(46) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(47) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.
(48) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: Virtual screening using molecular fields. Application to the DUD data set. *J. Chem. Inf. Model.* **2008**, *48*, 2108–2117.
(49) *DUD A Directory of Useful Decoys*; University of California; San Francisco, CA, 2006; http://dud.docking.org/r2 (release date October 22, 2006).
(50) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions 1: The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
(51) Mooij, W. T. M.; Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins* **2005**, *61*, 272–287.
(52) Nicholls, A. What do we know and when do we know it. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
(53) Brandt, F. A Parallel Framework for High-Throughput Automated Docking. Diploma thesis, TU Dresden, 2007.

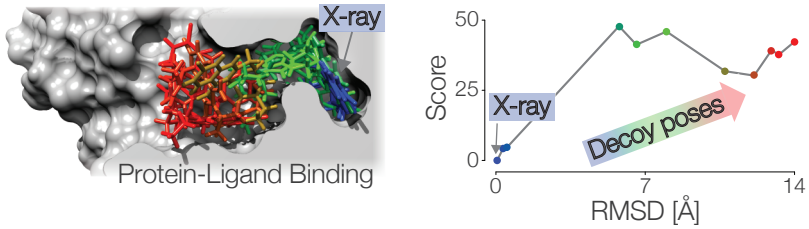
CI900467X

4.2 The SQM/COSMO filter: Reliable native pose identification based on the quantum-mechanical description of protein-ligand interactions and implicit COSMO solvation

Adam Pecina, René Meier, Jindřich Fanfrlík, Martin Lepšík, Jan Řezáč, Pavel Hobza, Carsten Baldauf

The SQM/COSMO Filter: Reliable Native Pose Identification Based on the Quantum-Mechanical Description of Protein-Ligand Interactions and Implicit COSMO Solvation

Chem. Comm. **2016** (52), 3312-3315.
DOI: 10.1039/C5CC09499B



ChemComm

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2016, 52, 3312

Received 16th November 2015,
Accepted 11th January 2016

DOI: 10.1039/c5cc09499b

www.rsc.org/chemcomm

The SQM/COSMO filter: reliable native pose identification based on the quantum-mechanical description of protein–ligand interactions and implicit COSMO solvation†

Adam Pecina,‡[a] René Meier,‡[b] Jindřich Fanfrlík,[a] Martin Lepšík,[a] Jan Řezáč,[a] Pavel Hobza*[ac] and Carsten Baldauf*[d]

Current virtual screening tools are fast, but reliable scoring is elusive. Here, we present the 'SQM/COSMO filter', a novel scoring function featuring a quantitative semiempirical quantum mechanical (SQM) description of all types of noncovalent interactions coupled with implicit COSMO solvation. We show unequivocally that it outperforms eight widely used scoring functions. The accuracy and chemical generality of the SQM/COSMO filter make it a perfect tool for late stages of virtual screening.

Despite the enormous advances in method development for structure-based *in silico* drug design, reliable predictions of the structures (docking) and affinities (scoring) of protein–ligand (P–L) complexes still remain an unsolved task.[1] A plethora of scoring functions (SFs) have been devised by utilising experimental data for regression analyses, by constructing knowledge-based potentials, or based on physical laws.[2,3] As none of the SFs is general enough to perform equally strongly for a diverse set of P–L complexes, utilising several SFs at once (consensus scoring) holds promise.[4] Regression analysis and knowledge-based approaches to scoring are trained on a set of P–L complexes and rely on variable master equation terms. Their validity is limited to complexes similar to the training set. In principle, this problem has been overcome in physics-based methods. Because of computational cost, preference has been given to molecular mechanics (MM) methods, such as the combination of MM interaction energies with implicit solvation free energy terms (generalised Born, GB, or Poisson–Boltzmann, PB) to estimate affinities.[2] Additionally, the wide coverage of organic chemical space in the GAFF (general AMBER force field)[5] has made the parameterisation of ligands for MM straightforward. However, an explicit description of quantum mechanical (QM) effects in P–L interactions, such as charge transfer, polarisation, covalent-bond formation or σ-hole bonding, was missing. QM methods, which describe these effects qualitatively better than the energy functions used in MM-based SFs, were thus introduced into computational drug design.[6,7] Recent developments in QM methods and algorithms as well as the availability of a powerful computing infrastructure have paved the way to apply them for P–L complexes in numerous setups: linear scaling or efficient parallelisation of semi-empirical QM (SQM) methods,[7–10] QM/MM,[7,8,11,12] DFT-D3 on truncated P–L complexes[13] or various fragmentation methods.[11,14] Specifically, AM1, RM1, PM6 or DF–TB SQM methods have been used[7–9,12,15] as such or with empirical corrections for dispersion, hydrogen- and halogen-bonding[16] to describe the P–L noncovalent interactions. Merz *et al.* pioneered this area by introducing a QM-based SF (QMScore), a combination of the AM1 SQM method with an empirical dispersion (D) and the PB implicit solvent [eqn (1)].[17] The method was useful for describing metalloprotein–ligand binding, but further corrections were needed, especially for a quantitative treatment of dispersion and hydrogen bonding.[10]

$$\text{Score} = \Delta E_{\text{int}} + \Delta\Delta G_{\text{solv}} + \Delta G^{\text{rw}}_{\text{conf}} - T\Delta S \qquad (1)$$

The above equation is a general physics-based SF. The terms are the gas-phase interaction energy (ΔE_{int}), the change of solvation free energy upon complex formation ($\Delta\Delta G_{\text{solv}}$), the change of conformational 'free' energy ($\Delta G^{\text{rw}}_{\text{conf}}$) and the change of entropy upon ligand binding ($-T\Delta S$).

Our approach is systematic. Using accurate calculations in small model systems as a benchmark, we developed corrections for SQM methods that provide reliable and accurate description of a wide range of noncovalent interactions including dispersion, hydrogen- and halogen-bonding.[16] Coupled with the PM6 SQM method,[18] the resulting PM6-D3H4X approach is applicable to a wide chemical space and does not require any

[a] *Institute of Organic Chemistry and Biochemistry (IOCB) and Gilead Sciences and IOCB Research Center, Flemingovo nám. 2, 16610 Prague 6, Czech Republic. E-mail: hobza@uochb.cas.cz*
[b] *Institut für Biochemie, Fakultät für Biowissenschaften, Pharmazie und Psychologie, Universität Leipzig, Brüderstrasse 34, D-04109 Leipzig, Germany*
[c] *Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Palacký University, 77146 Olomouc, Czech Republic*
[d] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany. E-mail: baldauf@fhi-berlin.mpg.de*
† Electronic supplementary information (ESI) available. See DOI: 10.1039/c5cc09499b
‡ These authors have contributed equally to this work.

system-specific parameterisation. We use it here to calculate the ΔE_{int} term. Subsequently, we compared MM-based (PB or GB) and QM-based (COSMO[19] or SMD) implicit solvent models and found the latter group to be more accurate.[20] These are therefore used for the $\Delta\Delta G_{solv}$ term. These two dominant terms, ΔE_{int} and $\Delta\Delta G_{solv}$, are at the heart of our SQM-based SF.[15] We have demonstrated its generality in various noncovalent P–L complexes, such as aldose reductase or carbonic anhydrase and moreover extended it to treat covalent inhibitor binding (ref. 15, 21 and 22).

In this work, we adapt our SQM-based SF to make it usable in virtual screening on the basis of our previous experience. By taking the two dominant terms only, ΔE_{int} and $\Delta\Delta G_{solv}$, we define the 'SQM/COSMO filter' energy. Its performance is tested here against eight widely used SFs. GlideScore XP (GlideXP),[23] PLANTS PLP (PLP),[24] AutoDock Vina (Vina),[25] Chemscore (CS),[26] Goldscore (GS)[27] and ChemPLP[24] are empirical, regression-based functions which use different terms to describe vdW contacts, lipophilic surface coverage, hydrogen bonding, ligand strain, and desolvation. The Astex Statistical Potential (ASP)[28] is a knowledge-based potential. The classical physics-based AMBER/GB SF combines the ff03-GAFF MM force fields with the GB implicit solvent.[5,29]

The goal is 'cognate docking',[30] *i.e.* the ability to identify sharply the known native X-ray P–L binding pose from a set of decoy structures generated by docking (Fig. 1). To understand our results in detail, we have not opted for treating them in a statistical manner[31] as in the pose decoy test sets available.[32] Instead we cautiously selected four unrelated difficult-to-handle P–L systems, which comply with strict criteria for the selection of crystallographic structures for docking (details in the ESI†).[33] These systems are acetylcholine esterase (AChE, PDB: 1E66),[34] TNF-α converting enzyme (TACE, PDB: 3B92),[35] aldose reductase (AR, PDB: 2IKJ)[36] and HIV-1 protease (HIV PR; PDB: 1NH0).[37] For the latter, the protonation of the active site is inferred from ultra-high resolution X-ray crystallography. Based on these P–L crystal structures, we have created a set of non-redundant poses (2865 in total) by docking with four popular docking programs (Glide, PLANTS, AutoDock Vina and GOLD) coupled to seven widely used SFs[23–28] (Fig. 1, Table S2, ESI†).

All the poses were re-scored by all nine SFs. For the seven regression- and knowledge-based SFs, we used the recommended protocols. For the two physics-based SFs, only hydrogen atoms and close contacts were relaxed by the AMBER/GB method. RMSD values of the poses relative to the crystal were measured (details in S1.6, ESI†). The scores were normalised and are shown relative to the score of the crystal pose.

The identification of the X-ray pose as the minimum-free-energy structure is an unambiguous criterion for the performance of any SF. The ideal behaviour of such a score *vs.* RMSD curve (Fig. 2) is characterised by the positive values of energies for the decoy poses. Small deviations (negative energies for very small RMSD values) are acceptable and might be explained by inaccuracies of the crystal structure. These conditions are met by the SQM/COSMO filter, unlike the other SFs (Fig. 2). The numbers of false-positive solutions as well as the maximum RMSD ($RMSD^{max}$) from the X-ray pose within a defined interval of the normalised score quantify the virtually ideal behaviour of the SQM/COSMO filter in comparison to the other SFs.

The number of false-positives is lowest for the SQM/COSMO filter, even zero for three P–L systems (Table 1). CS and ASP perform slightly worse. AMBER/GB performs satisfyingly well for three systems but yields 171 false-positives for TACE. For AChE, all the SFs perform satisfyingly well. For AR and HIV PR, GlideXP generates the highest number of false-positive solutions and also shape-wise the free energy landscape looks ill-defined (Fig. 2). In the case of AR, a plateau of negative relative scores is observed for GlideXP. The hardest case is the TACE metalloprotein. Here, all the SFs produce false-positive solutions but to a different extent. The SQM/COSMO filter performs best, followed by CS. This example in particular shows the strength of an electronic-structure theory description of P–L binding. The presence of the metal cation in this protein and the associated charge-transfer effects between the ligand and the cation are not adequately described by classical force-fields

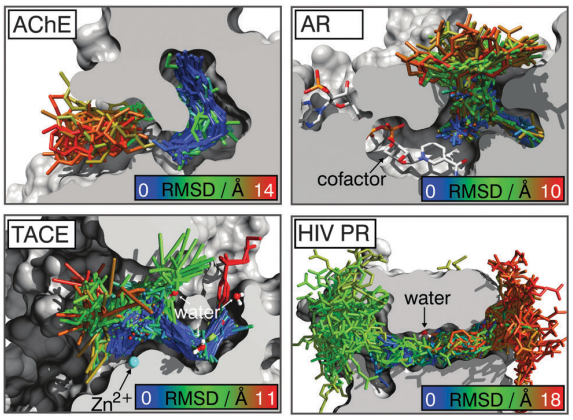


Fig. 1 The ligand poses generated by the four docking programs. Ligand poses are color-coded by RMSD.

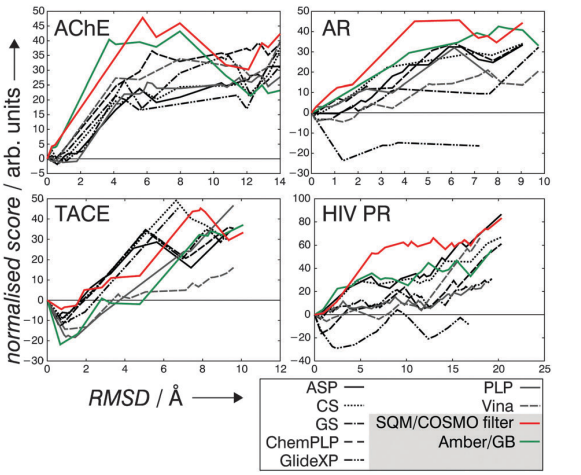


Fig. 2 The plots of normalised scores against RMSD values for all four P–L systems.

Table 1 The numbers of false-positive solutions, *i.e.* solutions that are scored better than the X-ray pose and have RMSD > 0.5 Å

	Scoring function								
			Glide	PLANTS	AutoDock	Gold			
	SQM/COSMO	AMBER/GB	XP	PLP	Vina	ASP	CS	GS	ChemPLP
AChE	0	0	4	12	0	2	3	0	0
AR	0	1	67	0	10	1	0	1	0
TACE	39	171	181	294	63	56	49	78	111
HIV PR	0	0	98	0	7	0	2	1	8
Total	**39**	**172**	**350**	**306**	**80**	**59**	**54**	**80**	**119**

Table 2 The maximum RMSD [Å] within all the poses in the defined range of the relative normalised score

	Scoring function								
			Glide	PLANTS	AutoDock	Gold			
	SQM/COSMO	AMBER/GB	XP	PLP	Vina	ASP	CS	GS	ChemPLP
Maximal RMSD within a window of 5 of the normalised score									
AchE	0.47	0.57	2.13	0.78	0.78	1.78	1.43	1.14	0.78
AR	0.19	0.19	7.54	1.14	3.54	2.32	1.15	2.21	1.49
TACE	1.91	4.76	3.02	2.91	7.13	2.01	1.54	2.44	2.40
HIV PR	0.94	0.94	17.26	12.60	11.62	1.00	1.01	12.60	11.62
Average	**0.88**	**1.62**	**7.49**	**4.61**	**5.77**	**1.78**	**1.28**	**4.60**	**4.55**

or statistical potentials, but they are well represented by the SQM/COSMO filter.

The second criterion, $RMSD^{max}$, is shown for the interval of the normalised relative scores below 5 (Table 2). The SQM/COSMO filter shows the lowest $RMSD^{max}$ of 0.88 Å on average. CS follows with 1.28 Å on average. ASP and AMBER/GB satisfy the conditions of an averaged $RMSD^{max}$ up to 2 Å. AMBER/GB, however, fails in the difficult case of TACE with $RMSD^{max}$ of 4.76 Å. Analogous analyses at greater intervals have revealed a similar ordering of the SFs (Table S4, ESI†).

The SQM/COSMO filter enables us not only to recognise the correct binding pose (RMSD below 2 Å) but also to go beyond this limit and evaluate even small changes in the geometry of the ligand binding.

The price for such a high accuracy is the increased computational time requirements. The SQM/COSMO filter is *ca.* 100-times slower than the statistics- and knowledge-based SFs and about 10-times slower than the classical physics-based AMBER/GB. However, compared to the standard SQM-based SF, it is *ca.* 100-times faster. The speed can be further enhanced by parallelisation.

To summarise, we have pushed the limits of the accuracy of SFs to judge the energetics of P–L noncovalent interactions. Based on our development and the extensive experience with SQM-based scoring functions,[3,21] the SQM/COSMO filter has been introduced. It features two dominant terms to describe P–L interaction, namely the ΔE_{int} term at the PM6-D3H4X level for gas-phase noncovalent interactions and the $\Delta\Delta G_{solv}$ term at the COSMO level for implicit solvation. We showed previously that both these methods are very accurate at a reasonable speed.[16,20] The SQM/COSMO energy is calculated in four unrelated P–L complexes. The SQM/COSMO filter is compared to eight widely used SFs, which are statistics-, knowledge- or force-field-based. The SQM/COSMO scheme exhibits a superior performance as judged by two criteria, the number of false positives and $RMSD^{max}$. In contrast to standard SFs, no fitting against data sets has been involved. Furthermore, it offers generality and comparability across the chemical space and no system-specific parameterisations have to be performed. The time requirements allow for calculations of thousands of docking poses as we have demonstrated in this pilot study. We propose the SQM/COSMO filter as a tool for accurate medium-throughput refinement in later stages of virtual screening or as a reference method for judging the performance of other scoring functions. The proof of concept that reliable QM calculations can now be performed for tens of thousands of large biochemical entities opens a way to progress in closely related disciplines such as materials design.

This work was supported by research projects RVO 61388963 awarded to the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic. We acknowledge the financial support of the Czech Science Foundation (grant number P208/12/G016). The authors acknowledge the support by the project L01305 of the Ministry of Education, Youth and Sports of the Czech Republic. The computations were performed at the Center for Information Services and High Performance Computing (ZIH) at TU Dresden.

References

1 G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevis, S. F. Semus and S. Senger, *et al.*, *J. Med. Chem.*, 2006, **49**, 5912; A. R. Leach, B. K. Shoichet and C. E. Peishoff, *J. Med. Chem.*, 2006, **49**, 5851.
2 H. Gohlke and G. Klebe, *Angew. Chem., Int. Ed.*, 2002, **41**, 2644.
3 R. Meier, M. Pippel, F. Brandt, W. Sippl and C. Baldauf, *J. Chem. Inf. Model.*, 2010, **50**, 879.
4 P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, *J. Med. Chem.*, 1999, **42**, 5100; R. Wang and S. J. Wang, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1422.

5 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157.
6 K. Raha, M. B. Peters, B. Wang, N. Yu, A. M. Wollacott, L. M. Westerhoff and K. M. Merz, *Drug Discovery Today*, 2007, **12**, 725; M. Xu and M. A. Lill, *Drug Discovery Today: Technol.*, 2013, **10**, 411.
7 D. Mucs and R. A. Bryce, *Expert Opin. Drug Discovery*, 2013, **8**, 263.
8 S. A. Hayik, R. Dunbrack and K. M. Merz, *J. Chem. Theory Comput.*, 2010, **6**, 3079.
9 M. Hennemann and T. Clark, *J. Mol. Model.*, 2014, **20**, 2331.
10 H. S. Muddana and M. K. Gilson, *J. Chem. Theory Comput.*, 2012, **8**, 2023; P. Mikulskis, S. Genheden, K. Wichmann and U. Ryde, *J. Comput. Chem.*, 2012, **33**, 1179.
11 P. Soderhjelm, J. Kongsted and U. Ryde, *J. Chem. Theory Comput.*, 2010, **6**, 1726.
12 K. Wichapong, A. Rohe, C. Platzer, I. Slynko, F. Erdmann, M. Schmidt and W. Sippl, *J. Chem. Inf. Model.*, 2014, **54**, 881; P. Chaskar, V. Zoete and U. F. Röhrig, *J. Chem. Inf. Model.*, 2014, **54**, 3137; S. K. Burger, D. C. Thompson and P. W. Ayers, *J. Chem. Inf. Model.*, 2011, **51**, 93.
13 J. Antony, S. Grimme, D. G. Liakos and F. Neese, *J. Phys. Chem. A*, 2011, **115**, 11210.
14 M. S. Gordon, D. G. Fedorov, S. R. Pruitt and L. V. Slipchenko, *Chem. Rev.*, 2012, **112**, 632; J. Antony and S. Grimme, *J. Comput. Chem.*, 2012, **33**, 1730.
15 M. Lepšík, J. Řezáč, M. Kolář, A. Pecina, P. Hobza and J. Fanfrlík, *ChemPlusChem*, 2013, **78**, 921.
16 J. Řezáč, J. Fanfrlík, D. Salahub and P. Hobza, *J. Chem. Theory Comput.*, 2009, **5**, 1749; J. Řezáč and P. Hobza, *Chem. Phys. Lett.*, 2011, **506**, 286; J. Řezáč and P. Hobza, *J. Chem. Theory Comput.*, 2012, **8**, 141.
17 K. Raha and K. M. Merz, *J. Am. Chem. Soc.*, 2004, **126**, 1020; K. Raha and K. M. Merz, *J. Med. Chem.*, 2005, **48**, 4558.
18 J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173.
19 A. Klamt and G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799.
20 M. Kolář, J. Fanfrlík, M. Lepšík, F. Forti, F. J. Luque and P. Hobza, *J. Phys. Chem. B*, 2013, **117**, 5950.
21 J. Fanfrlík, A. K. Bronowska, J. Řezáč, O. Přenosil, J. Konvalinka and P. Hobza, *J. Phys. Chem. B*, 2010, **114**, 12666; P. Dobeš, J. Řezáč, J. Fanfrlík, M. Otyepka and P. Hobza, *J. Phys. Chem. B*, 2011, **115**, 8581; A. Pecina, O. Přenosil, J. Fanfrlík, J. Řezáč, J. Granatier, P. Hobza and M. Lepšík, *Collect. Czech. Chem. Commun.*, 2011, **76**, 457; A. Pecina, M. Lepšík, J. Řezáč, J. Brynda, P. Mader, P. Řezáčová, P. Hobza and J. Fanfrlík, *J. Phys. Chem. B*, 2013, **117**, 16096; J. Fanfrlík, M. Kolář, M. Kamlar, D. Hurn, F. X. Ruiz, A. Cousido-Siah, A. Mitschler, J. Řezáč, E. Munusamy and M. Lepšík, *et al.*, *ACS Chem. Biol.*, 2013, **8**, 2484; J. Fanfrlík, F. X. Ruiz, A. Kadlčíková, J. Řezáč, A. Cousido-Siah, A. Mitschler, S. Haldar, M. Lepšík, M. H. Kolář and P. Majer, *et al.*, *ACS Chem. Biol.*, 2015, **10**, 1637.
22 J. Fanfrlík, P. S. Brahmkshatriya, J. Řezáč, A. Jílková, M. Horn, M. Mareš, P. Hobza and M. Lepšík, *J. Phys. Chem. B*, 2013, **117**, 14973.
23 R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin and D. T. Mainz, *J. Med. Chem.*, 2006, **49**, 6177.
24 O. Korb, T. Stützle and T. E. Exner, *J. Chem. Inf. Model.*, 2009, **49**, 84.
25 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455.
26 M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 425.
27 G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol.*, 1997, **267**, 727.
28 W. T. M. Mooij and M. L. Verdonk, *Proteins*, 2005, **61**, 272.
29 Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo and R. T. Lee, *J. Comput. Chem.*, 2003, **24**, 1999; V. Tsui and D. A. Case, *Biopolymers*, 2001, **56**, 275.
30 A. Nicholls and A. N. Jain, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 133.
31 B. Liu, S. Wang and X. Wang, *Sci. Rep.*, 2015, **50**, 15479.
32 E. Perola, W. P. Walters and P. S. Charifson, *Proteins*, 2004, **56**, 235; J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole and R. Taylor, *Proteins*, 2002, **49**, 457; P. Ferrara, H. Gohlke, D. J. Price, G. Klebe and C. L. Brooks, *J. Med. Chem.*, 2004, **47**, 3032.
33 G. Klebe, *Drug Discovery Today*, 2006, **11**, 580.
34 H. Dvir, D. M. Wong, M. Harel, X. Barril, M. Orozco, F. J. Luque, D. Munoz-Torrero, P. Camps, T. L. Rosenberry and I. Silman, *et al.*, *Biochemistry*, 2002, **41**, 2970.
35 U. K. Bandarage, T. Wang, J. H. Come, E. Perola, Y. Wei and B. G. Rao, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 44.
36 H. Steuber, A. Heine and G. Klebe, *J. Mol. Biol.*, 2007, **368**, 618.
37 J. Brynda, P. Řezáčová, M. Fábry, M. Hořejší, R. Štouračová, J. Sedláček, M. Souček, M. Hradílek, M. Lepšík and J. Konvalinka, *J. Med. Chem.*, 2004, **47**, 2030.

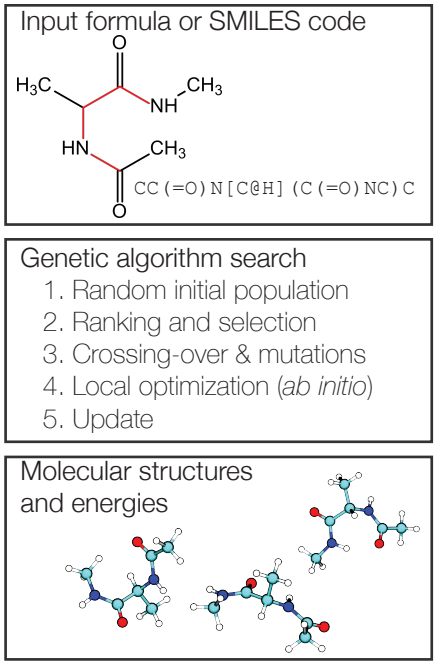
4.3 First-principles molecular structure search with a genetic algorithm

Supady, A.; Blum, V.; Baldauf, C.

First-principles molecular structure search with a genetic algorithm

J. Chem. Inf. Model. **2015** (55), 2338-2348.

DOI:10.1021/acs.jcim.5b00243



First-Principles Molecular Structure Search with a Genetic Algorithm

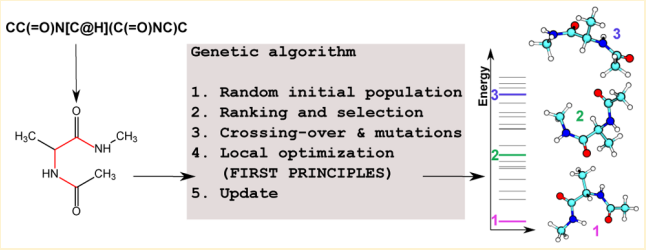
Adriana Supady,*[,†] Volker Blum,[†,‡] and Carsten Baldauf*[,†]

[†]Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany

[‡]Department of Mechanical Engineering & Materials Science, Duke University, Durham, North Carolina 27708, United States

Supporting Information

ABSTRACT: The identification of low-energy conformers for a given molecule is a fundamental problem in computational chemistry and cheminformatics. We assess here a conformer search that employs a genetic algorithm for sampling the low-energy segment of the conformation space of molecules. The algorithm is designed to work with first-principles methods, facilitated by the incorporation of local optimization and blacklisting conformers to prevent repeated evaluations of very similar solutions. The aim of the search is not only to find the global minimum but to predict all conformers within an energy window above the global minimum. The performance of the search strategy is (i) evaluated for a reference data set extracted from a database with amino acid dipeptide conformers obtained by an extensive combined force field and first-principles search and (ii) compared to the performance of a systematic search and a random conformer generator for the example of a drug-like ligand with 43 atoms, 8 rotatable bonds, and 1 cis/trans bond.



INTRODUCTION

One of the fundamental problems in cheminformatics and computational chemistry is the identification of three-dimensional (3D) conformers that are energetically favorable and likely to be encountered in experiment at given external conditions.[1] Conventionally, these conformers are often characterized by specific, fixed sets of nuclear coordinates or ensembles thereof, and their potential energy is given by the electronic degrees of freedom in a Born−Oppenheimer picture of the chemical bond. A variety of conformations can be adopted by flexible organic molecules as the multidimensional potential-energy surface (PES) usually contains multiple local minima, with a global minimum among them. Only when the relevant conformers are known, one can predict and evaluate chemical and physical properties of the molecules (e.g., reactivity, catalytic activity, or optical properties). In many practical applications, the PES minima are taken as starting points to explore the free-energy surface (FES). Generating conformers is an integral part of methods such as protein−ligand docking[2−5] or 3D pharmacophore modeling.[6] The propensity to adopt a certain conformation strongly depends on the environment and possible interactions with other compounds. It has been shown that the bioactive conformation of drug-like molecules can be higher in energy than the respective global minimum[7] and that different 3D conformations may be induced by specific interactions with other molecules.[8] Thus, it is crucial to focus not just on a single, global minimum of the PES, but instead to provide a good coverage of the accessible conformational space of a molecule yielding diverse low-energy conformers.

The exploration of a high-dimensional PES is challenging. A selection of popular sampling approaches utilized in conformer generation is summarized in Table 1. We focus specifically on genetic algorithms (GAs),[29−31] a sub-group of the evolutionary algorithms (EAs), that are frequently used for global structure optimization of chemical compounds.[3,4,32−56] GAs for chemical structure searches implement a "survival of the fittest" concept and adopt evolutionary principles starting from a population of, most commonly, random solutions. GAs use the accumulated information to explore the most promising regions of the conformational space. With this, the number of unhelpful evaluations of physically implausible high-energy solutions can be reduced. Examples of GA-based structure prediction applications include the following: (i) conformational searches for molecules like of unbranched alkanes[36] or polypeptide folding;[37] (ii) molecular design;[38,39] (iii) protein−ligand docking;[3,4] (iv) cluster optimization;[40−49] (v) predictions of crystal structures;[50−53] (vi) structure and phase diagram predictions.[54] Further, Neiss and Schoos[55] proposed a GA including experimental information into the global search process by combining the energy with the experimental data in the objective function. Since GAs typically rely on internal, algorithmic parameters that control the efficiency of a search, a meta-GA for optimization of a GA search for conformer searches was proposed by Brain and Addicoat.[56]

Aside from the search algorithm itself, the choice of the mathematical model for the PES is critical to ensure results that reliably reflect the experimental reality. Among the available atomistic simulation approaches, "molecular mechanics" models, i.e., so-called force fields, are especially fast from a computational point of view and therefore often employed.

Received: April 28, 2015
Published: October 20, 2015

Table 1. Popular Sampling Approaches[a]

method	description	implemented, e.g., in
grid-based	based on grids of selected Cartesian or internal coordinates (e.g., grids of different torsional angle values of a molecule)	CAESAR,[9] **Open Babel**,[10] **Confab**,[11] MacroModel,[12] MOE[13]
rule/knowledge-based	use known (e.g., from experiments) structural preferences of compounds	**ALFA**,[14] **CONFECT**,[15] CORINA and ROTATE,[16,17] **COSMOS**,[18,19] OMEGA[20]
population-based metaheuristic	improve candidate solutions in a guided search	**Balloon**,[21] **Cyndi**[22]
distance geometry	based on a matrix with permitted distances between pairs of atoms	**RDKit**[23]
basin-hopping[24]/ minima hopping[25]	based on moves across the PES combined with local relaxation	**ASE**,[26] **GMIN**,[27] **TINKER SCAN**[28]

[a]Names of freely available programs are highlighted in boldface.

However, the resulting predictions depend on the initial parametrization of a particular force field and can lead to considerable rearrangements of the true PES for molecules that were not included in the parametrization procedure.[57−59] On the other end of the spectrum of approaches, the PES can be faithfully represented based on the "first-principles" of quantum mechanics. Indeed, benchmark quality approaches such as coupled cluster theory at the level of singles, doubles and perturbative triples (CCSD(T)) are almost completely trustworthy for closed-shell molecules, but still prohibitively expensive toward larger systems and/or large-scale screening of energies of many conformers. Density-functional theory (DFT) approximations are an attractive alternative to balance accuracy and computational cost. The choice of the approximation is critical when using first-principles methods like DFT. It has been shown that it is necessary to incorporate dispersion effects for (bio)organic molecules and their complexes.[57,60−62] The challenge of including long-range interactions has been met for example by the dispersion correction schemes described by Grimme[63,64] or by Scheffler and Tkatchenko,[65−67] but validating the DFT approximation employed is critical. In fact, subtle energy balances of competing conformers can require relatively high-level DFT approximations for reliable predictions.[58,68]

The aim of our work is to develop and test an approach to sample the PES of small to medium sized (bio)organic molecules without relying on empirical force fields, utilizing instead electronic-structure methods for the entire search. With the molecular structure problem in mind, we define the following requirements for the search strategy and implementation:

- Global search based on user-curated torsional degrees of freedom (bond rotations).
- Local optimization based on full relaxation of Cartesian coordinates and avoidance of recomputing too similar structures to ensure both efficient sampling and economic use of a computationally demanding energy function.
- Design of the program in a way to use an external and easily exchangeable electronic structure code (in our case FHI-aims[69,70]).
- Simple input of molecules (composition and configuration) by means of SMILES codes.[71]
- A robust and simple metaheuristic that ideally identifies the complete ensemble of low-energy conformers.
- Free availability with a flexible open-source license model.
- Support for parallel architectures.

Based on these requirements, we present in this work a conformational search strategy based on a genetic algorithm. We provide a detailed description of our approach and a software implementation Fafoom (flexible algorithm for optimization of molecules) that is available under an open-source license (GNU Lesser General Public License[72]) for use by others. For simplicity, we abbreviate "potential energy" with energy and "minima of the potential-energy surface" with energy minima.

METHODS

In the following, we first motivate and explain assumptions that we met for handling 3D structures of molecules. Further, we outline the algorithm's implementation and describe its technical details. Finally we introduce a data set that we use as a reference for evaluating the performance of our implementation. Our work focuses on both the ability to reliably predict the global minimum and to provide a good conformational coverage with a computationally feasible approach. To achieve that, we formulate some specific algorithmic choices at the outset: (i) only sterically feasible conformations are accepted for local optimization; (ii) a geometry optimization to the next local minimum is performed for every generated conformation; (iii) an already evaluated conformation will not be evaluated again.

Choice of Coordinates. In computational chemistry, at least two ways of representing a molecule's 3D structure are commonly used, either Cartesian or internal coordinates. The simplest internal coordinates are based on the "Z-matrix coordinates", which include bond lengths, bond angles and dihedral angles (torsions) and can also be referred to as "primitive internal coordinates". These coordinates reflect the actual connectivity of the atoms and are well suited for representing curvilinear motions such as rotations around single bonds.[73] Bond lengths and bond angles possess usually only one rigid minimum, i.e. the energy increases rapidly if these parameters deviate from equilibrium. In contrast, torsions can change in value by an appreciable amount without a dramatic change in energy. Similar to the work of Damsbo et al.,[37] we use Cartesian coordinates for the local geometry optimizations while internal coordinates, in this work only torsional degrees of freedom (TDOFs), i.e. freely rotatable bonds and, if present, cis/trans bonds, are used for the global search. We consider only single, non-ring bonds between non-terminal atoms to be rotatable bonds after excluding bonds that are attached to methyl groups that carry three identical substituents. Further we allow for treating selected bonds in a cis/trans mode, i.e. allowing only for two different relative positions of the substituents. In cases in which the substituents are oriented in the same direction we refer to it as to cis, whereas, when the

substituents are oriented in opposite directions, we refer to it as to trans.

Handling of Molecular Structures. Figure 1 shows different chemical representations of a molecule, here for the example of 3,4-dimethylhex-3-ene. Figure 1A and B depict the standard 3D and 2D representation of the compound together with marked cis/trans and rotatable bonds. A SMILES (simplified molecular-input line-entry system) string is shown in Figure 1C. A SMILES representation[71] of a chemical compound encodes the composition, connectivity, and bond order (single, double, triple), as well as stereochemical information in a one-line notation. Finally, a vector representation (Figure 1D) can be created if the locations of cis/trans and rotatable bonds are known. The vector will store the corresponding torsion angle values. Our implementation takes as input a SMILES representation of a molecule, while vectors of angles are used to internally encode different structures in the genetic algorithm below.

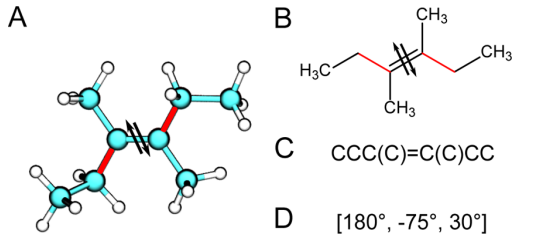


Figure 1. Different chemical representations of 3,4-dimethylhex-3-ene. (A) 3D structure with rotatable bonds marked in red and the cis/trans bond marked with double arrows. (B) 2D structure. (C) SMILES string. (D) Vector representation of the molecule. The first value encodes the torsion angle value for the cis/trans bond and the two remaining position store the torsion angle values of the rotatable bonds.

Frequently Used Terms. Several terms need to be defined prior to describing the structure of the algorithm. In the following, the parameters of the search are highlighted in boldface. These parameters are input parameters to the algorithm and need to be defined in the input file.

A *sensible geometry* meets two constraints. First, the atoms are kept apart, i.e. none of the distances between nonbonded atoms can be shorter than a defined threshold (**distance_cutoff_1**, default = 1.3 Å). Second, it is fully connected, i.e., none of the distances between bonded atoms can be longer than a defined threshold (**distance_cutoff_2**, default = 2.15 Å). The attribute *sensible* can be used further to describe any operation that outputs sensible geometries.

The *blacklist* stores all structures that (i) were starting structures for the local optimization and (ii) resulted from local optimization, as they may have changed significantly during the optimization. In case of achiral molecules (**chiral**, default = False) also the corresponding mirror images are created and stored.

A structure is *unique* if none of the root-mean-square deviation (RMSD) values calculated for the structure paired with any of the structures stored in the blacklist is lower than a defined threshold (**rmsd_cutoff_uniq**, default = 0.2 Å). We consider only non-hydrogen atoms for the calculation of the RMSD.

Basic Outline of the Search Algorithm. We implemented the genetic algorithm (GA) using the Python language (version 2.7) and employ the RDKit library.[23] An overview is presented in Algorithm 1.

```
# initialization
while i < popsize:
    x = random_sensible_geometry
    blacklist.append(x)
    x = DFT_relaxation(x)
    blacklist.append(x)
    population.append(x)
    i+=1
# iteration
while j < iterations:
    population.sort(index=energy)
    (parent1, parent2) = population.select_candidates(2)
    (child1, child2) = sensible_crossover(parent1, parent2)
    (child1, child2) = mutation(child1, child2)
    repeat
        (child1, child2) = mutation(child1, child2)
    until child1 and child2 are sensible and are not in the blacklist

    blacklist.append(child1, child2)
    (child1, child2) = DFT_relaxation(child1, child2)
    blacklist.append(child1, child2)
    population.append(child1, child2)
    population.sort(index=energy)
    population.delete_high_energy_candidates(2)
    if convergence criteria met:
        break
    else:
        j+=1
```

Algorithm 1: Genetic algorithm for sampling the conformational space of molecules.

Initialization of the Population. First, a random 3D structure is generated with RDKit directly from the SMILES code. This structure serves as a template for the upcoming geometries. Next, two lists of random values are generated: one for the rotatable bonds and one for the cis/trans values. If the resulting 3D geometry is sensible, the structure is then subjected to local optimization. To generate an initial population of size N (**popsize**), N sensible geometries with randomly assigned values for torsion angles need to be built and locally optimized. The optimized geometries constitute the initial population. Due to the fact that the geometries are created one after another, all randomly built structures can but do not have to be made unique in order to increase the diversity of the initial population.

Iteration of the GA. Our GA follows the established generation-based approach, i.e., the population evolves over subsequent generations. After completion of the initialization, the first iteration can be performed. For this purpose, the population is sorted and ranked based on the total energy values E_i of the different conformers i = 1, ..., N. For each individual the fitness F_i is being calculated according to

$$F_i = \frac{E_{\max} - E_i}{E_{\max} - E_{\min}} \quad (1)$$

$E_{\max}$ is the highest energy and $E_{\min}$ is the lowest energy among the energies of the conformers belonging to the current population. As a result, F = 1 for the "best" conformer and F = 0 for the "worst" conformer. In the case of a population with low variance in energy values ($E_{\max} - E_{\min}$ < **energy_var**, default = 0.001 eV), all individuals are assigned a fitness of 1.0.

Selection. Two individuals need to be selected prior to the genetic operations. We implemented three mechanisms for the selection.

(i) In the (energy-based) *roulette wheel*,[31] the probability p_i for selection of a conformer i is given by

$$p_i = \frac{F_i}{\sum_{n=1}^{N} F_i} \quad (2)$$

Table 2. GA Parameters for Isoleucine Dipeptide

	parameter	value
molecule	SMILES	CC(=O)N[C@H](C(=O)NC)[C@H](CC)C
	distance_cutoff_1	1.2 Å
	distance_cutoff_2	2.0 Å
	rmsd_cutoff_uniq	0.2 Å
	chiral	true
run settings	max_iter	10
	iter_limit_conv	10
	energy_diff_conv	0.001 eV
GA settings	popsize	5
	energy_var	0.001 eV
	selection	roulette wheel
	fitness_sum_limit	1.2
	prob_for_crossing	0.95
	cross_trial	20
	prob_for_mut_cistrans	0.5
	prob_for_mut_rot	0.5
	max_mutations_cistrans	1
	max_mutations_torsions	2
	mut_trial	100

With this, the probabilities of the conformers are mapped to segments of a line of length one. Next, two random numbers between zero and one are generated and the conformers whose segments contain these random numbers are selected. In the case when the sum of the fitness values is lower than a defined threshold near one (**fitness_sum_limit**, default = 1.2) the best and a random individual are selected.

(ii) The *reverse roulette wheel* proceeds similarly to the roulette wheel mechanism with the difference that the fitness values are swapped, i.e. new fitness F_i^* is assigned to each conformer:

$$F_i^* = F_{N+1-i} \tag{3}$$

Analogously, the probability p_i for selection of a conformer i is given by

$$p_i = \frac{F_i^*}{\sum_{n=1}^{N} F_i^*} \tag{4}$$

(iii) In the *random selection* mechanism all individuals have the same chance to be selected.

In all selection mechanisms the selected individuals must be different from each other so that the crossing-over has a chance to produce unique conformers.

Crossing-over. Crossing-over is considered to be the main feature distinguishing evolutionary algorithms from Monte Carlo techniques where only a single solution can evolve. Crossing-over allows the algorithm to take big steps in exploration of the search space.[37] In our algorithm, a crossing-over step is performed if a generated random number (between zero and one) is lower than a defined threshold (**prob_for_crossing**, default = 0.95). Between the selected individuals, parts of the representing vectors are then exchanged. To that end, the vectors characterizing the structure of both individuals are "cut" at the same single position (determined at random). The first part of the first individual is then combined with the second part of the second, and vice versa (a scheme explaining the crossing-over procedure is provided in Figure S1). Crossing-over is successful only when the resulting vectors can be used for generating sensible geometries. Otherwise the crossing-over is repeated until sensible geometries are generated or a maximum number of attempts (**cross_trial**, default = 20) is exceeded. In the latter case, exact copies of the selected conformers are used for the following step.

Mutations are performed independently for the values of cis/trans bonds and of the rotatable bonds and if randomly generated numbers exceed corresponding thresholds (**prob_for_mut_cistrans**, default = 0.5; **prob_for_mut_rot**, default = 0.5). For each, the number of mutation events is determined by a randomly picked integer number not higher than the user-defined maximal number of allowed mutations (**max_mutations_cistrans** and **max_mutations_torsions**). For each mutation, a random position of the vector is determined and the mutation is chosen to affect the value of that variable. In case of cis/trans bonds, the selected value is changed to 0° if it was above 90° or below −90°, else it is changed to 180°. A selected rotatable bond is changed to a random integer between −179° and 180°. A mutation step is only successful if the geometry built after the mutation of the vector is sensible and unique. Otherwise the entire set of mutations in a mutation step is repeated until a sensible and unique structure is generated or a maximum number of attempts (**mut_trial**, default = 100) is exceeded. In this case, the algorithm terminates. The mutation is performed for both vectors generated via crossing-over.

Local Optimization and Update. As the computational cost of the local optimization is significantly higher than all of the other operations,[54,74] only unique and sensible structures are subject to local optimization. The structures are transferred to an external program for local geometry optimization (here: FHI-aims,[69,70] see section DFT Calculations). The application of local optimization was shown to facilitate the search for minima by reducing the space the GA has to search.[24,43] Thus, the implemented GA is closer to Lamarckian than to Darwinian evolution, as the individuals evolve and pass on acquired and not inherited characteristics. Afterward, the population is extended by the newly optimized structures and, after ranking,

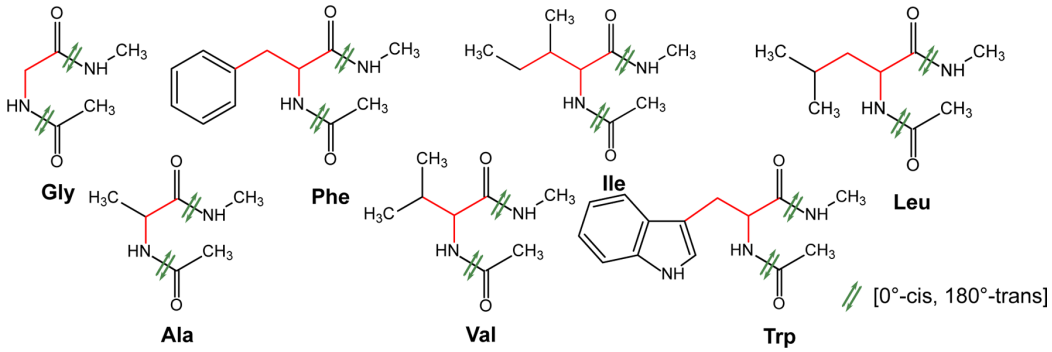


Figure 2. Chemical structures of the amino acid dipeptides. Rotatable bonds are single, nonring bonds between nonterminal atoms that are not attached to methyl groups that carry three identical substituents and are marked in red. Double arrows mark the cis/trans bonds.

the two individuals with highest energy are removed in order to keep the population size constant.

Termination of the algorithm is reached if one of the convergence criteria is met: (i) the lowest energy has not changed more than a defined threshold (**energy_diff_conv**, default = 0.001 eV) during a defined number of iterations (**iter_limit_conv**, default = 10), (ii) the lowest energy has reached a defined value (**energy_wanted**), or (iii) the maximal number of iterations (**max_iter**, default = 10) has been reached. The convergence criteria are checked only after a defined number of iterations (**iter_limit_conv**, default = 10).

We are interested not only in finding the global minimum but also in finding low-energy local minima as many of them may be relevant. Thus, all of the generated conformers are saved and are available for final analysis even if only a subset of them constitutes the final population. Table 2 summarizes practical GA parameters that were employed for one of the reference systems (isoleucine dipeptide).

The parameters listed in Table 2 can be taken as indicative of settings that will work for many small to midsize molecules. A few exceptions apply. Specifically, the **max_iter** and the **popsize** parameters are set to low values in Table 2, covering only a small set of structures within an individual GA run. This choice would be appropriate for an ensemble of many short independent GA runs to generate a broad structural ensemble with a bias toward the low-energy solution space. For larger and more complex molecules, and/or for runs designed to identify the global minimum in a single shot, **max_iter** could be increased significantly, and **popsize** could be increased somewhat (to 10−20 individuals) as well. Likewise, the mutation probabilities **prob_for_mut_cistrans** and **prob_for_mut_rot** are here set to relatively high values of 0.5, instilling a signicant amount of randomness into the search process. For a more "deterministic" search process, somewhat smaller values (e.g., 0.2) might be chosen. Finally, the **distance_cutoff** criteria are chosen to be appropriate for light elements (first and second row); adjustments may be appropriate if heavier covalently bonded atoms are included in the search.

DFT Calculations. For the tests presented below, all DFT calculations are performed with the FHI-aims code.[69,70] We employed the PBE functional[75] with a correction for van der Waals interactions (pairwise[65] for the amino acid dipeptides calculations and MBD[67] for the drug-like ligands) and with *light* computational settings and *tier*1 basis set.[69] For the local optimization, we use a trust radius enhanced version of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm[76] initialized by the model Hessian matrix of Lindh.[77] This is the default choice in the FHI-aims code and was implemented by Jürgen Wieferink. The local optimization is set to terminate when the maximum residual force component per atom is equal to 5×10^{-3} eV/ Å. Density functional, basis set, and numerical settings (e.g., integration grids) are user choices of the underlying density-functional theory code and must be set appropriately outside of Fafoom. The settings for numerical convergence (including basis set) must be chosen converged enough to not introduce artifacts in the landscape of minima found. The choice of the density-functional approximation (DFA) to the exact Born−Oppenheimer potential-energy surface needs to reproduce the local energy minima of the PES faithfully, as discussed in the Introduction. We here only note that costs for different electronic structure approximation can vary by orders of magnitude. In practice, and strictly speaking, the scope of our algorithm is to find the PES minima for a given DFT functional, while the physical choice of the "right" DFA is not the focus of this paper. We do show, however, that we can use our algorithm in practice with one specific functional, the PBE functional with a correction for van der Waals interactions, that has yielded very reliable results in the past.

Parallelization. Parallel computational resources can be utilized in two ways in order to speed up the computation. First, multiple GA runs can be started in parallel and the blacklist can be shared between different and subsequent runs. Sharing the blacklist increases diversity of solutions with already a few GA runs. Second, the time needed for the individual energy evaluations can be decreased if the molecular simulation package allows calculations across distributed nodes and is efficiently parallelized (e.g., in FHI-aims[78]). Our code supports both modes of computation.

Availability of the Code. The code is distributed as a python package named Fafoom (flexible algorithm for optimization of molecules) under the GNU Lesser General Public License.[72] It is available from following Web sites:

- https://aimsclub.fhi-berlin.mpg.de/aims_tools.php
- https://github.com/adrianasupady/fafoom

Although designed for usage with a first-principles method (e.g., FHI-aims, NWChem[79]), Fafoom can also be used with a force field (MMFF94,[80] accessed within RDKit[23,81]) for testing purposes. It is in principle possible to use any molecular simulation package which outputs optimized geometries together with their energies. Nevertheless, this requires

adjusting a part of the program to the specific needs of the used software. Details are provided with the program's documentation.

Reference Data. In order to evaluate the several aspects of the performance of the implemented algorithm we use two sets of reference data. The first reference data set (**Amino acid dipeptides**) was extracted from a database of computational data for the amino acid dipeptides. The second reference data set (**Mycophenolic acid**) contains conformers of a drug-like ligand that were obtained with three different search techniques.

Amino Acid Dipeptides. The first reference data set contains conformers of seven amino acid dipeptides[82] (Figure 2) and was extracted from a large database for amino acid dipeptide structures generated in a combined basin-hopping/multi-tempering based search. In that search (published in detail in ref 83), the framework of the reference search can be divided into a global search step and a refinement step. In the global search, the basin hopping search technique together with an empirical force field OPLS-AA was employed to perform the initial scan of the PES. The identified energy minima were relaxed at the PBE+vdW level with light computational settings in FHI-aims. In the refinement step, ab initio replica-exchange molecular dynamics runs were performed to locally explore the conformational space and to alleviate a potential bias of the initial search of a force field PES. The resulting minima were again optimized at the PBE+vdW level with tight computational settings and with the tier 2 basis set.[69] In order to compare to our data, they were reoptimized with the same functional with light computational settings, and the tier 1 basis set.[69] After this procedure, duplicates were removed from the set used for the comparison with the GA results. For benchmarking the performance of our search strategy for conformers predictions, we consider all structures with a relative energy up to 0.4 eV. These conformers define the reference energy hierarchy for each of the selected dipeptides. We summarize some characteristics and the number of conformers that were considered in Table 3.

Table 3. Reference Data Set: Seven Amino Acid Dipeptides

amino acid dipeptide	abbr	no. of atoms	no. of rotatable bonds + no. of cis/trans bonds	no. of conformers (below 0.4 eV ≈ 38.6 kJ/mol)
glycine	Gly	19	2 + 2	15 (15)
alanine	Ala	22	2 + 2	28 (17)
phenylalanine	Phe	32	4 + 2	64 (37)
valine	Val	28	3 + 2	60 (40)
tryptophan	Trp	36	4 + 2	141 (77)
leucine	Leu	31	4 + 2	183 (103)
isoleucine	Ile	31	4 + 2	176 (107)

Mycophenolic Acid. From the Astex Diverse Set,[84] a collection of X-ray crystal structures of complexes containing ligands from the Protein Data Bank (PDB), one example for a drug-like ligand was selected. This molecule, mycophenolic acid (target protein: 1MEH) has 43 atoms, 8 rotatable bonds, and 1 cis/trans bond (Figure 3).

Mycophenolic acid is a very flexible molecule. Even a coarse systematic search with a grid of only 60° for the freely rotatable torsions and 2 values (cis/trans) for the double bond and the X−X−O−H torsions yields already $6^6 \times 2 \times 2 \times 2 = 373\,248$ conformations to test. This makes this molecule a challenging

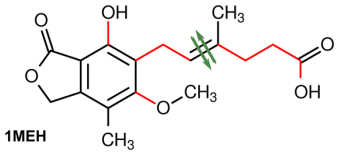


Figure 3. Chemical structure of the selected ligand together with the PDB-ID of the respective X-ray structure of the target protein. Rotatable bonds are marked in red and the cis/trans bond is marked with double arrows.

example to test the performance of three search techniques (A−C below) in combination with first-principles methods.

(A) Genetic Algorithm. 50 independent GA runs with following settings, **max_iter** = 30, **iter_limit_conv** = 20, and **popsize** = 10, were performed with Fafoom. A total of 3208 structures were generated.

(B) Random Search. 3200 random and clash-free structures were generated with Fafoom and further relaxed with DFT.

(C) Systematic Search with Confab.[11] First, 293 conformers were generated with Confab (assessed via Open Babel, used settings: **RMSD cutoff** = 0.65 and **Energy cutoff** = 15 kcal/mol). In order to account for two different values for the cis/trans bond and the X−X−O−H torsions (0° and 180°), eight starting structures per each of the conformers generated with Confab were considered. This procedure yields overall 2344 structures. After removing geometries with clashes, **2094** structures were subjected to DFT optimization.

Finally, all DFT optimized structures were merged to a common pool and the duplicates were removed. For this, a two-step criterion was used. First, the compared structures need to have a torsional RMSD (tRMSD) lower than 0.1π rad.[85] Second, the energy difference between the compared structures cannot exceed 10 meV. If both criteria are met, the structure that is higher in energy is labeled as "duplicate" and is removed from the pool. In total, 1436 unique structures were found. Table S1 shows the number of the obtained unique structures depending on the applied energy cutoff.

RESULTS AND DISCUSSION

The performance of the GA search is evaluated by the ability to reproduce the reference energy hierarchies and to find the global minimum. We performed multiple GA runs for the test systems to test the impact of varying search settings.

Amino Acid Dipeptides. For each of the amino acid dipeptides we performed 50 independent GA runs with 10 iterations (**max_iter**) each and a population size of 5 (**popsize**). One GA run with such settings requires **popsize** + 2 × **iterations** = 25 geometry optimizations at the PBE+vdW level and yields 25 conformers.

Finding the Global Minimum. First we assess the probability to find the global minimum (known from the reference energy hierarchy) among them. We check how many of the GA runs succeed in finding the global minimum and subsequently calculate the probability for finding the global minimum in one GA run and present the results in Table 4.

Table 4 illustrates how the magnitude of the sampling problem does not only depend on the dimensionality, i.e. here the number of TDOFs, but also on the chemical structure. Phenylalanine and isoleucine are two interesting cases, both have the same number of TDOFs and are of similar size, but the probability of finding the global minimum with a single run drops dramatically. The drop in probability is, of course,

Table 4. Average (from 50 GA Runs) Probability for Finding the Energy Global Minimum in a Given Run with 25 Locally Optimized Conformers

molecule	Gly	Ala	Phe	Val	Trp	Leu	Ile
TDOFs	4	4	6	5	6	6	6
probability for global minimum (/1 run)	0.82	0.79	0.53	0.60	0.22	0.20	0.10

correlated with the overall number of conformers listed in Table 3.

Conformational Coverage. A key point in our approach is to reproduce the known energy hierarchies of the reference systems. For each of the investigated compounds, we randomly choose 5, 10, 15, 20, and 25 runs (from the pool of 50 runs), merge the results, and check how many structures have been found. We repeat this procedure 10 000 times and present the results in Figure 4.

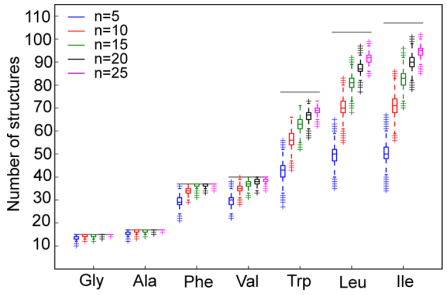


Figure 4. Number of minima found by the GA for seven amino acid dipeptides. The horizontal lines depict the total number of minima for the given compound as predicted by Ropo et al.[83] From a total of 50 GA runs, 5, 10, 15, 20, and 25 GA runs were randomly selected and the found structures were counted. This procedure was repeated 10 000 times and the resulting distributions are summarized in box plots. The line inside the box is the median, and the bottom and the top of the box are given by the lower ($Q_{0.25}$) and upper ($Q_{0.75}$) quartile. The length of the whisker is given by $1.5 \times (Q_{0.75} - Q_{0.25})$. Outliers (any data not included between the whiskers) are plotted as a cross.

It is evident that for dipeptides with a small number of reference minima (alanine and glycine) we obtain very good results, i.e. a very good coverage of conformational space, already with five repeats of the GA runs. For dipeptides with a slightly higher number of minima (phenylalanine and valine) at least 10 runs of the GA are needed to obtain a good result. For the remaining dipeptides, the GA is not able to find all of the reference minima, even with 25 GA runs. However, the coverage of the reference hierarchy with 20 GA runs is always higher than 80%. We next inspect in more detail which of the amino acid dipeptides' reference minima were missed. To this end we investigate the actual difference between the reference hierarchy and the hierarchy obtained from the 50 GA runs; see Figure 5. Although our search strategy misses a few of the reference structures even when 50 repeats of the GA search are performed, the first missed structure has a relative energy higher than 0.2 eV. This in turn means that no low-energy structures are being missed. Furthermore, there are multiple newly predicted structures that were not present in the reference data set (Figure 5). It should be noted that,

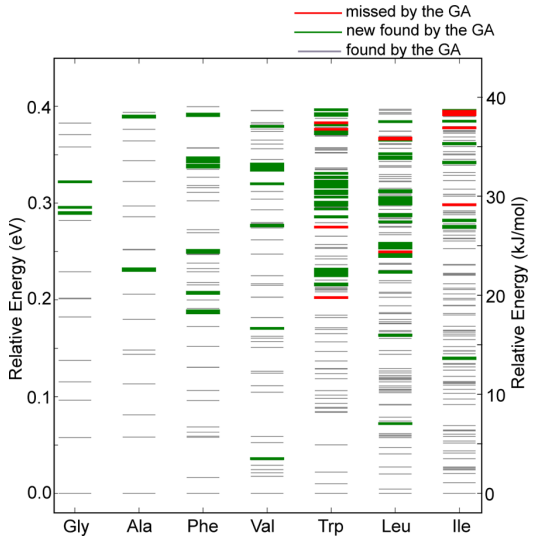


Figure 5. Difference hierarchies for the amino acid dipeptides. Red lines depict structures from the reference data set that were not found by the GA. Green lines depict structures found by the GA that were absent in the reference data set. Gray lines depict structures from the reference data set that were found by the GA. The results from all 50 GA runs for each dipeptide were taken into account.

considering the fact that the investigated GA runs are rather short, the random component of the search (randomly initialized populations) contributes to the good results of the search.

Parameter Sensitivity. In order to check the robustness of the default run parameters, several alternative settings were tested for the isoleucine dipeptide. The tested parameters include (i) the impact of the selection mechanism (roulette wheel, reverse roulette wheel, random), (ii) the effect of decreasing the cutoff for blacklisting from the default value of 0.2 to 0.05 Å, and (iii) the increase of the maximal number of iterations from the default 10 to 15, 20, and 25. For cases (i) and (ii), 100 GA runs were performed for each of the settings. In order to assess the effect of the number of iterations, 100 runs with a maximal number of iterations equal to 25 have been performed and subsequently only considered up to a maximum of 15, 20, and 25 maximum iterations. Additionally, 50 GA runs with a maximal number of iterations equal to 100 were performed. In all mentioned cases convergence criteria were evaluated after each iteration, starting from the **iter_limit_conv** = 10th iteration.

We find that none of the three selection mechanisms has a distinct impact on the probability for finding the global minimum or quality of the conformational coverage. Similarly, no substantial change was observed upon the decrease of the blacklisting cutoff. The probability value for finding the global minimum as well as the number of found reference minima increases with increased number of iterations. This is simply due to the increased number of trials for sampling the conformational space. Table 5 summarizes the probability to find the global minimum in one run with different settings. Detailed data about the conformational coverage is given in Figure S2.

Evaluation of the Computational Performance. The accuracy of a search/sampling strategy is its most crucial feature. Nevertheless, its computational cost plays a significant

Table 5. Probability of Finding the Global Minimum of Isoleucine in One Run for Different Setups[a]

setup		probability of finding the global minimum (per run)
default		0.17
selection mechanism	roulette wheel	0.18
	reverse	
	random	0.13
max. number of iterations	15 (13)	0.20
	20 (15)	0.25
	25 (16)	0.25
	100 (22)	0.46
cutoff for blacklisting	0.05 Å	0.14

[a]The default settings include roulette wheel selection mechanism, 0.2 Å cut-off for the blacklisting and maximal number of iteration equal to 10. The numbers in brackets denote the mean number of iterations needed for convergence.

role in practical applications. To this end, we quantify the total cost of the GA runs in terms of force evaluations required in the local geometry optimizations. The number of force evaluations, i.e. most expensive steps in the algorithm, is a suitable measure for the computational cost. One force evaluation requires approximately between 1 (glycine) to 3 (tryptophan) CPU minutes. We quantify the number of force evaluations required by the GA for reproducing 85% of the reference hierarchy and present the results in Table 6. The table also includes the

Table 6. Comparison of the Computational Cost: Amino Acid Dipeptides[a]

	total number of force evaluations [$\times 10^3$]						
	Gly	Ala	Phe	Val	Trp	Leu	Ile
GA (at least 85% reproduction of the reference hierarchy)	11	12	29	24	60	68	61
reference	380	400	480	440	500	460	460

[a]The cost is given in the total number of force evaluations [$\times 10^3$].

number of force evaluations required only in the replica-exchange MD refinement step of the reference search (the number of force evaluations required for the geometry optimizations is not even included).

Mycophenolic Acid. In the following we utilize as reference a set of structures that is a result of merging all structures found by three techniques: 3208 structures from 50 GA runs, 3200 random structures, and 2094 structures generated with Confab. We define the following subsets: (i) "GA" is a random selection of 25 GA runs (approximately 1600 structures); (ii) "SYS (CONFAB)" is a set of all 2094 structures generated in the systematic search; and (iii) "RANDOM" is a random selection of 1600 structures generated in the random search. For the performance evaluation we count how many of the reference structures can be found by the respective search technique. This procedure was repeated 1000 times for each of the energy cutoffs. The results are shown in Figure 6. More details can be found in Table S1.

All of the search techniques found the same global minimum several times. In case if no energy cutoff is applied, none of the searches is able to find all local minima in the conformational space (i.e., more calculation would be needed). With a decreasing energy cutoff, an improved coverage of the

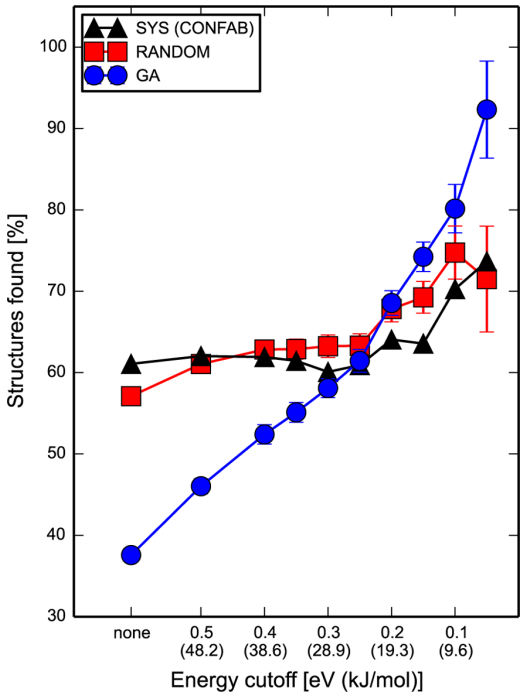


Figure 6. Share of the reference number of structures found by three search techniques: GA (blue circles), random search (red squares), and systematic search with Confab (black triangles) as a function of the applied energy cutoff. Energy values are given in electronvolts and, in parentheses, also in kilojoules per mole.

conformational space can be observed. The fact that the GA is a global optimization techniques is clearly visible as it performs better in the low-energy (<0.2 eV) region, whereas the random and systematic search perform uniformly but not perfectly independent of the energy cutoff used for the evaluation.

In order to show the wide and routine applicability of our first-principles structure search approach, we have performed short exploratory structure searches (only three GA runs each) to eight drug-like ligands from the Astex Diverse Set, which is widely utilized to assess the performance of, for example, conformer generators. The molecules vary in the size (15−32 heavy atoms) and number of rotatable bonds (6−13). A detailed analysis of this study is shown in the Supporting Information. In brief we find that in all eight instances a diverse pool of conformers can be generated. In each case, a conformer is found that is similar to the protein-bound ligand from the X-ray structure with an RMSD of 1.5 Å. In six of the eight instances, they are similar to RMSD values of less than 0.9 Å. Exploratory first-principles structure searches have a potential application in in silico protein−ligand studies:[59] the comparison of the structural space of the isolated ligand and the structure realized by the protein-bound ligand might reveal details about the binding process, for example whether the binding mechanism follows more the conformational-selection or induced-fit type. In contrast to many of the quicker (but simpler) established conformer generators, the first-principles energetics that we obtain here are not dependent on initial parametrizations and thus the method is in principle applicable throughout chemical space. It is important to note that, in this test, our goal was not to provide a converged GA search for

Table 7. Comparison of Parameters and Schemes That Are Used in Search Approaches Proposed in Four Selected Publications with the Approach Presented in This Work

parameter	Damsbo 2004[37]	Vainio 2007[21]	Nair 1998[36]	Brain 2011[56]	this work
algorithm type	EA	MO-GA	GA	GA	GA
population size	30	20	2−20	10−15	5
selection		tournament	roulette	rank	roulette
crossing-over probability		0.9	1.0	0.0−1.0	0.95
mutation probability		0.05	0.4	0.3−0.5	0.5

each molecule but rather to explore the GA's potential to provide approximate conformational coverage with a fixed computational budget. Our investigation of mycophenolic acid indicates that searches for each of these molecule could be reliably converged albeit at significantly higher computational expense.

Literature Context. In order to put the algorithm's parameters into perspective, we compare it to four selected applications of EA or GA to the conformational search of molecules in the following. In all considered algorithms, the initial populations are generated randomly and the conformational space of the respective molecules is represented and sampled (by mutation and crossing-over) by means of torsion angles, i.e. rotations around bonds. Table 7 summarizes a few parameters that illustrate the range over which the parameters that are characteristic to these kinds of evolutionary or genetic algorithms can vary. The approaches differ in the energy functions that are employed: Damsbo et al.[37] employ the CHARMM force field; Vainio and Johnson[21] use the torsional and the vdW term of the MMFF94 force field separately in a multiobjective genetic algorithm (MO-GA), while Nair and Goodman[36] use the MM2* force field. The study on optimizing the GA parameters for molecular search with a meta-GA, presented by Brain and Addicoat,[56] uses, similar to our work, a first-principles energy functions. Two choices in the algorithm highlight the difference between theirs and our aim: in order "to reliably find the already known a priori correct answer with minimum computational resources", the selection criterion "rank" focuses on the generation's best solution. Furthermore, crossing-over is considered as not helpful. In contrast, the aim of our work is to provide a GA implementation that ensures broad conformational coverage, i.e. the prediction of an energy hierarchy and not only the reproduction of a global optimum. For that we found it useful to employ random or roulette-wheel selection that also accepts less-optimal structures for genetic operations and a high probability for crossing-over. Both choices (accompanied by blacklisting) can be interpreted as means to increase diversity during the search.

CONCLUSIONS

We aimed at designing a user-friendly framework with an implementation of the genetic algorithm for searches in molecular conformational space that is particularly suitable for flexible organic compounds. A SMILES code for the selected molecule is the only required input for the algorithm. Furthermore, a wide selection of parameters (e.g., torsion definition, blacklist cutoff) allows for customizing the search. With minor changes, the code can be interfaced to external packages for molecular simulations that output optimized geometries together with corresponding energies. Besides its adaptability and ease of use, a further advantage of the implementation is the fact that it allows for using first-principles methods. With this, a potential bias resulting from the parametrization of a particular force-field can be avoided and makes the search applicable to a broad selection of problems. We examined the performance of the implementation in terms of efficiency and accuracy of the sampling. The algorithm is capable of reproducing the reference data with a high accuracy. For a set of amino acid dipeptides, we show that this conformational coverage is achieved much more efficiently than in an earlier, ab initio replica-exchange MD based search in our group. For a larger molecule (mycophenolic acid), we show that the low-energy conformational space coverage of the GA surpasses the coverage of two competing methods significantly at similar effort.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00243.

Scheme explaining the crossing-over procedure, detailed data about the conformational coverage for isoleucine and mycophenolic acid, and results of GA runs for a selection of eight ligands from the Astex Diverse Set (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: supady@fhi-berlin.mpg.de (A.S.).
*E-mail: baldauf@fhi-berlin.mpg.de (C.B.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Matthias Scheffler (FHI Berlin) is kindly acknowledged for support of this work and scientific discussions.

REFERENCES

(1) Schwab, C. H. Conformations and 3D pharmacophore searching. *Drug Discovery Today: Technol.* **2010**, *7*, e245−e253.
(2) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(3) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(4) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(5) Meier, R.; Pippel, M.; Brandt, F.; Sippl, W.; Baldauf, C. ParaDockS: a framework for molecular docking with population-based metaheuristics. *J. Chem. Inf. Model.* **2010**, *50*, 879−889.
(6) Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. *J. Chem. Inf. Model.* **2005**, *45*, 461−476.

(7) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422−430.
(8) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.* **2007**, *47*, 1067−1086.
(9) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmair, J. CAESAR: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *J. Chem. Inf. Model.* **2007**, *47*, 1923−1932.
(10) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(11) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic generation of diverse low-energy conformers. *J. Cheminf.* **2011**, *3*, 8.
(12) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. Macromodel - An integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.* **1990**, *11*, 440−467.
(13) *MOE (Molecular Operating Environment)*; Chemical Computing Group, Inc.: Montreal, Canada, 2008.
(14) Klett, J.; Cortés-Cabrera, A.; Gil-Redondo, R.; Gago, F.; Morreale, A. ALFA: Automatic Ligand Flexibility Assignment. *J. Chem. Inf. Model.* **2014**, *54*, 314−323.
(15) Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M. CONFECT: Conformations from an Expert Collection of Torsion Patterns. *ChemMedChem* **2013**, 1690−1700.
(16) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Model.* **1994**, *34*, 1000−1008.
(17) Renner, S.; Schwab, C. H.; Gasteiger, J.; Schneider, G. Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J. Chem. Inf. Model.* **2006**, *46*, 2324−2332.
(18) Andronico, A.; Randall, A.; Benz, R. W.; Baldi, P. Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress. *J. Chem. Inf. Model.* **2011**, *51*, 760−776.
(19) Sadowski, P.; Baldi, P. Small-molecule 3D structure prediction using open crystallography data. *J. Chem. Inf. Model.* **2013**, *53*, 3127−3130.
(20) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(21) Vainio, M. J.; Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.
(22) Liu, X.; Bai, F.; Ouyang, S.; Wang, X.; Li, H.; Jiang, H. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinf.* **2009**, *10*, 101.
(23) RDKit: Cheminformatics and Machine Learning Software. http://www.rdkit.org/.
(24) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111−5116.
(25) Goedecker, S. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **2004**, *120*, 9911−9917.
(26) Bahn, S.; Jacobsen, K. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.* **2002**, *4*, 56−66.
(27) Wales, D. J. GMIN: A program for finding global minima and calculating thermodynamic properties from basin-sampling. http://www-wales.ch.cam.ac.uk/GMIN/.
(28) Ponder, J. W. Tinker - Software Tools for Molecular Design. http://dasher.wustl.edu/tinker/.
(29) Holland, J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*; University of Michigan Press: Ann Arbor, MI, 1975.
(30) Fogel, D. B., Ed. *Evolutionary Computation: The Fossil Record*; IEEE Press: Piscataway, NJ, 1998.
(31) Goldberg, D. E. *Genetic algorithms in search, optimization, and machine learning*; Addison-Wesley: Reading, MA, 1989.
(32) Clark, D. E.; Westhead, D. R. Evolutionary algorithms in computer-aided molecular design. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 337−358.
(33) Wales, D. J. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*; Cambridge University Press: Cambridge, 2003.
(34) Wales, D. J.; Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **1999**, *285*, 1368−1372.
(35) Johnston, R. L., Ed. *Applications of Evolutionary Computation in Chemistry*; Structure and Bonding; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004.
(36) Nair, N.; Goodman, J. M. Genetic Algorithms in Conformational Analysis. *J. Chem. Inf. Model.* **1998**, *38*, 317−320.
(37) Damsbo, M.; Kinnear, B. S.; Hartings, M. R.; Ruhoff, P. T.; Jarrold, M. F.; Ratner, M. A. Application of evolutionary algorithm methods to polypeptide folding: comparison with experimental results for unsolvated Ac-$(Ala\text{-}Gly\text{-}Gly)_5$-$LysH^+$. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 7215−7222.
(38) Carstensen, N. O.; Dieterich, J. M.; Hartke, B. Design of optimally switchable molecules by genetic algorithms. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2903−2910.
(39) Carlotto, S.; Orian, L.; Polimeno, A. Heuristic approaches to the optimization of acceptor systems in bulk heterojunction cells: a computational study. *Theor. Chem. Acc.* **2012**, *131*, 1191.
(40) Hartke, B. Global geometry optimization of clusters using genetic algorithms. *J. Phys. Chem.* **1993**, *97*, 9973−9976.
(41) Deaven, D. M.; Ho, K. M. Molecular geometry optimization with a genetic algorithm. *Phys. Rev. Lett.* **1995**, *75*, 288−291.
(42) Hartke, B. Global cluster geometry optimization by a phenotype algorithm with Niches: Location of elusive minima, and low-order scaling with cluster size. *J. Comput. Chem.* **1999**, *20*, 1752−1759.
(43) Johnston, R. L. Evolving better nanoparticles: Genetic algorithms for optimizing cluster geometries. *Dalt. Trans.* **2003**, 4193−4207.
(44) Blum, V.; Hart, G. L. W.; Walorski, M. J.; Zunger, A. Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *72*, 165113.
(45) Schönborn, S. E.; Goedecker, S.; Roy, S.; Oganov, A. R. The performance of minima hopping and evolutionary algorithms for cluster structure prediction. *J. Chem. Phys.* **2009**, *130*, 144108.
(46) Sierka, M. Synergy between theory and experiment in structure resolution of low-dimensional oxides. *Prog. Surf. Sci.* **2010**, *85*, 398−434.
(47) Bhattacharya, S.; Levchenko, S. V.; Ghiringhelli, L. M.; Scheffler, M. Stability and Metastability of Clusters in a Reactive Atmosphere: Theoretical Evidence for Unexpected Stoichiometries of Mg_MO_x. *Phys. Rev. Lett.* **2013**, *111*, 135501.
(48) Hartke, B. Global optimization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 879−887.
(49) Heiles, S.; Johnston, R. L. Global optimization of clusters using electronic structure methods. *Int. J. Quantum Chem.* **2013**, *113*, 2091−2109.
(50) Hart, G. L. W.; Blum, V.; Walorski, M. J.; Zunger, A. Evolutionary approach for determining first-principles hamiltonians. *Nat. Mater.* **2005**, *4*, 391−394.
(51) Abraham, N. L.; Probert, M. I. J. A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *73*, 224104.

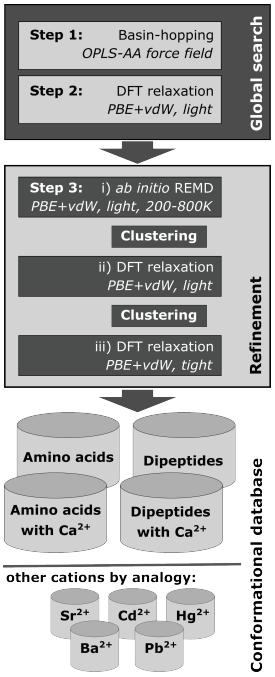
(52) Oganov, A. R.; Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J. Chem. Phys.* **2006**, *124*, 244704.
(53) Woodley, S. M.; Catlow, R. Crystal structure prediction from first principles. *Nat. Mater.* **2008**, *7*, 937−946.
(54) Tipton, W. W.; Hennig, R. G. A grand canonical genetic algorithm for the prediction of multi-component phase diagrams and testing of empirical potentials. *J. Phys.: Condens. Matter* **2013**, *25*, 495401.
(55) Neiss, C.; Schooss, D. Accelerated cluster structure search using electron diffraction data in a genetic algorithm. *Chem. Phys. Lett.* **2012**, *532*, 119−123.
(56) Brain, Z. E.; Addicoat, M. A. Optimization of a genetic algorithm for searching molecular conformer space. *J. Chem. Phys.* **2011**, *135*, 174106.
(57) Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Koksch, B.; Blum, V.; Scheffler, M. How cations change peptide structure. *Chem. - Eur. J.* **2013**, *19*, 11224−11234.
(58) Rossi, M.; Chutia, S.; Scheffler, M.; Blum, V. Validation Challenge of Density-Functional Theory for Peptides-Example of Ac-Phe-Ala_5-$LysH^+$. *J. Phys. Chem. A* **2014**, *118*, 7349−7359.
(59) Avgy-David, H. H.; Senderowitz, H. Toward Focusing Conformational Ensembles on Bioactive Conformations: A Molecular Mechanics/Quantum Mechanics Study. *J. Chem. Inf. Model.* **2015**, *55*, 2154−2167.
(60) Wu, Q.; Yang, W. Empirical correction to density functional theory for van der Waals interactions. *J. Chem. Phys.* **2002**, *116*, 515.
(61) Sedlak, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. The accuracy of quantum chemical methods for large noncovalent complexes. *J. Chem. Theory Comput.* **2013**, *9*, 3364−3374.
(62) Tkatchenko, A.; Rossi, M.; Blum, V.; Ireta, J.; Scheffler, M. Unraveling the Stability of Polypeptide Helices: Critical Role of van der Waals Interactions. *Phys. Rev. Lett.* **2011**, *106*, 118102.
(63) Grimme, S.; Antony, J.; Schwabe, T.; Mück-Lichtenfeld, C. Density functional theory with dispersion corrections for supramolecular structures, aggregates, and complexes of (bio)organic molecules. *Org. Biomol. Chem.* **2007**, *5*, 741−758.
(64) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(65) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.
(66) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.
(67) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **2014**, *140*, 18A508.
(68) Schubert, F.; Rossi, M.; Baldauf, C.; Pagel, K.; Warnke, S.; von Helden, G.; Filsinger, F.; Kupser, P.; Meijer, G.; Salwiczek, M.; Koksch, B.; Scheffler, M.; Blum, V. Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala_{19}-Lys+H^+vs. Ac-Lys-Ala_{19}+H^+ and the current reach of DFT. *Phys. Chem. Chem. Phys.* **2015**, *17*, 7373.
(69) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175−2196.
(70) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. Efficient integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **2009**, *228*, 8367−8379.
(71) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(72) GNU Lesser General Public License. https://www.gnu.org/licenses/lgpl.html.
(73) Schlegel, H. B. Geometry optimization. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 790−809.
(74) Cheng, J.; Fournier, R. Structural optimization of atomic clusters by tabu search in descriptor space. *Theor. Chem. Acc.* **2004**, *112*, 7−15.
(75) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(76) Nocedal, J.; Wright, S. J. *Numerical optimization*; Springer: New York, 2006.
(77) Lindh, R.; Bernhardsson, A.; Karlström, G.; Malmqvist, P.-Å. On the use of a Hessian model function in molecular geometry optimizations. *Chem. Phys. Lett.* **1995**, *241*, 423−428.
(78) Marek, A.; Blum, V.; Johanni, R.; Havu, V.; Lang, B.; Auckenthaler, T.; Heinecke, A.; Bungartz, H.-J.; Lederer, H. The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science. *J. Phys.: Condens. Matter* **2014**, *26*, 213201.
(79) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477−1489.
(80) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(81) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminf.* **2014**, *6*, 37.
(82) We use the term *dipeptide* for amino acids with an acetylated N terminus and an amino-methylated C terminus.
(83) Ropo, M.; Baldauf, C.; Blum, V. Energy/structure database of all proteinogenic amino acids and dipeptides without and with divalent cations. 2015, arXiv1504.03708, q-bio.BM.
(84) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−41.
(85) The value of 0.1π tRMSD corresponds to a 55° change of a single dihedral angle or to a change of 18° per each of nine the dihedral angles.

4.4 First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids

Ropo, M.; Schneider, M;
Baldauf, C.; Blum, V.

First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids

Sci. Data. **2016** (3), 16009.
DOI: 10.1038/sdata.2016.9



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Electronic structure of atoms and molecules
» Computational biophysics
» Infrared spectroscopy
» Density functional theory

Received: 13 April 2015
Accepted: 15 January 2016
Published: 16 February 2016

First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids

Matti Ropo[1,2,3], Markus Schneider[1], Carsten Baldauf[1] & Volker Blum[1,4]

We present a structural data set of the 20 proteinogenic amino acids and their amino-methylated and acetylated (capped) dipeptides. Different protonation states of the backbone (uncharged and zwitterionic) were considered for the amino acids as well as varied side chain protonation states. Furthermore, we studied amino acids and dipeptides in complex with divalent cations (Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, and Hg^{2+}). The database covers the conformational hierarchies of 280 systems in a wide relative energy range of up to 4 eV (390 kJ/mol), summing up to a total of 45,892 stationary points on the respective potential-energy surfaces. All systems were calculated on equal first-principles footing, applying density-functional theory in the generalized gradient approximation corrected for long-range van der Waals interactions. We show good agreement to available experimental data for gas-phase ion affinities. Our curated data can be utilized, for example, for a wide comparison across chemical space of the building blocks of life, for the parametrization of protein force fields, and for the calculation of reference spectra for biophysical applications.

Design Type(s)	observation design • data integration objective
Measurement Type(s)	protein conformation assessment
Technology Type(s)	data representational model
Factor Type(s)	amino acid • biochemical phenotype • peptidyl-amino acid modification
Sample Characteristic(s)	

[1]Fritz Haber Institute of the Max Planck Society, 14195 Berlin, Germany. [2]Department of Physics, Tampere University of Technology, 33720 Tampere, Finland. [3]COMP, Department of Applied Physics, Aalto University, 00076 Aalto, Finland. [4]Department of Mechanical Engineering and Materials Science, Duke University, Durham, North Carolina 27708, USA. Correspondence and requests for materials should be addressed to M.R. (email: matti.ropo@tut.fi) or to C.B. (email: baldauf@fhi-berlin.mpg.de) or to V.B. (email: volker.blum@duke.edu).

Background & Summary

Proteins are the machinery of life. We here present a first-principles study of the conformational preferences of their basic building blocks—specifically, as summarized in Fig. 1: 20 proteinogenic amino acids and dipeptides, with different possible protonation states, and the conformational space changes resulting from attaching six divalent cations, i.e., Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, and Hg^{2+}. In past studies, a wide range of different approximate electronic structure methods has been applied to some of these proteinogenic amino acids—see, for example, references[1–59]. These studies have deepened our understanding of the conformational basics of individual building blocks, but a systematic comparison of properties of the different building blocks is complicated when relying on data from different sources. On the one hand this is due to the molecular models that may differ in protonation states and backbone capping. On the other, the simulations can differ in several ways:

- Different sampling strategies or methods to generate conformers may have been used. Search-dependent settings, like energy cut-offs, can also have a significant impact on the results.
- The levels of theory that have been applied range from semi-empirical to Hartree-Fock (HF) to density-functional theory (DFT) up to coupled-cluster calculations[1–59].
- Numerical settings, e.g., basis sets, can differ substantially and might lead to different results.

A further point that limits a quantitative comparison is the accessibility of the data from different studies. Energies, for example, often have to be extracted from table footnotes and/or the structural data is not always accessible in the Supplementary Information of the respective articles, sometimes even only accessible as figures in the manuscript. The data set presented here overcomes such limitations by covering a comprehensive segment of chemical space exhaustively, using a large scale computational effort. This study treats 20 proteinogenic amino acids, their dipeptides and their interactions with the divalent cations Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, and Hg^{2+} (see Fig. 1 for an overview) on the same theoretical footing. The importance of peptide cation interactions may be highlighted by the fact that about 40% of all proteins bind cations[60–62]. Especially Ca^{2+} is important in a multitude of functions, ranging,

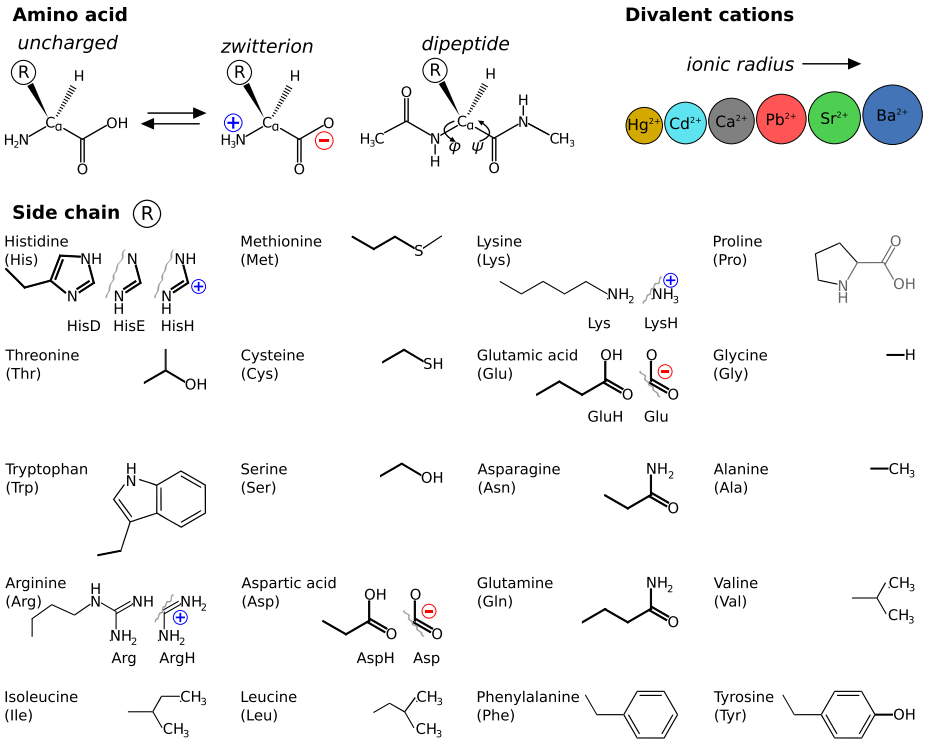


Figure 1. Molecular systems covered in this study. Top left and center: Schematic depiction of the backbone conformations of uncharged, zwitterionic, and dipeptide forms of the aminoacids considered in this work. Side chains are indicated by the letter ***R***. Top right: Divalent ions considered for complexation with the 20 proteinogenic amino acids. Lower five rows: Side chains, including different protonation states where applicable, of the 20 proteinogenic amino acids considered in this work.

for example, from blood clotting[63] to cell signaling to bone growth[64]. Such calcium mediated functions can be disturbed by the presence of alternative divalent heavy metal cations like Pb^{2+}, Cd^{2+}, and Hg^{2+} (refs 62,65,66).

The conformations and total energies of each molecular system are calculated from first principles in the framework of density-functional theory (DFT)[67,68] using the PBE generalized-gradient exchange-correlation functional[69]. Energies are corrected for van der Waals interactions using the Tkatchenko-Scheffler formalism[70]. In this formalism, pairwise $C_6[n]/r^6$ terms are computed and summed up for all pairs of atoms. r is the interatomic distance, a cut-off for short interatomic distances is applied, and $C_6[n]$ coefficients are obtained from the self-consistent electron density. The combined approach is referred to as 'PBE+vdW' throughout this work. This level of theory is robust for potential-energy surface (PES) sampling of peptide systems[71–78]. The curated data is provided as basis for comparative studies across chemical space to reveal conformational trends and energetic preferences. It can, for example, further be used for force-field development, theoretical studies at higher levels of theory, and as a starting point for theoretical calculations of spectra for biophysical applications.

Methods

Molecular models

This study covers a total of 280 molecular systems (summarized in Fig. 1). The number is the product of the following chemical degrees of freedom that were considered in our study:

20 proteinogenic amino acids. In case of (de)protonatable side chains, all protomers (different protonations states) were considered as well.

2 different backbone types, either free termini (considered in uncharged or zwitterionic form) or capped (N-terminally acetylated or C-terminally amino-methylated).

7 reflecting that the respective amino acid or dipeptide was considered either in isolation or with one of six different cation additions: Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, or Hg^{2+}.

Conformational search and energy functions

For the initial scan of the PES, the empirical force field OPLS-AA[79] was employed, followed by DFT-PBE+vdW relaxations of the energy minima identified in the force field. The identified set of structures was then subjected to a further first-principles refinement step, *ab initio* replica-exchange molecular dynamics (REMD). An overview of the procedure is given in Fig. 2 and the steps are described in more detail below.

Force-field based (OPLS-AA)[79] **global conformational searches (Step 1)** were performed for all dipeptides and amino acids (i) without a coordinating cation and (ii) with Ca^{2+}. These searches employed a basin hopping search strategy[80,81] as implemented in the tool 'scan', distributed with the TINKER molecular simulation package[82,83]. We here use an in-house parallelized version of the TINKER scan utility that was first used in reference[74]. In this search strategy, input structures for relaxations are generated by projecting along normal modes starting from a local minimum. The number of search directions from a local minimum was set to 20. Conformers were accepted within a relative energy window of 100 kcal/mol and if they differ in energy from already found minima by at least 10^{-4} kcal/mol. The search terminates when the relaxations of input structures do not result in new minima.

After that, **PBE+vdW relaxations (Step 2)** were performed with the program FHI-aims[84–86]. FHI-aims employs numeric atom-centered orbital basis sets as described in reference 84 to discretize the Kohn-Sham orbitals. Different levels of computational defaults are available, distinguished by choice of the basis set, integration grids, and the order of the multipole expansion of the electrostatic (Hartree) potential of the electron density. For the chemical elements relevant to this work, 'light' settings include the so-called *tier1* basis sets and were used for initial relaxations. 'Tight' settings include the larger *tier2* basis sets and ensure converged conformational energy differences at a level of few meV (ref. 84). Unless noted otherwise, all energies discussed here are results of PBE+vdW calculations with a *tier2* basis and 'tight' settings. Relativistic effects were taken into account by the so-called atomic zero-order regular approximation (atomic ZORA)[87,88] as described in reference[84]. Previous comparisons to high-level quantum chemistry benchmark calculations at the coupled-cluster level, CCSD(T), demonstrated the reliability of this approach for polyalanine systems[72,76], alanine, phenylalanine, and glycine containing tripeptides[76], and alanine dipeptides with Li^+ (ref. 73). Further benchmarks at the MP2 level of theory are reported below in the section Technical Validation.

The **refinement (Step 3)** by *ab initio* REMD[89,90] is intended to alleviate the potential effects of conformational energy landscape differences between the force field and the DFT method. In REMD, multiple molecular dynamics trajectories of the same system are independently initialized and run in a range of different temperatures. Based on a Metropolis criterion, configurations are swapped between trajectories of neighboring temperatures. Thus, the simulations can overcome barriers and provide an enhanced conformational sampling in comparison to classical molecular dynamics (MD)[90,91]. The simulations were carried out employing a script-based REMD scheme that is provided with FHI-aims and that was first used in reference[92]. Computations were performed at the PBE+vdW level with 'light' computational settings. The run time for each REMD simulation was 20 ps with an integration time step of 1 fs. The frequent exchange attempts (every 0.04 or 0.1 ps) ensure efficient sampling of the

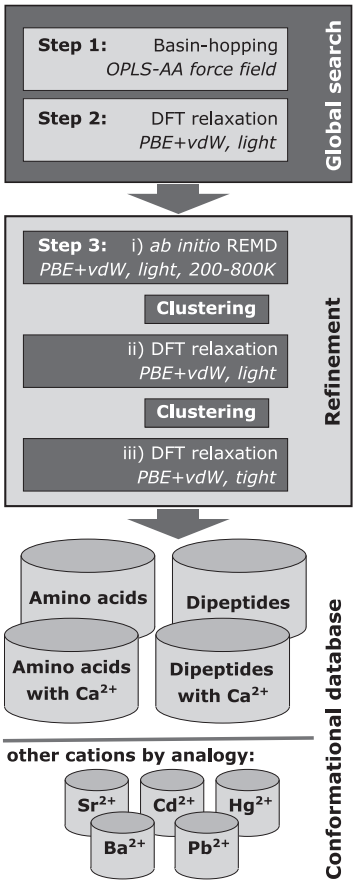


Figure 2. Schematic representation of the workflow employed to locate stationary points on the potential-energy surfaces of the respective molecular systems.

potential-energy surface as shown by Sindhikara *et al.*[93]. The velocity-rescaling approach by Bussi *et al.*[94] was used to sample the canonical distribution. Starting geometries for the replicas were taken from the lowest energy conformers resulting from the PBE+vdW relaxations in Step 2. REMD parameters for the individual systems, i.e. the number of replicas, acceptance rates for exchanges between replicas, the frequency for exchange attempts, and the temperature range, are summarized in Supplementary Table S1 of the Supplementary Information. Conformations were extracted from the REMD trajectories every 10th step, i.e. every 10 fs of simulation time. In order to generate a set of representative conformers, these structures were clustered using a *k*-means clustering algorithm[95] with a cluster radius of 0.3 Å as provided by the MMSTB package[96]. The resulting arithmetic-mean structures from each cluster were then relaxed using PBE+vdW with 'light' computational settings. The obtained conformers were again clustered and cluster representatives were relaxed with PBE+vdW ('tight' computational settings) to obtain the final conformation hierarchies. The refinement step by REMD is essential, as shown in Fig. 3, which separately identifies the number of distinct conformers found in Step 2 and, subsequently, the number of additional conformers found in Step 3.

After step 2, a total of 17,381 stationary points was found for the amino acids and dipeptides in isolation and in complex with Ca^{2+}. The refinement procedure in Step 3 increases this number to a total of 21,259 structures. Initial structures for the Ba^{2+}, Cd^{2+}, Hg^{2+}, Pb^{2+} and Sr^{2+} binding amino acid and dipeptide systems were then obtained by replacing the Ca^{2+} cation in the amino acid and dipeptide structures binding a Ca^{2+} cation. These structures were subsequently relaxed with PBE+vdW employing 'tight' computational settings and a tier-2 basis set. This procedure results in 24,633 further conformers with bound cations. Altogether, we thus provide information on 45,892 stationary points of the PBE+vdW PES for all systems studied in this work.

The numbers of conformers identified in the searches are also given in Supplementary Table S2 of the Supplementary Information. Supplementary Tables S3 and S4 provide detailed accounts of how many structures were found for which amino acid/dipeptide in isolation or with attached cations.

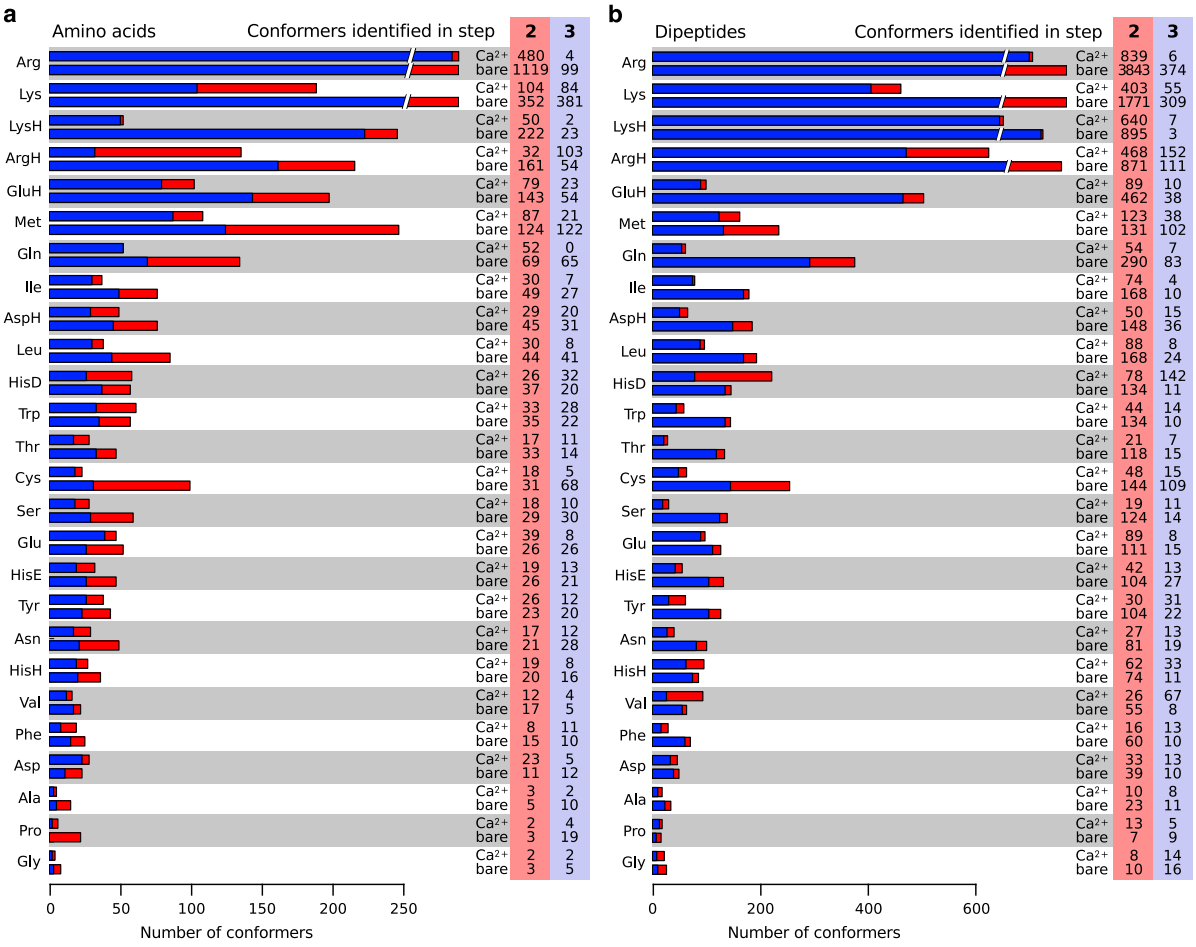


Figure 3. Numbers of stationary points of the PBE+vdW potential-energy surface (PES) at the 'tight'/tier-2 level of accuracy that were found for the different **a)** uncapped amino acids or **b)** dipeptides in isolation ('bare') or with a Ca^{2+} cation. Blue segments of the bars and blue shaded numbers give the number of stationary points ('conformers') located in Step 2 of the search procedure detailed in Fig. 2. Red bar segments and red shading highlight the number of conformers that were additionally found during Step 3 of the search. The total number of conformers found for each system is the sum of the numbers found in steps two and steps three.

Data Records

The curated data, consisting of the Cartesian coordinates of 45,892 stationary point geometries of the PBE+vdW PES (the main outcome of our work) and their potential energies computed at the 'tight'/tier-2 level of accuracy in the FHI-aims code, is provided as plain text files sorted in directories (see Fig. 4). The PBE+vdW total energies are included since they are an integral part of the construction of our geometry data sets. Importantly, the stationary point geometries could be used as starting points to refine the total energy accuracy by higher-level methods, e.g., those discussed in 'Technical Validation' below. The folder structure is hierarchic and straightforward. The naming scheme is explained in the following:

Description of the file types:

conformer.(...).xyz coordinates in standard xyz format in Å, readable by a wide range of molecule viewers, e.g. VMD[97], Jmol (http://www.jmol.org/), etc.

conformer.(...).fhiaims coordinate file in FHI-aims geometry input format: for each atom of the particular system, the Cartesian coordinates are given in Å (`atom [x] [y] [z] [element]`). The electronic total energy (in eV) at the PBE+vdW level is given there as a comment.

control.in FHI-aims input file with technical parameters for the calculations. Please note that these files also include the exact specifications of the 'tight' numerical settings for all included elements.

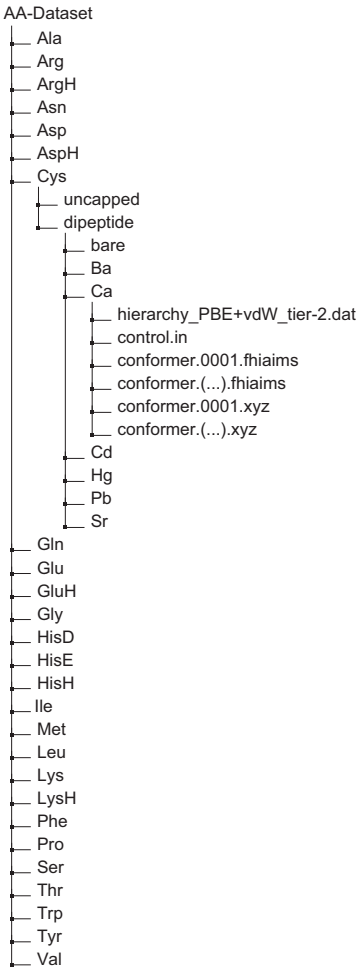


Figure 4. Schematic representation of folder organization of the data. Each folder, as exemplified for the Ca^{2+}-coordinated cysteine dipeptide, contains coordinate files in two formats (standard XYZ and FHI-aims input), the computational settings file for FHI-aims (control.in), and the energy hierarchies (PBE+vdW, 'tight'/tier-2 level) per system.

hierarchy_PBE+vdW_tier-2.dat in each final subfolder, contains three columns: number of the conformer, total energy (in eV, PBE+vdW, tier-2 basis set, 'tight' numerical settings, computed with FHI-aims version 031011), and relative energy (in eV, relative to the respective global minimum).

The curated data is publicly available from two sources:

1. A website dedicated to this data set has been set up (http://aminoaciddb.rz-berlin.mpg.de) and allows users to browse and download the data and to visualize molecular structures online.
2. From the NOMAD repository (http://nomad-repository.eu) the data is available via the DOI 10.17172/NOMAD/20150526220502 [Data citation 1].

Technical Validation

The conformational coverage for the amino acid alanine is validated by comparing to a recent study by Maul *et al.*[12]. In that reference[10], low energy conformers of alanine were reported, spanning an energy range of approximately 0.26 eV between the reported lowest and highest energy conformers. The level of theory used by Maul *et al.* was DFT in the generalized gradient approximation by means of the Perdew-Wang 1991 functional[98]. In our case, the force field based search step with subsequent PBE+vdW relaxations yields 5 conformers. The following *ab initio* REMD simulations increase the number of conformers to 15 within an energy range of 0.43 eV. The respective conformational energy hierarchies

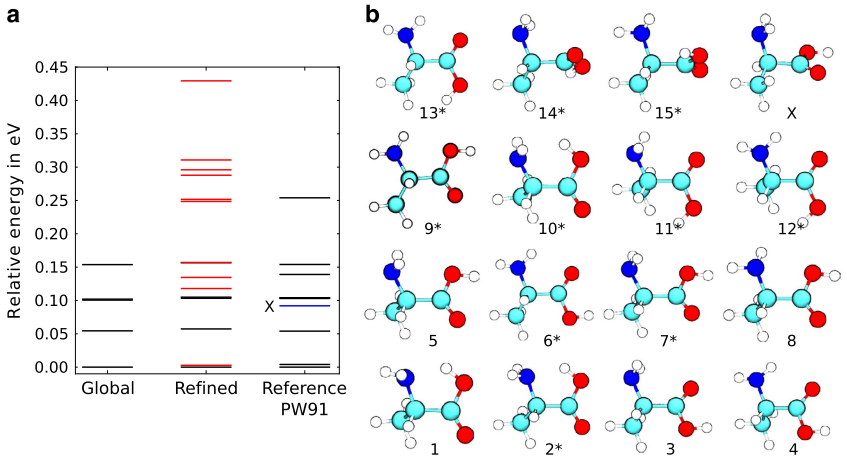


Figure 5. Comparison of search strategies. (**a**) The conformational energy hierarchies for alanine after the global search and the local refinement together with the reference hierarchy at the DFT-PW91 level that was published by Maul *et al.*[12]. Conformers indicated by black lines were found in the global search, the conformers in red were located only after the local refinement step. The blue line in the reference conformational hierarchy represents a minimum not found in our search and not present at the PBE+vdW level. (**b**) Conformations of the alanine molecule. Conformers marked with an asterisk (*) were found in the local refinement step of our search strategy. Atoms are color-coded as follows: Cyan (C), blue (N), red (O), white (H). The conformer labeled with X was found by Maul *et al.* in PW91 calculations[12] but is unstable at the PBE+vdW level.

after global search and after REMD-refinement are shown in Fig. 5a. The results of our search (with the refinement step) are in good agreement with the data from reference[12] that is also shown in Fig. 5a. Structures are shown in Fig. 5b. Nine of the ten conformers identified by Maul *et al.* can be confirmed. The single conformer that is missing (highlighted by an X in Fig. 5a) is not a stationary point of the PBE+vdW potential energy surface. Conformers 14 and 15 are classified as saddle points by analysis of the vibrational modes.

In order to further quantify the reliability of the DFT-PBE+vdW level of theory for peptides, beyond earlier benchmark work[72,73,76] and especially with divalent cations, benchmark calculations were performed at the level of Møller-Plesset second-order perturbation theory (MP2)[99,100] using the electronic structure program package ORCA[101]. Single-point energy calculations were performed for all fixed stationary-point DFT-PBE+vdW geometries in our data base for the amino acids alanine (Ala) and phenylalanine (Phe) with neutral N and C termini in isolation as well as in complex with a Ca^{2+} cation. Phe was selected to represent a 'difficult' example, i.e., the interaction of the cation with a larger aromatic side chain. The MP2 calculations did not include any frozen-core treatment (including semicore states is essential for Ca^{2+}) and were performed using Dunning's correlation-consistent polarized core-valence basis sets (cc-pCVnZ), with $n = \mathrm{T/Q/5}$ denoting the triple-zeta, quadruple-zeta, and quintuple-zeta basis sets respectively[102]. The calculated SCF (Hartree-Fock) and MP2 correlation energies were then individually extrapolated to the complete basis set (CBS) limit as follows: For SCF energies, we used the extrapolation strategy proposed by Karton and Martin[103]:

$$E_{SCF}^{n} = E_{SCF}^{CBS} + Ae^{-\alpha\sqrt{n}}. \tag{1}$$

A, α, and the CBS-extrapolated energy E_{SCF}^{CBS} are parameters determined from a least-squares fitting algorithm applied individually for each conformer. For the MP2 correlation energies, an extrapolation scheme proposed by Truhlar[104] was applied:

$$E_{corr}^{n} = E_{corr}^{CBS} + Bn^{-\beta}. \tag{2}$$

Again, B, β, and the CBS-extrapolated energy E_{corr}^{CBS} are parameters determined from a least-squares fitting algorithm as before. A detailed account of all numbers is given in the Supplementary Information (Supplementary Table S5). Mean absolute errors between the density-functional approximation (DFA) relative energies and the basis-set extrapolated MP2 relative energies were calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\Delta E_i^{DFA} - \Delta E_i^{MP2} + c\right|, \tag{3}$$

where the index i runs over all N conformations of a given data set. ΔE_i in principle denotes the energy

difference between conformer i and the lowest-energy conformer of the set. The adjustable parameter c is used to shift the MP2 and DFA conformational hierarchies versus one another to obtain the lowest possible MAE, rendering the reported MAE value independent of the choice of any reference structure. Fig. 6a shows the corresponding obtained mean absolute errors (MAE) and maximal errors ($max_i|\Delta E_i^{DFA} - \Delta E_i^{MP2}+c|$) of different DFA calculations—performed with the FHI-aims code—with respect to benchmarks on the MP2 level obtained as described above. Within FHI-aims, the accuracy of integration grids and of the electrostatic potential was also verified by comparing 'tight' and 'really_tight' numerical settings, giving virtually identical results. The DFA level of theory of PBE+vdW shows a MAE well within *chemical accuracy* of ~ 1 kcal/mol ≈ 43 meV for both structural sets of Ala and Phe; for Phe, the maximal error is ~ 2 kcal/mol. We next applied a different long-range dispersion treatment, a recent

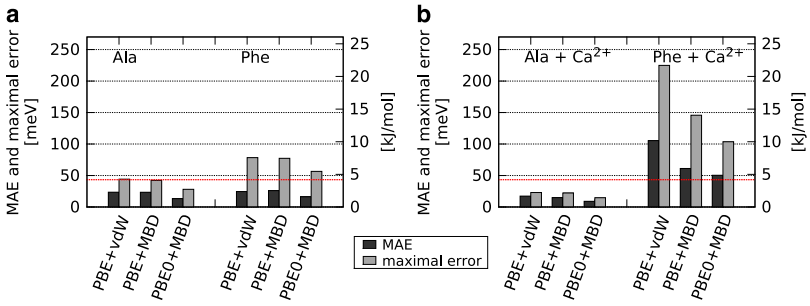


Figure 6. Comparison of different DFAs to MP2 energies. Mean absolute error (MAE) and maximal error (in meV) between different relative energies at the DFA (PBE+vdW, PBE+MBD*, and PBE0+MBD*) and MP2 level of theory, using structures of obtained minima at the PBE+vdW level from the database for the systems of Ala and Phe with neutral end caps, both (**a**) in isolation and (**b**) in complex with a Ca^{2+} cation. Computational details are given in the text. Exact numbers are summarized in Table 1.

System	MAE [meV]	Maximal error [meV]
Ala		
PBE+vdW	24 (0.5)	44 (1.0)
PBE+MBD*	23 (0.5)	44 (1.0)
PBE0+MBD*	13 (0.3)	28 (0.6)
Phe		
PBE+vdW	25 (0.6)	78 (1.8)
PBE+MBD*	26 (0.6)	77 (1.8)
PBE0+MBD*	16 (0.4)	57 (1.3)
Ala+Ca^{2+}		
PBE+vdW	17 (0.4)	23 (0.5)
PBE+MBD*	15 (0.3)	22 (0.5)
PBE0+MBD*	9 (0.2)	15 (0.3)
Phe+Ca^{2+}		
PBE+vdW	105 (2.4)	225 (5.2)
PBE+MBD*	61 (1.4)	146 (3.4)
PBE0+MBD*	50 (1.2)	104 (2.4)

Table 1. Mean absolute error (MAE) and maximal error (in meV; in parentheses: in kcal/mol) between different relative energies at the DFA (PBE+vdW, PBE+MBD*, and PBE0+MBD*) and MP2 level of theory, using structures of obtained minima at the PBE+vdW level from the database for the systems of Ala and Phe with neutral end caps, both in isolation and in complex with a Ca^{2+} cation. Computational details are given in the text.

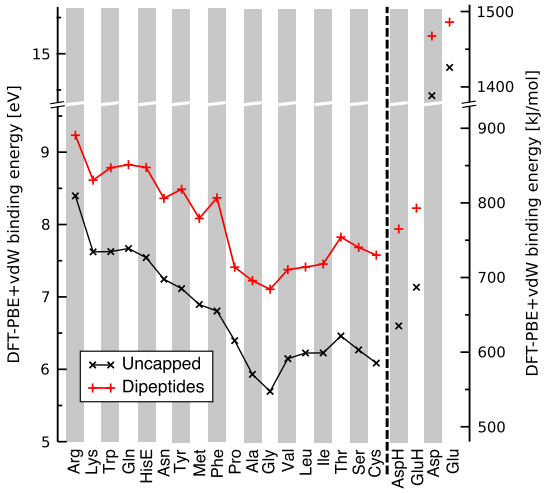


Figure 7. Comparison of the gas-phase binding energies of Ca^{2+} to different amino acids calculated in this work (*y* axis) to the experimentally inferred hierarchy of gas-phase binding energies of Ca^{2+} to different amino acids by Ho *et al.*[110] The amino acids are ordered along the *x* axis from the highest to lowest experimental Ca^{2+} binding energy. Protonated and deprotonated Asp and Glu are not included among the experimental data and are here shown as predictions. $E_{binding}$ is high for deprotonated Asp and Glu since these forms of the amino acid would carry a negative charge.

many-body dispersion model based on interacting quantum harmonic oscillators denoted as MBD*[105], showing no significant improvement for the isolated amino acids. In line with ref. 76, applying the more expensive PBE0 (ref. 105) hybrid exchange correlation functional reduces the maximum deviation for Phe to ~57 meV, i.e., 1.3 kcal/mol. For Ala and Phe with neutral end caps in complex with a Ca^{2+} cation, Fig. 6b compares the same set of DFAs to MP2 benchmark energy hierarchies. However, obtaining basis-set converged total energies of the same accuracy as for the isolated peptides by straightforward CBS extrapolation proved remarkably more difficult when Ca^{2+} was involved. The reason is traced to the significant and slow-converging correlation contribution of the Ca^{2+} semicore electrons, which leads to large and conformation dependent basis set superposition errors (BSSE). This problem was verified for MP2 calculations in the FHI-aims and ORCA codes, with several different basis set prescriptions[107], and for CCSD(T) calculations. Standard DFAs, if sufficiently accurate, have a significant advantage in this respect since they are not subject to comparable numerical convergence problems. To yet arrive at reliable CBS-extrapolated MP2 conformational energy differences, we thus subjected the SCF and correlation energies of each Ca^{2+} coordinated conformation to a counterpoise correction[108,109] to minimize the effect of BSSE on the Ca^{2+} correlation energy contribution, prior to performing CBS extrapolation as described above. For the example of Ala+Ca^{2+} and assuming rigid conformers, the BSSE is estimated as:

$$\begin{aligned} E_{BSSE} = \quad & E_{BSSE}(Ala) + E_{BSSE}(Ca^{2+}), \text{ with} \\ & E_{BSSE}(Ala) = E^{Ala+Ca^{2+}}(Ala) - E^{Ala}(Ala), \text{ and} \\ & E_{BSSE}(Ca^{2+}) = E^{Ala+Ca^{2+}}(Ca^{2+}) - E^{Ca^{2+}}(Ca^{2+}). \end{aligned} \tag{4}$$

$E^{Ala+Ca^{2+}}(Ala)$ represents the energy of Ala evaluated in the union of the basis sets on Ala and Ca^{2+}, $E^{Ala}(Ala)$ represents the energy of Ala evaluated in the basis set on Ala, *etc.* The individual BSSE errors are then subtracted from the SCF and correlation energy respectively. Phe+Ca^{2+} is treated equivalently. Complete numerical details are given in the Supplementary Information (Supplementary Table S6). Following this procedure, the MAE and maximal error values of various DFAs compared to MP2 are well within 1 kcal/mol for Ala+Ca^{2+}. The PBE+vdW MAE for Phe+Ca^{2+} amounts to just above ~2 kcal/mol. The contributions from both the many-body dispersion and the hybrid PBE0 functional improve the MAE to just above 1 kcal/mol at to PBE0+MBD* level of theory. The maximum errors in the energy hierarchies between individual conformers are correspondingly larger. Overall, this assessment shows that our data base of conformer geometries constitutes, e.g., an excellent starting point for more exhaustive future benchmark work of new electronic structure methods for cation-peptide systems. For example, it would be very interesting to explore how F12 approaches, which address the correlation energy convergence problem explicitly, fare for a broad range of different Ca^{2+} containing conformations of our peptides.

As a final validation, we compare the correlation of calculated gas-phase amino acid-Ca^{2+} binding energies to the binding energy hierarchy found experimentally in a study by Ho *et al.*[110]. We calculate

binding energy at the PES level as

$$E_{binding} = E_{amino\ acid} + E_{cation} - E_{complex}. \quad (5)$$

Energies E denote the PBE+vdW Born-Oppenheimer potential energies, including $E_{amino\ acid}$ of the lowest-energy conformers of the isolated amino acid and $E_{complex}$ of the same amino acid in complex with a Ca^{2+} ion. Experimentally[110], the gas-phase Ca^{2+} affinities of 18 proteinogenic amino acids were determined by fragmenting Ca^{2+} complexes with a combinatoric library of tripeptides at $T \approx 330$ K, recording the mass spectrometric peak intensities of different fragmentation products. Quantitative average relative binding energies of Ca^{2+} to different amino acids were thus inferred and can be compared to our findings, albeit with several important experiment-theory differences: (i) Entropy effects[73,75,111] should affect the specific complexes probed experimentally but cannot be included into the calculated numbers in the exact same way, (ii) structural differences (e.g., protonation, dimerized amino acids) between the fragments recorded in experiment and the amino acids covered here, (iii) experimental Ca^{2+} affinities are not given for Asp and Glu because their gas-phase acidities, needed for data conversion, are not known. Fig. 7 compares the experimentally and theoretically inferred Ca^{2+} binding affinities qualitatively. The x-axis reflects the experimental binding affinity energy hierarchy, arranging amino acids from left to right in order of decreasing binding affinity. The y axis shows calculated binding energies according to equation (5). Perfect correlation of the experimental and calculated hierarchies would imply a strictly monotonic decrease of calculated $E_{binding}$ values from left to right. This monotonic trend is not obeyed exactly; however, in view of the significant differences (i) and (ii) above, the qualitative agreement is quite striking. Normalized correlation coefficients between the experimental (1) and calculated (2) binding affinity data were calculated following the formula:

$$r_{12} = s_{12}/(s_1 s_2), \quad (6)$$

with s_{12} being the covariance of data sets and s_i being the standard deviations of data sets $i = 1,2$. The result, correlation coefficients of $r_{12} = 0.979$ or 0.909 for uncapped amino acids or dipeptides, respectively, also point to an overall remarkably good agreement. Finally, Fig. 7 also gives predicted $E_{binding}$ values for protonated (overall system charge +2) and deprotonated (overall system charge +1) Asp and Glu, reflecting the significant electrostatic attraction between cations and negatively charged (deprotonated) Asp and Glu side chains. The binding energy data sets are included as Supplementary Table S5.

Usage Notes

The present data contains stationary-point geometries (mainly minima, but also saddle points since no routine normal-mode analysis was performed) on the potential energy surface of the 20 proteinogenic amino acids and dipeptides, either isolated or in complex with a divalent cation (Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, Hg^{2+}). The users of this dataset may find openbabel[112] (www.openbabel.org) to be a useful tool to convert FHI-aims and xyz files to other common file formats in chemistry.

References

1. Yu, W. *et al.* Extensive conformational searches of 13 representative dipeptides and an efficient method for dipeptide structure determinations based on amino acid conformers. *J. Comput. Chem.* **30,** 2105–2121 (2009).
2. Kishor, S., Dhayal, S., Mathur, M. & Ramaniah, L. M. Structural and energetic properties of α-amino acids: A first principles density functional study. *Mol. Phys.* **106,** 2289–2300 (2008).
3. Császár, A. G. & Perczel, A. Ab initio characterization of building units in peptides and proteins. *Prog. Biophys. Mol. Biol.* **71,** 243–309 (1999).
4. Bouchoux, G. Gas phase basicities of polyfunctional molecules. Part 3: Amino acids. *Mass Spectrom. Rev.* **31,** 391–435 (2012).
5. Matta, C. F. & Bader, R. F. W. Atoms-in-molecules study of the genetically encoded amino acids. II. Computational study of molecular geometries. *Proteins: Struct., Funct., Bioinf* **48,** 519–538 (2002).
6. Schlund, S., Muller, R., Grassmann, C. & Engels, B. Conformational analysis of arginine in gas phase-a strategy for scanning the potential energy surface effectively. *J. Comput. Chem.* **29,** 407–415 (2008).
7. Császár, A. G. On the structures of free glycine and α-alanine. *J. Mol. Struct.* **346,** 141–152 (1995).
8. Császár, A. G. Conformers of gaseous glycine. *J. Am. Chem. Soc.* **114,** 9568–9575 (1992).
9. Riffet, V., Frison, G. & Bouchoux, G. Acid-base thermochemistry of gaseous oxygen and sulfur substituted amino acids (Ser, Thr, Cys, Met). *Phys. Chem. Chem. Phys.* **13,** 18561–18580 (2011).
10. Kabelac, M., Hobza, P. & Spirko, V. The ab initio assigning of the vibrational probing modes of tryptophan: Linear shifting of approximate anharmonic frequencies vs. multiplicative scaling of harmonic frequencies. *Phys. Chem. Chem. Phys.* **11,** 3921–3926 (2009).
11. Kaschner, R. & Hohl, D. Density functional theory and biomolecules: A study of glycine, alanine, and their oligopeptides. *J. Phys. Chem. A* **102,** 5111–5116 (1998).
12. Maul, R., Ortmann, F., Preuss, M., Hannewald, K. & Bechstedt., F. DFT studies using supercells and projector-augmented waves for structure, energetics, and dynamics of glycine, alanine, and cysteine. *J. Comput. Chem.* **28,** 1817–1833 (2007).
13. Selvarengan, P. & Kolandaivel, P. Potential energy surface study on glycine, alanine and their zwitterionic forms. *J. Mol. Struct.: THEOCHEM* **671,** 77–86 (2004).
14. Cao, M., Newton, S. Q., Pranata, J. & Schafer, L. *J. Mol. Struct. THEOCHEM* **332,** 251 (1995).
15. Jaeger, H. M., Schaefer, H. F., Demaison, J., Császár, A. G. & Allen, W. D. Lowest-lying conformers of alanine: Pushing theory to ascertain precise energetics and semiexperimental re structures. *J. Chem. Theory Comput.* **6,** 3066–3078 (2010).
16. Beachy, M. D., Chasman, D., Murphy, R. B., Halgren, T. A. & Friesner, R. A. Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *J. Am. Chem. Soc.* **119,** 5908–5920 (1997).
17. Baek, K. Y., Fujimura, Y., Hayashi, M., Lin, S. H. & Kim, S. K. Density functional theory study of conformation-dependent properties of neutral and radical cationic l-tyrosine and l-tryptophan. *J. Phys. Chem. A* **115,** 9658–9668 (2011).
18. Chen, M. & Lin., Z. Ab initio studies of aspartic acid conformers in gas phase and in solution. *J. Chem. Phys.* **127,** 154314 (2007).

19. Floris, F. M., Filippi, C. & Amovilli., C. A density functional and quantum monte carlo study of glutamic acid in vacuo and in a dielectric continuum medium. *J. Chem. Phys.* **137,** 075102 (2012).
20. Heaton, A. L., Moision, R. M. & Armentrout, P. B. Experimental and theoretical studies of sodium cation interactions with the acidic amino acids and their amide derivatives. *J. Phys. Chem. A* **112,** 3319–3327 (2008).
21. Armentrout, P.B., Gabriel, A. & Moision., R.M. An experimental and theoretical study of alkali metal cation/methionine interactions. *Int. J. Mass Spectrom.* **283,** 56–68 (2009).
22. Nguyen, D. T. *et al.* A density functional study of the glycine molecule: Comparison with post-hartree-fock calculations and experiment. *J. Comput. Chem.* **18,** 1609–1631 (1997).
23. Espinoza, C., Szczepanski, J., Vala, M. & Polfer, N. C. Glycine and its hydrated complexes: A matrix isolation infrared study. *J. Phys. Chem. A* **114,** 5919–5927 (2010).
24. Boeckx, B., Nelissen, W. & Maes, G. Potential energy surface and matrix isolation ft-ir study of isoleucine. *J. Phys. Chem. A* **116,** 3247–3258 (2012).
25. Close, D. M. Calculated vertical ionization energies of the common alpha-amino acids in the gas phase and in solution. *J. Phys. Chem. A* **115,** 2900–2912 (2011).
26. Baek, K. Y., Hayashi, M., Fujimura, Y., Lin, S. H. & Kim, S. K. Investigation of conformation-dependent properties of l-phenylalanine in neutral and radical cations by using a density functional taking into account noncovalent interactions. *J. Phys. Chem. A* **114,** 7583–7589 (2010).
27. Sahai, M. A. *et al.* First principle computational study on the full conformational space of l-proline diamides. *J. Phys. Chem. A* **109,** 2660–2679 (2005).
28. Linder, R., Seefeld, K., Vavra, A. & Kleinermanns, K. Gas phase infrared spectra of nonaromatic amino acids. *Chem. Phys. Lett.* **453,** 1–6 (2008).
29. Vyas, N. & Ojha, A. K. Investigation on transition states of [alanine+m2+] (m = ca, cu, and zn) complexes: A quantum chemical study. *Int. J. Quant. Chem* **112,** 1526–1536 (2012).
30. Lavrich, R. J. *et al.* Experimental studies of peptide bonds: Identification of the c[sub 7][sup eq] conformation of the alanine dipeptide analog n-acetyl-alanine n[sup [prime]]-methylamide from torsion-rotation interactions. *J. Chem. Phys.* **118,** 1253–1265 (2003).
31. Zhang, M., Huang, Z. & Lin, Z. Systematic ab initio studies of the conformers and conformational distribution of gas-phase tyrosine. *J. Chem. Phys.* **122,** 134313 (2005).
32. Dokmaisrijan, S., Lee, V. S. & Nimmanpipug., P. The gas phase conformers and vibrational spectra of valine, leucine and isoleucine: An ab initio study. *J. Mol. Struct.: THEOCHEM* **953,** 28–38 (2010).
33. Ceci, M. L. *et al.* Exploratory conformational analysis of n-acetyl-l-tryptophan-n-methylamide. an ab initio study. *J. Mol. Struct.: THEOCHEM* **631,** 277–290 (2003).
34. Chen, M., Huang, Z. & Lin., Z. *J. Mol. Struct. THEOCHEM* **719,** 153 (2005).
35. Rassolian, M., Chass, G. A., Setiadi, D. H. & Csizmadia., I. G. Asparagine-ab initio structural analyses. *J. Mol. Struct.: {THEOCHEM}* **666–667,** 273–278 (2003).
36. Zamora, M. A. *et al.* An exploratory ab initio study of the full conformational space of n-acetyl-l-cysteine-n-methylamide. *J. Mol. Struct.: {THEOCHEM}* **540,** 271–283 (2001).
37. Rai, A. K., Song, C. & Lin., Z. An exploration of conformational search of leucine molecule and their vibrational spectra in gas phase using ab initio methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **73,** 865–870 (2009).
38. von Helden, G., Compagnon, I., Blom, M. N., Frankowski, M., Erlekam, U., Oomens, J., Brauer, B., Gerber, R. B. & Meijer, G. Mid-ir spectra of different conformers of phenylalanine in the gas phase. *Phys. Chem. Chem. Phys.* **10,** 1248–1256 (2008).
39. Riffet, V. & Bouchoux, G. Gas-phase structures and thermochemistry of neutral histidine and its conjugated acid and base. *Phys. Chem. Chem. Phys.* **15,** 6097–6106 (2013).
40. Shemesh, D., Sobolewski, A. L. & Domcke, W. Role of excited-state hydrogen detachment and hydrogen-transfer processes for the excited-state deactivation of an aromatic dipeptide: N-acetyl tryptophan methyl amide. *Phys. Chem. Chem. Phys.* **12,** 4899–4905 (2010).
41. Gabor, P., Perczel, A., Vass, E., Magyarfalvi, G. & Tarczay., G. A matrix isolation study on ac-gly-nhme and ac-l-ala-nhme, the simplest chiral and achiral building blocks of peptides and proteins. *Phys. Chem. Chem. Phys.* **9,** 4698–4708 (2007).
42. Bakker, J. M., Aleese, L. M., Meijer, G. & von Helden, G. Fingerprint ir spectroscopy to probe amino acid conformations in the gas phase. *Phys. Rev. Lett.* **91,** 203003 (2003).
43. Blanco, S., Sanz, M. E., López, J. C. & Alonso, J. L. Revealing the multiple structures of serine. *Proc. Natl. Acad. Sci. USA* **104,** 20183–20188 (2007).
44. Szidarovszky, T., Czakó, G. & Császár, A. *Mol. Phys.* **107,** 761 (2009).
45. Boeckx, B. & Maes, G. Experimental and theoretical observation of different intramolecular h-bonds in lysine conformations. *J. Phys. Chem. B* **116,** 12441–12449 (2012).
46. Meng, L. & Lin, Z. Comprehensive computational study of gas-phase conformations of neutral, protonated and deprotonated glutamic acids. *Computational and Theoretical Chemistry* **976,** 42–50 (2011).
47. Shankar, R., Kolandaivel, P. & Senthilkumar, L. Interaction studies of cysteine with li+, na+, k+, be2+, mg2+, and ca2+ metal cation complexes. *Journal of Physical Organic Chemistry* **24,** 553–567 (2011).
48. Fleming, G. J., McGill, P. R. & Idriss, H. Gas phase interaction of l-proline with be2+, mg2+ and ca2+ ions: a computational study. *Journal of Physical Organic Chemistry* **20,** 1032–1042 (2007).
49. Hu, C.-H., Shen, M. & Schaefer., H. F. Glycine conformational analysis. *J. Am. Chem. Soc.* **115,** 2923 (1993).
50. Barone, V., Biczysko, M., Bloino, J. & Puzzarini, C. Characterization of the elusive conformers of glycine from state-of-the-art structural, thermodynamic, and spectroscopic computations: Theory complements experiment. *J. Chem. Theory Comput.* **9,** 1533–1547 (2013).
51. Ai, H. Q., Bu, Y. X., Li, P. & Zhang., C. The regulatory roles of metal ions (m+/2+ = li+, na+, k+, be2+, mg2+, and ca2+) and water molecules in stabilizing the zwitterionic form of glycine derivatives. *New J. Chem.* **29,** 1540–1548 (2005).
52. Baldauf, C. & Hofmann., H.-J. Ab initio mo theory-an important tool in foldamer research: Prediction of helices in oligomers of ω-amino acids. *Helvetica Chimica Acta* **95,** 2348–2383 (2012).
53. Ramek, M., Kelterer, A.-M. & Nikolić, S. Ab initio and molecular mechanics conformational analysis of neutral l-proline. *Int. J. Quant. Chem* **65,** 1033–1045 (1997).
54. Czinki, E. & Császár, A. G. Conformers of gaseous proline. *Chem. Eur. J* **9,** 1008–1019 (2003).
55. Kang., Y. K. *Ab initio* molecular orbital calculations on low-energy conformers of N-Acetyl-N‘-methylprolineamide. *J. Phys. Chem.* **100,** 11589 (1996).
56. Xu, S., Ke-Dong, W. & Peng-Fei, M. Conformation effects on the molecular orbitals of serine. *Chinese Physics B* **20,** 33102 (2011).
57. Yuan, Y., Mills, M. J. L., Popelier, P. L. A. & Jensen, F. Comprehensive analysis of energy minima of the 20 natural amino acids. *J. Phys. Chem. A* **118,** 7876–7891 (2014).

58. Karton, A., Yu, L.-J., Kesharwani, M. & Martin, J. M. L. Heats of formation of the amino acids re-examined by means of w1-f12 and w2-f12 theories. *Theoretical Chemistry Accounts* **133,** 1483 (2014).
59. Kesharwani, M. K., Karton, A. & Martin, J. M. L. Benchmark ab initio conformational energies for the proteinogenic amino acids through explicitly correlated methods. assessment of density functional methods. *J. Chem. Theory Comput.* **12,** 444–454 (2015).
60. Holm, R. H., Kennepohl, P. & Solomon, E. I. Structural and functional aspects of metal sites in biology. *Chemical Reviews* **96,** 2239–2314 (1996).
61. Tainer, J. A., Roberts, V. A. & Getzoff, E. D. Protein metal-binding sites. *Current Opinion in Biotechnology* **3,** 378–387 (1992).
62. Kirberger, M. & Yang, J. J. Structural differences between Pb^{2+}- and Ca^{2+}-binding sites in proteins: Implications with respect to toxicity. *Journal of Inorganic Biochemistry* **102,** 1901–1909 (2008).
63. Zhou, M. *et al.* and Jianping Ding. A novel calcium-binding site of von Willebrand factor A2 domain regulates its cleavage by ADAMTS13. *Blood* **117,** 4623–4631 (2011).
64. Cheng, R. & Zhorov, B. Docking of calcium ions in proteins with flexible side chains and deformable backbones. *European Biophysics Journal* **39,** 825–838 (2010).
65. Sadiq, S., Ghazala, Z., Chowdhury, A. & Büsselberg, D. Metal toxicity at the synapse: Presynaptic, postsynaptic, and long-term effects. *Journal of Toxicology* **2012,** 132671 (2012).
66. Sharma, S. K., Goloubinoff, P. & Christen, P. Heavy metal ions are potent inhibitors of protein folding. *Biochemical and Biophysical Research Communications* **372,** 341–345 (2008).
67. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev* **136,** B864–B871 (1964).
68. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev* **140,** A1133–A1138 (1965).
69. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77,** 3865–3868 (1996).
70. Tkatchenko, A. & Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102,** 073005 (2009).
71. Rossi, M. *et al.* Secondary structure of Ac-Ala_n-$LysH^+$ polyalanine peptides (n = 5,10,15) in vacuo: Helical or not? *J. Phys. Chem. Lett.* **1,** 3465–3470 (2010).
72. Tkatchenko, A., Rossi, M., Blum, V., Ireta, J. & Scheffler, M. Unraveling the stability of polypeptide helices: Critical role of van der Waals interactions. *Phys. Rev. Lett.* **106,** 118102 (2011).
73. Baldauf, C. *et al.* How cations change peptide structure. *Chemistry-A European Journal* **19,** 11224–11234 (2013).
74. Chutia, S., Rossi, M. & Blum, V. Water adsorption at two unsolvated peptides with a protonated lysine residue: From self-solvation to solvation. *J. Phys. Chem. B* **116,** 14788–14804 (2012).
75. Rossi, M., Scheffler, M. & Blum, V. Impact of vibrational entropy on the stability of unsolvated peptide helices with increasing length. *J. Phys. Chem. B* **117,** 5574–5584 (2013).
76. Rossi, M., Chutia, S., Scheffler, M. & Blum, V. Validation challenge of density-functional theory for peptides-example of Ac-Phe-Ala_5-$LysH^+$. *J. Phys. Chem. A* **118,** 7349–7359 (2014).
77. Schubert, F. *et al.* Native like helices in a specially designed [small beta] peptide in the gas phase. *Phys. Chem. Chem. Phys.* **17,** 5376–5385 (2015).
78. Schubert, F. *et al.* Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala19-Lys+H^+ vs. Ac-Lys-Ala19+H^+ and the current reach of DFT. *Phys. Chem. Chem. Phys.* **17,** 7375–7385 (2015).
79. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118,** 11225–11236 (1996).
80. Wales, D. J. & Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **101,** 5111–5116 (1997).
81. Wales, D. J. & Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **285,** 1368–1372 (1999).
82. Ponder, J. W. & Richards., F. M. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.* **8,** 1016–1024 (1987).
83. Ren, P. & Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **107,** 5933–5947 (2003).
84. Blum, V. *et al.* Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* **180,** 2175–2196 (2009).
85. Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient integration for all-electron electronic structure calculation using numeric basis functions. *Journal of Computational Physics* **228,** 8367–8379 (2009).
86. Ren, X. *et al.* Resolution-of-identity approach to Hartree-Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. *New Journal of Physics* **14,** 053020 (2012).
87. van Lenthe, J. H., Faas, S. & Snijders, J. G. Gradients in the ab initio scalar zeroth-order regular approximation (ZORA) approach. *Chem. Phys. Lett.* **328,** 107–112 (2000).
88. van Wullen, C. Molecular density functional calculations in the regular relativistic approximation: Method, application to coinage metal diatomics, hydrides, fluorides and chlorides, and comparison with first-order relativistic calculations. *J. Chem. Phys.* **109,** 392–399 (1998).
89. Swendsen, R. H. & Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **57,** 2607–2609 (1986).
90. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314,** 141–151 (1999).
91. Earl, D. J. & Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **7,** 3910 (2005).
92. Beret, E. C., Ghiringhelli, L. M. & Scheffler, M. Free gold clusters: beyond the static, monostructure description. *Faraday Discuss.* **152,** 153–167 (2011).
93. Sindhikara, D., Meng, Y. & Roitberg, A. E. Exchange frequency in replica exchange molecular dynamics. *J. Chem. Phys.* **128,** 024103 (2008).
94. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 014101 (2007).
95. Hartigan, J. A. & Wong, M. A. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28,** 100–108 (1979).
96. Feig, M., Karanicolas, J. & Brooks III, C. L. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling* **22,** 377–395 (2004).
97. Humphrey, W., Dalke, A. & Schulten, K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics* **14,** 33–38 (1996).
98. Perdew., J. P. Unified Theory of Exchange and Correlation Beyond the Local Density Approximation. in *Electronic Structure of Solids '91-Proceedings of the 75. WE-Heraeus-Seminar and 21st Annual International Symposium on Electronic Structure of Solids; Gaussig, Germany*; 11–15 March 1991 (Akademie Verlag, Berlin, 1991).
99. Møller, C. & Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev* **46,** 618–622 (1934).

100. Head-Gordon, M., Poplei, J. A. & Frisch., M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **153,** 503–506 (1988).
101. Neese., F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2,** 73–78 (2012).
102. Woon, D. E. & Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. V. Corevalence basis sets for boron through neon. *J. Chem. Phys.* **103,** 4572–4585 (1995).
103. Karton, A. & Martin, J. M. L. Comment on: "Estimating the Hartree-Fock limit from finite basis set calculations" [Jensen F (2005) Theor Chem Acc 113:267]. *Theor. Chem. Acc.* **115,** 330–333 (2006).
104. Truhlar, D.G. Basis-set extrapolation. *Chem. Phys. Lett.* **294,** 45–48 (1998).
105. Ambrosetti, A., Reilly, A. M., DiStasio, R. A. & Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **140,** 18A508 (2014).
106. Adamo, C. & Barone., V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110,** 6158–6170 (1999).
107. Zhang, I. Y., Ren, X., Rinke, P., Blum, V. & Scheffler, M. Numeric atom-centered-orbital basis sets with valence-correlation consistency from h to ar. *New Journal of Physics* **15,** 123033 (2013).
108. Jansen, H. B. & Ros, P. Non-empirical molecular orbital calculations on the protonation of carbon monoxide. *Chem. Phys. Lett.* **3,** 140–143 (1969).
109. Boys, S.F. & Bernardi., F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **19,** 553–566 (1970).
110. Ho, Y.-P., Yang, M.-W., Chen, L.-T. & Yang, Y.-C. Relative calcium-binding strengths of amino acids determined using the kinetic method. *Rapid Communications in Mass Spectrometry* **21,** 1083–1089 (2007).
111. Liwo, A., Khalili, M. & Scheraga., H. A. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. USA* **102,** 2362–2367 (2005).
112. O'Boyle, N. *et al.* Open babel: An open chemical toolbox. *Journal of Cheminformatics* **3,** 33 (2011).

Data Citation

1. Ropo, M., Baldauf, C. & Blum, V. *NOMAD repository* http://dx.doi.org/10.17172/NOMAD/20150526220502 (2015).

Acknowledgements

The authors are grateful to Matthias Scheffler (Fritz Haber Institute Berlin) for support of this work and stimulating discussions. Luca Ghiringhelli is gratefully acknowledged for his work on the script-based parallel-tempering scheme that is provided with FHI-aims and that was used in the present work. The authors thank Robert Maul and Karsten Hannewald for making available the original alanine geometries derived in their 2007 study for comparison with the present results. The authors further thank Mariana Rossi, Franziska Schubert, and Sucismita Chutia for sharing their extensive experience with all search methods employed in this work.

Author Contributions

M.R. performed the calculations to assemble all conformers. M.R. and C.B. curated the data. Validation calculations by DFAs and correlated methods other than PBE+vdW were carried out by M.S. M.R., C.B., V.B. designed the study and wrote the data descriptor.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/sdata

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Ropo, M. *et al.* First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data* 3:160009 doi: 10.1038/sdata.2016.9 (2016).

Curriculum Vitae

Carsten Baldauf

Desk address	Fritz-Haber-Institut der Max-Planck-Gesellschaft Faradayweg 4-6 D-14195 Berlin
Date of birth	December 23, 1977
Place of birth	Karl-Marx-Stadt
E-mail	baldauf@fhi-berlin.mpg.de

Teaching

WS 2015/16

"*Atombau und chemische Bindung* (PC 2)" for chemistry students (Bachelor) at FU Berlin

WS 2012/13, WS 2013/14, WS 2014/15

"Bioorganic chemistry II" for chemistry students (Master) at FU Berlin

August 2013, July 2015, May 2016

"Hands-on workshop DFT and beyond" (10 to 12 days) for PhD students and young PostDocs, organization and conduction of the workshops at ICTP Trieste, Harnackhaus Berlin, and Isfahan University of Technology

Research experience

Since 09/2013, Group leader

Research group *Ab Initio* Biomolecular Simulations, Prof. Matthias Scheffler's Theory Department at Fritz-Haber-Institute Berlin

Since 08/2010, Scientist

Research group *Ab Initio* Biomolecular Simulations, Prof. Matthias Scheffler's Theory Department at Fritz-Haber-Institute Berlin

05/2010-07/2010, Visiting scholar

Molecular Biomechanics Groups (Prof. Frauke Gräter), Heidelberg Institute for Theoretical Studies (HITS)

04/2010, Teaching "Physikalisch-chemisches Grundpraktikum für Pharmazeuten und Biochemiker"

Institute of Biochemistry, Fakultät für Biowissenschaften, Pharmazie und Psychologie, Universität Leipzig

02/2009-03/2009, Visiting scholar

Department of Materials Science and Engineering (DMSE, Prof. Alfredo Alexander-Katz), Massachusetts Institute of Technology (MIT)

01/2008-03/2010, Scientist, Feodor Lynen fellow of the Humboldt foundation

BioQuant, Universität Heidelberg, Germany and MPG/CAS Partner-Institute for Computational Biology, Shanghai Institutes for the Biological Sciences, Chinese Academy of the Sciences (CAS/MPG-PICB), Shanghai, China

11/2005-12/2007, Scientist

Biotechnologiezentrum der TU Dresden (Dr. M. Teresa Pisabarro)

04/2002-10/2005, Scientist, PhD student

Institut für Biochemie (Prof. Hans-Jörg Hofmann), Fakultät für Biowissenschaften, Pharmazie und Psychologie, Universität Leipzig

Education

06/2005, PhD in Biochemistry

Grade	*magna cum laude*, Dr. rer. nat.
Institution	Fakultät für Biowissenschaften, Pharmazie und Psychologie Universität Leipzig
Thesis title	Secondary Structure Formation in Homologous Peptides
Supervisor	Prof. Hans-Jörg Hofmann (Universität Leipzig)

03/2002, Diploma in Biochemistry

Grade	*sehr gut* (1,5), Diplom-Biochemiker
Institution	Fakultät für Biowissenschaften, Pharmazie und Psychologie Universität Leipzig
Thesis title	Sulfonamidopeptide und vinyloge Peptide als Foldamere
Supervisor	Prof. Hans-Jörg Hofmann (Universität Leipzig)

1998-2002, Studies in Biochemistry

Institution	Fakultät für Biowissenschaften, Pharmazie und Psychologie Universität Leipzig
Exams	Biochemistry (A. Beck-Sickinger): *gut* (1,7) Biophysical Chemistry (H.-J. Hofmann): *gut* (2,0) Molecular biology (U. Hahn): *sehr gut* (1,3) Chemistry of natural products (P. Welzel): *sehr gut* (1,3)

1997-1998, Studies in Biochemistry

Fakultät für Chemie und Biochemie
Ruhr-Universität Bochum

Scientific awards

2007 Travel stipend to give a talk at the American Peptide Symposium in Montreal, Canada

2008 Feodor-Lynen Fellowship of the Alexander-von-Humboldt foundation to stay at CAS/MPG Partner-Institute for Computational Biology in Shanghai, China

2010 '*Nachwuchsförderpreis*' of the German Society for Thrombosis and Hemostasis for research on the mechanics of von Willebrand factor

2010 JTH Mannucci Award for the best article by a junior researcher: Shear-Induced Unfolding Activates von Willebrand Factor A2 Domain for Proteolysis. *J. Thromb. Haemost.* **2009** (7), 2096-2105.

Publications

Sub Ropo M, Blum V, Baldauf C
Trends for isolated amino acids and dipeptides: Conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead
Sci Rep. **2016**; *submitted*, arXiv: 1606.02151.

---

37 Posch S, Aponte-Santamaría C, Schwarzl R, Karner A, Radtke M, Gräter F, Obser T, König G, Brehm MA, Gruber HJ, Netz RR, Baldauf C, Schneppenheim R, Tampé R, Hinterdorfer P
Mutual A domain interactions in the force sensing protein von Willebrand factor
J Struct Biol. **2016**; *accepted*, DOI: 10.1016/j.jsb.2016.04.012.

36 Pecina A, Meier R, Fanfrlík J, Lepšík M, Řezáč J, Hobza P, Baldauf C
The SQM/COSMO filter: Reliable native pose identification based on the quantum-mechanical description of protein-ligand interactions and implicit COSMO solvation
Chem Comm. **2016**;52:3312.

35 Ropo M, Schneider M, Baldauf C, Blum V
First-principles data set of 45,892 isolated and cation coordinated conformers of 20 proteinogenic amino acids
Sci Data. **2016**;3:160009.

34 Lippok S, Kolšek K, Löf A, Eggert D, Vanderlinden W, Müller JP, König G, Obser T, Röhrs K, Schneppenheim S, Budde U, Baldauf C, Aponte-Santamaría C, Gräter F, Schneppenheim R, Rädler JO, Brehm MA
Von Willebrand factor is dimerized by protein disulfide isomerase
Blood. **2016**;127:1183.

33 Chen P, Marianski M, Baldauf C
H-bond isomerization in crystalline cellulose III_I: Proton hopping versus hydroxyl flip-flop
ACS Macro Lett. **2016**;5:50.

32 Baldauf C, Rossi M
Going clean: Structure and dynamics of peptides in the gas phase and paths to solvation
J Phys Cond Matt. **2015**;27:493002.

31 Supady A, Blum V, Baldauf C
First-principles molecular structure search with a genetic algorithm
J Chem Inf Model. **2015**;55:2338.

30 Aponte-Santamaría C, Huck V, Posch S, Bronowska AK, Grässle S, Brehm MA, Obser T, Schneppenheim R, Hinterdorfer P, Schneider SW, Baldauf C, Gräter F
Force-sensitive autoinhibition of the von Willebrand factor mediated by inter-domain interactions
Biophys J. **2015**;108:2312.

29 Schubert F, Rossi M, Baldauf C, Pagel K, Warnke S, von Helden G, Filsinger F, Kupser P, Meijer G, Salwiczek M, Koksch B, Scheffler M, Blum V
Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala_{19}-Lys+H^+ vs. Ac-Lys-Ala_{19}+H^+ and the current reach of DFT
Phys Chem Chem Phys. **2015**;17:7373.

28 Mortier J, Nyakatura EK, Reimann O, Huhmann S, Daldrop JO, Baldauf C, Wolber G, Miettinen MS, Koksch B
Coiled-coils in phage display screening: Insight into exceptional selectivity provided by molecular dynamics
J Chem Inf Model. **2015**;55:495.

27 Schubert F, Pagel K, Rossi M, Warnke S, Salwiczek M, Koksch B, von Helden G, Blum V, Baldauf C, Scheffler M
Native like helices in a specially designed β peptide in the gas phase
Phys Chem Chem Phys. **2015**;17:5376.

26 Warnke S, Baldauf C, Bowers MT, Pagel K, von Helden G
Photodissociation of conformer-selected ubiquitin ions reveals site-specific cis/trans isomerization of proline peptide bonds
J Amer Chem Soc. **2014**;136:10308.

25 Nyakatura EK, Rezaei Araghi R, Mortier J, Wieczorek S, Baldauf C, Wolber G, Koksch B
An unusual interstrand H-bond stabilizes the heteroassembly of helical αβγ-chimeras with natural peptides
ACS Chem Biol. **2014**;9:613.

24 Grässle S, Huck V, Pappelbaum KI, Gorzelanny C, Aponte-Santamaría C, Baldauf C, Gräter F, Schneppenheim R, Obser T, Schneider SW
von Willebrand factor directly interacts with DNA from neutrophil extracellular traps
Arterioscler Thromb Vasc Biol. **2014**;34:1382.

23 Elisabeth K. Nyakatura JM, Vanessa S. Radtke, Sebastian Wieczorek, Raheleh Rezaei Araghi, Carsten Baldauf, Gerhard Wolber, Beate Koksch
β- and γ-amino acids at α-helical interfaces: Toward the formation of highly stable coldameric coiled coils
ACS Med Chem Lett. **2014**;5:1300.

22 Brehm MA, Huck V, Aponte-Santamaria C, Obser T, Grässle S, Oyen F, Budde U, Schneppenheim S, Baldauf, C., Gräter, F, Schneider, SW, Schneppenheim, R
von Willebrand disease type 2A phenotypes IIC, IID and IIE: A day in the life of shear-stressed mutant von Willebrand factor
Thromb Haemost. **2014**;112:96.

21 Baldauf C, Pagel K, Warnke S, von Helden G, Koksch B, Blum V, Scheffler M
How cations change peptide structure
Chem Eur J. **2013**;19:11224.

20 Chen J, Edwards SA, Gräter F, Baldauf C
On the *cis* to *trans* isomerization of prolyl-peptide bonds under tension
J Phys Chem B. **2012**;116:9346.

19 Baldauf C, Hofmann H-J
Ab initio MO theory - An important tool in foldamer research: Prediction of helices in oligomers of ω-amino acids
Helv Chim Acta. **2012**;95:2348.

18 Zhou M, Dong X, Baldauf C, Chen H, Zhou Y, Springer TA, Luo X, Zhong C, Gräter F, Ding J
A novel calcium-binding site of von Willebrand factor A2 domain regulates its cleavage by ADAMTS13
Blood. **2011**;117:4623.

17 Araghi RR, Baldauf C, Gerling UI, Cadicamo CD, Koksch B
A systematic study of fundamentals in α-helical coiled coil mimicry by alternating sequences of β-and γ-amino acids
Amino Acids. **2011**;41:733.

16 Rezaei Araghi R, Jäckel C, Cölfen H, Salwiczek M, Völkel A, Wagner SC, Wieczorek S, Baldauf C, Koksch B
A β/γ Motif to mimic α-helical turns in proteins
ChemBioChem. **2010**;11:335.

15 Meier R, Pippel M, Brandt F, Sippl W, Baldauf C
ParaDockS: a framework for molecular docking with population-based metaheuristics
J Chem Inf Model. **2010**;50:879.

14 Sharma GV, Babu BS, Ramakrishna KV, Nagendar P, Kunwar AC, Schramm P, Baldauf C, Hofmann H-J
Synthesis and structure of α/δ-hybrid peptides - Access to novel helix patterns in foldamers
Chem Eur J. **2009**;15:5552.

13 Baldauf C, Schneppenheim R, Stacklies W, Obser T, Pieconka A, Schneppenheim S, Budde U, Zhou J, Gräter F
Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis
J Thromb Haemost. **2009**;7:2096.

12 Baldauf C, Pisabarro MT
Stable hairpins with β-peptides: route to tackle protein-protein interactions
J Phys Chem B. **2008**;112:7581.

11 Scheike JA, Baldauf C, Spengler J, Albericio F, Pisabarro MT, Koksch B
Amide-to-ester substitution in coiled coils: The effect of removing hydrogen bonds on protein structure
Angew Chem Int Ed. **2007**;46:7766.

10 Lang M, De Pol S, Baldauf C, Hofmann H-J, Reiser O, Beck-Sickinger AG
Identification of the key residue of calcitonin gene related peptide (CGRP) 27-37 to obtain antagonists with picomolar affinity at the CGRP receptor
J Med Chem. **2006**;49:616.

9 Baldauf C, Günther R, Hofmann H-J
Theoretical prediction of the basic helix types in α, β-hybrid peptides
Biopolymers Pept Sci. **2006**;84:408.

8 Baldauf C, Günther R, Hofmann H-J
Helices in peptoids of α-and β-peptides
Phys Biol. **2006**;3:S1.

7 Baldauf C, Günther R, Hofmann H-J
Helix formation in α,γ- and β,γ-hybrid peptides: theoretical insights into mimicry of α- and β-peptides
J Org Chem. **2006**;71:1200.

6 Baldauf C, Günther R, Hofmann H-J
Side-chain control of folding of the homologous α-, β-, and γ-peptides into "mixed" helices (β-helices)
Biopolymers Pept Sci. **2005**;80:675.

5 Baldauf C, Günther R, Hofmann H-J
Control of helix formation in vinylogous γ-peptides by (E)-and (Z)-double bonds: a way to ion channels and monomolecular nanotubes
J Org Chem. **2005**;70:5351.

4 Baldauf C, Günther R, Hofmann H-J
Mixed helices - A general folding pattern in homologous peptides?
Angew Chem Int Ed. **2004**;43:1594.

3 Baldauf C, Günther R, Hofmann H-J
Conformational properties of sulfonamido peptides
J Mol Struct THEOCHEM. **2004**;675:19.

2 Baldauf C, Günther R, Hofmann H-J
δ-Peptides and δ-amino acids as tools for peptide structure design - A theoretical study
J Org Chem. **2004**;69:6214.

1 Baldauf C, Günther R, Hofmann H-J
Helix formation and folding in γ-peptides and their vinylogues
Helv Chim Acta. **2003**;86:2573.

Diwold K, Himmelbach D, Meier R, Baldauf C, Middendorf M
Bonding as a swarm: applying bee nest-site selection behaviour to protein docking
Proceedings of GECCO'11. **2011**:93. (Conference article)

Patent Baldauf C, Pisabarro MT (Technische Universität Dresden)
Modular scaffold for the design of specific molecules for the use as peptidomimetics and inhibitors of protein interaction. EP 2 105 433 A1.