Universität für Bodenkultur Wien
University of Natural Resources
and Life Sciences, Vienna

Master Thesis

Exploring the Structure and Dynamics of Proteins in Soil Organic Matter

Mathias Gotsmy

In Fulfilment of the Requirments for the academic degree
Diplom-Ingenieur (Dipl.-Ing.)

Master degree course: Biotechnology

Department: Material Sciences and Process Engineering
Institute: Molecular Modeling and Simulation
Supervisor: Univ.Prof. Dr. Chris Oostenbrink
Co-supervisor: Yerko Escalona, MSc.

Vienna, April 2020

Declaration

I hereby declare that I am the sole author of this work. No assistance other than that which is permitted has been used. Ideas and quotes taken directly or indirectly from other sources are identified as such.

Mathias Gotsmy

Acknowledgements

I want to thank the following people that supported me in my work for this thesis.

- Yerko Escalona, for being my co-supervisor, helping me with countless problems and introducing me to Bash and PYTHON programming.
- Drazen Petrov, for giving me scientific advice as well as providing me with a lot of helpful scripts.
- Matthias Diem, for helping me with statistical questions and questions related to SASA.
- Christoph Öhlknecht, for helping me with free energy calculations and charge corrections and providing me with helpful scripts.
- Chris Oostenbrink, for sparking my interest in molecular dynamics, being my supervisor, giving me scientific advise and enabling me to do my master thesis.
- All members of the MMS for answering a lot of my questions as well as creating an awesome work environment.
- My friends and family for their emotional support.

Thank you all!

Abstract

Soil organic matter (SOM) is an important component of soil. Organisms like plants, fungi and bacteria that live in soil excrete proteins into it. At the present, there is little research done on the stability and interaction of proteins in SOM at an atomistic level.

In this thesis, molecular dynamics simulations were used to investigate selected proteins in different soil models of different complexity. I found that the stability of proteins is not impaired if hydrophobic and hydrophilic functional groups are linked in the same SOM molecule However, spatially differentiation of these functional groups in different SOM molecules can result in unfolding of proteins. Additionally, I showed that different functional groups of SOM always order around the protein in a similar pattern, though the structure of the SOM molecules is different. The interaction energies between proteins and SOM were primarily governed by electrostatic interactions, mostly represented as hydrogen bonds. Van der Waals interactions were observed to be significantly weaker.

By computing Gibbs free energies of mutation of amino acids in water and SOM I showed that the interaction of proteins and SOM also greatly depended on the amino acid sequence of proteins. With these results SOMscore, a method that scores proteins on their interaction strength with SOM, was developed. By applying SOMscore on a set of 1190 proteins, I found that the majority of these proteins were predicted to be absorbed by SOM. However, several enzymes that are often found in soils were observed to be repulsive to SOM. This has biological meaning since enzymes that get absorbed by SOM cannot longer function correctly. Therefore, there appears to be an evolutionary benefit for soil enzymes to become repulsive to SOM.

Kurzfassung

Organische Bodensubstanzen (soil organic matter, SOM) sind ein wichtiger Bestandteil des Bodens. Organismen wie Pflanzen, Pilze und Bakterien, die im Boden leben, sekretieren Proteine in diesen. Allerdings sind die Stabilität und Interaktion von Proteinen in SOM auf atomistischer Ebene wenig erforscht.

In dieser Arbeit wurden molekulardynamische Simulationen verwendet, um exemplarische Proteine in verschiedenen SOM Modellen zu untersuchen. Ich fand heraus, dass die Stabilität von Proteinen nicht leidet, wenn hydrophobe und hydrophile funktionelle Gruppen in demselben SOM Molekül verbunden sind. Allerdings können sich Proteine entfalten, wenn diese funktionellen Gruppen in unterschiedlichen SOM Molekülen separiert sind. Außerdem konnte ich zeigen, dass die Anordnung verschiedener funktioneller Gruppen von SOM rund um das Protein sehr ähnlich ist, auch wenn sich die Struktur von SOM Molekülen stark unterscheidet. Die Interaktionsenergien zwischen Proteinen und SOM resultieren in erster Linie aus elektrostatischen Anziehungskräften, welche oft in Form von Wasserstoffbrückenbindungen auftraten. Van der Waals Kräfte traten signifikant geringer auf.

Mittels Berechnung von Gibbs-Energien von Mutationen von Aminosäuren in Wasser und SOM zeigte ich, dass die Interaktion von Proteinen und SOM ebenfalls stark von der Aminosäurensequenz abhängt. Mit diesen Resultaten wurde SOMscore, eine Methode Proteine nach ihrer Interaktionsenergie zu SOM zu bewerten, entwickelt. SOMscore wurde anschließend an einem Set von 1190 Proteinen angewandt und prognostizierte, dass der Großteil aller getesteten Proteine zur Absorption in SOM neigten. Allerdings wurden einige Enzyme, die häufig im Boden anzutreffen sind, auffällig abstoßend zu SOM bewertet. Dieser Umstand hat biologische Bedeutung, da es für die Aktivität von Enzymen wichtig ist, nicht in SOM absorbiert zu werden. Demzufolge scheint es einen evolutionären Vorteil für Bodenenzyme zu geben abstoßend zu SOM zu sein.

Contents

1. Introduction to Soil Sciences

Soil science seeks to describe the nature and properties of soil, which is the top layer of earth's crust. The thickness of the layer usually varies from few to multiple centimeters and, therefore, constitutes only a tiny part of the whole crust.[1] However, soil is a fundamental basis for life on earth as we know it and its development is deeply entangled with the development of life itself.[2] The capacity of soil to absorb a magnitude of compounds, such as water, organic and inorganic molecules, including pesticides and toxins, makes it an important part of earths ecosphere. Soil plays a key role in the carbon cycle and is, therefore, also linked to climate change[3]. Additionally, soil itself is a habitat for a multitude of organisms, for example animals, plant roots, fungi and microorganisms.[1] Therefore, soil is an important factor which has multiple influences on the environment and subsequently on wildlife and humanity.

Soil has been titled the most complex biomaterial on earth.[4] Rather than being homogeneous, it is highly variable in its composition, porosity, layer size, texture, pH and many more properties. Soil classification is challenging because the properties are dependent on location as well as depth. In addition, land use has an important influence.[5] Moreover, changes of soil properties occur on the macroscale but also on the microscale. Different classification models were developed for and by different groups of people (for example farmers or nations) which lead to a lot of confusion in the past. Nowadays, a comprehensive classification system (soil taxonomy) was devised, which bases on objectively measurable soil properties, such as moisture, temperature, color, texture and the structure of soil.[6] Soil consists of solids (organic and inorganic matter), liquids (water, dissolved molecules and ions) and gases (mostly air). A typical silt loam soil, which is a good basis for plant growth, is composed mostly of minerals and a small fraction of soil organic matter (SOM) (Figure 1.1).[7]

The most abundant elements in the earths crust, and therefore, in minerals are O, Si, Al, Fe, Mg, Ca, Na and K.[1] Soil minerals are either present as insoluble crystal particles or dissolved as ions. Soil particles are classified by size (clay < silt < sand). Clay particles are the finest particles present and their properties (like adsorption and cation-exchange capacity) play an important role since they affect soil chemical reactions and processes.[7]

The principal components of soil form aggregates which are categorized into micro- and macroaggregates.[8] Microaggregates mainly comprise organic molecules and clay minerals. Polyvalent cations such as Si^{4+}, Fe^{3+}, Al^{3+} and Ca^{2+} form salt bridges

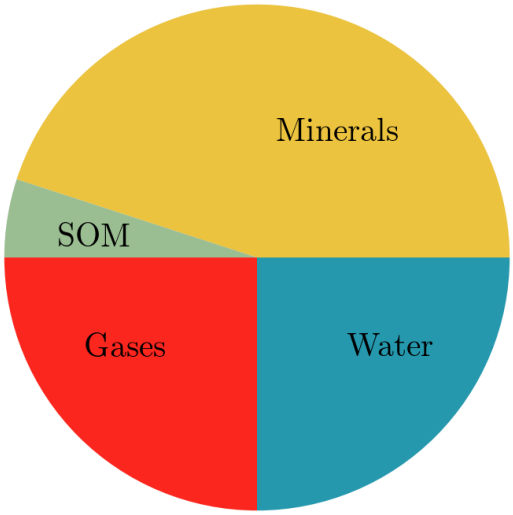

Figure 1.1: Soil component fractions for a typical silt loam soil: 45% minerals, 5% organic mater (SOM), 20-30% water and 20-30% gases such as air.[7]

between them.[9] Multiple microaggregates are bound together by temporary binding agents to form macroaggregates. Roots and hyphae can release organic molecules that act as binding agents and thus influence macroaggregate formation.[9, 10] The formation of aggregates stabilizes organic matter and protects it from microbial decay.[11]

1.1 Soil Organic Matter

Soil organic matter (SOM) is defined as the product of organic molecule degradation processes in soil. Its major components are decomposing plant parts, microbial remains, mineral bound organic matter, charcoal from forest fires as well as dissolved organic matter.[1] Even though SOM is a small fraction of soil, it has great influence on the carbon cycle.[1] The total amount of organic carbon stored in soil is estimated at 3,500 – 4,800 gigatons which exceeds what is stored in the global vegetation and atmosphere combined.[12, 13]

1.1.1 SOM Extraction Process

The study of SOM has a long history. The first report was published in 1786.[14] Since experimental possibilities were limited at that time, a relatively simple alkaline extraction technique was used to separate SOM from minerals and other soil components. Even though since then there have been many variations on the method the basic principle of extraction stayed.[13] A soil sample is mixed with NaOH for a specific time in

a specific temperature to make SOM molecules soluble. Subsequently, the mixture is centrifuged which results in an alkaline supernatant containing humic substances in solution and a non-soluble part containing organic compounds (humin) but also minerals or undecomposed plant material. The supernatant is then removed and acidified to a pH of 1 to 2 in which humic acids precipitate, fulvic, however, stay in solution. The last extraction step usually consists of desalting and washing out of nonhumic materials.[15] This operational definition of humic substances has been criticized because of the extraction process being incomplete and not representing the entirety of SOM which could lead to the creation of incorrect models.

1.1.2 SOM Models and the Definition of Humic Substances

To understand the origins of SOM, how and in which form it is stored in soil and how it is decomposed, several models have been developed. The classical humification model bases on the assumption that organic matter from vegetation and microorganisms is firstly degraded into small molecules such as phenols, phenylpropene units and others, which subsequently react with each other to form higher molecular weight substances which are referred to as humic substances.[16] Therefore, the classical humification model describes HS as being distinct from sole degradation products of organic compounds on a molecular level and differ from SOM by not comprising free, identifiable constituents such as amino acids, sugars and polysaccharides.[17] However, this model has recently been heavily criticized, for example by Lehmann and Kleber.[13] They argued that there is no experimental evidence for the existence of humic substances on a molecular level. They rather proposed the soil continuum model in which they completely dismiss the term 'humic'. Instead they propose a linear degradation pathway where complex organic molecules are gradually digested into smaller subunits until they are either assimilated by living cells or finally oxidized to CO_2. This model, however, has also been criticized by overstating the biological degradation processes without taking chemical polymerization reactions into account.[16]

Even though both models are under discussion there is agreement between most research groups on four central assumptions:[16] (1) main starting materials of SOM are various plants and soil microorganisms which are partially degraded. (2) the degradation products, for example phenols, phenylpropene units, amino acids and sugars, can polymerize. These reactions can be catalysed by soil oxidoreductases, mineral surfaces or soil microorgansims to form organic molecules that are usually referred to as humic substances. (3) Humic substances have high biochemical stability and are, therefore, highly resistant to microbial degradation. [18] (4) It is possible to deduct structural models from the investigation of chemical properties of humic substances.[19–22] Ultimately, this assumption lead to a more recent molecular dynamics model called VSOMM introduced by Sündermann et al. (2015)[23] (Section 1.1.5).

1.1.3 IHSS Standard Samples

Due to the considerable variability of SOM it is difficult for researchers to compare and reproduce their results around the world. A solution to this problem was the foundation of the International Humic Substance Society (IHSS), which goal is to provide access to the same humic substance samples.[24] They, therefore, defined a set of standard samples (Suwannee River, Elliot Soil, Pahokee Peat and Leonardite),[25] used in a multitude of studies.[23, 26–31] In this work we emphasize the Leonardite Humic Acid (LHA) standard sample which is extracted from lignite, a low-grade coal obtained from the Gascoyne Mine in Bowman County, North Dakota, USA.[25]

1.1.4 Chemical Properties of HS

Although soil science has been studied for centuries, there is little knowledge on the chemical properties and the molecular structure of HS. This is explained by the extreme complexity of humic substances[4, 17] as well as their previously mentioned extraction method. However, the introduction of IHSS standard samples in combination with the emergence of nuclear magnetic resonance (NMR) spectroscopy methods shed light on SOM functional group content. The most abundant groups in HS are carboxyl, aromatic, heteroaliphatic and aliphatic groups.[32] Due to a high content of carboxyl groups in the HS, molecules carry a negative charge at pH 7. By employing a comprehensive multiphase NMR technique Masoom et al. concluded that the NMR spectrum of soil is similar to one from a mixture of carbohydrates, proteins, lignin and lipids.[33] In addition to NMR other methods, for example MS,[34] FTIR,[35] X-ray[36] or DSC,[37] have been employed to study HS. Nevertheless, due to the fact that no HS molecule is identical to another, it is impossible to resolve a complete structure of humic substances.

1.1.5 VSOMM and VSOMM2

To understand SOM on a molecular level, a tool called Vienna Soil Organic Matter Modeler (VSOMM) was developed, which can be easily accessed over the internet (http://somm.boku.ac.at/).[23] It uses small organic fragments which are called building blocks to create molecular models. These building blocks contain different amounts of carbon, oxygen, nitrogen and sulfur and were designed to have a high diversity in the number and kind of functional groups. Following functional groups occur in one or more building blocks: carbonyl, carboxyl, O-aryl, aryl, di-O-alkyl, O-alkyl, methoxyl and alkyl. Experimentally obtained data can be set as input parameters, such as pH the carbon and nitrogen fraction as well as the fractions of previously mentioned functional groups. Additionally, the total number of building blocks and the number of building blocks per molecule can be set. There is an option to choose between Na^+, Ca^{2+} and Mg^{2+} as counter ions. The VSOMM then automatically selects building blocks and combines them to molecules which subsequently form a model system which matches the experimental data as closely as possible. The amount of building blocks per SOM molecule can be defined in the input[23]. Therefore, a large number of different soil samples can be studied using VSOMM by entering respective experimental data. Sev-

eral studies have successfully used VSOMM so far and experimental properties were reproduced.[26, 27, 38–40]

Currently the second generation of VSOMM, VSOMM2, was developed by Escalona et al.(to be submitted). Several improvements were made to the models. Firstly, new building blocks were added, which were derived from MS data to represent a greater chemical variety. Additionally, existing building blocks were altered to represent more realistic chemical structures. Moreover, a genetic algorithm was introduced to increase the chemical and geometric diversity of the models. Finally, the primary objective of the algorithm was changed from fulfilling the atomic fractions to fulfilling the functional group fractions.

1.2 Proteins in SOM

The presence of nitrogen in soil is explained by the encapsulation of proteins by organic matter.[41, 42] Originally, studies suggested that proteins are encapsulated by hydrophobic domains of humic acids, however, Tomaszewski et al. (2011) experientally observed that electrostatic forces drive the encapsulation of positively charged proteins with HS at pH 5 to 8. Moreover, they suggested that attractive forces, likely supported by additional hydrophobic forces, can overcompensate the electrostatic repulsion of negatively charged protein patches.[28] The association of proteins into parts of SOM/ HS increases the protein stability.[28, 41] Additionally, decomposition of proteins is significantly slowed down when mixed with humic polymers.[43] However, the encapsulation of protein can lead to structural changes or blocking of active sites, reducing enzyme activity. For example, a linear relationship between SOM content and decreased enzymatic activity of laccase and peroxidase was found.[44]

There is a scientific effort to understand the interactions of proteins in SOM, which includes the mycorrhizal symbiosis, degradation of cellulose and protein toxicity in soil. Proteins that can be found in soil and, therefore, also in SOM can be split up in two categories: proteins that are excreted into the soil on purpose to fulfill specific functions and proteins that enter soil after cell death.[45]

Proteins that are secreted into soil can have different functions, for example, defense, signalling, symbiosis and nutrient uptake.[46] There is also a class of proteins, glomalins, that are excreted by plants and which primary function is to go into soil and alter its structure. Glomalins are secreted by root hyphae in high concentrations and act as glue to hold soil particles together.[47] Another important class of secreted proteins are lytic polysaccharide monooxygenases (LPMOs) which are found in many fungi and bacteria. These proteins are copper-dependent enzymes which cleave glycosidic bonds of cellulose and chitin.[48] The active site comprises of two conserved histidine residues which coordinate a copper ion.[49] LPMOs are important players in the decay mechanisms of microbial degradation.[50]

The pool of proteins that come into contact with soil after cell death is enormous. The following two paragraphs give examples of proteins and what kind of impact they could have on soil properties and the environment:

The first example is the Cry toxin family. It is originally found in *Bacillus thuringiensis* and, therefore, often referred to as *Bt* toxin. Due to its known insecticidal properties, the Cry toxin is popular for introducing new resistances in genetically modified crops. In 2011 around 40% of all genetically modified crops expressed one or multiple copies of the protein.[51] The intended purpose of the toxin is to be present in the plant cells and to act as insects ingest plant material.[52] Nevertheless, there are several ways how Cry toxins can enter agricultural soils, most notably by decaying plant material and by excretion by plant roots.[53] It was found the the Cry toxin is strongly absorbed by SOM.[29] Additionally, the Cry proteins retain full insecticidal activity in SOM.[30] The crystal structure of the Cry1A toxin was resolved in 2018 showing a large three domain protein containing 591 amino acids.[54]

A second example for the relevance of studying proteins in SOM is the case of prion proteins. Prions are infectious misfolded proteins that lead to neurodegenerative diseases in various animals, for example, Creutzfeldt-Jakob disease in humans, scrapie in sheep and goat, chronic wasting disease in cervids (e.g. deer) and bovine spongioform encephalopathy in cattle.[55] The prion proteins can exist in two forms: the correctly folded form (PrP^{C}) and the infectious misfolded form (PrP^{Sc}). Upon death of an infected animal, misfolded prion proteins can shed into the soil and various soil compounds can interact with the protein. It has been shown that murine prion proteins are absorbed by humic substances, however, they stay structurally intact. The humic substances-prion protein complexes can, therefore, act as a natural reservoir for these infectious particles.[31, 56] Conversely, a study found that the humic acid - elk prion protein mixtures showed reduced infectivity when administered to transgenic mice.[57] Although the structure of PrP^{C} has been resolved, the structure of the misfolded prion protein (PrP^{Sc}) remains unknown, even though there has been significant effort also by means of molecular dynamics simulation to create models.[58]

There are several possible applications for proteins in soil, most notably the use of enzymes for bioremediation. Soil has a great potential of absorbing and storing pollutants for long times and removing said pollutants can pose a big challenge.[59] In 2008 scientists estimated a total of 80,000 polluted sites in Austria alone.[60] On the other hand, enzymes from bacteria and fungi have evolved to degrade many organic compounds. Enzymes that have been linked to detoxification of organic compounds are, for example, oxidoreductases (for phenolic compounds), oxygenases (for chlorinated aliphatics), monooxygenases (for wide range of alkanes to fatty acids), dioxygenases (for aromatic compunds), laccases (for phenolic and aromatice substrates) and peroxidases (for lignin) and many more.[61] A solution for cleaning of contaminated soils, therefore, could be the administration of said enzymes either in living organisms or directly in a cell-free manner onto the soil. However, many of potential enzymes only have been tested under

laboratory conditions and their activity may vary significantly when put into soil.[60, 62] Further understanding of soil protein interactions could facilitate the engineering of better enzymes for bioremediation purposes.

2. Aim

The properties of soil particularly influence agriculture and consequently humanity. Therefore, studying soil has a long tradition, however, investigation was mainly done by observation of macroscopic properties, meanwhile molecular level explanations remained unknown. Many effects cannot be explained on a macroscopic scale which lead to an increased interest in the molecular description of soil. This, however, is hard because of the vast complexity of soil components. Classic chemical analysis methods, for example NMR, are not able to recognize independent chemical fragments, just gross averages which do not reflect the molecular nature of soil. In recent years atomistic models were developed to shed light on the molecular interactions of soil components, for example the Vienna soil organic matter modeler (VSOMM). Molecular dynamics permit the simulation of these models and also proteins which can answer several unsolved questions regarding the molecular interactions of proteins with SOM and solving them could give more clues on how proteins behave upon contact with soil and on their respective environmental impact.

The aim of this work was to investigate the interactions between proteins and SOM on a molecular level. Two proteins that act as a reference were selected based on their simplicity and simulated in different SOM model conditions by means of molecular dynamics. Subsequently we analysed non-bonded interaction energies, especially focusing on the difference between electrostatic and van der Waals forces, as well as the formation of hydrogen bonds. Moreover, the change of protein stability and their secondary structure in the model systems was monitored. Additionally, the spacial arrangement of solvent molecules and functional groups around the proteins were studied. In a next step, we computed differences in free energies of mutation of amino acids between water and SOM systems. With the results we developed a scoring function that gives an indication about the interaction strength between proteins and SOM. Finally, several extracellular and cytoplasmic proteins were scored and results analysed.

3. Theory of Molecular Dynamics

Molecular dynamics simulations are a powerful tool to investigate molecule scale processes and interactions, especially if said processes are inaccessible to experiment. Molecular dynamics bases in the principles of classical and statistical mechanics as well as thermodynamics. The following sections provide an overview over the most important principles and how they are applied.[63]

The choice of the level of detail of a computer simulation is important since it affects accuracy as well as efficiency. Realistically, this choice is always a compromise. Matter consists out of atoms which themselves can further be subdivided into electrons and nuclei. Whereas it is possible to describe trajectories of nuclei with classical mechanics, the nature of electrons is different, quantum mechanics are necessary. The Born-Oppenheimer approximation implies that the movement of nuclei and electrons can be decoupled since the mass of an electron is several magnitudes lower than the mass of a nucleus.[64].Biomolecular systems are usually multi-atomic and complex systems and their interactions are governed by weak non-bonded energies. For molecular dynamics simulations of such systems the Born-Oppenheimer approximation is reasonable, and therefore, only atoms are considered explicitly (points in space). In addition, several more tricks can be applied to reduce the number of degrees of freedom, for example the summation of multiple atoms in coarse grained particles to further increase the efficiency of simulations.

3.1 Force Fields

Force fields define the potential energy landscape experienced by atoms in a molecular dynamics simulation as a function of the atom's position $\mathbf{r}$. Several different force fields are currently developed including AMBER, CHARMM and GROMOS.[65–67] Since they were used in this work, the GROMOS force fields 54A7 and 54A8 are subsequently described in more detail.[67, 68] The GROMOS force fields are united atom force fields; CH_n (where n is 1,2,3 or 4) groups are simulated as one particle to further simplify and accelerate the simulation. The potential energy is defined as the sum of the energy terms for bonded and non-bonded interactions (Equation 3.1).

$$\mathcal{V}^{(potential)}(\mathbf{r}) = \mathcal{V}^{(bonded)}(\mathbf{r}) + \mathcal{V}^{(non-bonded)}(\mathbf{r}) \tag{3.1}$$

Bonded interactions are further subdivided into bond stretching, angle bending, torsional dihedral angle and improper dihedral angle terms (Equation 3.2). Bond stretching is defined by a quartic potential, angle bending by a cosine harmonic potential and improper dihedral angles by a simple harmonic potential with their respective force constants (K_i^b, K_i^θ, K_i^ξ) and reference values (b_i^0, θ_i^0, ξ_i^0). The torsional dihedral angle term is approximated by a cosine series expansion with the multiplicity (m_i), the phase shift (δ_i) and the force constant (K_i^ϕ) as parameters.

$$\begin{aligned}\mathcal{V}^{(bonded)}(\mathbf{r}) = &\sum_i^{N_{bond}} \frac{1}{4} K_i^b \left[b_i(\mathbf{r})^2 - (b_i^0)^2\right]^2 \\ &+ \sum_i^{N_{angle}} \frac{1}{2} K_i^\theta \left[\cos\theta_i(\mathbf{r}) - \cos\theta_i^0\right]^2 \\ &+ \sum_i^{N_{torsion}} K_i^\phi \left[1 + \cos\left(m_i\phi_i(\mathbf{r}) + \delta_i\right)\right] \\ &+ \sum_i^{N_{improper}} \frac{1}{2} K_i^\xi \left[\xi_i(\mathbf{r}) - \xi_i^0\right]^2\end{aligned} \tag{3.2}$$

Non-bonded interactions are subdivided in a van der Waals term for the interaction of uncharged atoms and a Coulomb term for electrostatic interactions (Equations 3.3 to 3.5). A visual representation of the non-bonded interaction energy terms is shown in Figure 3.1.

$$\mathcal{V}^{non-bonded}(\mathbf{r}) = \mathcal{V}^{vdw} + \mathcal{V}^{coul} \tag{3.3}$$

$$\mathcal{V}^{(vdw)}(\mathbf{r}) = \sum_{pairs\ i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{3.4}$$

$$\mathcal{V}^{(coul)}(\mathbf{r}) = \sum_{pairs\ i<j} \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_i q_j}{r_{ij}} \tag{3.5}$$

In order to solve the previously mentioned equations of potential energy, multiple parameters have to be defined. The GROMOS force fields 54A7/54A8 were parameterized to reproduce free energies of solvation for a variety of functional groups in water and organic solvents. Force fields 54A7 and 54A8 only have minor differences, most notably the reparametrization of charged amino acid side chains. Since all sets of parameters have been designed to work as generally as possible the 54A7/54A8 force fields arguably allow realistic modeling of organic matter of soil.[23]

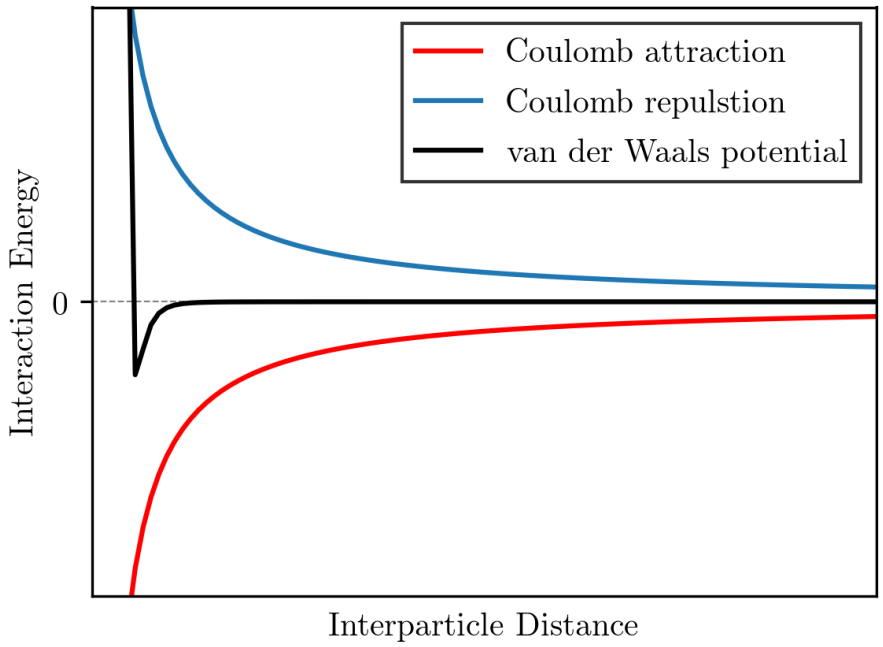

Figure 3.1: Comparison of van der Waals potential and Coulombic terms for two particles with equal and opposite charges (repulsion and attraction respectively).

3.2 Classical Mechanics

3.2.1 Laws of Motion

The origin of classical mechanics were set by Isaac Newton in the 17$^{\text{th}}$ century. He postulated three simple laws to describe the motion of objects.

(1) An object will either rest or travel along a straight line with a constant velocity $\mathbf{v}$ unless force is exerted upon it.

(2) The force $\mathbf{F}$ that acts on an object is equals to the object's mass m multiplied by its acceleration $\mathbf{a}$.

$$\mathbf{F} = m \times \mathbf{a} = m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} \tag{3.6}$$

(3) If a force is exerted from object A to B the opposite force is exerted from object B to A.

$$\mathbf{F}_{BA} = -\mathbf{F}_{AB} \tag{3.7}$$

Even though Newton had no atoms in mind when he postulated his laws they can be used to describe the movement of such. The force that acts on an atom depends on the derivative of the potential energy it experiences as a function of its position (Equation 3.8).

$$\mathbf{F}_i = -\frac{\partial \mathcal{V}(\mathbf{r})}{\partial \mathbf{r}} \tag{3.8}$$

3.2.2 Leap-Frog Algorithm

For molecular dynamics simulations several integration algorithms have been proposed based on Newton's laws of motion. Most algorithms employ a step-wise integration scheme with an equidistant time step size of Δt. A popular example of such an algorithm is the leap-frog algorithm which is here described in more detail.[69, 70] It is called leap-frog algorithm because with Equations 3.9 and 3.10 it is possible to alternatingly calculate the velocities and positions of all particles in the system.

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}(t - \frac{1}{2}\Delta t) + \Delta t \times \mathbf{a}(t) + O[(\Delta t)^3] \tag{3.9}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \times \mathbf{v}(t + \frac{1}{2}\Delta t) + O[(\Delta t)^3] \tag{3.10}$$

Only the first two terms of the Taylor expansion are calculated by the algorithm, further terms (here denoted as O) are assumed to be small enough to be ignored. It is important to note that the total energy of a system cannot be computed at the same time, since the kinetic energy is dependent on the velocities, whereas the potential energy is a function of the positions of atoms. Another important consideration to make is the size of Δt since it has a direct influence on the simulation time. As a rule of thumb the time step should be approximately a tenth of the fastest motion of a simulation which is the vibration of atomic bonds. However, bond vibrations are of no interest for molecular dynamics and can, therefore, be constrained by algorithms such as LINCS and SHAKE.[71, 72] This allows a maximum time step of $\Delta t = 2$ fs which is usually employed for maximum computational efficiency.

3.3 Statistical Thermodynamics

3.3.1 Thermodynamic Definitions

In thermodynamics there is an important distinction between systems and their surroundings. Usually we seek to describe a system, either by absolute values or by relative differences. A thermodynamic system can be defined by six central state variables, three of which are extensive and three that are intensive. Extensive variables are proportional

Table 3.1: Thermodynamic variables and the properties which they discribe.

Described Properties	Extensive Variable		Intensive Variable	
mechanical	volume	V	pressure	P
thermal	energy	E	temperature	T
chemical	number of particles	N	chemical potential	μ

to the system size whereas intensive variables are not. A pair of state variables as depicted in Table 3.1 can describe one property of a system. Both state variables of a pair can never be kept constant at the same time if their described property changes. As a consequence a thermodynamic system with changing properties can be described by a maximum of three constant state variables.

Initially the laws of thermodynamics have been designed to describe macroscopic features. For example the heat transfer and work generated or consumed by a system in regard to its surroundings. There are three central laws of thermodynamics.[63]

(1) The change of the inner energy (U) of a system is depending on the heat absorbed by the system (Q) and the work performed on it (W).

$$\mathrm{d}U = \delta W + \delta Q \tag{3.11}$$

(2) For any thermodynamic transformation the total entropy (S) of the universe must either increase or remain the same.

$$\mathrm{d}S_{tot} \geq 0 \tag{3.12}$$

(3) The entropy of a thermodynamic system in a temperature of 0 K is 0 J/(mol K).

The first two laws of thermodynamics can be combined to the equation for the characteristic state function. The characteristic state function of systems with constant S,V and N is give in Equation 3.13.

$$\mathrm{d}U = -P\mathrm{d}V + \sum_{i=1}^{N} \mu_i \mathrm{d}n_i + T\mathrm{d}S \tag{3.13}$$

The characteristic state functions of thermodynamic systems change depending on which state variables are constant. The most important examples are the definition of enthalphy H (constant SPN), the Helmholtz free energy A (constant NVT) and the Gibbs free energy G (constant NPT) (Equations 3.14 to 3.16).

$$\mathrm{d}H = V\mathrm{d}P + \sum_{i=1}^{N} \mu_i \mathrm{d}n_i + T\mathrm{d}S \tag{3.14}$$

$$\mathrm{d}A = -P\mathrm{d}V + \sum_{i=1}^{N} \mu_i \mathrm{d}n_i - S\mathrm{d}T \tag{3.15}$$

$$\mathrm{d}G = V\mathrm{d}P + \sum_{i=1}^{N} \mu_i \mathrm{d}n_i - S\mathrm{d}T \tag{3.16}$$

Thermodynamic systems can have various interactions with their surroundings. Either the system is open, heat, work and particle transfer are possible, the system is closed, so only heat transfer can happen or the system is isolated, there is no interaction with its surrounding at all. The universe is an isolated system, the number of particles N, the volume V and the energy E are constant. This kind of systems are called micro-canonical ensembles. Other important ensembles are the canonical ensemble (constant NVT), the grand canonical ensemble (constant µVT), the isothermal-isobaric ensemble (constant NPT) and the isenthalpic-isobaric ensemble (constant NPH).

3.3.2 Statistical Mechanics

Matter consists of microscopic constituents and, therefore, it must be possible to describe the laws of thermodynamics based on microscopic mechanical laws.[63] This means that all macroscopic properties of thermodynamic systems can be described as a result of microscopic states of all particles in the same system which is done by the field of study of statistical mechanics. Assuming there are N particles in a system which can be described by their positions $(\mathbf{q}_1, ..., \mathbf{q}_\mathrm{N})$ and momenta $(\mathbf{p}_1, ..., \mathbf{p}_\mathrm{N})$. One frame in this systems constitutes one microstate, whereas the sum of all microstates is regarded as the phase space. A priori every microstate of an NVT ensemble is equally likely (Equation 3.17).[73]

$$\sum_{k}^{\infty} P(k) = \sum_{k}^{\infty} Q^{-1} e^{-\beta E_k} = 1 \tag{3.17}$$

$$P_{NVT} = Q^{-1} e^{-\beta \mathcal{H}(\mathbf{q},\mathbf{p})} \tag{3.18}$$

Where Q is the partition function, $\mathcal{H}$ is the Hamiltonian and β is the reciprocal of Boltzmann constant multiplied by the temperature. The partition function is a dimensionless number that defines how many microstates are accessible. For the NVT ensemble it is defined as in Equation 3.19.

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \iint e^{-\beta \mathcal{H}(\mathbf{q},\mathbf{p})} \mathrm{d}q \; \mathrm{d}p \tag{3.19}$$

Where h is Planck constant and N is the total numbers of particles. The Hamiltonian, $\mathcal{H}$, describes the total energy of the system as a function of positions ($\mathbf{q}$) and momenta ($\mathbf{p}$) of its particles. The Hamiltonian, $\mathcal{H}$, can be written as the sum of kinetic and potential energy (Equation 3.20).[63]

$$\mathcal{H}(\mathbf{q},\mathbf{p}) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + \mathcal{V}(\mathbf{q}) \tag{3.20}$$

The probability, P, for the isothermal-isobaric ensemble (NPT) is calculated with Equation 3.21, where $\mathcal{Z}$ is the partition function as defined as in Equation 3.22.

$$P_{NPT}(p,q) = \mathcal{Z}^{-1} e^{-\beta[\mathcal{H}(\mathbf{q},\mathbf{p})+PV]} \tag{3.21}$$

$$\mathcal{Z}_{NPT} = \frac{1}{N!}\frac{1}{h^{3N}} \iiint e^{-\beta[\mathcal{H}(\mathbf{q},\mathbf{p})+PV]} \mathrm{d}q \; \mathrm{d}p \; \mathrm{d}V \tag{3.22}$$

Ludwig Boltzmann postulated that the thermodynamic entropy is a function of the partition function.

$$S = -k_B \sum_{k}^{\infty} P(k) \ln P(k) = \frac{U}{T} + k_B \ln Q \tag{3.23}$$

Where U is the inner energy of a system, T the temperature and k_B the Boltzmann constant.

3.4 Free Energy Calculations

Free energy is an important thermodynamic property. It describes if chemical reactions happen spontaneously (negative values) or not (positive values). The calculation of the free energy, therefore, can elucidate how strongly molecules react with each other. The free energy is denoted either as Helmholtz free energy, ΔA, for NVT ensembles or Gibbs free energy, ΔG, for NPT ensembles. In statistical thermodynamics the free energy can be calculated using the partition function as following:

$$A = -k_B T \times \ln Q_{NVT} \tag{3.24}$$

$$G = -k_B T \times \ln \mathcal{Z}_{NPT} \tag{3.25}$$

The absolute free energy of a systems is not only practically impossible to calculate, since the whole phase space needs to be sampled, it additionally has no chemical meaning. However, the free energy differences between two states 1 and 2 are of much more interest. They can be calculated as following.

$$\Delta A_{21} = A_2 - A_1 = -k_B T \times \ln \frac{Q_{2(NVT)}}{Q_{1(NVT)}} \tag{3.26}$$

$$\Delta G_{21} = G_2 - G_1 = -k_B T \times \ln \frac{\mathcal{Z}_{2(NPT)}}{\mathcal{Z}_{1(NPT)}} \tag{3.27}$$

Free energy calculation with molecular dynamics uses the fact that the free energy is a state function, it is path independent. That means that the difference in free energy from two states, ΔG, is always constant, regardless of the path going from one state to another even if there are unphysical intermediates. As a consequence of this property a thermodynamic cycle as drawn as in Figure 3.2 can be used[74].

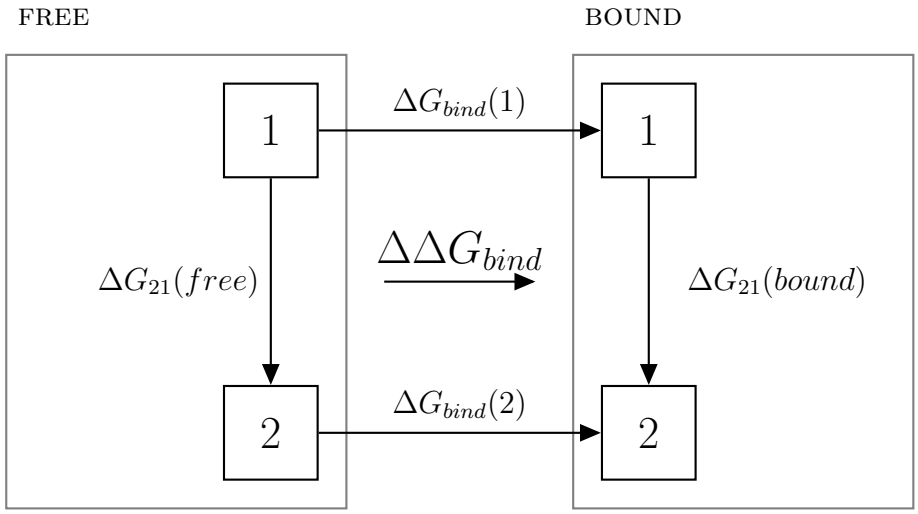

Figure 3.2: A thermodynamic cycle that can be employed to calculate free energy differences of binding. The left side of the cycle represents a ligand in an unbound (free) state, whereas the right side represents a ligand bound to a protein.

It is relatively easy to compute the free energy of changing molecule 1 to molecule 2 in bound as well as unbound state ($\Delta G_{21}(free)$ and $\Delta G_{21}(bound)$ respectively) compared to the computation of the binding free energies directly ($\Delta G_{bind}(1)$ and $\Delta G_{bind}(2)$). By closing the thermodynamic cycle the relative free energy changes of binding of molecule 1 and 2 ($\Delta\Delta G$) can be calculated using Equation 3.28.

$$\Delta\Delta G_{bind} = \Delta G_{bind}(2) - \Delta G_{bind}(1) = \Delta G_{21}(bound) - \Delta G_{21}(free) \tag{3.28}$$

Zwanzig (1954) proposed the following energy perturbation formula:[75]

$$\Delta A_{21} = A_2 - A_1 = -k_B T \times \ln \frac{\mathcal{Z}_{2(NVT)}}{\mathcal{Z}_{1(NVT)}} \tag{3.29}$$

$$= -k_B T \times \ln \langle e^{-(\mathcal{H}_2 - \mathcal{H}_1)/k_B T} \rangle_1 \tag{3.30}$$

The ergodic hypothesis implies that the molecular properties observed in a simulation over a period of time approach experimentally measured ensemble averages.[76] Therefore, the ensemble average of Hamiltonians, $\langle \mathcal{H} \rangle$, of molecular dynamics simulations can be used to calculated free energy differences between two states. This, however, only works well if both conformational ensembles have enough overlap, which is usually not the case. However, tricks can be applied to enhance sampling. On method is the use of several λ states, which are usually alchemical states that lie in between two end states (Figure 3.3 A). They are defined in a way that sufficient conformational overlap is reached and the free energies of going from each step to the next can be summed up. Figure 3.3 B shows a second method which employs the definition of a reference state (R), which is designed to have enough conformational overlap with all end states. This method is especially advantageous when more than two end states are investigated, since only one molecular dynamics simulation needs to be done.

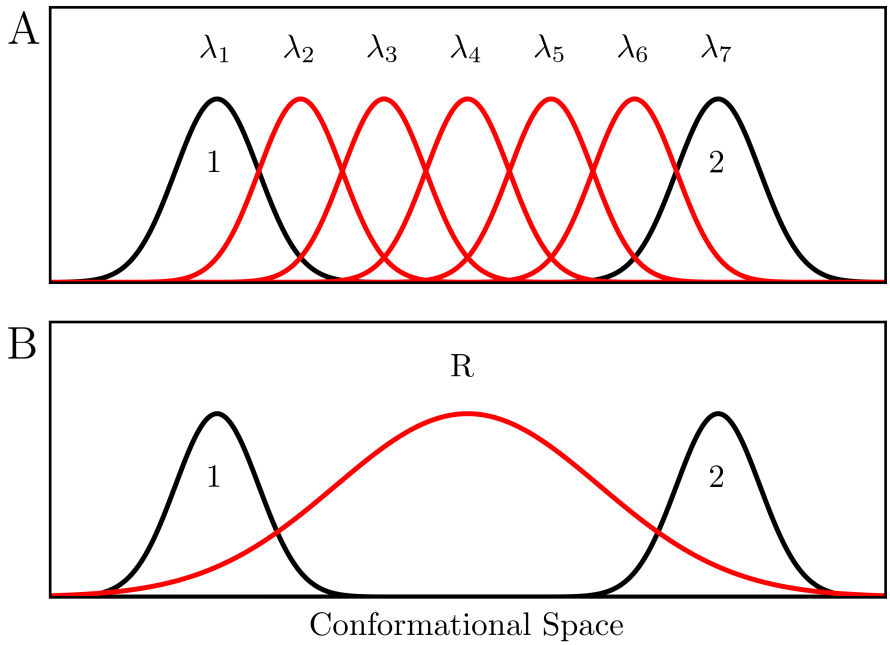

Figure 3.3: Methods for overcoming bad sampling in free energy calculations: the use of (A) multiple λ states and (B) a single reference state.

Several free energy calculation methods have been developed for molecular dynamics, for example LIE, LRA, WHAM, (extended) TI, OSP and TPF. In this work one-step perturbation (OSP) and third power fitting (TPF) were used and are, therefore, described in more detail in the following sections.

3.4.1 One-Step Perturbation

The idea behind one-step perturbation is to define a reference state that has enough overlap between its conformational ensemble and the ensamble of a multitude of other end states (Figure 3.3 B).[77] Therefore, only one molecular dynamics simulation of the reference state needs to be done for this method and Zwanzig's equation (Equation 3.30) is directly applicable. The reference state can then be used as a proxy to compare free energy differences of several end states as shown in Figure 3.4 and Equation 3.31.

$$\Delta A_{1\to 2} = \Delta A_{R\to 2} - \Delta A_{R\to 1} \tag{3.31}$$

The reference state usually consist out of dummy or soft atoms which are both unphysical particles. A dummy atom is an atom where all interaction parameters are set to zero, it is, therefore, not seen by other atoms in the simulation. Soft atoms still have some interaction with their adjacent particles, however, their force field parameters have been modified, for example it is possible for other atoms to pass through them. Soft atoms are used to enhance sampling.[78] One disadvantage of the one-step perturbation method, however, is the fact that it performs best when no polarity changes happens. High polarity changes unfortunately result in poor overlaps of conformational ensembles of reference and end states.[79]

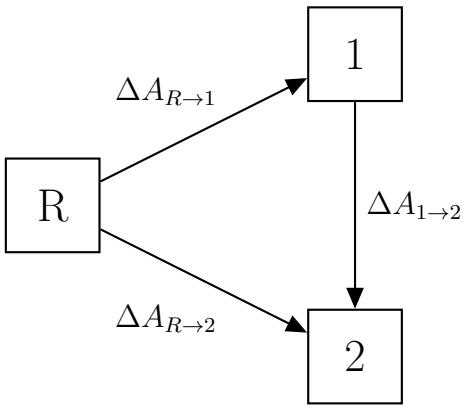

Figure 3.4: Thermodynamic cycle which can be employed to calculate free energy differences between physical states (1,2) by going over a non-physical reference state R.

3.4.2 Third Power Fitting

The third power fitting method (TPF) is a relatively new approach proposed in 2011.[80] TPF is a method that can be used for calculating free energies of uncharging of molecules

and atoms. It relies on two end state MD simulations of the perturbed molecule: one state with all atoms with full or partial charges and one without charges. The principle of TPF is similar to the principle of the linear response approximation (LRA) method. It assumes an easy mathematical relationship of free energy change going from the charged to the uncharged state. For, LRA this relationship is assumed linear which gives unsatisfactory results. By introducing a third-order polynomial, however, as done by TPF, the precision of the results can be improved without the need to perform additional molecular dynamics simulations.[80] A visualization of both methods is given in Figure 3.5.

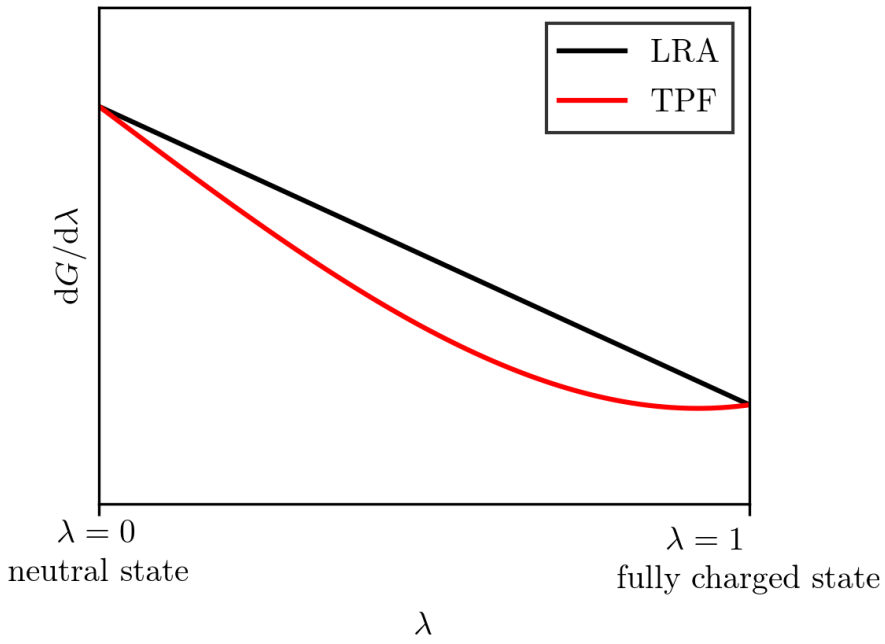

Figure 3.5: Comparison of the LRA and TPF method for calculation of free energies of charging.

Since intermediate states between the charged and uncharged state are referred to as λ states (Figure 3.3 A) the free energy change going from state 1 to state 2 be can be written as $\frac{dG}{d\lambda}$. By integration of this curve from state 1 to state 2 the free energy of uncharging is calculated as described in Equations 3.32 and 3.33. The parameters a, b, c and d of Equation 3.33 are obtained by measuring $\frac{dG}{d\lambda}$ and $\frac{d^2G}{d^2\lambda}$ at $\lambda = 0$ and $\lambda = 1$ according to Equations 3.34 and 3.35 where V_{ls}^{el} is the electrostatic interaction energy between a molecule and its surroundings and the angular brackets $\langle\ \rangle$ indicate an ensemble property.

$$\Delta G^{TPF}_{A\to B} = \int_0^1 \frac{dG}{d\lambda}\, d\lambda = \int_0^1 f(\lambda) d\lambda \tag{3.32}$$

$$f(\lambda) = a\lambda^3 + b\lambda^2 + c\lambda + d \tag{3.33}$$

$$\left.\frac{dG}{d\lambda}\right|_\lambda = \langle V^{el}_{ls}\rangle_\lambda \tag{3.34}$$

$$\left.\frac{d^2G}{d\lambda}\right|_\lambda = \frac{1}{k_B T} \times (\langle V^{el}_{ls}\rangle^2_\lambda - \langle (V^{el}_{ls})^2\rangle_\lambda) \tag{3.35}$$

The free energy calculations of uncharging are artifacted if the total charge of a charge group changes, for example by calculating the free energy of going from a protonated lysine to a neutral reference state. In order to correct for these artifacts a correction free energy (ΔG_{cor}) can be added to the raw result (ΔG_{raw}) (Equation 3.36). The correction free energy consist out of three terms: ΔG_{pol}, for spurious solvent polarization, ΔG_{dsm}, for the impracticality of calculating the zero of the potential under periodic boundary conditions using discrete solvent molecules and ΔG_{dir}, for artifacted direct interactions between two molecules (Equation 3.37).[81]

$$\Delta G = \Delta G_{raw} + \Delta G_{cor} \tag{3.36}$$

$$\Delta G_{cor} = \Delta G_{pol} + \Delta G_{dir} + \Delta G_{dsm} \tag{3.37}$$

3.4.3 Combining free energy methods

To overcome disadvantages of single free energy methods they can be combined. The combination of OSP and TPF has been tested to perform well[80, 82] and, therefore, was applied in this work. With TPF the free energy of uncharging of molecules can be calculated efficiently. Subsequently, the uncharged molecules can be perturbed into a reference state by OSP. Both free energy terms are added to get the free energy of going from charged state to a reference state according to Equation 3.38 where Q, N and R are the charged, neutral and reference state respectively.

$$\Delta G_{Q\to R} = \Delta G^{TPF}_{Q\to N} + \Delta G^{OSP}_{N\to R} \tag{3.38}$$

A thermodynamic cycle similar to what is shown in Figure 3.4 can subsequently be used to calculate free energy differences between two charged states.

4. Methodology

4.1 Selection of Proteins

In order to study the interactions between protein and humic substances, proteins were selected from a subset of well known and previously characterized proteins.[83, 84] The selection considered following criteria: (1) small size, to ensure short simulation times, (2) different protein net charges, to study possible differences of interactions between the protein and the strongly negatively charged humic acids, (3) different secondary structures, to investigate different protein structures. By using the mentioned criteria two proteins were selected as reference proteins: the villin headpiece and the EGF domain of spitz (Figure 4.1).

The thermostable subdomain from chicken (*Gallus gallus*) villin, also called villin headpiece, is a 36 amino acids (389 atoms) long protein consisting of three α helices according to its protein database entry (PDB code: 1VII).[85] It contains a number of positively charged amino acids which leads to 2+ net charge at pH 7. For simplicity reasons the villin headpice is just referred to as villin in this thesis.

The second reference protein was the EGF domain of spitz from *Drosophila melanogaster*.[86] Although the EGF domain of spitz has a similar size as villin (50 amino acids, 524 atoms) their properties differ. The protein database structure (PDB code: 3CA7) resolves not only an α helix but also two antiparallel β sheets.[87] Additionally this protein has three S–S cystein bridges and a 2− net charge at pH 7. For simplicity reasons, again, the EGF domain of spitz is just referred to as spitz in this thesis.

4.2 Experimental Design

In this master thesis multiple molecular dynamics simulations were performed to understand the structure and dynamics of proteins in humic substances and to get insights on the forces that command this interaction. The work was split up into two parts. The first part comprised molecular dynamics simulations of the two reference proteins in simple solvent environments as well as more complex HS models. The intended purpose of simple solvent systems was to investigate possible interactions between individual functional groups of SOM and the proteins. Additionally, only solvent simulations for

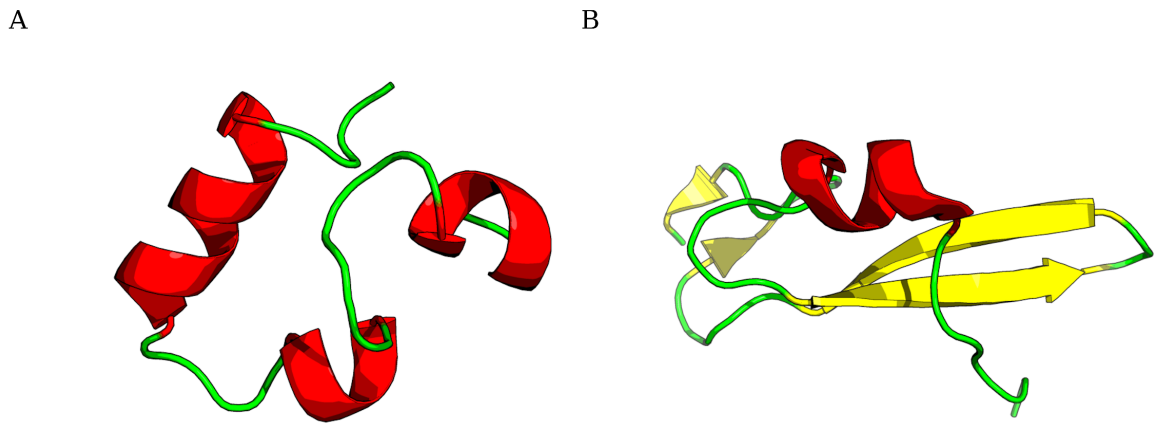

Figure 4.1: Protein database structure of (A) villin (1VII) and (B) spitz (3CA7). The proteins are colored according to their secondary structure: α helix = red, β sheet = yellow, loop = green. The figure was rendered with PyMol[88].

the estimation of salting-in/out effects were conducted. To proceed we studied our reference proteins by putting them into complex leonardite humic acid systems created by the Vienna Soil Organic Matter Modeler 2 (VSOMM2). In the second part we calculated the free energy of mutation of amino acid side chains to investigate which amino acids drive protein – humic acids interactions. With these results we subsequently developed a scoring method, SOMscore, to rank proteins according to their potential interaction strength with humic substances. Several extracellular proteins were scored and compared.

4.2.1 Simple Solvents

4.2.1.1 Simulation Setup

Part one of this work comprises several molecular dynamics simulations of small solvent molecule systems (Table 4.1). To start, eight replicates of each reference protein were simulated in water. Then, four replicates were simulated in five different simple solvent molecule systems which represented properties of humic substances. The simple solvent systems were: (1) calcium chloride ($CaCl_2$), to see high salt concentration effects, (2) calcium acetate ($CaAc_2$), to see how organic acids behave, (3) calcium benzoate ($CaBenz_2$), to see changes by adding aromatic groups, (4) calcium acetate with benzene ($CaAc_2$ + Bz), to investigate how interactions change when the carboxyl and aryl functional groups are not part of the same molecule and, finally, (5) realistic concentration (Real. Conc.), to see what happens if previously used small molecules are mixed to represent realistic leonardite humic acid. In Figure 4.2 the chemical structure of the multiatomic solvent molecules is shown. Solute topologies were parametrised using the GROMOS united atom force field 54A7.[67] For the creation of simple solvent molecule topology files, building block templates for benzoate and acetate were downloaded from

the Automated Topology Builder (ATB) and Repository[89, 90] and subsequently manually revised using the `bb_editor`. The building block of benzoate was used as a template to create the topology file of benzene. All created building block files are listed in the appendix Figure 8.1 and Listings 8.1, 8.2 and 8.3.

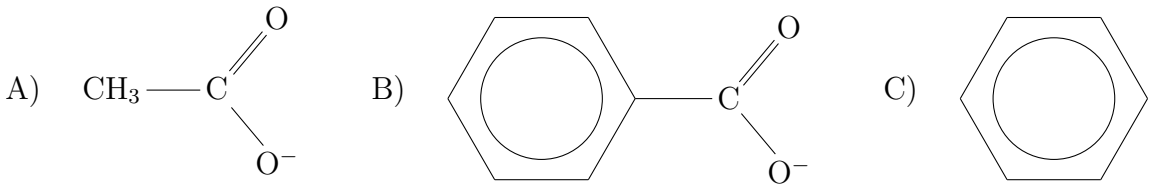

Figure 4.2: Molecular structures of acetate (A), benzoate (B) and benzene (C).

4.2.1.2 Molecular Dynamcis

Molecular dynamics simulations were performed with the GROMOS11 simulation package.[91] All systems were simulated in rectangular periodic boxes using time steps of 2 fs. The system charges were neutralized with Cl^- and Ca^{2+} counter ions for villin and spitz respectively for water conditions. All other simulations were normalized to approximately 1 mol/L Ca^{2+} ion concentration. The exact concentrations of ions and carbon fractions of all simulated systems are depicted in Table 4.1. The temperature and pressure were coupled to 300 K and 1 bar using weak coupling[92] with a coupling constant of 0.1 ps and 0.5 ps respectively. The isothermal compressability was set to 4.575×10^{-4} mol nm^3/kJ. The SPC water model[93] was used for water molecules. Bond length constraints for SPC waters were imposed by the SETTLE algorithm,[94] for all other molecules SHAKE[72] was used. The non-bonded interactions were estimated using a twin range cutoff based on a pairlist, for short-range interactions up to 0.8 nm the interactions were computed every time step and for intermediate-range up to 1.4 nm the interactions and the pairlist were updated every 5 time steps. Coulomb interactions were calculated with an additional reaction-field with a relative dielectric permittivity of 61. The system equilibration was performed in five steps. The first step started at 60 K where velocities were randomly assigned according to a Maxwell – Boltzmann distribution. Solute atoms were positionally restrained with a force constant of 2.5 $\times$ 10^4 kJ/mol. The molecular dynamics simulation was performed for 20 ps. In the four following steps the temperature was increased by 60 K and the force constant decreased by a factor of 10 respectively leading to a fifth equilibration step at 300 K and without position restraining. The molecular dynamics production run was subsequently performed for 50 ns of converged potential energy for solvent and protein systems.

Table 4.1: Solvent concentrations of simulated systems with villin, spitz and solvent only. Abbreviations: Ac^- = acetate, $Benz^-$ = benzoate, Bz = benzene, LHA = leonardite humic acids. As a comparison, the last line shows data given by the International Humic Substance Society (IHSS) for the concentrations of LHA carbon fractions.[32] Only the fractions of carboxyl, aryl and alkyl groups were taken into account and normalized to sum up to 1.

		Concentrations (mol/L)					Carbon Fractions		
Protein	System	Ca^{2+}	Cl^-	Ac^-	$Benz^-$	Bz	Carboxyl	Aryl	Alkyl
Villin	H_2O	-	0.02	-	-	-	-	-	-
	$CaCl_2$	1.04	2.11	-	-	-	-	-	-
	$CaAc_2$	1.00	-	2.02	-	-	0.50	-	0.50
	$CaBenz_2$	0.90	-	-	1.83	-	0.14	0.86	-
	$CaAc_2$ + Bz	1.06	-	2.15	-	2.12	0.13	0.75	0.13
	Real. Conc.	1.04	-	1.98	0.13	1.24	0.17	0.67	0.16
Spitz	H_2O	0.01	-	-	-	-	-	-	-
	$CaCl_2$	1.07	2.11	-	-	-	-	-	-
	$CaAc_2$	1.02	-	2.02	-	-	0.50	-	0.50
	$CaBenz_2$	0.92	-	-	1.82	-	0.14	0.86	-
	$CaAc_2$ + Bz	1.12	-	2.23	-	2.23	0.12	0.75	0.12
	Real. Conc.	1.06	-	1.97	0.12	1.25	0.17	0.67	0.16
-	H_2O	-	-	-	-	-	-	-	-
	$CaCl_2$	1.06	2.12	-	-	-	-	-	-
	$CaAc_2$	1.02	-	2.03	-	-	0.50	-	0.50
	$CaBenz_2$	0.94	-	-	1.87	-	0.14	0.86	-
	$CaAc_2$ + Bz	0.96	-	1.91	-	1.91	0.12	0.75	0.12
	Real. Conc.	0.99	-	1.87	0.11	1.21	0.17	0.67	0.16
IHSS LHA Sample[32]		-	-	-	-	-	0.17	0.67	0.16

4.2.2 VSOMM2 Systems

The simulation of reference proteins in VSOMM2 systems was performed using different molecular dynamics software (GROMACS) and thus with different cutoff parameters as the simple solvents. Therefore, to be able to compare protein stability, villin was simu-

lated in the Realistic Concentration condition with VSOMM2 parameters as well. The solvent concentrations of these simulations are depicted in Table 4.2. Four replicates were made.

Table 4.2: Concentration of the three villin simple solvent system replicates that were simulated with GROMACS. Abbreviations: Ac^- = acetate, $Benz^-$ = benzoate, Bz = benzene.

	Concentrations (mol/L)				Carbon Fractions		
System	Ca^{2+}	Ac^-	$Benz^-$	Bz	Carboxyl	Aryl	Alkyl
Real. Conc.	1.01	1.91	0.13	1.20	0.17	0.67	0.17

VSOMM2 models representing LHA were used as basis in which reference proteins were introduced. All models contained the same total number of building blocks (200) that comprised the humic acid molecules, however the number of building blocks per molecule, hence the number of HS molecules changed. The systems were neutralized with Ca^{2+} ions. In Table 4.3 the carbon fractions of all simulated replicates is depicted. As a comparison, the last line shows the data of carbon fractions of experimental LHA samples provided by the IHSS. All systems were simulated in an aequous solution with H_2O mass fractions of 0.74 to 0.77, which are well above what has been previously reported as minimum for a water activity of 1.[27]

4.2.2.1 Insertion of Protein into LHA systems

The procedure used to insert the reference proteins into the LHA systems is similar to the `InflateGro` method proposed by Kandt et al.[95] In `InflateGro` an amphiphilic bilayer was first inflated until the distances between lipids is big enough to accommodate a protein in it. Then, the system is again compressed in small steps, where energy minimization simulations were performed in between to permit lipids to reassemble around the inserted protein. However, since there were some differences between inserting a membrane peptide into a bilayer and proteins to soil organic matter, several adjustments were made. Firstly, SOM systems are not forming membrane bilayers, therefore, the inflation needed to be applied in three dimensions. Furthermore, `InflateGro` inflation increased the distance between all membrane molecules, however only one gap is needed for the insertion of a protein. Lastly, `InflateGro` inflation worked by multiplying distances with a scalar. For this reasons I wrote a more specific PYTHON script that was optimized to work for VSOMM2 systems (Appendix Listing 8.5). Rectangular periodic boxes containing the soil systems were prepared for the insertion process by removing water and Ca^{2+} molecules. To insert the reference proteins into the prepared systems, firstly, the systems were inflated by a summand of +1 nm and the protein was added. The systems subsequently were deflated in ten steps (summand −0.1 nm) with the same script each followed by an energy minimization simulation back to its original box size. The energy minimization was done by the steepest descent algorithm and

Table 4.3: Functional group concentrations of the VSOMM2 systems. The same systems were used for both reference proteins. Abbreviations: BB/Mol = number of building blocks per molecule, Arom. = aromatic, Het.Al. = heteroaliphatic. A table with the elemental fractions can be found in the appendix (Table 8.3). As a comparison the last line shows data given by IHSS for the concentrations of LHA carbon fractions[32].

			Carbon Fractions					
Name	BB/Mol	Rep	Carbonyl	Carboxyl	Arom.	Acetal	Het.Al.	Aliphatic
BB2_1	2	1	0.076	0.143	0.588	0.039	0.01	0.142
BB2_2	2	2	0.080	0.143	0.589	0.038	0.01	0.137
BB2_3	2	3	0.077	0.143	0.587	0.042	0.01	0.141
BB5_1	5	1	0.076	0.144	0.593	0.039	0.01	0.137
BB5_2	5	2	0.078	0.144	0.591	0.041	0.01	0.135
BB5_3	5	3	0.077	0.144	0.584	0.038	0.01	0.145
BB10_1	10	1	0.081	0.143	0.582	0.039	0.01	0.142
BB10_2	10	2	0.076	0.145	0.593	0.039	0.01	0.134
BB10_3	10	3	0.077	0.144	0.584	0.041	0.01	0.142
BB20_1	20	1	0.077	0.145	0.584	0.039	0.01	0.142
BB20_2	20	2	0.080	0.144	0.582	0.040	0.01	0.142
BB20_3	20	3	0.082	0.143	0.581	0.042	0.01	0.142
IHSS LHA Sample[32]			0.08	0.15	0.58	0.04	0.01	0.14

by positionally restraining the protein with a force constant of 10^5 kJ/(mol nm^2). A visual representation of the script and its application for protein insertion is shown in Figure 4.3. The step size of deflation is crucial for the success of this protein insertion protocol. If the deflation step size is too big, molecules can overlap, which will lead to errors during energy minimization. If the step size is too small the amount of work and computational power spent is unnecessarily high. In general the step size has to be adjusted to work for the molecule density in the used system.

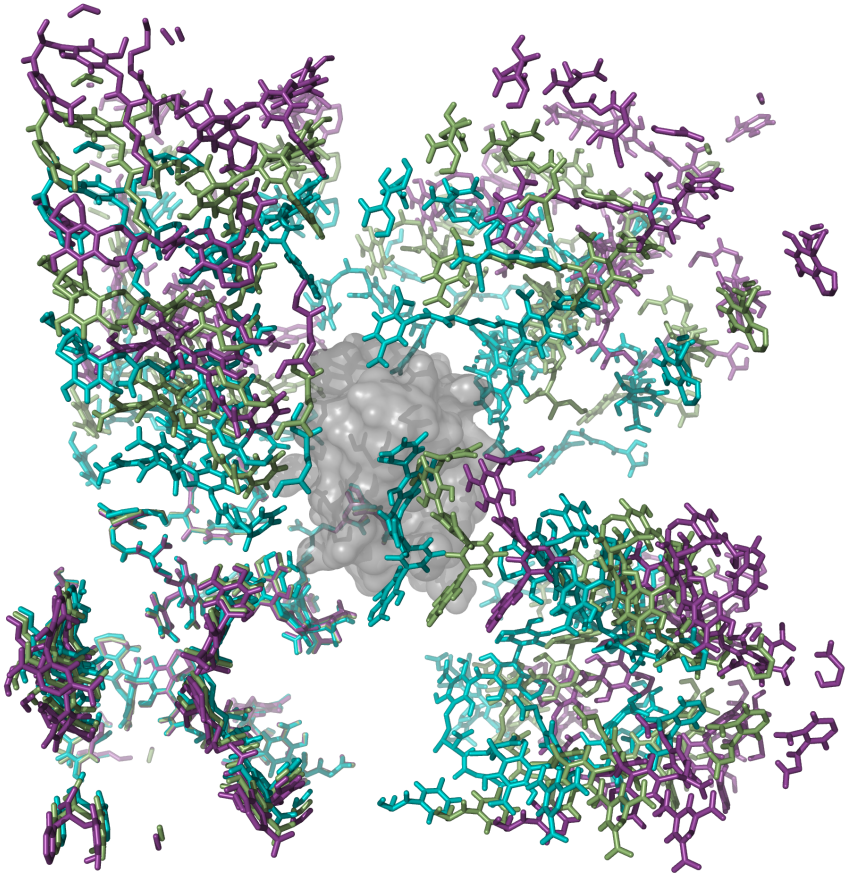

Figure 4.3: Three frames of the VSOMM2 systems protein insertion protocol. The protein is represented in gray. The most inflated system state is shown in violet purple. Half way and complete deflation to the original box size are showen in smudge green and teal respectively. The figure was rendered with PyMol[88].

4.2.2.2 Molecular Dynamics

After protein insertion the deflated boxes were subsequently processed with tools of GROMACS version 2019.1.[96–103] The molecular topology files were created with the 54A7 GROMOS force field.[67] The systems were solvated using default van der Waals radii[104] and subsequently neutralised by the replacement of water molecules with Ca^{2+} ions. A energy minimization step was performed using the steepest descent algorithm to a force lower than 10^3 kJ/(mol nm). An atomistic cutoff-scheme was used for all molecular dynamics simulations with a cut-off for electrostatic and van der Waals forces at 1.4 nm. An additional reaction-field with a dielectric permittivity of 61 was applied. Equilibration was performed in two distinct steps of 100 ps simulation each, starting with an NVT simulation. The leap-frog algorithm was used for integration with a step size of 2 fs. The protein was positionally restrained with a force of 10^3 kJ/(mol nm^2). All bonds were constrained with the LINCS algorithm.[71] The temperature was restrained at 300 K with a weak coupling thermostat[92] with three different temperature

groups (protein, SOM, water + ions) and a coupling time of 0.1 ps. The velocities were initially assigned according to a Maxwell – Boltzmann distribution. The second step of equilibration is a NPT molecular dynamics simulation. Simulation parameters stayed the same, except for the addition of a isotropic weak coupling barostat.[92] The coupling parameter was set to 0.5 ps and the isothermal compressibility of water to 4.5 $\times 10^{-5}$ bar^{-1}. The molecular dynamics run was performed with the same settings except for no positional restraining of the protein for 100 ns. After 20 ns we observed equilibrium by convergence of the potential energy and the last 80 ns of each run were used as production run.

4.2.3 Trajectory Analysis

Several analyses were performed with GROMOS and GROMACS trajectories using their respective analysis tools[103, 105]. In the following sections the underlying principles of these analysis methods and the settings are outlined. Unless stated otherwise, methods, equations and settings are the same for both simulation packages.

4.2.3.1 RMSD

To measure protein stability the positional root-mean-square deviation (RMSD) was calculated from backbone atoms of the protein using Equation 4.1 where $r_{ref,i}$ is the position of atom i in a reference structure and r_i is the position of an atom in the current frame of a trajectory. Both reference and current frame proteins need to be aligned in order to assure correct RMSD values. In this work the last frame of the equilibration trajectory was used as reference structure.

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_{ref,i} - r_i)^2} \tag{4.1}$$

4.2.3.2 Hydrogen Bonds

The hydrogen bonds created by the reference protein were tracked over the trajectories. The interactions were treated separately depending on the partner molecule, either itself or several solvent molecules. However no distinction was made regarding the donor and acceptor of the hydrogen bond. The default values for hydrogen bond definition were used in both analysis packages. For GROMOS++ `hbond` the maximal distance between acceptor and the hydrogen was defined as 0.25 nm and the minimal angle between donor – hydrogen – acceptor as 135°. For the GROMACS analysis the maximal distance between donor and acceptor was set to 0.35 nm and the maximal angle between hydrogen – donor – acceptor to 30° (Figure 4.4).

Figure 4.4: Representation of the angles used for definition of hydrogen bonds for A) GROMOS++ and B) GROMACS. Abbreviations: A = acceptor, D = donor, H = hydrogen.

4.2.3.3 MDF

The minimum distance function (MDF) is the distance between the closest two atoms from two previously specified groups of atoms for every frame of a MD trajectory. The output over the trajectories of similar molecules was averaged and converted into a histogram. For GROMOS trajectories I used a modified version of the `mdf` tool which is able to deal with more than one atom for both selected groups. This tool was kindly provided by Yerko Escalona.

4.2.3.4 SASA

The solvent accessible surface area (SASA) of proteins is the surface area which can be accessed by the solvent and, therefore, interacts the most with it. The SASA of single amino acids or whole molecules can be estimated with the algorithm of Lee and Richards[106]. The SASA was calculated with the GROMOS++ program `sasa`. The probe radius was set to 0.14 nm.

4.2.3.5 DSSP

DSSP is an algorithm developed to identify secondary structures of proteins[107]. Its implementation to GROMOS++, `dssp`, was used for trajectory analysis. It can identify seven different secondary structures: bends, turns, β bridges, β strands as well as 3-, 4- and 5-helices. In order to simplify the secondary structure analysis, in this work 3-, 4- and 5-helices were grouped as helix and β bridges and β strands were grouped as β strands.

4.2.3.6 Widom Method

Additionally, solvent only simulations of all conditions were performed in quadruplicates for subsequent calculations of solvation free energies of methane. The exact carbon fractions are depicted in Table 4.1. All molecular dynamics settings were kept as described above, however, the production run was shortened to 10 ns of converged potential energy. Solvation free energies of methane were calculated with the Widom method.[108] At every specified trajectory frame a methane (CH_4) united atom was inserted 10^4 times at a random position. The non-bonded interaction energy $\mathcal{V}^{nbd}$ was calculated for every insertion. By using the ensemble average and Equations 4.2 to 4.4 the free energy of solvation and its enthalpic and entropic terms were computed. Since no volume change

takes place upon insertion of CH_4 ΔU equals ΔH. Methane insertion was done every 10^{th} snapshot. Data convergence was verified for each simulation.

$$\Delta G_S = -k_B T \times \ln\left(\frac{\langle V\ e^{-\mathcal{V}^{nbd}}\rangle}{\langle V\rangle}\right) \tag{4.2}$$

$$\Delta U_{uv} = \frac{\langle \mathcal{V}^{nbd} V\ e^{-\mathcal{V}^{nbd}/k_B T}\rangle}{\langle V\ e^{-\mathcal{V}^{nbd}/k_B T}\rangle} = \Delta H_{uv} \tag{4.3}$$

$$T\Delta S_{uv} = \Delta G - \Delta H_{uv} \tag{4.4}$$

Where $\mathcal{V}^{nbd}$ is the non-bonded interaction energy, k_B is Boltzmann's constant, V is the system volume and T is the system temperature.

4.2.4 Free Energy Calculations

The knowledge of the free energy of solvation of a protein could explain the strength of absorption of a protein to humic substances compared to water. However, proteins are too big for efficient free energy calculations and simplifications were necessary. Firstly, not whole proteins but single amino acid sidechains were analysed, which additionally allowed us to investigate which amino acids drive protein – humic substances interactions. Secondly, no solvation free energies but mutational free energies of going from Ala to another amino acid and going from SPC water to LHA systems were computed. The calculations were performed analogously to Jandova et al. (2018).[82] A scheme of the applied thermodynamic cycle is depicted in Figure 4.5. An arbitrary amino acid sidechain (except for Gly and Pro) is represented by X, whereas A stands for alanine and R for a reference state. In a first step the free energies of going from an amino acid sidechain to a reference state ($\Delta G_{X\to R}$) were calculated. For non-polar amino acid sidechains (Ala, Leu, Ile, Val) this was done only with the OSP method. For polar amino acid sidechains (all other except Gly and Pro) the free energy estimation was split up into two steps: TPF (Section 4.2.4.1) and OSP (Section 4.2.4.2). Third power fitting is used to calculate the free energy of residues going from a charged state Q to a neutral one N ($\Delta G_{Q\to N}$). The one-step perturbation approach was then applied to get free energies of going from the neutral state N to a reference state R ($\Delta G_{N\to R}$). By simple addition of both terms the free energy of going from a charged to a reference state is calculated (Equation 4.5).

$$\Delta G_{Q\to R} = \Delta G^{TPF}_{Q\to N} + \Delta G^{OSP}_{N\to R} \tag{4.5}$$

Since there were polar as well as non-polar amino acid sidechains, in Figure 4.5 the state of an arbitrary amino acid is depicted as X instead of Q or N. The free energy of mutation in the same solvent was subsequently calculated as shown in Equation 4.6.

Finally, the free energy difference of mutation of alanine to an arbitrary amino acid going from SPC water to LHA VSOMM2 systems ($\Delta\Delta G_{A\to X}^{SPC\to LHA}$) for every amino acid was calculated using Equation 4.7.

$$\Delta G_{A\to X}^{SPC} = \Delta G_{A\to R}^{SPC} - \Delta G_{X\to R}^{SPC} \tag{4.6}$$

$$\Delta\Delta G_{A\to X}^{SPC\to LHA} = (\Delta G_{A\to R}^{LHA} - \Delta G_{X\to R}^{LHA}) - (\Delta G_{A\to R}^{SPC} - \Delta G_{X\to R}^{SPC}) \tag{4.7}$$

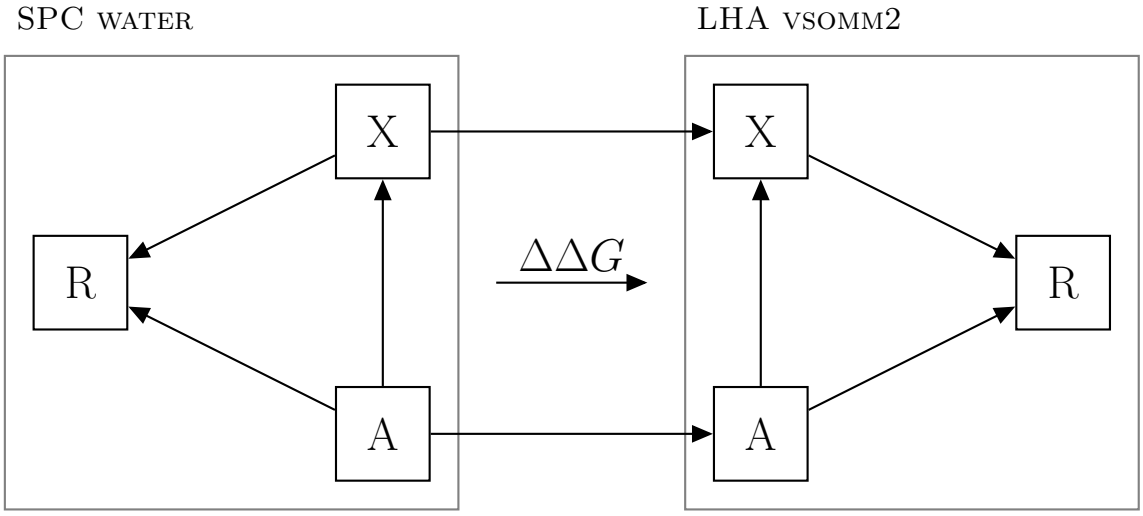

Figure 4.5: Thermodynamic cycle used for free energy calculations where X can stand for Arg, Asn, Asp, Cys, Gln, Glu, His, Ile, Leu, Lys, Met, Phe, Ser, Thr, Trp, Tyr or Val. A = Ala, R = reference state.

4.2.4.1 Third Power Fitting (TPF)

Third power fitting was done for all amino acid sidechains which contain charged atoms: Asn, Arg, Asp, Cys, Gln, Glu, Hisa, Hisb, Lys, Met, Phe, Ser, Thr, Trp and Tyr. Two topologies were created for each of these amino acids; one where sidechain atoms contain their original 54A8 charge and one were all sidechain atom charges were set to 0. The tripeptide coordinate files were taken from Jandova et al.[82] and were either directly solvated with SPC water or first combined with modified LHA VSOMM2 54A8 systems and subsequently solvated. Molecular dynamics simulations were performed for uncharged and charged sidechains as described in Section 4.2.1.2. To analyse the electrostatic interaction energy of the charged and uncharged trajectory the GROMOS++ program `ener` was used. The output was subsequently processed by the GROMOS++ program `tcf` which calculated, among others, the ensemble average and an error estimate of that energy. By solving Equations 3.32 to 3.35 the free energy of uncharging was computed.

Charge corrections are necessary if the combined charge of a charge group changes during perturbation which is the case for TPF calculations for charged amino acids (Arg, Lys, Glu, Asp). Charge corrections consist of three terms that can be calculated separately and need to be added to the raw values (ΔG_{raw}) to result real free energies (Equations 4.8 and 4.9).[81]

$$\Delta G = \Delta G_{raw} + \Delta G_{cor} \tag{4.8}$$

$$\Delta G_{cor} = \Delta G_{pol} + \Delta G_{dir} + \Delta G_{dsm} \tag{4.9}$$

For the calculation of ΔG_{pol} and ΔG_{dir} equidistant frames were taken from the uncharged and charged trajectories of net charged amino acids. Since both terms are independent on solvent all SPC water molecules were removed. Initially three frames were used but since values for Lys and Glu did not converge two additional frames were taken into account. To remove periodic boundaries all frames were gathered with the GROMOS++ program `frameout`. To estimate ΔG_{pol} the newly rewritten GROMOS++ program `dGslv_pbsolv` was applied for all non-water molecules. The probe integer atom code was set to 5 and and radii calculation mode was set to rmin. The non-bonded interaction parameters were set to values corresponding the ones used for molecular dynamics simulations. ΔG_{dir} was calculated with the GROMOS++ program `ener` concerning only atoms with charge changes. ΔG_{dsm} was calulated by solving an analytical equation.[81] Several scripts were kindly provided by Christoph Öhlknecht.

4.2.4.2 One-Step Perturbation (OSP)

Two reference states which were proposed by Jandova et al.[82] were used, R4 and R5 (Figure 4.6). The reason for that was the good performance of reference state R5 for small and medium sized amino acid sidechains and R4 for big ones. The reference state MD simulations in SPC water were taken from Jandova et al., LHA VSOMM2 systems, however, were performed exclusively. From the 100 ns trajectory the last 80 ns were used for further analysis. Van der Waals interactions were calculated for the reference states over the trajectory with the GROMOS++ program `ener`. By superimposing the five most popular sidechain conformations sampled by Jandova et al. and utilizing the GROMOS++ program `fit_ener` it was possible to calculate the van der Waals interaction energies for every frame and for said conformations. Non-bonded interaction parameters were set as done during the MD simulations. Only the sidechain atoms of reference and superimposed structures contributed to free energy calculations. By using the GROMOS++ program `dg_ener` the free energy difference of both states was calculated for every frame of the trajectory and every amino acid sidechain. With the output Equations 4.10 to 4.12 were subsequently solved.

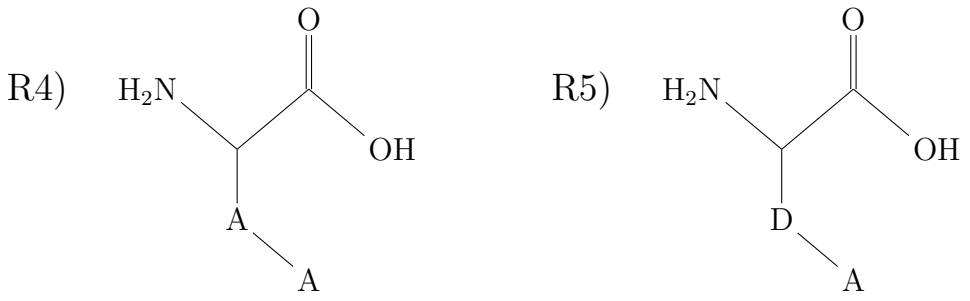

Figure 4.6: Reference states R4 and R5. D = dummy atom, A = soft atom.

$$\Delta G^{OSP}_{R\to N} = G_N - G_R = -k_B T \times \ln\langle e^{-(H_N - H_R)/(k_B T)}\rangle_R \tag{4.10}$$

$$\Delta G^{conf}_i = -k_B T \times \ln(P_i) \tag{4.11}$$

$$\Delta G^{OSP}_{R\to N} = -k_B T \times \ln\left(\frac{\sum e^{-(\Delta G^{OSP}_{R\to N_i} + \Delta G^{conf}_i)/k_B T}}{5}\right) \tag{4.12}$$

The fitness of the two reference states was scored by counting the number of contributing frames to the free energy term. If the number of contributing frames for reference state R5 was bigger than 1% the free energy calculated was directly used for comparison. However, for amino acids where less than 1% of frames of the simulated trajectory contributed to the free energy calculated the free energy calculated by going to R4 was used. Thus, in order to compare free energies of amino acids going to R4 and R5 a correction term was needed. This correction term was determined by linear regression of the most robust sidechain values. Robustness in this case means that only amino acids were selected that were not too big or too small (not: Trp, Arg, Lysh and Leu, Ala, Thr, Ser) since these would be a lot more favorable for one reference state than the other. Additionally, Ile was not taken into consideration since it was reported to be problematic using OSP.[82] Moreover, Cys was not used because the calculated free energy values varied a lot and thus it was identified as an outlier. These considerations left a set of 11 amino acids used for linear regression (Asn, Asp, Gln, Glu, Hisa, Hisb Ile, Met, Phe, Tyr and Val).

4.2.4.3 Molecular Dynamics

Similarly to the paper of Jandova et al., methylated tripeptides of the form Ala-X-Ala, where X can stand for all amino acids, except glycin and proline, were used. Protonation states were picked according to the most probable state in pH 7, however, for histidine both neutral tautomers were used: protonation of Nϵ and Nδ which from now on will be called Hisa and Hisb respectively according to the GROMOS nomenclature.

Simulations were performed with the GROMOS11 simulation package in SPC water and LHA VSOMM2 systems with similar LHA building block concentrations as applied in previous VSOMM2 simulations of this work. The H_2O mass fractions in VSOMM2 simulations averaged around 0.80. Since Jandova et al. used the GROMOS 54A8 force field, the VSOMM2 building block topologies had to be modified since there are changes in the charge distribution of carboxyl groups from force field 54A7 to 54A8. Therefore, for all carboxy groups the charges were altered accordingly and the charge groups were extended to the adjacent C atom. However, this was not possible for building block HS13 which was deprecated and replaced by building block, HS37, which contains an additional carbon atom to solve the problem of overlapping charge groups (Appendix Figure 8.2 and Listing 8.4). The equilibration setup was changed into a six step process starting from an initial simulation temperature of 40 K and constant restraining of 2.5×10^4 kJ/mol for the protein. In the following steps the temperature of the system is increased to 80, 120 and subsequently to 300 K in 40 and 60 K steps respectively. Simultaneously the force constant was lowered by a factor of 10 each step leading to a force constant of 2.5 kJ/mol at step five and 0 kJ/mol at step 6. All other parameters concerning the MD simulations were kept the same as in previous GROMOS simulations (Section 4.2.1.2). The molecular dynamics production run was then performed for 20 and 40 ns for the TPF runs in SPC water and LHA VSOMM2 systems respectively. The simulation of reference states for OSP was done for 100 ns. For systems containing LHA we observed an equilibrium of potential energy after 20 ns and, therefore, the first 20 ns of simulation were not used for further analysis.

4.2.4.4 SOMscore

With the results of the free energy calculations a scoring function was developed, SOMscore, that was able to score proteins on their interaction strength to LHA. SOMscore relies on $\Delta\Delta G_{A\rightarrow X}^{SPC\rightarrow LHA}$ values of every residue number r and amino acid X as well as its solvent accessible surface area (Equation 4.13). An additional fitness value gave an indication on how much of the surface of a protein was described with this scoring method (Equation 4.14).

$$\text{SOMscore} = \frac{\sum_r \Delta\Delta G_{A\rightarrow X}^{SPC\rightarrow LHA} \frac{SASA_r}{SASA_{max,r}}}{\sum_r \frac{SASA_r}{SASA_{max,r}}} \tag{4.13}$$

$$\text{Fitness} = \frac{\sum_r SASA_r}{SASA_{protein}} \tag{4.14}$$

r stands for every amino acid sidechain of a protein that can be scored, $SASA_{max,r}$ stands for the maximum SASA that an amino acid sidechain can have, $SASA_r$ stands for the current SASA observed for the respective sidechain r and $SASA_{protein}$ stands

for the total protein solvent accessible surface area, including backbone and residues which are not scored.

SOMscore was implemented in a PYTHON script that requires protein structure files (.pdb) as input (Appendix Listing 8.6). The SASA of the sidechains was calculated with the freeSASA PYTHON package.[109] For the definition of maximal SASA values for various side chains a simulation average of single amino acid simulations was calculated with freeSASA. The single amino acid simulation data was kindly provided by Matthias Diem. The $\Delta\Delta G_{A\rightarrow X}^{SPC\rightarrow LHA}$ for both tautomeric histidines were averaged to a single histidine value weighted with their respective Boltzmann probability (Equation 4.15).

$$\frac{p_{\mathrm{Hisa}}}{p_{\mathrm{Hisb}}} = e^{-\Delta\Delta\Delta G/k_B T} \tag{4.15}$$

Subsequently, to test our scoring function several proteins with a complete PDB structure from the UniprotKB database[110] were scored. Firstly, two sample groups were selected. The first sample comprised proteins from bacteria and fungi that were classified as extracellular (368 proteins). The second sample group, which was used as a reference, comprised bacterial and fungal proteins that were classified as cytoplasmic (822 proteins). Subsequently, the distributions of SOMscores of both groups were compared.

5. Results and Discussion

Villin is a popular protein for molecular dynamics studies, several papers on its folding have been published.[111, 112] Additionally it has been found that the proximity of three phenylalanines that comprise the hydrophobic core is crucial for the correct structure.[84] Although spitz is not as popular as villin, it has been used in several MD studies because of its simplicity and an existing high quality structure.[83, 113]

5.1 Simple Solvents

The initial simulations of the proteins in water were done as a reference to which changes in the measured properties can be compared to. To get more reliable statistics the H_2O simulations were done with eight replicates. The simple solvent molecule structures were selected to represent the most prominent functional groups in humic substances (alkyl, carboxyl and aryl). These three groups were represented by acetate (alkyl and carboxyl), benzoate (aryl and carboxyl) and benzene (aryl). By splitting up a magnitude of properties in the simpler systems it was easier to explore causes of effects on the protein. Ca^{2+} ions have been identified to play a significant role in SOM structure and stability,[23, 114–116] therefore, all conditions were normalized to approximately 1 mol/L of Ca^{2+} ions in the system.

5.1.1 Protein Stability

The root-mean-square deviation (RMSD) was calculated for both reference proteins to investigate the change of protein backbone stability in different simple solvent conditions. The average RMSD over the simulation time is shown in Figure 5.1 A for villin and B for spitz. Strikingly, there are big differences between both proteins. Villin had a high structural variability in water which resulted not only in an increased average RMSD of 0.33 nm but especially an increased standard deviation of 0.15 nm. The high deviation resulted from one water simulation replicate where villin unfolded drastically to an average RMSD of 0.67 nm over the trajectory. Similar unfolding events happened consistently in all replicates of conditions were benzene molecules are present ($CaAc_2$ + Bz and Real. Conc., the latter represented by the red line in Figure 5.1). Therefore, significant increases of RMSD in this conditions are observed. However, this is just the case for villin. Additionally, the RMSD analysis showed that increasing the Ca^{2+} concentration to 1 mol/L had no significant influence on protein stability. Interestingly, when the functional aryl groups were part of benzoate rather than its own molecule

(condition $CaBenz_2$) the RMSD of villin was not affected. Spitz was not only very stable in water resulting in an average RMSD over all eight replicates of 0.19 ± 0.04 nm, also in no other solvent conditions significant deviations from the RMSD measured in water were observed (Figure 5.1 B). The highest measured RMSD as an average over all replicates was found in $CaCl_2$ (0.31 ± 0.13 nm). The exact values of all RMSD averages are listed in the appendix in Table 8.1.

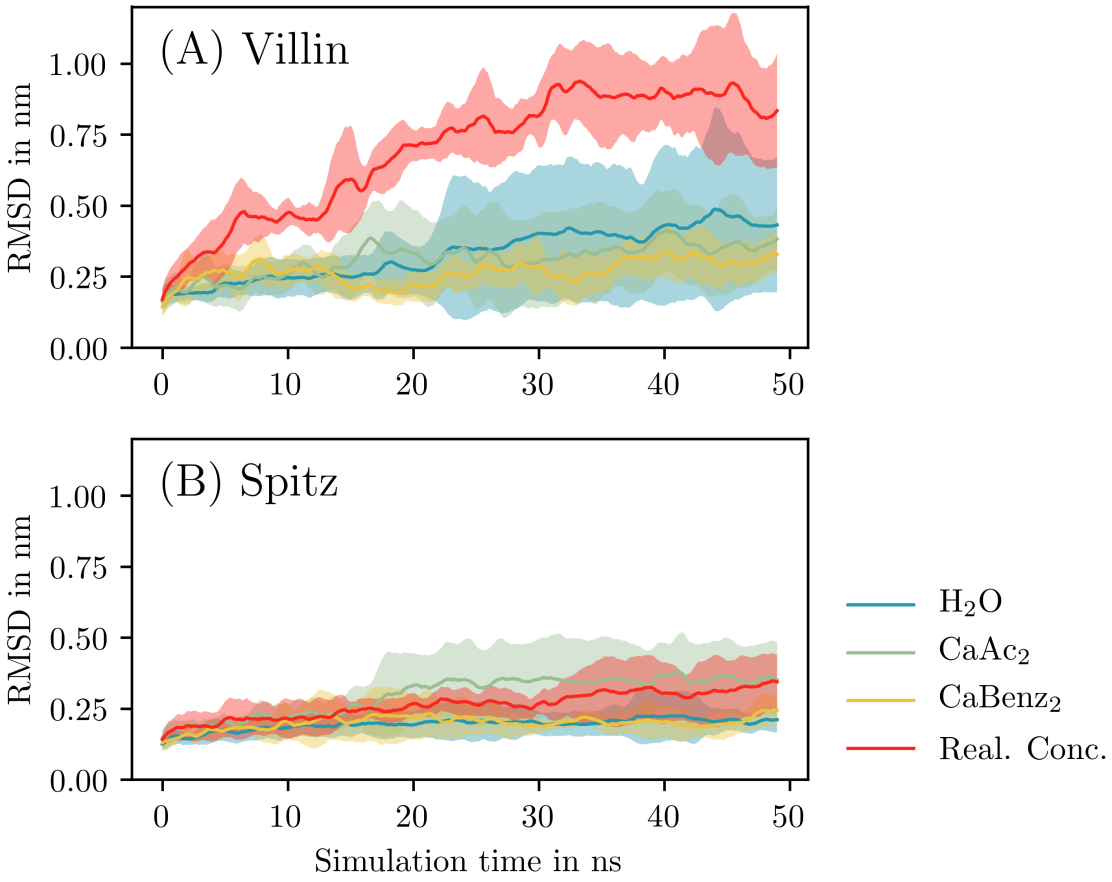

Figure 5.1: Running average (1 ns) RMSD timeseries of (A) villin and (B) spitz in selected conditions (dark lines). The lighter colored area represents the standard deviation of RMSD between different replicates.

The RMSD results of villin in water were slightly higher than what has been reported in literature (0.33 ± 0.15 and 0.27 ± 0.12 nm respectively).[84] However, concerning the standard deviation we concluded that both results are comparable. The unfolding of villin in conditions where benzene molecules were present was probably due to disruptions of the hydrophobic core of the protein. Stability studies done on villin found that the positions of three phenylalanine residues that comprise the hydrophobic core are crucial for the villin headpiece stability.[84] By introducing benzene molecules to the environment around the protein, the phenylalanine residues started to interact more

with the solvent than with each other, which lead to the unfolding of the protein. However, this did not happen for simulations where benzoate instead of benzene was used, which indicates that the atomic bond between the carboxyl and aryl group had drastic impact on protein stability. The unfolding event of villin sampled in an H_2O simulation might be an outlier, however, it emphasised the higher conformational variability of villin compared to spitz. In all simulations where villin unfolded (all replicates of conditions $CaAc_2$ + Bz and Realistic Concentration as well as one H_2O replicate) the periodic box was too small to ensure that the protein was not interacting with its own periodic copy. Therefore, numbers obtained from these simulations were interpreted very carefully. However, they are included in this project work because they can still elucidate protein solvent interactions. The average RMSD for spitz in water agreed with Setz (2018).[83] The high stability of spitz can be explained by the presence of stabilizing cystein bonds which might make the core less susceptible for disruption by hydrophobic solvent molecules. Interestingly, despite its cystein bridges, spitz is susceptible for destabilisation by high salt concentrations, where the highest average RMSD value ($CaCl_2$: 0.52 nm) was observed.

Subsequently, the secondary structure of reference proteins in the simple solvent conditions was analysed. In Figure 5.2 the secondary structure of each residue as defined by the DSSP algorithm (Section 4.2.3.5) is depicted over time for villin (A, C, E) and spitz (B, D, F). Figure 5.2 A clearly shows that villin formed three distinct α helices (depicted in red) when simulated in SPC water. The helices were separated by few amino acids that were defined either as bends, turns or coils. The same held true for the simulation in $CaBenz_2$ as well as Real. Conc. condition (Figure 5.2 C and E). The average occurrences of helices in all simple solvent conditions were compared to H_2O conditions and no significant differences were found. In Figure 5.2 B it can be seen that the secondary structure of spitz is more diverse than for villin. It contained not only one section that was defined as α helix but also several sections that comprised both β sheets of the protein. Similarly to spitz a statistical analysis of the average occurrences of helices as well as β sheets showed no significant changes in different conditions. The average occurrences observed can be found in the appendix in Table 8.2.

The results of the secondary structure analysis were interesting, especially in the context of the unfolding events seen in the RMSD analysis. Even in conditions where there were significant RMSD changes for villin ($CaAc_2$ + Bz and Real. Conc.) the secondary structure was kept intact. This indicates that even though the protein structure was more flexible there was still a degree of organisation kept within the protein and the unfolding event described by RMSD did not affect the secondary structure.

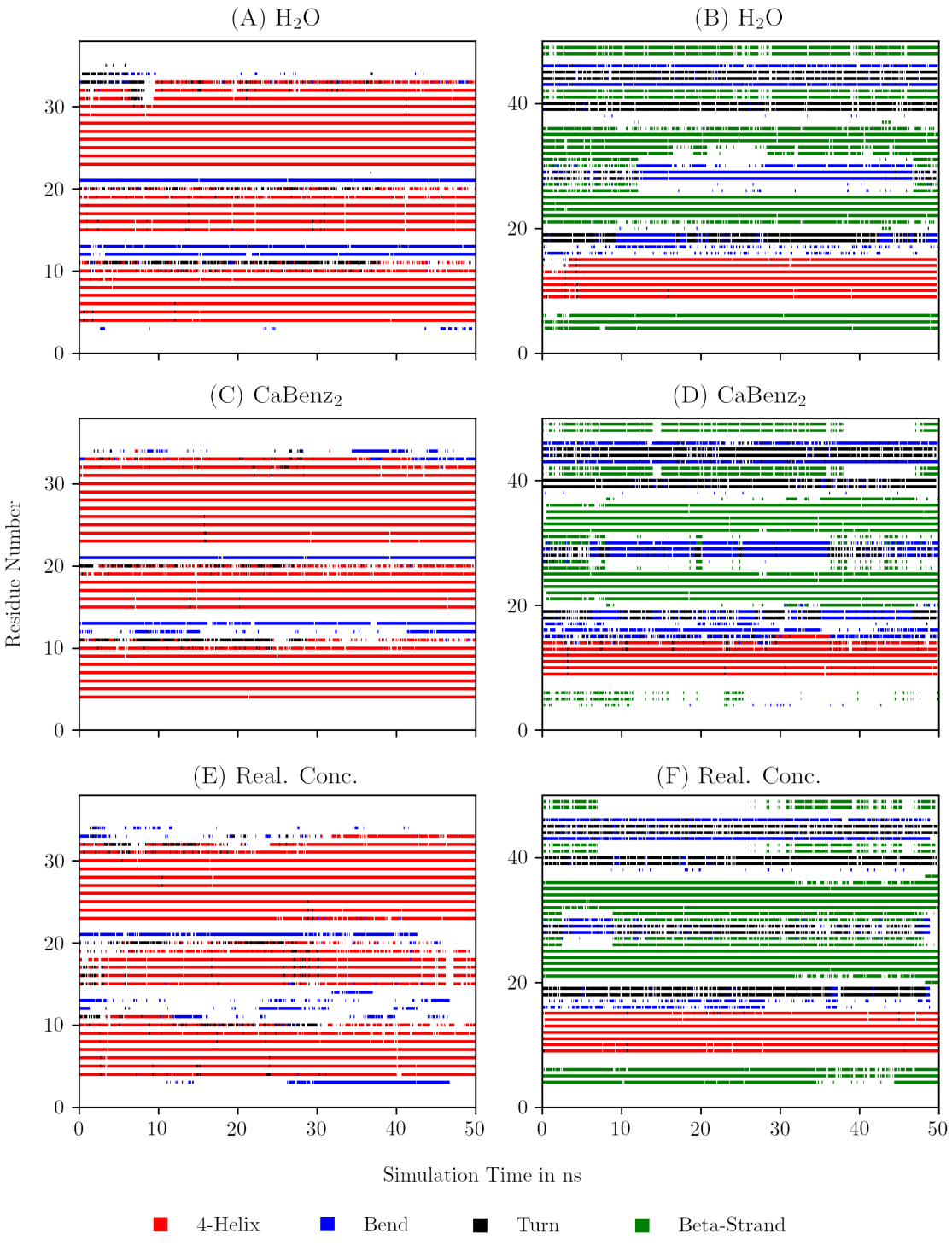

Figure 5.2: The secondary structure of selected simulation replicates of villin (A, C, E) and spitz (B, D, F) in the conditions H_2O, $CaBenz_2$ and Realistic Concentration along the time. On the y axis the residue number and on the x axis the simulation time is plotted. Notice that villin does not have any β strands.

5.1.2 Salting in and salting out

Solvents can possess salting in or salting out effects on proteins. Salting in is defined as a protein solubility increase upon addition of a co-solvent, salting out, conversely, is defined as a solubility decrease. Since proteins are relatively hydrophobic molecules, salting in and salting out effects of co-solvents can be deduced from the free energy of solvation of methane.[117] By calculating these free energies in different conditions and comparing them to water it is possible to predict how the solubility of proteins would change.

The free energies of solvation of methane (ΔG_s) are depicted in Figure 5.3. The free energy of solvation of CH_4 in water was 8.53 ± 0.07 kJ/mol. By the addition of purely charged co-solvents (conditions $CaCl_2$, $CaAc_2$ and $CaBenz_2$) the energy demand for solvation increased significantly going up as high as 10.09 ± 0.19 kJ/mol in 1 mol/L $CaCl_2$. 9.87 ± 0.19 and 9.65 ± 0.21 kJ/mol were measured for $CaAc_2$ and $CaBenz_2$ respectively. Interestingly, there were big differences between ΔG_s in $CaBenz_2$ and $CaAc_2$ + Bz (7.26 ± 0.13 kJ/mol) which demonstrated how much molecular interactions changed upon having a chemical bond between carboxyl and aryl groups. The values observed for Realistic Concentration conditions score in between both extremes, close to water at 8.88 ± 0.26 kJ/mol. In the $CaAc_2$ + Bz systems a phase separation of water (containing the ions) and benzene was observed.

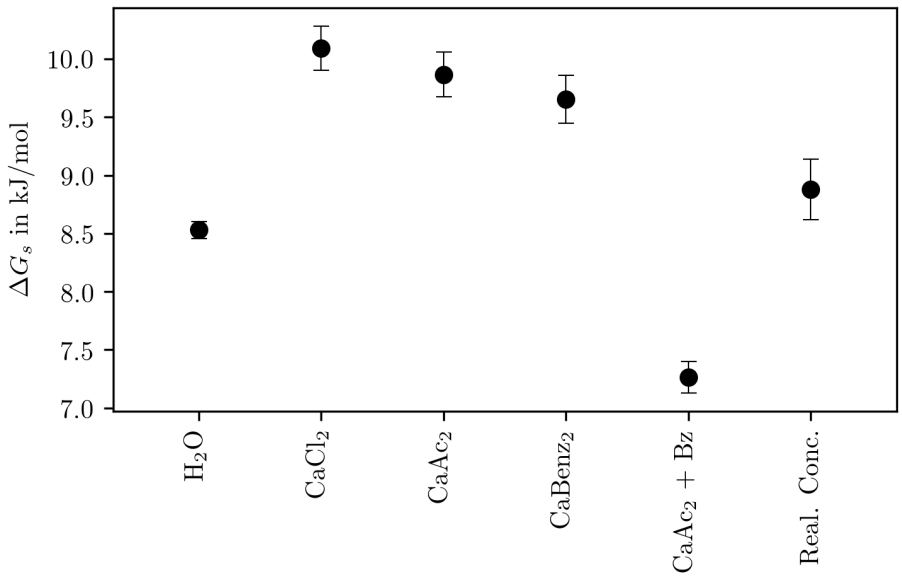

Figure 5.3: Free energies of solvation of methane in different conditions. The respective enthalpy and entropy terms can be found in the appendix Figure 8.3.

The solvation free energy of methane calculated in this work is reasonably close to the experimentally measured 8.4 kJ/mol[118] and can, therefore, be used as a reference for other results. ΔG_s values higher than water indicate salting out properties of the

solvents. This can have a stabilizing effect on the protein, since hydrophobic parts of the protein tend to stick together rather than interact with such solvents. Contrarily, ΔG_s values lower than water can lead to salting in effects increasing the hydrophobic interactions between solvents and solute. This is the case for $CaAc_2$ + Bz condition simulations. The free energy data obtained from these simulations agrees with RMSD data gained for villin, where the standard deviation between replicate averages was reduced for conditions were high ΔG_s values were measured, for Real. Conc., however, the RMSD increased significantly. Interestingly the free energy of solvation in the Realistic Concentration condition scores similarly to the water, which is surprising since solvent concentrations are close to concentrations in $CaAc_2$ + Bz and the measured RMSD behaves similarly to $CaAc_2$ + Bz. This could explain why villin was not as stable in water either which was demonstrated by one H_2O condition replicate that unfolded.

5.1.3 Non-bonded Protein Interactions

5.1.3.1 Interaction Energies

To further understand which forces govern the interaction of proteins and their solvents the non-bonded interaction energies were investigated. They were grouped according to their respective solvent components and interaction term to gather more insights. In Figure 5.4 the non-bonded interaction energies of both reference proteins with their solvent are depicted. The cumulative bars per condition with their respective error bars represent the total interaction energies and its deviation over the replicates. The color scheme represents how different solvent molecules contribute to these energies. There was a significant increase of total non-bonded interaction energies for $CaAc_2$, $CaBenz_2$ for both proteins compared to their respective H_2O values. This significant increase resulted from increased contributions of Coulomb as well as van der Waals energies. Simultaneously, the contribution of anions and cations (dark and light green bars) to the interaction energies increased, whereas the contribution of water (brown bars) decreased. The presence of benzene molecules in solution is responsible for a big part of the van der Waals energy (blue bars). However, Coulombic forces exceeded all van der Waals forces by approximately one order of magnitude. A further observation was that the fraction of contributions of cations and anions to the non-bonded interaction depended on the charge of the protein. As expected, the interaction energies between the positively net charged villin and Ca^{2+} were small. Compared to that the negatively net charged spitz interacted more strongly with the anions.

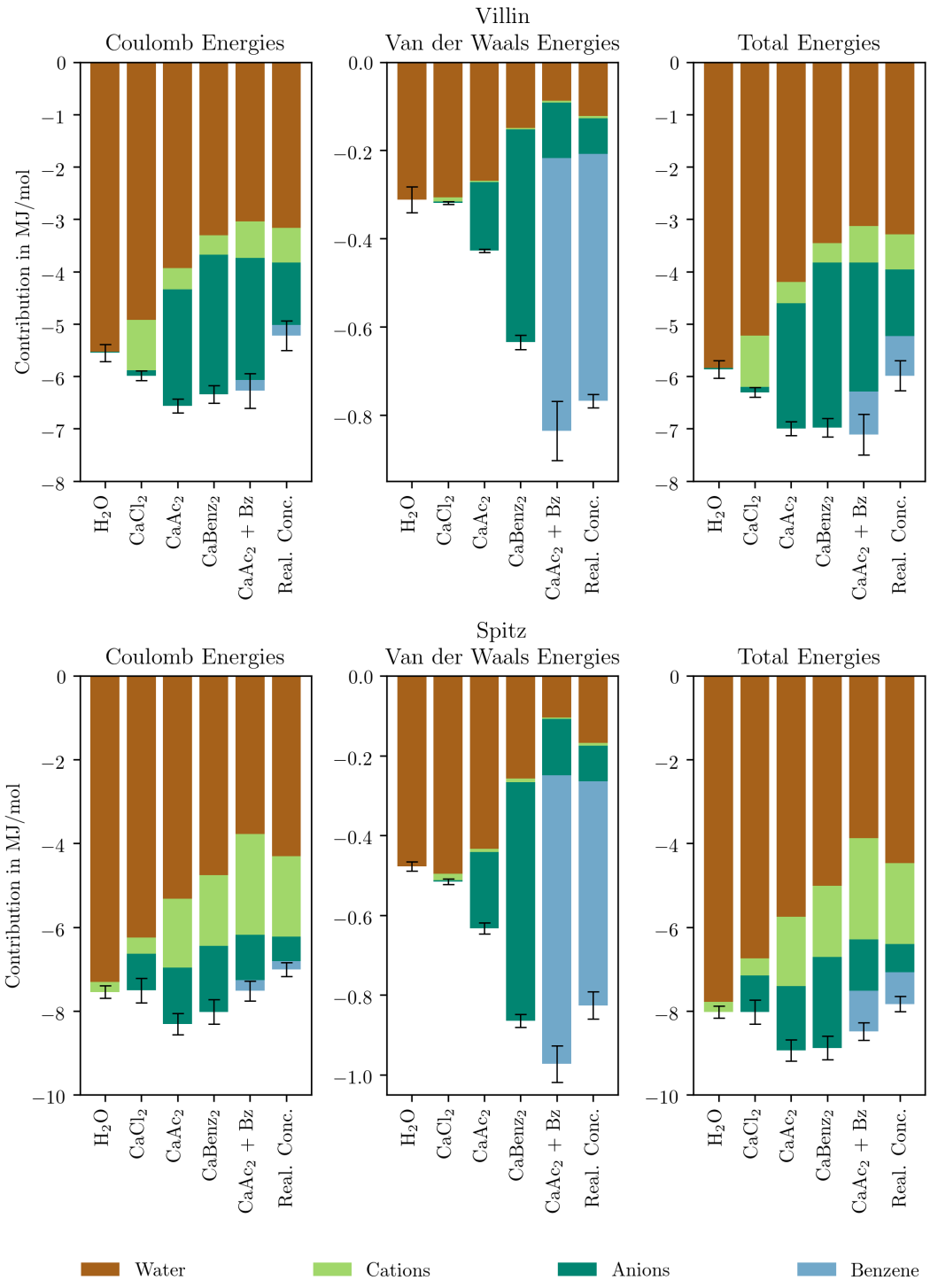

Figure 5.4: Non-bonded interaction energies of both reference proteins in all tested conditions (top row: villin, bottom row: spitz). The different colors symbolize different contributions of solvent molecules to the energy term. The following total energies differ from the H_2O reference significantly ($\alpha < 0.05$): $CaAc_2$, $CaBenz_2$ and $CaAc_2$ + Bz for villin and $CaAc_2$ and $CaBenz_2$ for spitz.

The differences in Coulombic energy contributions between villin and spitz were expected considering their respective net charge. However, both proteins also carry charges of the opposite type than their net charge, which explains why there are energy contributions of both positively and negatively charged ions for both proteins. An explanation of the overall higher contribution of anions to interaction energies is their higher complexity. The reason for relatively low van der Waals energies even in systems with benzene was that there were also always a lot of ions in the systems as well. This lead to the assumption that SOM protein interactions are mostly governed by electrostatic interactions.

5.1.3.2 Spacial Arrangement of Solvent Molecules

To understand which kind of molecular interface a protein is experiencing in different solvents it is of interest to investigate how close different ions or molecules get to the protein. For visualization a minimum distance function (MDF) was calculated and the normalized frequency of the distance was plotted for all simple solvent conditions (Figure 5.5 for villin and 5.6 for spitz). It is important to note that the minimum distance was always calculated between two physical atoms and not the center of geometry of a group of atoms. To begin with the water simulations (A) only few counter ions to neutralize the system were monitored. In the figure for spitz (Figure 5.6 A) the single Ca^{2+} ion (black line) had a distinct peak at 0.44 nm distance to the protein. Since Ca^{2+} was present in all simulated systems analogous peaks were visible in all other simulations (B – F) as well with its maximum ranging from 0.44 to 0.46 nm. Interestingly for both reference proteins the charged carboxy groups (blue line, Figures 5.5 and 5.6 C – F) showed peaks very close to the protein with maxima ranging from 0.18 to 0.19 nm. A second distinct peak of carboxyl groups was found between 0.43 and 0.44 nm. The aromatic peak (green line, Figures 5.5 and 5.6 D – F) always scored in between the peak for carboxyl groups and Ca^{2+} with maxima between 0.34 and 0.36 nm. Strikingly, even in systems where benzoate was present and thus both carboxyl and aryl groups were on the same molecule the maxima of the peaks did not differ from other simulations.

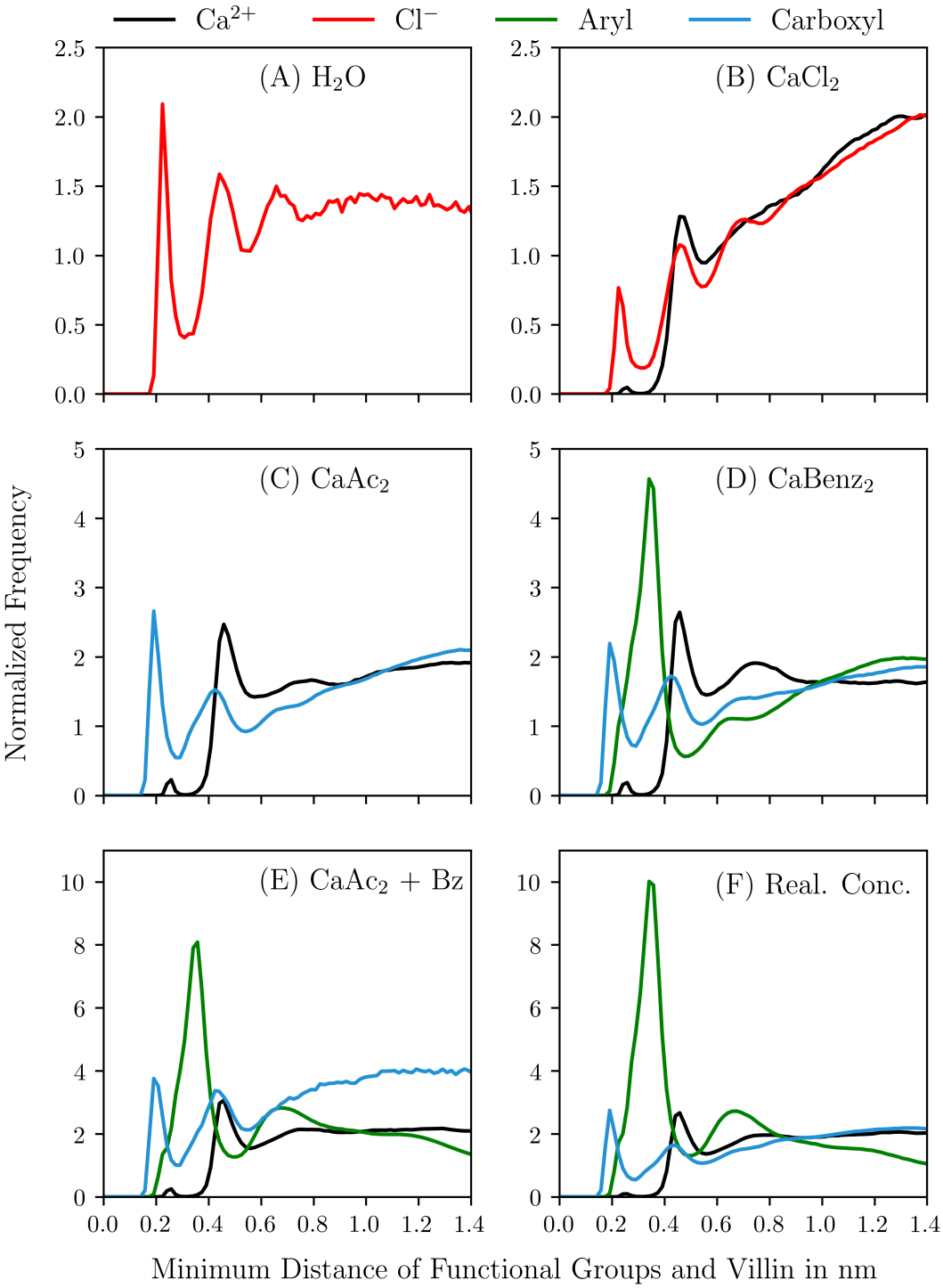

Figure 5.5: Frequency of the minimum distances of selected ions and functional groups to villin. The frequency is depicted in the number of snapshots $\times 10^{-3}$ sampled over all replicates of the simulation. Normalization was done over the number of ions/ functional groups present in the systems.

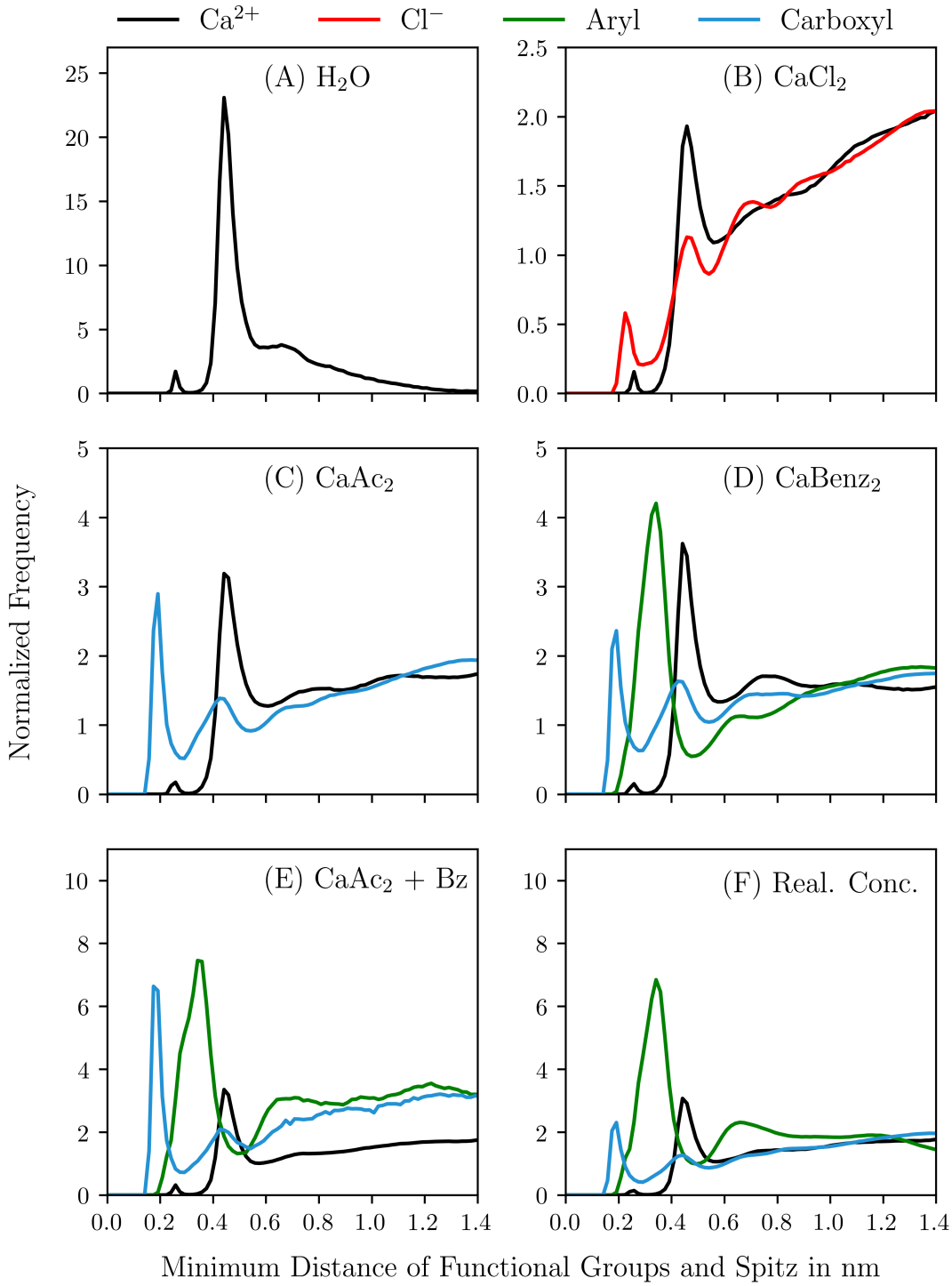

Figure 5.6: Frequency of the minimum distances of selected ions and functional groups to spitz. The frequency is depicted in the number of snapshots $\times 10^{-3}$ sampled over all replicates of the simulation. Normalization was done over the number of ions/ functional groups present in the systems.

The MDF analysis can elucidate how functional groups and ions arrange around the protein. For example, one can imagine a protein which is completely surrounded by a shell of molecules of type one and just further distant by molecules of type two. In the MDF analysis one would expect peaks for both molecule types, the closer peak corresponding to molecule type one and the more distant peak corresponding to molecule type two. However, if peaks have big overlaps, as observed in simple solvent conditions, this would indicate that several different molecule types are close to the protein. As amino acids have a big variety of properties, it would be expected to see more hydrophobic solvent molecules to interact with more hydrophobic residues and polar solvent molecules with more polar residues. However, the order of the peaks still has meaning on how tight the protein – solvent molecule interactions are, going from close carboxyl interactions to more distant aryl interactions and to relatively loose Ca^{2+} interactions. Especially the close proximity of negatively charged carboxyl groups to the net negatively charged protein spitz was interesting. This was probably due to positively charged residues that cause attractive forces. Another reason might be the possibility of the formation of hydrogen bonds between protein and solvent molecules. This is in accordance with the fact that the first peak of carboxyl groups is approximately 0.2 nm apart from the protein which is also the average distance of hydrogen bonds.[119] The proximity of carboxyl groups to the protein emphasizes the importance of them for the interaction of protein and solvent. This effect is even more pronounced by the fact that van der Waals interactions are more distance dependent than electrostatic interactions in the GROMOS force fields.

5.1.3.3 Hydrogen Bonds

Since the MDF analysis indicated that carboxyl groups of solvent molecules form hydrogen bonds with the protein, the hydrogen bonds were monitored and averaged over the simulated trajectories. Figure 5.7 depicts the number of hydrogen bonds formed by villin (A) and spitz (B). The total bar heights and their respective error bars represent all hydrogen bonds formed by the protein and their deviation over the replicates. The different colors indicate which molecules contribute as hydrogen bond forming partners. In H_2O simulations the average numbers of hydrogen bonds were 109.8 ± 1.2 and 154.9 ± 1.4 for villin and spitz respectively. There were significant deviations of the total number of hydrogen bonds to H_2O in conditions $CaAc_2$ + Bz and $CaAc_2$ for villin and $CaAc_2$ + Bz, $CaAc_2$ and Real. Conc. for spitz ($\alpha < 0.05$). The number of hydrogen bonds only within the protein (black bar) was significantly reduced for spitz in condition $CaAc_2$. For all other conditions and all villin simulations no significant changes of hydrogen bonds within the protein were measured. When carboxylate anions were introduced to the systems they formed a considerable amount of hydrogen bonds with the protein (dark green bars). However, except for condition $CaAc_2$ with spitz, the introduction of hydrogen bond forming anions did not interfere with number of hydrogen bonds within the protein, they rather formed new hydrogen bonds or replaced SPC waters.

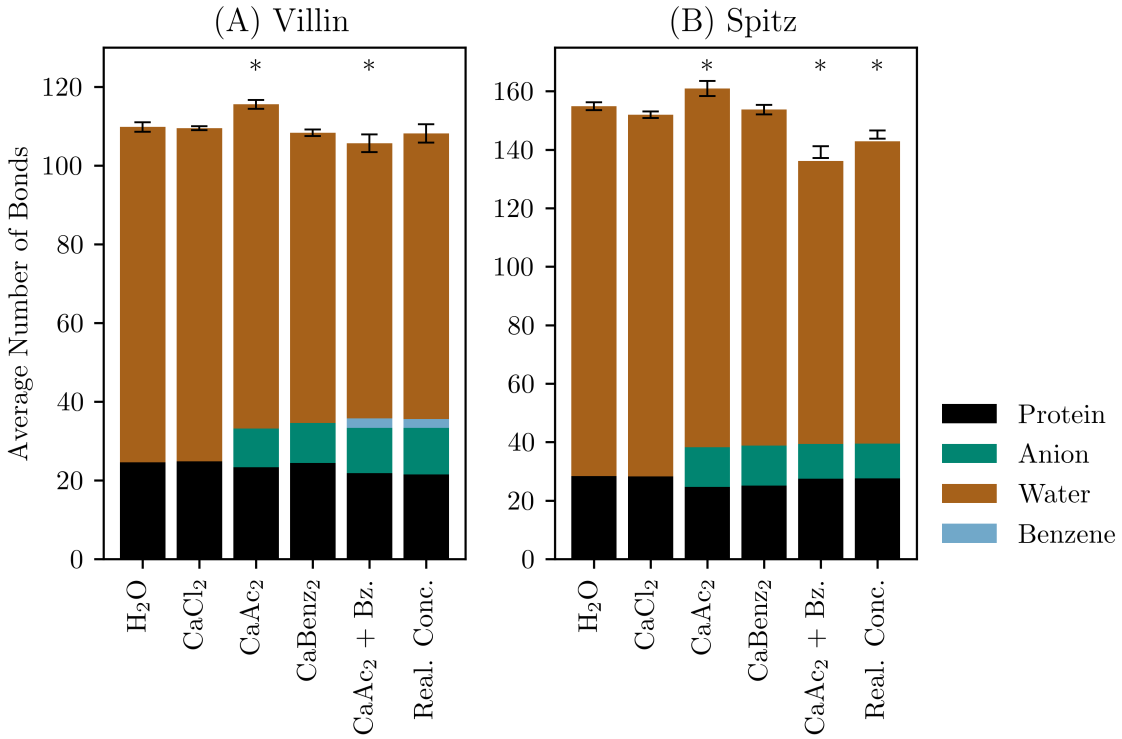

Figure 5.7: Average number of hydrogen bonds formed by the reference proteins (villin (A) and spitz (B)) over the simulation time. The errorbars represent the standard deviation of total numbers between replicates. The color code symbolizes different hydrogen bond partners. No differentiation was made between hydrogen donors and acceptors. Significant changes in the total number of hydrogen bonds is marked with an asterisk.

The total number of hydrogen bonds formed is bigger for spitz than for villin, which makes sense since spitz is a larger protein. However, when compared relatively results for both proteins were similar. $CaAc_2$ increased the number of total hydrogen bonds that are formed compared to $CaBenz_2$. This difference might be due to the aryl group carried by the benzoate, which is relatively big and could hinder the formation of additional hydrogen bonds. When benzene was introduced to the systems the number of hydrogen bonds is reduced ($CaAc_2$ + Bz) which can be explained by the shielding of hydrogen bond acceptors and donors on the protein surface by a layer of benzene molecules. Some benzene molecules are observed to form hydrogen bonds, which was probably an artifact of the broad definition of hydrogen bonds in the algorithm, however, the amount of benzene hydrogen bonds is in the range of the standard error, so it had no significant influence. It was again visible that the Realistic Concentration conditions were more similar to water conditions even though the functional group concentrations was relatively close to $CaAc_2$ + Bz The high number of hydrogen bonds formed by the anions and the protein are an explanation why there were considerably higher non-bonded interaction energies between protein and anions compared to protein and cations

(Figure 5.4). The secondary structure of a protein is stabilized by hydrogen bonds within the protein. Although some simple solvent molecules liked to form hydrogen bonds with the protein they apparently did not disrupt the secondary structure of the proteins as shown previously. This was true for both major secondary structure classes, α helices as observed in villin and spitz and β sheets as observed in spitz.

5.2 VSOMM2 Systems

After dissecting protein solvent interactions in more controlled simple solvent environments we investigated more realistic SOM models, provided by the VSOMM2 tool. We focused on two properties of these models that were previously not represented: (1) the influence of a bigger number of functional groups on proteins and (2) the influence of humic substance molecule length on protein - solvent interaction.

5.2.1 Protein Stability

To investigate the backbone stability of protein the average root-mean-square deviation was calculated. Table 5.1 shows the RMSD of the reference proteins. No significant changes were measured between the different sizes of LHA molecules. There was also no significant difference between the average RMSD of the Realistic Concentration and the LHA VSOMM2 systems for villin. However, there was a significant difference between the variance of the Real. Conc. replicates compared to the variance of the VSOMM2 systems.

As recent studies suggest proteins are more stable when simulated with atomistic cutoffs compared to group based cutoffs.[120] Since the simple solvent models were simulated with a group based cutoff and VSOMM2 systems with an atomistic cutoff it is difficult to compare protein stability. To solve this problem villin in Real. Conc. conditions replicates were simulated with both settings. As expected the average RMSD value for atomistic cutoffs was lower (0.44 ± 0.21 nm compared to 0.69 ± 0.06 nm). There were no significant differences in the mean of the RMSDs measured between Real. Conc. and VSOMM2 systems, which is most likely due to the high variance of the Real. Conc. RMSDs. This theory is supported by the fact that there are significant differences in variances between Real. Conc. and VSOMM2 replicates which indicates that there is higher variability in the protein structure in Real. Conc. systems. It is, therefore, save to assume that proteins are more stable in VSOMM2 systems compared to Real. Conc. systems.

Table 5.1: Average RMSD values obtained from all simulations of part two of this work in nm. Note that for the Real. Conc. conditions four replicates were made only for villin. Abbreviations: Rep. = replicates, Real. Conc. = Realistic Concentration, Std. Dev. = standard deviation.

	Rep.	BB2	BB5	BB10	BB20	Real. Conc.
Villin	1	0.26	0.30	0.20	0.32	0.43
	2	0.25	0.20	0.24	0.21	0.72
	3	0.29	0.32	0.38	0.23	0.22
	4	-	-	-	-	0.40
Mean		0.27	0.27	0.27	0.25	0.44
Std. Dev.		0.02	0.06	0.09	0.06	0.21
Spitz	1	0.22	0.20	0.18	0.16	-
	2	0.34	0.31	0.37	0.15	-
	3	0.25	0.18	0.52	0.20	-
Mean		0.27	0.23	0.36	0.17	-
Std. Dev.		0.06	0.07	0.17	0.03	-

5.2.2 Non-bonded Protein Interactions

5.2.2.1 Interaction Energies

In addition to protein stability investigation, the protein - solvent interaction energies were monitored and grouped. Figure 5.8 depicts the non-bonded interaction energies of both reference proteins. The whole bars and the black error bars show the total interaction energy acting on the respective protein and its variability over the replicates, respectively. The colours symbolize different solvent components contributing to the energies. The left and middle graphs show the Coulombic and van der Waals forces respectively, whereas the right graph depicts the sum of both. For Coulombic interactions it is visible that the contribution of water (brown bars) increased as the number of building blocks per molecule increased. However, the overall Coulombic interaction energy did not change significantly for both proteins. In contrast the van der Waals energies of simulations with two building blocks per LHA molecule (BB2) differed significantly to ones from other LHA molecule sizes ($\alpha < 0.05$). Additionally, the contribution of humic acids (dark green bars) to the van der Waals energies was significantly increased. Similarly to what was demonstrated for simple solvent molecules in a previous part of this work the van der Waals interactions were approximately ten times smaller than Coulombic interactions which translates to no significant differences

in the total interaction energies for both proteins (right graphs).

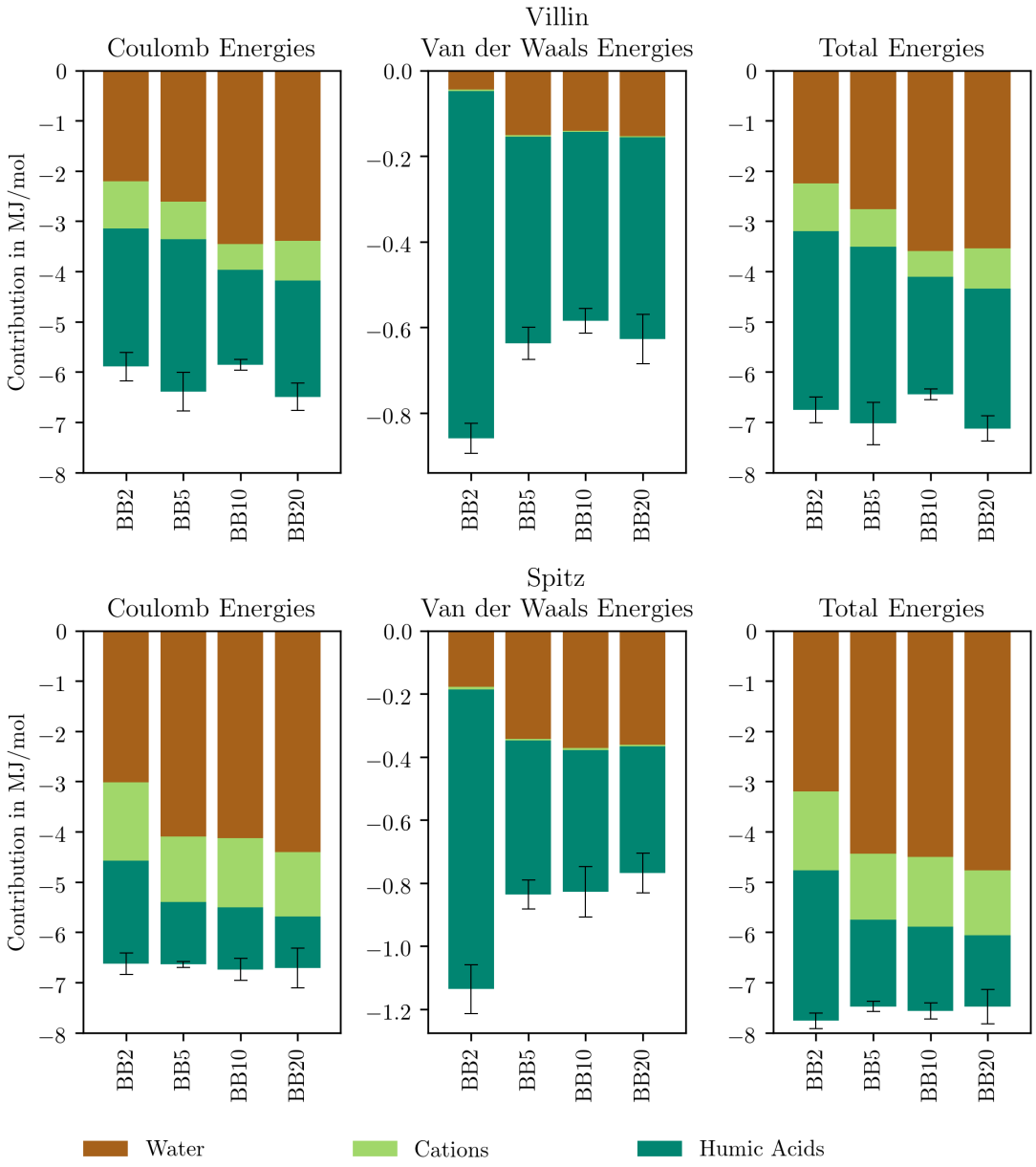

Figure 5.8: Non-bonded interaction energies of the reference proteins to their solvent (villin: top row, spitz: bottom row). The colour code represents the average contribution of different types of solvent to the energy term. There are significant differences between van der Waals interactions of protein at condition BB2 compared to the others.

Overall VSOMM2 sytems contained similar amounts of functional groups, only the size of humic substance molecules was altered. Therefore, it was expected that all interaction energies and their respective contributions would be comparable. However, this was not the case, especially striking with the increased van der Waals interaction energies of BB2

replicates. Therefore, it should be evident that the size of humic substance molecules causes this differences. BB2 molecules were still relatively small with an average of 24 atoms per molecule. It was possible for them to align to the protein in a way that polar and hydrophobic groups interact with respective amino acids of the protein. BB20 molecules, however, were ten times larger than BB2 ones. They could not align as well to the protein and, since the Coulomb energies were significantly higher than van der Waals energies, optimized polar alignment which lead to a further decrease of van der Waals energies. Moreover, the results of non-bonded energies for VSOMM2 systems were similar to the non-bonded interaction energies measured in $CaAc_2$ and $CaBenz_2$ systems (Section 5.1.3.1). Despite of the different complexity between VSOMM2 and simple solvent systems the forces that act on the protein are very similar in size and origin. As can be seen in the following chapter this observation also holds true for the spacial arrangement of LHA molecules around the protein.

5.2.2.2 Spacial Arrangement of LHA

To investigate how functional groups of the humic substances order around proteins a minimum distance function analysis was made. We observed a similar pattern for different LHA molecule lengths (Appendix Figure 8.4 and 8.5). For easy comparison to previous results Figure 5.9 shows the results of the $CaBenz_2$ simple solvent condition and of the shortest and longest HS molecule conditions. The left column of the figure refers to villin, the right to spitz. Similarly as found in simple solvent conditions Ca^{2+} ions showed a distinct peak with a maximum ranging from 0.44 to 0.45 nm distance to the reference proteins. Interestingly, also the MDF of carboxyl groups were similarly close to the protein with a peak maximum at 0.19 to 0.20 nm for villin and 0.18 to 0.19 nm for spitz. In between the relatively distant Ca^{2+} peak and the relatively close carboxyl peak there were less distinct aryl peaks with maxima ranging from 0.32 to 0.37 nm and 0.32 to 0.35 nm for villin and spitz respectively.

The big distinct peak of Ca^{2+} can be used as reference to compare results from simple solvent and VSOMM2 conditions, since the properties of Ca^{2+} have not been changed. Both peaks were at a very similar distance (approximately 0.45 nm) to the protein. However, even though the composition of solvent molecules of part one and part two of this work have changed drastically the characteristic peaks of the functional groups stay at constant distances from the reference proteins. This gave an additional indication that even though there was high variability in the arrangement of functional groups in the solvent molecules the protein was experiencing a relatively similar environment around itself.

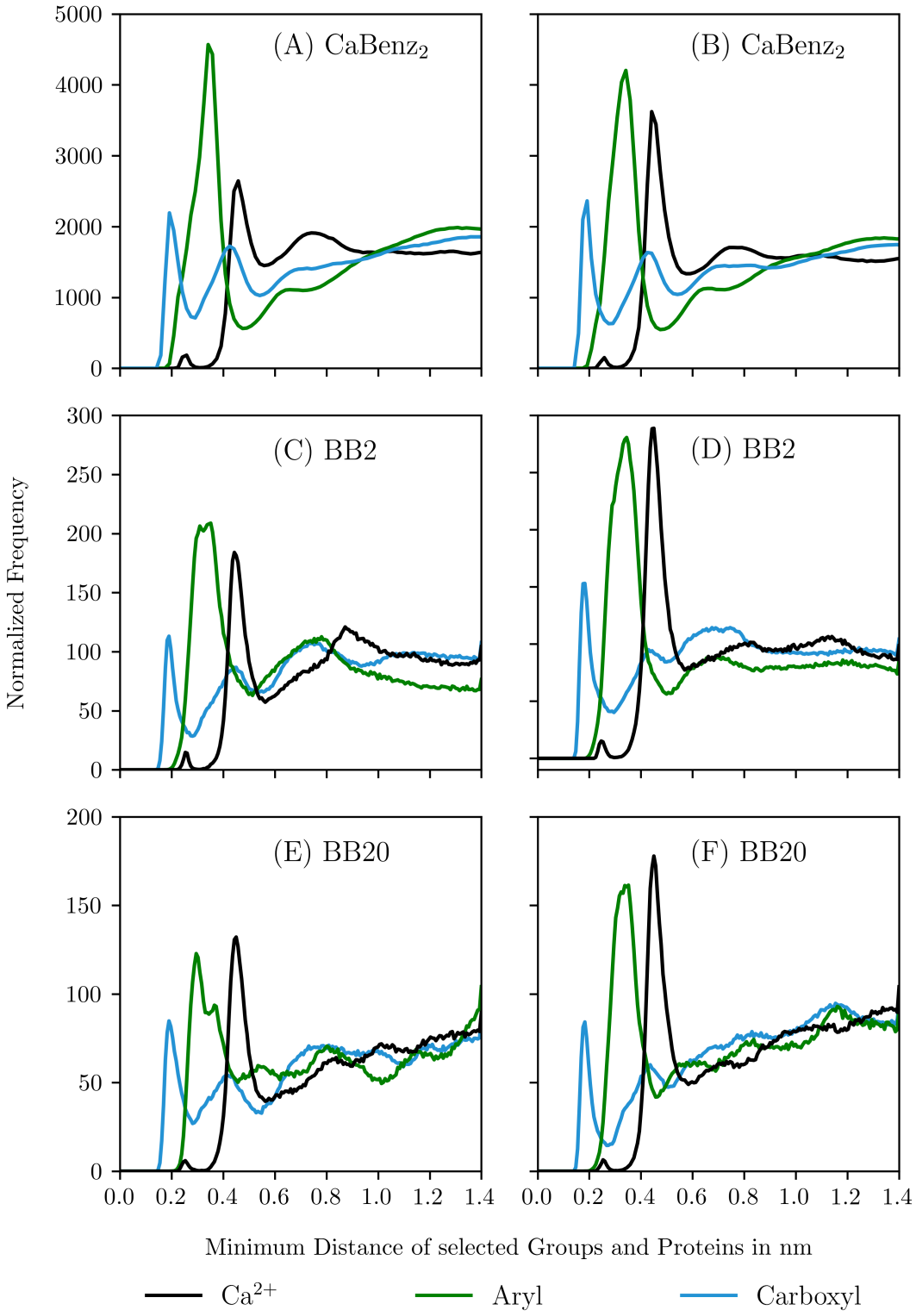

Figure 5.9: A comparison of MDF analyses. The left column (A, C, E) refers to villin, the right (B, D, F) to spitz. The frequency is depicted in the number of snapshots sampled over all replicates of the simulation. Normalization was done over the number of ions/ functional groups present in the system.

An attraction between humic substance molecules and proteins was observed in all simulations. To quantify this behaviour a cluster analysis was done. Two molecules were considered a cluster if at least one hydrogen bond was connecting them. The hydrogen bond definition for GROMACS was used as described in Section 4.2.3.2. Figure 5.10 A depicts the number of clusters made of humic substance molecules and villin at a certain time of simulation. After an initial random distribution of molecules the number of clusters quickly decreased until a lower limit is reached where the number of clusters stayed roughly constant. Figure 5.10 B shows the average number of clusters of the last 50 ns of simulation. There was a non-linear correlation of humic substance molecule size and the number of clusters observed.

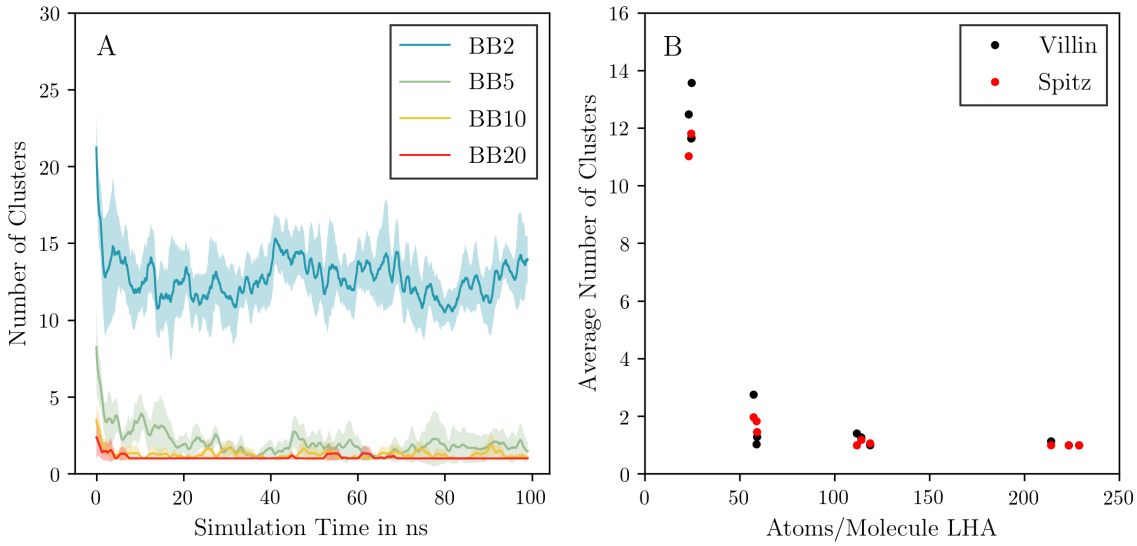

Figure 5.10: A) Running average (1 ns) time series of the number of clusters formed by humic substance molecules and villin. The colors represent the different sizes of the LHA molecules and the filled area represents the standard error of the respective three replicates. B) The average number of clusters of the last 50 ns of the simulation is plotted against the average number of united atoms per molecule for both reference proteins.

The cluster analysis showed the structure of humic acids during the simulation. They started at random positions but it is clearly visible that in the first 20 ns they associated with the protein and each other to form fewer and larger clusters. The number of clusters which is actually formed on average depends heavily on the size of LHA molecules, which can easily be explained by the fact that if there are less humic substance molecules present, less hydrogen bonds need to be formed to create one big cluster. The quick formation of protein – HS clusters indicates that proteins are likely absorbed by SOM in soil.

5.2.2.3 Hydrogen Bonds

The preliminary results with simple solvent systems indicated that hydrogen bonds were potentially formed between protein and humic substance molecules. Therefore, the hydrogen bonds were monitored over the simulation trajectories. Figure 8.6 in the appendix shows the average number of hydrogen bonds formed by both reference proteins in different VSOMM2 systems. No significant differences of the total number of hydrogen bonds between the conditions were observed. Moreover, no significant differences were measured regarding the average number of hydrogen bonds within the proteins. However, BB2 conditions showed significantly more hydrogen bonds formed with HS than BB10 and BB20 for villin and all other conditions and spitz. Figure 5.11 depicts the total non-bonded interaction energies between HS and protein against the number of hydrogen bonds formed. A linear trend with coefficients of determination of 0.83 and 0.99 for villin and spitz respectively was observed.

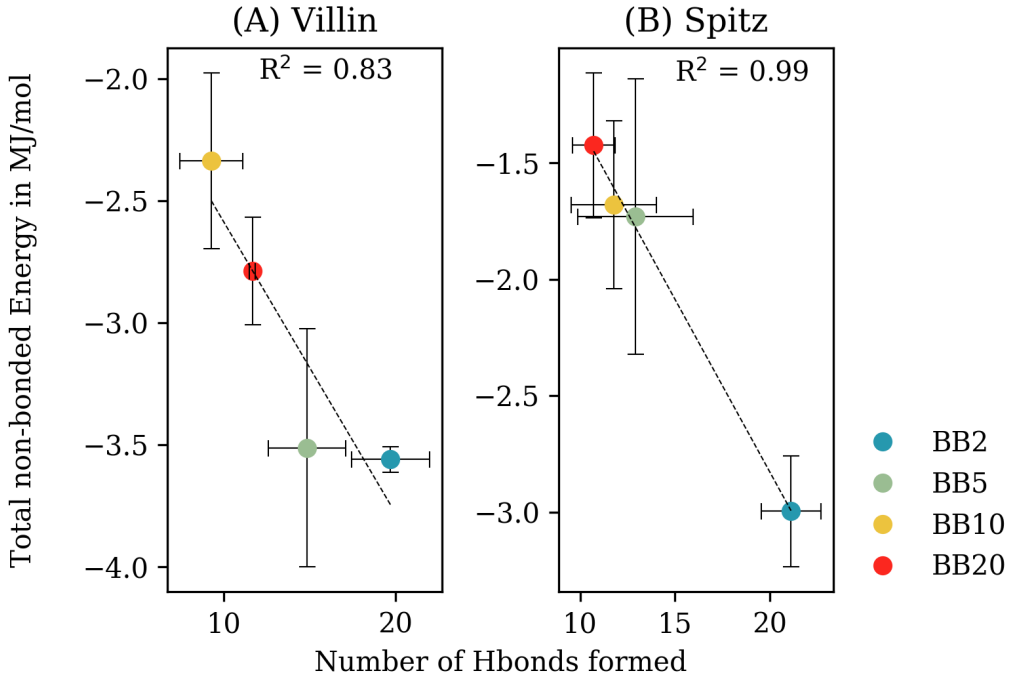

Figure 5.11: The total non-bonded interaction energies between HS and protein are plotted against the number of hydrogen bonds formed for villin (A) and spitz (B).

The increased number of hydrogen bonds matched the observations made for non-bonded interaction energies. At simulations where the LHA molecules were small, for example 2 building blocks per molecule, the multiple molecules could occupy more space around the protein and, therefore, could replace SPC water to form more hydrogen bonds. However, when the LHA molecules reached a certain length it became impossible for them to align as perfectly to the protein and, therefore, some hydrogen bonds were

rather formed by the solvent SPC water than by the LHA molecules.

5.3 Free Energy Calculations

In the previous sections the interactions of SOM with proteins was described. Interestingly the non-bonded interaction energies as well as the spacial arrangement of solvent molecules around proteins observed were very similar in most conditions, even though the solvent molecule properties were diverse. Additionally, many relevant soil proteins have an unknown structure, are very large or have multiple chains, which makes them hard to simulate with MD. Therefore, we concluded that it is more feasible to predict the strength of protein – HS interaction by investigation of interaction energies of each amino acid separately and subsequent analysis of the amino acid sequence of a protein of interest. The investigation of interaction energies of separate amino acids was done by calculation of free energies of mutation going from Ala to an arbitrary amino acid X and going from SPC water to LHA VSOMM2 sytems. The raw results of the free energy calculations are depicted in Tables 8.4 and 8.5 in the appendix. Since two different reference states (R4 and R5) were used for the OSP method they could not be directly compared. Since for most amino acids reference state R5 is more preferable, all values where R4 was more preferable were converted, which was done by using a linear equation. The equation was obtained by linear regression of the free energy calculation results of selected amino acids (Figure 5.12). The exact procedure is described in the methods chapter (Section 4.2.4.2). In the top left corner of both graphs the linear equation which was used for the conversion of free energies is shown.

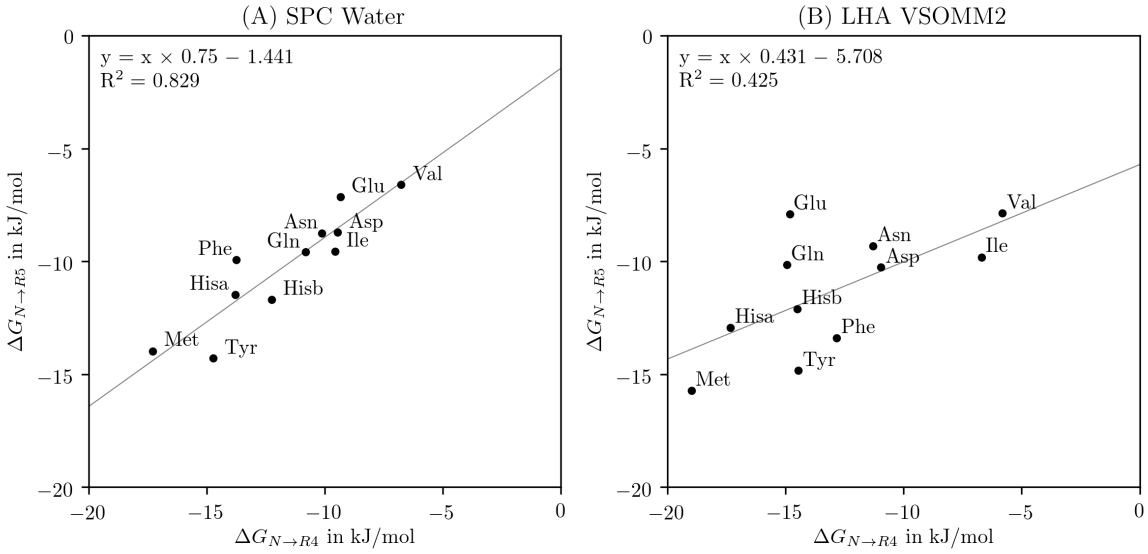

Figure 5.12: Correlation of free energies of mutation to reference states R4 and R5 in SPC water (A) and LHA VSOMM2 (B) systems. Only selected amino acids were used for the calculation of the regression line. y corresponds to $\Delta G_{N\rightarrow R5}$ and x to $\Delta G_{N\rightarrow R4}$.

In Figure 5.12 the correlation between both reference states was fitted with a linear regression. Physically, the free energy difference between two reference states should be constant, and therefore a linear regression is already overfitting the values. However, the coefficient of determination is generally low in In Figure 5.12 B.

In Figure 5.13 the end results of the free energy calculations are depicted. The range of $\Delta\Delta G_{A\to X}^{SPC\to LHA}$ was between -13.7 kJ/mol for Arg and 6.9 kJ/mol for Asp. Interestingly, positively charged amino acids like Arg and Lys had the lowest values measured, whereas negatively charged amino acids had the highest values. Polar and hydrophobic amino acids were not as clearly grouped, however, it is clearly visible that the majority of amino acid residues tested scored lower free energy values than Ala.

The low values of positively charged and high values of negatively charged amino acids made sense since previous results showed that electrostatics have a high influence on the interaction of proteins and SOM. As expected hydrogen bonds that could be formed by the negatively charged amino acids are not enough to counteract strong repulsive forces. This indicated that SOM acts mainly as hydrogen bond acceptors which in this case had little influence on a single negatively charged amino acid. It is important to note that error bars are missing in this figure. In the paper of Jandova et al.[82] the error was estimated with bootstrap replicates. However, this method could not be completely adopted for VSOMM2 systems which had two reasons. Firstly, since for the OSP method a lot of simulated frames did not contribute to the free energy (up to 99%), it takes big samples to get converged results. Additionally, when there were several SOM molecules in the system the number of frames until convergence was expected was even more increased. Secondly, there is intended variation between several VSOMM2 systems with the same input parameters. This variability cannot be reflected by creating bootstrap replicates. Therefore, to get realistic errors it is necessary to create replicates with different VSOMM2 systems, which unfortunately exceeds the scope of this master thesis. A previous study done by Moon et al. (2016) found that the relative concentration of positively charged amino acids (Arg, Lys and protonated His) increases with the age of the tested soil, whereas the relative concentration of negatively charged amino acids drops[42]. Our results could explain this observations: it is easy to imagine that positively charged amino acids which stick tighter to SOM (low values of $\Delta\Delta G_{A\to X}^{SPC\to LHA}$ in Figure 5.13) are less prone to degradation and washing out, whereas the opposite is true for negatively charge amino acids.

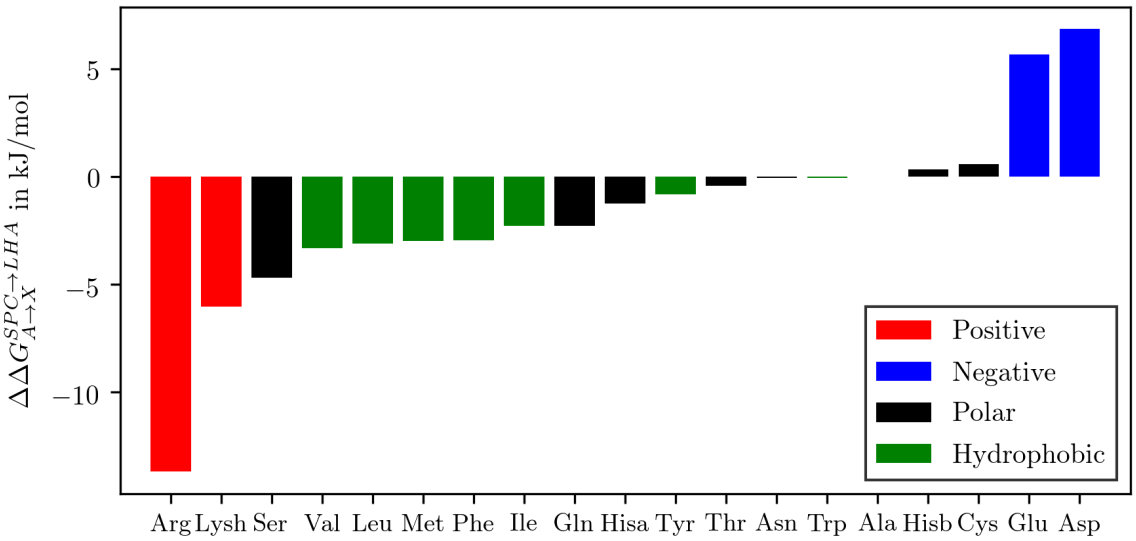

Figure 5.13: Free energies of going from Ala to a desired amino acid and going from SPC water to LHA systems. The color code represents the amino acid type.

5.4 SOMscore

With the results of the previous chapter and the solvent accessible surface area the SOMscore of proteins but also distinct amino acid residues was calculated as described in Section 4.2.4.4. The SOMscore is a scoring function which can estimate the strength of protein – humic substances interactions based on free energy differences. It is furthermore capable to predict which parts of a protein are more likely to be affected by SOM, since the score can be calculated for every amino acid separately. High scoring values would indicate that for the protein it is less favorable to be surrounded by HS compared to water, low values conversely would indicate that protein HS interactions are preferred. The free energy and maximum SASA input data for the SOMscore script as well as the code is listed in the appendix (Table 8.6 and Listing 8.6). In Figure 5.14 the surface of the scored reference proteins is shown. Negative values are colored in red whereas positive values are colored in blue. The highest occurrence of SOM molecules over a 80 ns trajectory is depicted in green. The total score of villin (A) was -1.96 kJ/mol and of spitz was -0.61 kJ/mol with a fitness of 0.77 and 0.74 respectively. It is visible that both proteins exhibited positive (blue) as well as negative (red) patches. Some strongly negative patches are deeply covered by green SOM molecules, however there are also negative patches that point to water. Conversely, several positive patches were pointing away from SOM molecules but some were still covered.

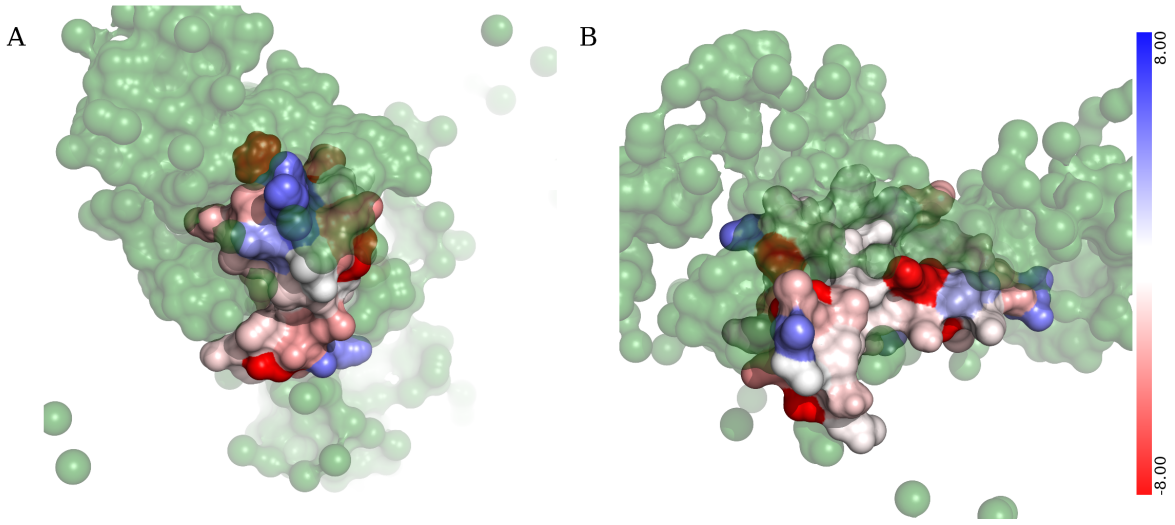

Figure 5.14: Visualization of SOMscore for villin (A) and spitz (B). Proteins were colored according to the SOMscore of each amino acid. Low values are red, high values are blue. The green volume comprises the most popular residence locations of humic substance molecules. Renderings were made with Pymol.[88]

With its implementation in a PYTHON script and the possibility to process PDB files it is an easy way to quickly score proteins. However, these estimations are still very rough with a fitness factor of 70 to 80% for most proteins. This means that between 20 to 30% of the proteins surface are not represented in the score at all. This surface area is comprised of basically two components, amino acid residues that could not be calculated with the used free energy estimation method, namely proline and the protein backbone. An additional disadvantage of this method results from the use of the solvent accessible surface area for residue influence normalization. Even though this method allows to asses if an amino acid residue is contributing to possible SOM interaction or if it is deeply buried inside the protein it makes the SOMscore very sensitive to structural changes. This can be shown by comparing the top clustered structure of villin in a VSOMM2 simulation resulting in a score of -1.96 kJ/mol whereas its PDB structure 1VII only scores at -1.60 kJ/mol. The same, however not as extreme, is true for spitz with a score of -0.61 kJ/mol and -0.54 kJ/mol respectively. For this reasons it is necessary to understand what SOMscore is capable to predict and what its limitations are in order to use it correctly.

By scoring multiple proteins of a database, for example UniProt,[110] with SOMscore it was possible to systematically investigate how selected protein groups differ from each other. In Figure 5.15 the SOMscore distribution of extracellular (red line) and cytoplasmic (black line) proteins is shown. The averages of both curves were similar (-1.37 kJ/mol and -1.33 kJ/mol respectively). Interestingly for both curves there were two peaks close to the center and additional, smaller more peripheral peaks were visible (single and double arrows in Figure 5.15 respectively). The range of SOMscores predicted went from -4.89 kJ/mol to 0.63 kJ/mol in extracellular proteins and from

-5.70 kJ/mol to 0.92 kJ/mol in cytoplasmic proteins. The range of fitness was 0.62 to 0.95 and 0.01 to 0.93 respectively. The extremely low value of fitness is an outlier, since it results from an structure were the protein is associated to DNA, which is not scored by SOMscore.

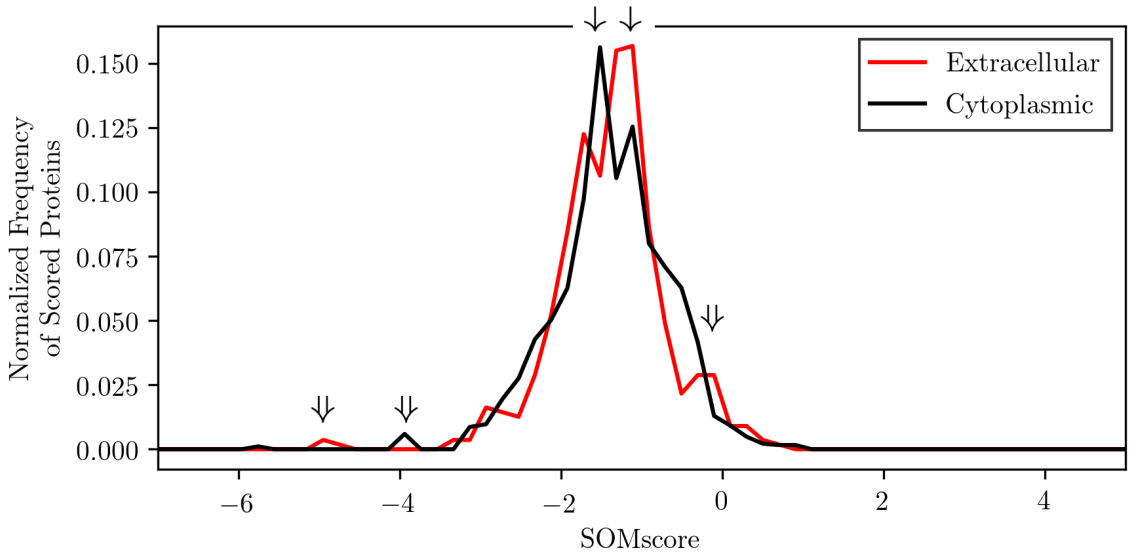

Figure 5.15: The SOMscore distribution of selected UniProtKB proteins. Black line = cytoplasmic proteins, red line = extracellular proteins. Normalization was done by the number of proteins scored. 1190 proteins were scored in total (368 extracellular, 822 cytoplasmic).

To further investigate the protein distribution the ten highest and the ten lowest scored extracellular proteins were monitored (Table 5.2). Out of the ten extracellular proteins with the lowest scores eight were derived from four potentially pathogenic organisms (*Mycobacterium tuberculosis, Streptococcus pneumoniae, Escherichia coli* and *Lecanicillium psalliotae*) which are less present in soil. Only two proteins from soil organisms were found. Contrarily, among the then highest scored proteins eight were derived from organisms that survive in soil (*Neosartorya fumigata*,[121] *Aspergillus niger*,[122] *Saccharomyces cerevisiae*[123], *Actinomadura sp.*,[124] *Bacillus subtilis*[125] and *Salipaludibacillus agaradhaerens*[126]). Additionally, six proteins out of the eight soil organisms were enzymes. The single highest and lowest scored proteins are depicted in Figure 5.16.

Table 5.2: List of the ten lowest and highest scored proteins of the UniProtKB search of extracellular proteins. The SOMscore is given in kJ/mol.

Protein Type	PDB Code	Organism	SOM Score	Fitness	Ref.
Signaling Protein	6CJ8	*Streptococcus pneumoniae*	-4.89	0.86	[127]
Signaling Protein	6COT	*Streptococcus pneumoniae*	-4.80	0.86	[127]
Signaling Protein	2A1C	*Streptococcus pneumoniae*	-4.79	0.84	[128]
Antifungal Protein	1AFP	*Aspergillus giganteus*	-3.28	0.79	[129]
RNA binding protein	4PT4	*Mycobacterium tuberculosis*	-3.23	0.74	[130]
Hydrolase	4CGE	*Mycobacterium tuberculosis*	-3.13	0.70	[131]
Toxin	2F1N	*Escherichia coli*	-3.00	0.72	[132]
Thiol Peroxidase	1XVQ	*Mycobacterium tuberculosis*	-3.00	0.72	[133]
Ribonuclease	1BUJ	*Bacillus intermedius*	-2.87	0.75	[134]
Protease	3F7M	*Lecanicillium psalliotae*	-2.87	0.71	[135]
Protein G	1EM7	*Streptococcus sp.*	-0.01	0.80	[136]
Elastase Inhibitor	3W0D	*Neosartorya fumigata*	-0.01	0.76	[137]
Ferulic Acid Esterase	1USW	*Aspergillus niger*	-0.01	0.74	[138]
Xylanase	2UWF	*Bacillus halodruans*	0.00	0.77	[139]
Endopolygalacturonase	1NHC	*Aspergillus niger*	0.01	0.77	[140]
Pathogen Related Protein	5JYS	*Saccharomyces cerevisiae*	0.11	0.70	[141]
Peptidase	1W79	*Actinomadura sp.*	0.27	0.69	[142]
Glutamyl Transferase	2V36	*Bacillus subtilis*	0.33	0.72	[143]
Pathogen Related Protein	3Q4H	*Mycolicibacterium segmantis*	0.59	0.75	[144]
Endoglucanase	1A3H	*Salipaludibacillus agaradhaerens*	0.63	0.75	[145]

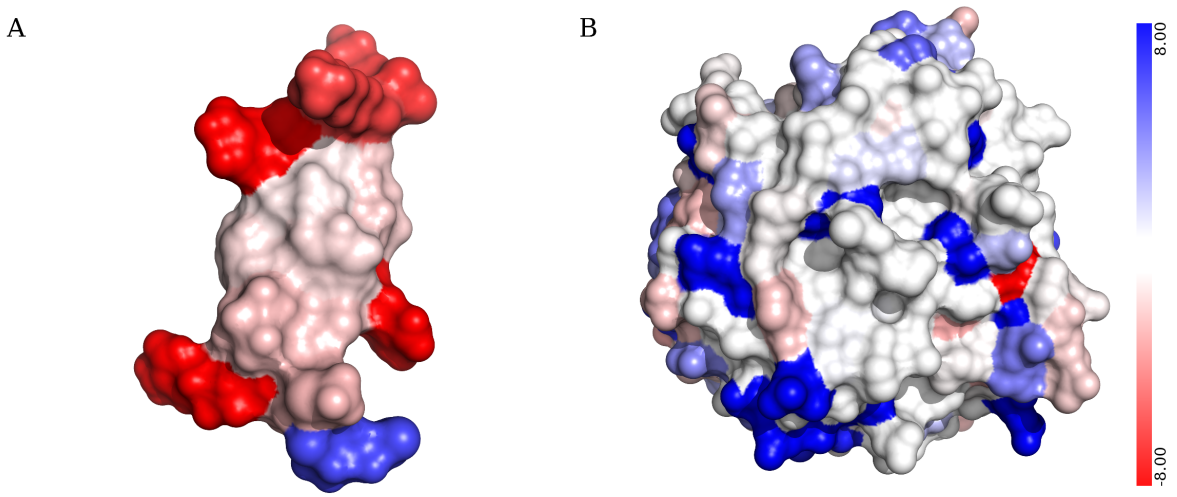

Figure 5.16: Visualization of SOMscore for to highest and lowest ranked protein of the UniProtKB analysis of extracellular proteins 6CJ8 (A) and 1A3H (B). Proteins were colored according to the SOMscore of each amino acid. Low values are red, high values are blue. Renderings were made with Pymol[88].

The UniProtKB analysis had no statistical significance. The average SOMscore of extracellular and cytoplasmic proteins are very similar. However, this did not come as a surprise. Since proteins have so many different functions, it is not hard to imagine that there are proteins that evolved to associate to SOM whereas others did not. Taking a very broad approach and averaging over all extracellular proteins, therefore, might be too coarse to elucidate the properties of different protein families. What is striking, however, is the fact that almost all proteins scored negative SOMscores, which means that almost all tested proteins tend to get absorbed by SOM. Additionally, the study of highest and lowest scored proteins indicated that enzymes connected to biomass degradation score high, thus interact less with SOM. This would make evolutionary sense since the proteins might not be able to show full enzymatic activity if they are completely associated with and immobilized by SOM, which has been shown for laccase and peroxidase.[44] Contrarily, most of the lowest scoring proteins were either no enzymes or found in organisms which usually live in different environments. However, this analysis is very speculative as just twenty proteins were examined. To get statistical evidence further investigations need to be done, which unfortunately exceeds the scope of this master thesis. The SOMscore PYTHON script which is provided in the appendix (Listing 8.6) is an easy tool to do so.

6. Conclusion

6.1 Simulation of Proteins in SOM models

Experimental data suggested that protein – humic substance interaction is particularly governed by electrostatic interactions with weak contributions of van der Waals forces.[28] however, the exact molecular mechanisms stayed unknown. In this work we investigated the influence of two major functional groups of humic substances (carboxyl and aryl) on the stability and interaction with proteins. With simple solvent systems we found that both groups have different effects on protein stability. Whereas carboxyl groups stabilized the hydrophobic interactions within the proteins, aryl groups have the potential to interfere and to significantly destabilize proteins. Additionally, we found that there is difference if aryl and carboxyl groups are linked by atomic bonds in one molecule or if they are separate. The analysis of non-bonded energies showed that Coulombic interactions are significantly higher than van der Waals interactions. We traced the root of Coulombic interactions and found that protein and carboxyl groups of solvent molecules formed hydrogen bonds. However, they did not decrease the number of hydrogen bonds the proteins formed with themselves. Additionally, we found that even though high concentrations of benzene molecules are present in simulations, carboxyl groups stay close to the protein. The secondary structure of proteins was not significantly disrupted by the addition of several solvent molecules.

By introducing proteins into SOM models made by VSOMM2 we observed an association of HS molecules and proteins to clusters. This could suggest that soil proteins are absorbed by clusters of humic substances. The formation of clusters did not differ depending on the net charge of the protein. The non-bonded interaction energies in this more complex systems still were governed by Coulombic interactions agreeing with what was found experimentally. Interestingly, differences in the van der Waals energies experienced by the protein were found when the size of HS molecules changed. However, taking also electrostatic energies into account no significance was left. We verified that also for this more complex SOM models multiple hydrogen bonds were formed between protein and HS molecules. Therefore, we concluded that simple solvent systems can resemble more complex models when it comes to protein – SOM interactions.

We tested multiple different conditions and observed similar results regarding the spacial arrangement of solvent molecules and non-bonded interaction energies, therefore, we concluded that even though soil and HS are very variable, it is possible to assume that

proteins in SOM experience very similar interactions with their environment. However, not only soil is very variable, also proteins are. We subsequently used this observation to create a scoring method which is able to predict the interaction strength of proteins in SOM.

6.2 SOMscore

To measure how each of the proteinogenic amino acids (minus glycine and proline) drive protein – humic substance interactions free energies were calculated. We applied a combinatorial approach using two different methods, TPF and OSP, analogously to Jandova et al.[82] The resulting free energy of mutation of going from alanine to an arbitrary amino acid and going from SPC water to an LHA VSOMM2 model showed that postively charged amino acids increased protein – humic substance interactions, whereas negative charges decreased them. Polar and hydrophobic amino acids scored between both extremes. This information was subsequently used to create a scoring method, SOMscore. In this work we publish the PYTHON3 code for scoring proteins with SOMscore, the only additional input needed is a PDB file of the three dimensional structure of a protein of interest. The script outputs not only a total score of protein but also writes out the score for each amino acid, therefore, it can elucidate which part of a protein is more likely to interact with humic substances than another.

We subsequently scored two sets of proteins (extracellular and cytoplasmic) containing more than 1000 proteins combined. This analysis showed that averaging large numbers of proteins lead to very similar SOMscore distributions. However, by examining the edges of the distribution of extracellular proteins we showed that enzymes produced by organisms that are commonly found in soil tend to have high scores. We suggest that this property could have evolved in order for enzymes to stay in solution and to be better protected against absorption to and immobilization by soil organic matter.

In conclusion, SOMscore (and its PYTHON3 script) was shown to be a powerful tool that can be used for the quick analysis of many proteins. We hope that this lays a good foundation for further studies regarding the interaction of proteins and humic substances. It could especially help to investigate big proteins and protein complexes, like the Cry toxins, and explain why they associate so closely to humic substances without losing their activity, whereas enzymes like laccase and peroxidase are less functional upon absorption. In addition, with SOMscore enzyme candidates for bioremediation can be assessed and possible strengths or weaknesses can be elucidated.

7. References

[1] Wulf Amelung, Hans-Peter Blume, Heiner Fleige, Rainer Horn, Ellen Kandeler, Ingrid Kögel-Knabner, Ruben Kretzschmar, Karl Stahr, and Berndt-Michael Wilke. *Scheffer/Schachtschabel Lehrbuch der Bodenkunde*. Springer-Verlag, 2018.

[2] Walter L Kubiëna et al. *Entwicklungslehre des Bodens*. Springer-Verlag, 1948.

[3] Winfried EH Blum. Boden und Klimawandel. In *Boden und globaler Wandel*, pages 95–101. Springer, 2019.

[4] Iain M Young and John W Crawford. Interactions and self-organization in the soil-microbe complex. *Science*, 304(5677):1634–1637, 2004.

[5] Heide Gotsmy. Einfluß von Ackeraufforstungen auf Bodenentwicklung, Kohlenstoffhaushalt und Mineralstoffhaushalt am Beispiel niederösterreichischer Windschutzanlagen, 1997.

[6] Nyle C Brady, Ray R Weil, and Ray R Weil. *The nature and properties of soils*, volume 13. Prentice Hall Upper Saddle River, NJ, 2008.

[7] Donald L Sparks. *Environmental soil chemistry*. Elsevier, 2003.

[8] AP Edwards and JM Bremner. Use of sonic vibration for separation of soil particles. *Canadian Journal of Soil Science*, 44(3):366–366, 1964.

[9] Carol Jean Bronick and Rattan Lal. Soil structure and management: a review. *Geoderma*, 124(1-2):3–22, 2005.

[10] Kai Uwe Totsche, Wulf Amelung, Martin H Gerzabek, Georg Guggenberger, Erwin Klumpp, Claudia Knief, Eva Lehndorff, Robert Mikutta, Stephan Peth, Alexander Prechtel, et al. Microaggregates in soils. *Journal of Plant Nutrition and Soil Science*, 181(1):104–136, 2018.

[11] I Koegel-Knaber and M Kleber. Mineralogical, physicochemical and microbiological controls on soil organic matter stabilization and turnover. *Resource Management and Environmental Impacts*, 2011.

[12] P Ciais, C Sabine, G Bala, L Bopp, V Brovkin, J Canadell, A Chhabra, R DeFries, J Galloway, M Heimann, et al. Carbon and other biogeochemical cycles. climate change 2013: the physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change. *Cambridge University Press Cambridge United Kingdom and New York NY USA*, pages 465–570, 2013.

[13] Johannes Lehmann and Markus Kleber. The contentious nature of soil organic matter. *Nature*, 528(7580):60–68, 2015.

[14] FK Achard. Chemische Untersuchung des Torfs. *Crell's Chem. Ann*, 2:391–403, 1786.

[15] P MacCarthy, RL Malcolm, CE Clapp, and PR Bloom. An introduction to soil humic substances. *Humic substances in soil and crop sciences: Selected readings*, pages 1–12, 1990.

[16] Jörg Gerke. Concepts and misconceptions of humic substances as the stable part of soil organic matter: A review. *Agronomy*, 8(5):76, 2018.

[17] Patrick MacCarthy. The principles of humic substances. *Soil Science*, 166(11):738–751, 2001.

[18] Konrad Haider. Problems related to the humification processes in soils of temperate climates. *Soil biochemistry*, 7:55–94, 1992.

[19] HR Schulten and M Schnitzer. A state of the art structural concept for humic substances. *Naturwissenschaften*, 80(1):29–30, 1993.

[20] RA Alvarez-Puebla, C Valenzuela-Calahorro, and JJ Garrido. Theoretical study on fulvic acid structure, conformation and aggregation: a molecular modelling approach. *Science of the total environment*, 358(1-3):243–254, 2006.

[21] Lawrence T Sein, James M Varnum, and Susan A Jansen. Conformational modeling of a new building block of humic acid: Approaches to the lowest energy conformer. *Environmental science & technology*, 33(4):546–552, 1999.

[22] Christian Nyrop Albers, Gary Thomas Banta, OS Jacobsen, and Poul Erik Hansen. Characterization and structural modelling of humic substances in field soil displaying significant differences from previously proposed structures. *European Journal of Soil Science*, 59(4):693–705, 2008.

[23] Axel Sündermann, Roland Solc, Daniel Tunega, Georg Haberhauer, Martin H Gerzabek, and Chris Oostenbrink. Vienna soil-organic-matter modeler—generating condensed-phase models of humic substances. *Journal of Molecular Graphics and Modelling*, 62:253–261, 2015.

[24] IHSS samples - a retrospective. Viewed on 11.11.2019. URL: `http://humic-substances.org/ihss-samples-a-retrospective/`.

[25] Source materials for IHSS samples. Viewed on 11.11.2019. URL: `http://humic-substances.org/source-materials-for-ihss-samples/`.

[26] Drazen Petrov, Daniel Tunega, Martin H Gerzabek, and Chris Oostenbrink. Molecular modelling of sorption processes of a range of diverse small organic molecules in leonardite humic acid. *European Journal of Soil Science*, 2019.

[27] Drazen Petrov, Daniel Tunega, Martin H Gerzabek, and Chris Oostenbrink. Molecular dynamics simulations of the standard leonardite humic acid: Microscopic analysis of the structure and dynamics. *Environmental science & technology*, 51(10):5414–5424, 2017.

[28] Jeanne E Tomaszewski, René P Schwarzenbach, and Michael Sander. Protein encapsulation by humic substances. *Environmental science & technology*, 45(14):6003–6010, 2011.

[29] Michael Sander, Jeanne E Tomaszewski, Michael Madliger, and René P Schwarzenbach. Adsorption of insecticidal cry1ab protein to humic substances. 1. experimental approach and mechanistic aspects. *Environmental science & technology*, 46(18):9923–9931, 2012.

[30] Jeanne E Tomaszewski, Michael Madliger, Joel A Pedersen, René P Schwarzenbach, and Michael Sander. Adsorption of insecticidal cry1ab protein to humic substances. 2. influence of humic and fulvic acid charge and polarity characteristics. *Environmental science & technology*, 46(18):9932–9940, 2012.

[31] Gabriele Giachin, Ridvan Nepravishta, Walter Mandaliti, Sonia Melino, Alja Margon, Denis Scaini, Pierluigi Mazzei, Alessandro Piccolo, Giuseppe Legname, Maurizio Paci, et al. The mechanisms of humic substances self-assembly with biological molecules: The case study of the prion protein. *PloS one*, 12(11):e0188308, 2017.

[32] Kevin A Thorn, Daniel W Folan, and Patrick MacCarthy. Characterization of the international humic substances society standard and reference fulvic and humic acids by solution state carbon-13 (13c) and hydrogen-1 (1h) nuclear magnetic resonance spectrometry. *Water-Resources Investigations Report*, 89(4196):1–93, 1989.

[33] Hussain Masoom, Denis Courtier-Murias, Hashim Farooq, Ronald Soong, Brian P Kelleher, Chao Zhang, Werner E Maas, Michael Fey, Rajeev Kumar, Martine Monette, et al. Soil organic matter in its native state: unravelling the most complex biomaterial on earth. *Environmental science & technology*, 50(4):1670–1680, 2016.

[34] Alexander Zherebker, Irina V Perminova, Yury Kostyukevich, Alexey S Kononikhin, Oleg Kharybin, and Eugene Nikolaev. Structural investigation of coal humic substances by selective isotopic exchange and high-resolution mass spectrometry. *Faraday discussions*, 2019.

[35] WM Davis, CL Erickson, CT Johnston, JJ Delfino, and JE Porter. Quantitative fourier transform infrared spectroscopic investigation humic substance functional group composition. *Chemosphere*, 38(12):2913–2928, 1999.

[36] Olga Bezuglova. Molecular structure of humus acids in soils. *Journal of Plant Nutrition and Soil Science*, 182(4):676–682, 2019.

[37] Begoña Mayans, Javier Pérez-Esteban, Consuelo Escolástico, Enrique Eymar, and Alberto Masaguer. Evaluation of commercial humic substances and other organic amendments for the immobilization of copper through 13C CPMAS NMR, FT-IR, and DSC analyses. *Agronomy*, 9(11):762, 2019.

[38] Daniel Tunega, Martin H Gerzabek, Georg Haberhauer, Hans Lischka, Roland Solc, and Adelia JA Aquino. Adsorption process of polar and nonpolar compounds in a nanopore model of humic substances. *European Journal of Soil Science*, 2019.

[39] Hongru Feng, Haiyan Zhang, Huiming Cao, Yuzhen Sun, Aiqian Zhang, and Jianjie Fu. Application of a novel coarse-grained soil organic matter model in the environment. *Environmental science & technology*, 52(24):14228–14234, 2018.

[40] Yuzhen Liang, Yang Ding, Pei Wang, Guining Lu, Zhi Dang, and Zhenqing Shi. Molecular characteristics, proton dissociation properties, and metal binding properties of soil organic matter: A theoretical study. *Science of The Total Environment*, 656:521–530, 2019.

[41] Xu Zang, Jasper DH van Heemst, Karl J Dria, and Patrick G Hatcher. Encapsulation of protein in humic acid from a histosol as an explanation for the occurrence of organic nitrogen in soil and sediment. *Organic Geochemistry*, 31(7-8):679–695, 2000.

[42] Jinyoung Moon, Li Ma, Kang Xia, and Mark A Williams. Plant–microbial and mineral contributions to amino acid and protein organic matter accumulation during 4000 years of pedogenesis. *Soil Biology and Biochemistry*, 100:42–50, 2016.

[43] L Verma, JP Martin, and K Haider. Decomposition of carbon-14-labeled proteins, peptides, and amino acids; free and complexed with humic polymers 1. *Soil Science Society of America Journal*, 39(2):279–284, 1975.

[44] Liliana Gianfreda and Jean-Marc Bollag. Effect of soils on the behavior of immobilized enzymes. *Soil Science Society of America Journal*, 58(6):1672–1681, 1994.

[45] Matthias C Rillig, Bruce A Caldwell, Han AB Wösten, and Philip Sollins. Role of proteins in soil carbon and nitrogen storage: controls on persistence. *Biogeochemistry*, 85(1):25–44, 2007.

[46] Clelia De-la Peña and Jorge M Vivanco. Proteins in the rhizosphere: Another example of plant-microbe exchange. *Ecological Aspects of Nitrogen Metabolism in Plants*, pages 95–116, 2011.

[47] MJ Haddad and Dibyendu Sarkar. Glomalin, a newly discovered component of soil organic matter: Part i—environmental significance. *Environmental Geosciences*, 10(3):91–98, 2003.

[48] Finn L Aachmann, Morten Sørlie, Gudmund Skjåk-Bræk, Vincent GH Eijsink, and Gustav Vaaje-Kolstad. NMR structure of a lytic polysaccharide monooxygenase provides insight into copper binding, protein dynamics, and substrate interactions. *Proceedings of the National Academy of Sciences*, 109(46):18779–18784, 2012.

[49] R Jason Quinlan, Matt D Sweeney, Leila Lo Leggio, Harm Otten, Jens-Christian N Poulsen, Katja Salomon Johansen, Kristian BRM Krogh, Christian Isak Jørgensen, Morten Tovborg, Annika Anthonsen, et al. Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components. *Proceedings of the National Academy of Sciences*, 108(37):15079–15084, 2011.

[50] Jane W Agger, Trine Isaksen, Anikó Várnai, Silvia Vidal-Melgosa, William GT Willats, Roland Ludwig, Svein J Horn, Vincent GH Eijsink, and Bjørge Westereng. Discovery of LPMO activity on hemicelluloses shows the importance of oxidative processes in plant cell wall degradation. *Proceedings of the National Academy of Sciences*, 111(17):6287–6292, 2014.

[51] Clive James. *Global status of commercialized biotech/GM crops, 2011*, volume 44. ISAAA Ithaca, NY, 2011.

[52] Liliana Pardo-Lopez, Mario Soberon, and Alejandra Bravo. Bacillus thuringiensis insecticidal three-domain cry toxins: mode of action, insect resistance and consequences for crop protection. *FEMS microbiology reviews*, 37(1):3–22, 2013.

[53] Susanne Baumgarte and Christoph C Tebbe. Field studies on the environmental fate of the cry1ab bt-toxin produced by transgenic maize (mon810) and its effect on bacterial communities in the maize rhizosphere. *Molecular Ecology*, 14(8):2539–2551, 2005.

[54] Cunxi Wang, Wenze Li, Colton R Kessenich, Jay S Petrick, Timothy J Rydel, Eric J Sturman, Thomas C Lee, Kevin C Glenn, and Thomas C Edrington. Safety of the bacillus thuringiensis-derived cry1a. 105 protein: Evidence that domain exchange preserves mode of action and safety. *Regulatory Toxicology and Pharmacology*, 99:50–60, 2018.

[55] Muhammad Imran and Saqib Mahmood. An overview of animal prion diseases. *Virology journal*, 8(1):493, 2011.

[56] Gabriele Giachin, Joanna Narkiewicz, Denis Scaini, Ai Tran Ngoc, Alja Margon, Paolo Sequi, Liviana Leita, and Giuseppe Legname. Prion protein interaction with soil humic substances: environmental implications. *PloS one*, 9(6):e100016, 2014.

[57] Alsu Kuznetsova, Catherine Cullingham, Debbie McKenzie, and Judd M Aiken. Soil humic acids degrade cwd prions and reduce infectivity. *PLoS pathogens*, 14(11):e1007414, 2018.

[58] Giovanni Spagnolli, Marta Rigoli, Simone Orioli, Alejandro M Sevillano, Pietro Faccioli, Holger Wille, Emiliano Biasini, and Jesus R Requena. Full atomistic model of prion structure and conversion. *PLoS pathogens*, 15(7), 2019.

[59] Chi-Yuan Fan and S Krishnamurthy. Enzymes for enhancing bioremediation of petroleum-contaminated soils: a brief review. *Journal of the Air & Waste Management Association*, 45(6):453–460, 1995.

[60] MA Rao, R Scelza, R Scotti, and L Gianfreda. Role of enzymes in the remediation of polluted environments. *Journal of soil science and plant nutrition*, 10(3):333–353, 2010.

[61] Chandrakant S Karigar and Shwetha S Rao. Role of microbial enzymes in the bioremediation of pollutants: a review. *Enzyme research*, 2011, 2011.

[62] R Boopathy. Factors limiting bioremediation technologies. *Bioresource technology*, 74(1):63–67, 2000.

[63] Mark Tuckerman. *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2010.

[64] Max Born and Robert Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[65] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[66] Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya, Sibsankar Kundu, Shijun Zhong, Jihyun Shim, Eva Darian, Olgun Guvench, P Lopes, Igor Vorobyov, et al. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of computational chemistry*, 31(4):671–690, 2010.

[67] Nathan Schmid, Andreas P Eichenberger, Alexandra Choutko, Sereina Riniker, Moritz Winger, Alan E Mark, and Wilfred F van Gunsteren. Definition and testing of the GROMOS force-field versions 54a7 and 54b7. *European biophysics journal*, 40(7):843, 2011.

[68] Maria M Reif, Philippe H Hünenberger, and Chris Oostenbrink. New interaction parameters for charged amino acid side chains in the GROMOS force field. *Journal of chemical theory and computation*, 8(10):3705–3723, 2012.

[69] J.W Eastwood R.W Hockney. *Computer simulation using particles*. A. Hilger, 1988.

[70] Wilfred F Van Gunsteren and Herman JC Berendsen. A leap-frog algorithm for stochastic dynamics. *Molecular Simulation*, 1(3):173–185, 1988.

[71] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[72] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics*, 23(3):327–341, 1977.

[73] Josiah Willard Gibbs. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundation of thermodynamics*. C. Scribner's sons, 1902.

[74] Bhalachandra L Tembre and J Andrew Mc Cammon. Ligand-receptor interactions. *Computers & Chemistry*, 8(4):281–283, 1984.

[75] Robert W Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954.

[76] PT Ehrenfest. Begriffliche Grundlagen der statistischen Auffassung in der Mechanik. In *Mechanik*, pages 773–860. Springer, 1907.

[77] Haiyan Liu, Alan E Mark, and Wilfred F van Gunsteren. Estimating the relative free energy of different molecular states with respect to a single reference state. *The Journal of Physical Chemistry*, 100(22):9485–9494, 1996.

[78] Heiko Schäfer, Wilfred F Van Gunsteren, and Alan E Mark. Estimating relative free energies from a single ensemble: hydration free energies. *Journal of computational chemistry*, 20(15):1604–1617, 1999.

[79] Chris Oostenbrink and Wilfred F Van Gunsteren. Single-step perturbations to calculate free energy differences from unphysical reference states: Limits on size, flexibility, and character. *Journal of computational chemistry*, 24(14):1730–1739, 2003.

[80] Anita de Ruiter and Chris Oostenbrink. Efficient and accurate free energy calculations on trypsin inhibitors. *Journal of chemical theory and computation*, 8(10):3686–3695, 2012.

[81] Christoph Öhlknecht, Bettina Lier, Drazen Petrov, Julian Fuchs, and Chris Oostenbrink. Correcting electrostatic artifacts due to net-charge changes in the calculation of ligand binding free energies. *Journal of Computational Chemistry*, 2020.

[82] Zuzana Jandova, Daniel Fast, Martina Setz, Maria Pechlaner, and Chris Oostenbrink. Saturation mutagenesis by efficient free-energy calculation. *Journal of chemical theory and computation*, 14(2):894–904, 2018.

[83] Martina Setz. Molecular dynamics simulations of biomolecules : from validation to application, May 2018.

[84] Drazen Petrov and Bojan Zagrovic. Are current atomistic force fields accurate enough to study proteins in crowded environments? *PLoS computational biology*, 10(5):e1003638, 2014.

[85] C James McKnight, Paul T Matsudaira, and Peter S Kim. NMR structure of the 35-residue villin headpiece subdomain. *Nature structural biology*, 4(3):180–184, 1997.

[86] E Wieschaus, Ch Nüsslein-Volhard, and Gerd Jürgens. Mutations affecting the pattern of the larval cuticle indrosophila melanogaster. *Wilhelm Roux's archives of developmental biology*, 193(5):296–307, 1984.

[87] Daryl E Klein, Steven E Stayrook, Fumin Shi, Kartik Narayan, and Mark A Lemmon. Structural basis for egfr ligand sequestration by argos. *Nature*, 453(7199):1271, 2008.

[88] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[89] Alpeshkumar K Malde, Le Zuo, Matthew Breeze, Martin Stroet, David Poger, Pramod C Nair, Chris Oostenbrink, and Alan E Mark. An automated force field topology builder (ATB) and repository: version 1.0. *Journal of chemical theory and computation*, 7(12):4026–4037, 2011.

[90] Martin Stroet, Bertrand Caron, Koen M Visscher, Daan P Geerke, Alpeshkumar K Malde, and Alan E Mark. Automated topology builder version 3.0: Prediction of solvation free enthalpies in water and hexane. *Journal of chemical theory and computation*, 14(11):5834–5845, 2018.

[91] Nathan Schmid, Clara D Christ, Markus Christen, Andreas P Eichenberger, and Wilfred F van Gunsteren. Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation. *Computer Physics Communications*, 183(4):890–903, 2012.

[92] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[93] Herman JC Berendsen, James PM Postma, Wilfred F van Gunsteren, and Jan Hermans. Interaction models for water in relation to protein hydration. In *Intermolecular forces*, pages 331–342. Springer, 1981.

[94] Shuichi Miyamoto and Peter A Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry*, 13(8):952–962, 1992.

[95] Christian Kandt, Walter L Ash, and D Peter Tieleman. Setting up and running molecular dynamics simulations of membrane proteins. *Methods*, 41(4):475–488, 2007.

[96] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[97] Szilárd Pall, Mark James Abraham, Carsten Kutzner, Berk Hess, and Erik Lindahl. Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In *International Conference on Exascale Applications and Software*, pages 3–27. Springer, 2014.

[98] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David Van Der Spoel, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[99] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, 4(3):435–447, 2008.

[100] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. GROMACS: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.

[101] Erik Lindahl, Berk Hess, and David Van Der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*, 7(8):306–317, 2001.

[102] Herman JC Berendsen, David van der Spoel, and Rudi van Drunen. GROMACS: a message-passing parallel molecular dynamics implementation. *Computer physics communications*, 91(1-3):43–56, 1995.

[103] Lindahl, Abraham, Hess, and van der Spoel. Gromacs 2019.1 source code, February 2019. URL: `https://doi.org/10.5281/zenodo.2564764`, `doi:10.5281/zenodo.2564764`.

[104] A Bondi. van der Waals volumes and radii. *The Journal of physical chemistry*, 68(3):441–451, 1964.

[105] Andreas P Eichenberger, Jane R Allison, Jozica Dolenc, Daan P Geerke, Bruno AC Horta, Katharina Meier, Chris Oostenbrink, Nathan Schmid, Denise Steiner, Dongqi Wang, et al. GROMOS++ software for the analysis of biomolecular simulation trajectories. *Journal of chemical theory and computation*, 7(10):3379–3390, 2011.

[106] Byungkook Lee and Frederic M Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4, 1971.

[107] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[108] Ben Widom. Some topics in the theory of fluids. *The Journal of Chemical Physics*, 39(11):2808–2812, 1963.

[109] Simon Mitternacht. Freesasa: An open source c library for solvent accessible surface area calculations. *F1000Research*, 5, 2016.

[110] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018. URL: `https://doi.org/10.1093/nar/gky1049`, `arXiv:https://academic.oup.com/nar/article-pdf/47/D1/D506/27437297/gky1049.pdf`, `doi:10.1093/nar/gky1049`.

[111] Hongxing Lei, Chun Wu, Haiguang Liu, and Yong Duan. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 104(12):4925–4930, 2007.

[112] Yong Duan, Lu Wang, and Peter A Kollman. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proceedings of the National Academy of Sciences*, 95(17):9897–9902, 1998.

[113] Matthias Diem and Chris Oostenbrink. Hamiltonian reweighing to refine protein backbone dihedral-angle parameters in the GROMOS force field. *Journal of Chemical Information and Modeling*, 2019.

[114] Xiang Xu, Andrey G Kalinichev, and R James Kirkpatrick. 133cs and 35cl NMR spectroscopy and molecular dynamics modeling of cs+ and cl- complexation with natural organic matter. *Geochimica et Cosmochimica Acta*, 70(17):4319–4331, 2006.

[115] Eugenia Iskrenova-Tchoukova, Andrey G Kalinichev, and R James Kirkpatrick. Metal cation complexation with natural organic matter in aqueous solutions: molecular dynamics simulations and potentials of mean force. *Langmuir*, 26(20):15909–15919, 2010.

[116] Nanci Kloster, Maximiliano Brigante, Graciela Zanini, and Marcelo Avena. Aggregation kinetics of humic acids in the presence of calcium ions. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 427:76–82, 2013.

[117] Nico FA van der Vegt and Wilfred F van Gunsteren. Entropic contributions in cosolvent binding to hydrophobic solutes in water. *The Journal of Physical Chemistry B*, 108(3):1056–1064, 2004.

[118] A Ben-Naim and Y Marcus. Solvation thermodynamics of nonionic solutes. *The Journal of chemical physics*, 81(4):2016–2027, 1984.

[119] E Espinosa, E Molins, and C Lecomte. Hydrogen bond strengths revealed by topological analyses of experimentally observed electron densities. *Chemical Physics Letters*, 285(3-4):170–173, 1998.

[120] Tomás FD Silva, Diogo Vila-Viçosa, Pedro BPS Reis, Bruno L Victor, Matthias Diem, Chris Oostenbrink, and Miguel Machuqueiro. The impact of using single atomistic long-range cutoff schemes with the GROMOS 54a7 force field. *Journal of chemical theory and computation*, 14(11):5823–5833, 2018.

[121] Céline M O'Gorman, Hubert T Fuller, and Paul S Dyer. Discovery of a sexual cycle in the opportunistic fungal pathogen aspergillus fumigatus. *Nature*, 457(7228):471–474, 2009.

[122] Nicole Nolard, Monique Detandt, and Hugues Beguin. Ecology of aspergillus species in the human environment. In *Aspergillus and aspergillosis*, pages 35–41. Springer, 1988.

[123] Paula Jouhten, Olga Ponomarova, Ramon Gonzalez, and Kiran R Patil. Saccharomyces cerevisiae metabolism in ecological context. *FEMS yeast research*, 16(7):fow080, 2016.

[124] OS Zakharova, GM Zenova, and DG Zvyagintsev. Some approaches to the selective isolation of actinomycetes of the genus actinomadura from soil. *Microbiology*, 72(1):110–113, 2003.

[125] Ashlee M Earl, Richard Losick, and Roberto Kolter. Ecology and genomics of bacillus subtilis. *Trends in microbiology*, 16(6):269–275, 2008.

[126] Vishnuvardhan Reddy Sultanpuram and Thirumala Mothe. Salipaludibacillus aurantiacus gen. nov., sp. nov. a novel alkali tolerant bacterium, reclassification of bacillus agaradhaerens as salipaludibacillus agaradhaerens comb. nov. and bacillus neizhouensis as salipaludibacillus neizhouensis comb. nov. *International Journal of Systematic and Evolutionary Microbiology*, 66(7):2747–2753, 2016.

[127] Yifang Yang, Gabriel Cornilescu, and Yftah Tal-Gan. Structural characterization of competence-stimulating peptide analogues reveals key features for comd1 and comd2 receptor binding in streptococcus pneumoniae. *Biochemistry*, 57(36):5359–5369, 2018.

[128] 2A1C PDB entry. Viewed on 17.01.2020. URL: `https://www.rcsb.org/structure/2a1c`.

[129] Ramon Campos-Olivas, Marta Bruix, Jorge Santoro, Javier Lacadena, Alvaro Martinez del Pozo, Jose G Gavilanes, and Manuel Rico. NMR solution structure of the antifungal protein from aspergillus giganteus: evidence for cysteine pairing isomerism. *Biochemistry*, 34(9):3009–3021, 1995.

[130] Tuhin Bhowmick, Soumitra Ghosh, Karuna Dixit, Varsha Ganesan, Udupi A Ramagopal, Debayan Dey, Siddhartha P Sarma, Suryanarayanarao Ramakumar, and Valakunja Nagaraja. Targeting mycobacterium tuberculosis nucleoid-associated protein hu with structure-based inhibitors. *Nature communications*, 5:4124, 2014.

[131] Daniela Mavrici, Daniil M Prigozhin, and Tom Alber. Mycobacterium tuberculosis rpfe crystal structure reveals a positively charged catalytic cleft. *Protein science*, 23(4):481–487, 2014.

[132] Jill S Hontz, Maria T Villar-Lecumberri, Belinda M Potter, Marilyn D Yoder, Lawrence A Dreyfus, and John H Laity. Differences in crystal and solution structures of the cytolethal distending toxin b subunit relevance to nuclear translocation and functional activation. *Journal of Biological Chemistry*, 281(35):25365–25372, 2006.

[133] Beom-Seop Rho, Li-Wei Hung, James M Holton, Dominico Vigil, Su-Il Kim, Min S Park, Thomas C Terwilliger, and Jean-Denis Pédelacq. Functional and structural characterization of a thiol peroxidase from mycobacterium tuberculosis. *Journal of molecular biology*, 361(5):850–863, 2006.

[134] M Ya Reibarkh, DE Nolde, LI Vasilieva, EV Bocharov, AA Shulga, MP Kirpichnikov, and AS Arseniev. Three-dimensional structure of binase in solution. *FEBS letters*, 431(2):250–254, 1998.

[135] Chunzi Liang, Antje Bruckbauer, and Michael B Zemel. Leucine modulation of sirtuins and ampk in adipocytes and myotubes, 2012.

[136] Pavel Strop, Andrei M Marinescu, and Stephen L Mayo. Structure of a protein g helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. *Protein Science*, 9(7):1391–1394, 2000.

[137] Mayuko Sakuma, Katsumi Imada, Yoshiyuki Okumura, Kei-ichi Uchiya, Nobuo Yamashita, Kenji Ogawa, Atsushi Hijikata, Tsuyoshi Shirai, Michio Homma, and Toshiaki Nikai. X-ray structure analysis and characterization of afuei, an elastase inhibitor from aspergillus fumigatus. *Journal of Biological Chemistry*, 288(24):17451–17459, 2013.

[138] Juan A Hermoso, Julia Sanz-Aparicio, Rafael Molina, Nathalie Juge, Ramon Gonzalez, and Craig B Faulds. The crystal structure of feruloyl esterase a from aspergillus niger suggests evolutive functional convergence in feruloyl esterase family. *Journal of molecular biology*, 338(3):495–506, 2004.

[139] Gashaw Mamo, Marjolein Thunnissen, Rajni Hatti-Kaul, and Bo Mattiasson. An alkaline active xylanase: insights into mechanisms of high ph catalytic adaptation. *Biochimie*, 91(9):1187–1196, 2009.

[140] Gertie van Pouderoyen, Harm J Snijder, Jacques AE Benen, and Bauke W Dijkstra. Structural insights into the processivity of endopolygalacturonase i from aspergillus niger. *FEBS letters*, 554(3):462–466, 2003.

[141] Rabih Darwiche, Alan Kelleher, Elissa M Hudspeth, Roger Schneiter, and Oluwatoyin A Asojo. Structural and functional characterization of the cap domain of pathogen-related yeast 1 (pry1) protein. *Scientific reports*, 6:28838, 2016.

[142] Eric Sauvage, Raphaël Herman, Stephanie Petrella, Colette Duez, Fabrice Bouillenne, Jean-Marie Frère, and Paulette Charlier. Crystal structure of the actinomadura r39 dd-peptidase reveals new domains in penicillin-binding proteins. *Journal of Biological Chemistry*, 280(35):31249–31256, 2005.

[143] 2V36 PDB entry. Viewed on 26.01.2020. URL: `https://www.rcsb.org/structure/2V36`.

[144] Mark A Arbing, Sum Chan, Liam Harris, Emmeline Kuo, Tina T Zhou, Christine J Ahn, Lin Nguyen, Qixin He, Jamie Lu, Phuong T Menchavez, et al. Heterologous expression of mycobacterial esx complexes in escherichia coli for structural studies is facilitated by the use of maltose binding protein fusions. *PLoS One*, 8(11), 2013.

[145] Gideon J Davies, Miroslawa Dauter, A Marek Brzozowski, Mads Eskelund Bjørnvad, Kim V Andersen, and Martin Schülein. Structure of the bacillus agaradherans family 5 endoglucanase at 1.6 å and its cellobiose complex at 2.0 å resolution. *Biochemistry*, 37(7):1926–1932, 1998.

8. Appendix

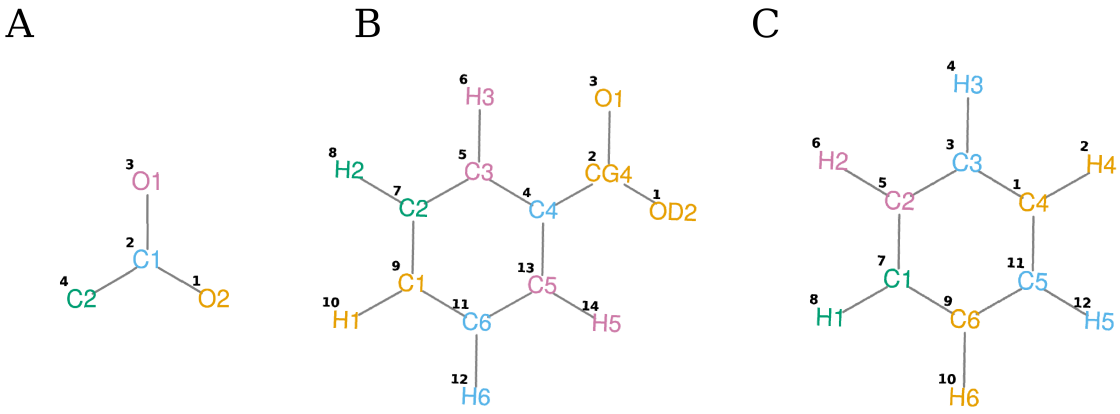

Figure 8.1: Atom nomenclature of simple solvent building blocks. (A) acetate, (B) benzoate, (C) benzene.

```
TITLE
Exported from new Building Block Editor (browser based)
END
FORCEFIELD
54A7
END
PHYSICALCONSTANTS
# FPEPSI: 1.0/(4.0*PI*EPS0); (EPS0 is the permittivity of vacuum);
  0.1389354E+03
# HBAR: Planck's constant HBAR = H/(2* PI);
  0.6350780E-01
# SPDL: Speed of light (in nm/ps);
  2.9979245800E05
# BOLTZ: Boltzmann's constant
  8.31441E-03
END
LINKEXCLUSIONS
#nearest neighbour exclusions when linking
#NRNE
    2
END
MTBUILDBLSOLUTE
# building block created using TopologyBuilder
# by
ACET
#
# number of atoms, number of preceding exclusions
# NMAT NLIN
    4    0
# preceding exclusions
#ATOM                                MAE MSAE
# atoms
#ATOM ANM  IACM MASS        CGMICGM MAE MSAE
    1 O2      2   16   -0.63500   1   3    2    3    4
    2 C1     12   12    0.27000   1   2    3    4
    3 O1      2   16   -0.63500   1   1    4
    4 C2     16    5    0.00000   1   0
# bonds
#  NB
    3
#  IB   JB  MCB
    1    2    6
    2    3    6
    2    4   27
# bond angles
# NBA
    3
#  IB   JB   KB  MCB
    1    2    3   38
    1    2    4   22
    3    2    4   22
# improper dihedrals
# NIDA
    1
```

```
#  IB    JB    KB    LB   MCB
    2     1     3     4     1
# dihedrals
# NDA
    0
#  IB    JB    KB    LB   MCB
# LJ exceptions
# NEX
    0
#  IB    JB   MCB   NCO   IND   CON
END

POSITIONS
4
    1          396.60          416.14
    2          366.38          398.49
    3          366.63          363.47
    4          335.96          415.76
END
```

Listing 8.1: Acetate 54A7 mtb file.

```
TITLE
Exported from new Building Block Editor (browser based)
END
FORCEFIELD
54A7
END
PHYSICALCONSTANTS
# FPEPSI: 1.0/(4.0*PI*EPS0); (EPS0 is the permittivity of vacuum);
  0.1389354E+03
# HBAR: Planck's constant HBAR = H/(2* PI);
  0.6350780E-01
# SPDL: Speed of light (in nm/ps);
  2.9979245800E05
# BOLTZ: Boltzmann's constant
  8.31441E-03
END
LINKEXCLUSIONS
#nearest neighbour exclusions when linking
#NRNE
    2
END
MTBUILDBLSOLUTE
# building block created using TopologyBuilder
# by
BENZ
#
# number of atoms, number of preceding exclusions
# NMAT NLIN
   14    0
# preceding exclusions
#ATOM                                MAE MSAE
# atoms
#ATOM ANM  IACM MASS        CGMICGM MAE MSAE
    1 OD2     2   16   -0.63500   0   3    2    3    4
    2 CG4    12   12    0.27000   0   8    3    4    5    6    7   11
                                          13   14
    3 O1      2   16   -0.63500   1   1    4
    4 C4     12   12    0.00000   1   9    5    6    7    8    9   11
                                          12   13   14
    5 C3     12   12   -0.14000   0   8    6    7    8    9   10   11
                                          13   14
    6 H3     20    1    0.14000   1   4    7    8    9   13
    7 C2     12   12   -0.14000   0   6    8    9   10   11   12   13
    8 H2     20    1    0.14000   1   3    9   10   11
    9 C1     12   12   -0.14000   0   5   10   11   12   13   14
   10 H1     20    1    0.14000   1   3   11   12   13
   11 C6     12   12   -0.14000   0   3   12   13   14
   12 H6     20    1    0.14000   1   2   13   14
   13 C5     12   12   -0.14000   0   1   14
   14 H5     20    1    0.14000   1   0
# bonds
#  NB
   14
#  IB   JB  MCB
```

```
    1     2     6
    2     3     6
    2     4    27
    4     5    16
    4    13    16
    5     6     3
    5     7    16
    7     8     3
    7     9    16
    9    10     3
    9    11    16
   11    12     3
   11    13    16
   13    14     3
# bond angles
# NBA
   21
#  IB    JB    KB   MCB
    1     2     3    38
    1     2     4    22
    3     2     4    22
    2     4     5    27
    2     4    13    27
    5     4    13    27
    4     5     6    25
    4     5     7    27
    6     5     7    25
    5     7     8    25
    5     7     9    27
    8     7     9    25
    7     9    10    25
    7     9    11    27
   10     9    11    25
    9    11    12    25
    9    11    13    27
   12    11    13    25
    4    13    11    27
    4    13    14    25
   11    13    14    25
# improper dihedrals
# NIDA
   13
#  IB    JB    KB    LB   MCB
    2     1     3     4     1
    4     2     5    13     1
    4     5     7     9     1
    5     4     6     7     1
    5     4    13    11     1
    5     7     9    11     1
    7     5     8     9     1
    7     9    11    13     1
    9     7    10    11     1
    9    11    13     4     1
   11     9    12    13     1
```

```
   13     4    11    14     1
   13     4     5     7     1
# dihedrals
# NDA
    1
#  IB    JB    KB    LB   MCB
    3     2     4     5    10
# LJ exceptions
# NEX
    0
#  IB    JB   MCB   NCO   IND   CON
END

POSITIONS
14
    1          257.87          386.96
    2          227.08          370.30
    3          226.16          335.31
    4          197.26          388.62
    5          166.47          371.95
    6          165.52          336.96
    7          136.65          390.27
    8          105.87          373.60
    9          137.60          425.26
   10          107.78          443.58
   11          168.39          441.93
   12          169.33          476.91
   13          198.21          423.61
   14          228.98          440.28
END
```

Listing 8.2: Benzoate 54A7 mtb file.

```
TITLE
Exported from new Building Block Editor (browser based)
END
FORCEFIELD
54A7
END
PHYSICALCONSTANTS
# FPEPSI: 1.0/(4.0*PI*EPS0); (EPS0 is the permittivity of vacuum);
  0.1389354E+03
# HBAR: Planck's constant HBAR = H/(2* PI);
  0.6350780E-01
# SPDL: Speed of light (in nm/ps);
  2.9979245800E05
# BOLTZ: Boltzmann's constant
  8.31441E-03
END
LINKEXCLUSIONS
#nearest neighbour exclusions when linking
#NRNE
    2
END
MTBUILDBLSOLUTE
# building block created using TopologyBuilder
# by
C6H6
#
# number of atoms, number of preceding exclusions
# NMAT NLIN
   12    0
# preceding exclusions
#ATOM                                MAE MSAE
# atoms
#ATOM ANM  IACM MASS        CGMICGM MAE MSAE
    1 C4     12   12   -0.14000   0  10    2    3    4    5    6    7
                                            9   10   11   12
    2 H4     20    1    0.14000   1   6    3    4    5    9   11   12
    3 C3     12   12   -0.14000   0   8    4    5    6    7    8    9
                                           11   12
    4 H3     20    1    0.14000   1   4    5    6    7   11
    5 C2     12   12   -0.14000   0   6    6    7    8    9   10   11
    6 H2     20    1    0.14000   1   3    7    8    9
    7 C1     12   12   -0.14000   0   5    8    9   10   11   12
    8 H1     20    1    0.14000   1   3    9   10   11
    9 C6     12   12   -0.14000   0   3   10   11   12
   10 H6     20    1    0.14000   1   2   11   12
   11 C5     12   12   -0.14000   0   1   12
   12 H5     20    1    0.14000   1   0
# bonds
#  NB
   12
#  IB   JB  MCB
    1    2    3
    1    3   16
    1   11   16
```

```
    3     4     3
    3     5    16
    5     6     3
    5     7    16
    7     8     3
    7     9    16
    9    10     3
    9    11    16
   11    12     3
# bond angles
# NBA
   18
#  IB    JB    KB   MCB
    2     1     3    25
    2     1    11    25
    3     1    11    27
    1     3     4    25
    1     3     5    27
    4     3     5    25
    3     5     6    25
    3     5     7    27
    6     5     7    25
    5     7     8    25
    5     7     9    27
    8     7     9    25
    7     9    10    25
    7     9    11    27
   10     9    11    25
    1    11     9    27
    1    11    12    25
    9    11    12    25
# improper dihedrals
# NIDA
   12
#  IB    JB    KB    LB   MCB
    1     3     5     7     1
    3     1     4     5     1
    3     1    11     9     1
    3     5     7     9     1
    1     3    11     2     1
    5     3     6     7     1
    5     7     9    11     1
    7     5     8     9     1
    7     9    11     1     1
    9     7    10    11     1
   11     1     9    12     1
   11     1     3     5     1
# dihedrals
# NDA
    0
#  IB    JB    KB    LB   MCB
# LJ exceptions
# NEX
    0
```

```
#   IB    JB   MCB   NCO   IND   CON
END

POSITIONS
12
     1          403.38          398.26
     2          433.20          379.94
     3          372.59          381.59
     4          371.64          346.60
     5          342.77          399.91
     6          311.99          383.24
     7          343.72          434.90
     8          313.90          453.22
     9          374.51          451.57
    10          375.45          486.55
    11          404.33          433.25
    12          435.10          449.92
END
```

Listing 8.3: Benzene 54A7 mtb file.

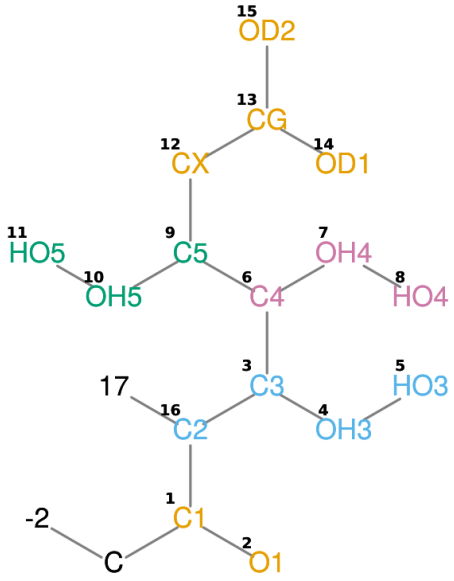

Figure 8.2: HS37 Building block.

```
TITLE
Exported from new Building Block Editor (browser based)
END
FORCEFIELD
54A8
END
PHYSICALCONSTANTS
# FPEPSI: 1.0/(4.0*PI*EPS0); (EPS0 is the permittivity of vacuum);
  0.1389354E+03
# HBAR: Planck's constant HBAR = H/(2* PI);
  0.6350780E-01
# SPDL: Speed of light (in nm/ps);
  2.9979245800E05
# BOLTZ: Boltzmann's constant
  8.31441E-03
END
LINKEXCLUSIONS
#nearest neighbour exclusions when linking
#NRNE
    2
END
MTBUILDBLSOLUTE
# building block created using TopologyBuilder
# by
HS13
#
# number of atoms, number of preceding exclusions
# NMAT NLIN
   16    1
# preceding exclusions
#ATOM                                MAE MSAE
    0                                  3    1    2   16
# atoms
#ATOM ANM  IACM MASS        CGMICGM MAE MSAE
    1 C1     12   12    0.45000   0   4    2    3   16   17
    2 O1      1   16   -0.45000   1   1   16
    3 C3     14    3    0.26600   0   7    4    5    6    7    9   16
                                          17
    4 OH3     3   16   -0.67400   0   3    5    6   16
    5 HO3    21    1    0.40800   1   0
    6 C4     14    3    0.26600   0   6    7    8    9   10   12   16
    7 OH4     3   16   -0.67400   0   2    8    9
    8 HO4    21    1    0.40800   1   0
    9 C5     14    3    0.26600   0   4   10   11   12   13
   10 OH5     3   16   -0.67400   0   2   11   12
   11 HO5    21    1    0.40800   1   0
   12 CX     15    4    0.16000   0   3   13   14   15
   13 CG     12   12    0.27000   0   2   14   15
   14 OD1     2   16   -0.71500   0   1   15
   15 OD2     2   16   -0.71500   1   0
# trailing atoms
#ATOM ANM  IACM MASS        CGMICGM
   16 C2     14    3    0.00000   1
# bonds
```

```
#  NB
   16
#  IB   JB  MCB
    1    2    5
    1   16   27
    3    4   18
    3    6    5
    3   16   27
    4    5    1
    6    7   18
    6    9   27
    7    8    1
    9   10   18
    9   12   27
   10   11    1
   12   13   27
   13   14    6
   13   15    6
   16   17   27
# bond angles
# NBA
   22
#  IB   JB   KB  MCB
    0    1    2   27
    0    1   16   27
    2    1   16   27
    4    3    6   13
    4    3   16   13
    6    3   16   13
    3    4    5   12
    3    6    7   13
    3    6    9   13
    7    6    9   13
    6    7    8   12
    6    9   10   13
    6    9   12   13
   10    9   12   13
    9   10   11   12
    9   12   13   13
   12   13   14   22
   12   13   15   22
   14   13   15   38
    1   16    3   13
    1   16   17   27
    3   16   17   27
# improper dihedrals
# NIDA
    6
#  IB   JB   KB   LB  MCB
    1    0    2   16    1
    3    4    6   16    2
    6    3    7    9    2
   12    9   10    6    5
   14   12   13   15    1
```

```
   16    1    3   17    2
# dihedrals
# NDA
   15
#  IB   JB   KB   LB  MCB
   -1    0    1   16   12
    2    1   16    3   40
    6    3    4    5   23
    4    3    6    7   34
    4    3   16    1   34
    3    6    7    8   23
    3    6    9   12   34
    7    6    9   12   34
    3    6    9   10   34
   12    9   10   11   23
    6    9   10   11   23
    6    9   12   13   34
   10    9   12   13   34
    9   12   13   14   40
    9   12   13   15   40
# LJ exceptions
# NEX
    0
#  IB   JB  MCB  NCO  IND  CON
END

XYPOSITION
19
   -1          306.55          476.08
    0          338.76          489.79
    1          366.72          468.55
    2          398.91          482.21
    3          389.94          412.40
    4          422.28          425.16
    5          449.93          403.75
    6          384.51          377.58
    7          411.46          355.64
    8          444.34          367.71
    9          351.60          365.58
   10          325.08          388.91
   11          292.07          377.34
   12          344.70          331.68
   13          369.75          307.51
   14          403.51          316.81
   15          361.21          273.41
   16          362.24          433.93
   17          329.87          420.44
END
```

Listing 8.4: HS37 54A7 mtb file.

Table 8.1: RSMD averages over the simulations and replicates in nanometers of both reference proteins in simulated conditions. Significant differences to H_2O conditions are labelled with an asterisk ($\alpha < 0.05$).

	H_2O		$CaCl_2$	$CaAc_2$	$CaBenz_2$	$CaAc_2$ + Bz	Real. Conc.
Villin	0.20	0.36	0.30	0.29	0.27	0.73	0.78
	0.36	0.67	0.28	0.27	0.25	0.83	0.69
	0.23	0.33	0.25	0.20	0.21	0.58	0.63
	0.29	0.19	0.21	0.46	0.33	0.57	0.66
Mean	0.33		0.26	0.31	0.27	0.68*	0.69*
Std. Dev.	0.15		0.03	0.10	0.04	0.11	0.06
Spitz	0.19	0.17	0.52	0.25	0.16	0.26	0.20
	0.18	0.21	0.20	0.34	0.16	0.19	0.26
	0.17	0.29	0.34	0.41	0.27	0.24	0.29
	0.17	0.17	0.19	0.19	0.21	0.27	0.29
Mean	0.19		0.31	0.30	0.20	0.24	0.26
Std. Dev.	0.04		0.13	0.08	0.05	0.03	0.04

Table 8.2: Average secondary structure occurrence of helices and β sheets for villin and spitz in %.

Helix	H_2O		$CaCl_2$	$CaAc_2$	$CaBenz_2$	$CaAc_2$+Bz	Real. Conc.
Villin	65.0	64.9	65.1	64.3	66.6	57.8	55.6
	63.3	46.7	65.5	63.7	66.6	63.9	55.5
	61.2	62.6	65.1	65.4	66.8	50.6	61.9
	61.3	50.4	65.7	62.0	66.2	49.8	53.5
Mean	59.4		65.4	63.9	66.6	55.5	56.6
Std. Dev.	6.9		0.3	1.4	0.3	6.6	3.6
Helix	H_2O		$CaCl_2$	$CaAc_2$	$CaBenz_2$	$CaAc_2$+Bz	Real. Conc.
Spitz	14.2	13.9	1.7	14.1	12.1	14.2	14.0
	11.8	11.0	15.1	14.7	14.3	13.9	13.3
	13.2	12.1	14.0	14.0	9.6	14.2	12.1
	14.3	13.6	13.8	10.4	13.3	14.8	14.6
Mean	13.0		11.2	13.3	12.3	14.3	13.5
Std. Dev.	1.2		6.3	2.0	2.0	0.4	1.1
β Sheet	H_2O		$CaCl_2$	$CaAc_2$	$CaBenz_2$	$CaAc_2$+Bz	Real. Conc.
Spitz	33.2	34.4	35.0	30.6	31.2	34.2	27.5
	27.9	34.0	34.1	25.8	31.5	30.3	33.1
	32.9	35.7	34.5	27.1	29.4	40.5	30.0
	35.7	34.7	29.1	32.3	33.8	31.6	36.2
Mean	33.6		33.2	29.0	31.5	34.2	31.7
Std. Dev.	2.5		2.7	3.0	1.8	4.5	3.8

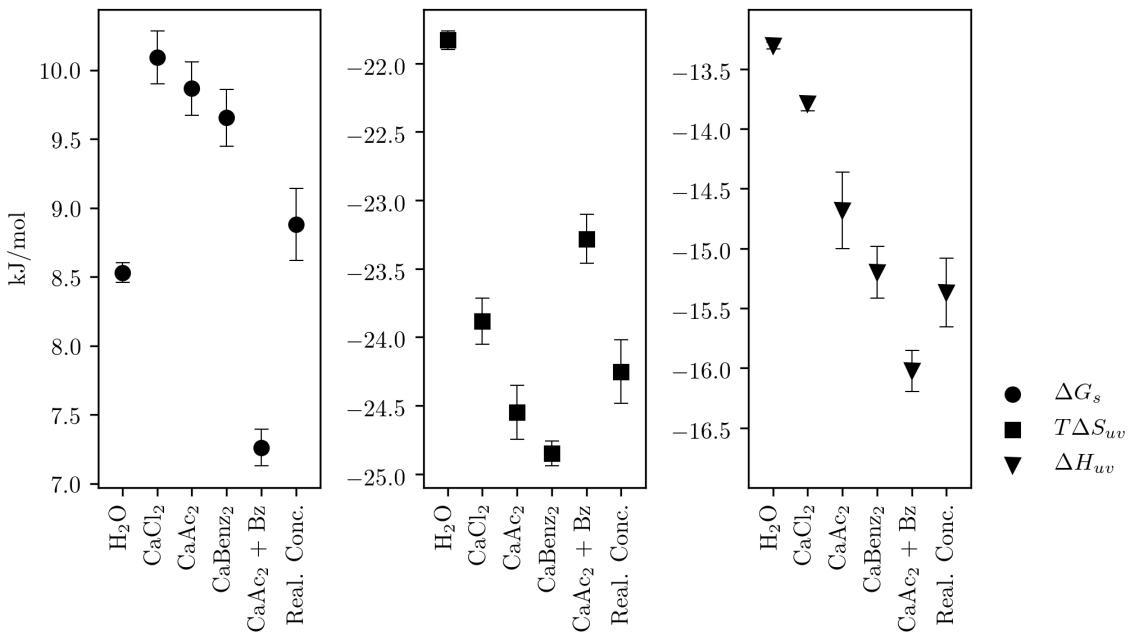

Figure 8.3: Free energies of solvation with their respective enthalpy and entropy terms calculated for all solvent conditions.

Table 8.3: Elemental fractions of used VSOMM2 systems and their respective IHSS experimental data.

	Elemental Fractions				
Name	C	H	O	N	S
BB2_1	0.536	0.033	0.401	0.014	0.016
BB2_2	0.531	0.032	0.395	0.014	0.028
BB2_3	0.544	0.034	0.400	0.013	0.009
BB5_1	0.524	0.028	0.409	0.014	0.025
BB5_2	0.534	0.028	0.404	0.014	0.019
BB5_3	0.557	0.030	0.394	0.015	0.005
BB10_1	0.547	0.028	0.396	0.015	0.015
BB10_2	0.530	0.026	0.403	0.014	0.027
BB10_3	0.557	0.027	0.401	0.012	0.002
BB20_1	0.536	0.027	0.404	0.014	0.020
BB20_2	0.533	0.026	0.402	0.015	0.025
BB20_3	0.558	0.027	0.394	0.014	0.006
IHSS LHA Sample[32]	0.638	-	-	0.012	-

```
#!/usr/bin/env python3

import numpy as np
import argparse

#-----INFORMATION-----------------------#

# version 1.1
# This script reads in .gro files specified by (-i) which contain HS
    with a specific amout (-s) of building blocks.
# It then inflates the system by adding a specified factor (-f).
# Finally, the program writes out a new coordinate file (-o).

#-----PARSING-ARGS----------------------#

parser = argparse.ArgumentParser()
parser.add_argument('-i',help = 'input.gro, required, only containing
    humic substances and without pbc', required = True)
parser.add_argument('-o',help = 'output.gro, default is output.gro',
    default = 'output.gro')
parser.add_argument('-f',help = 'float, required, factor that will be
    added to the coordinates', required = True)
parser.add_argument('-s',help = 'int, number of building blocks per
    humic substance, default is 5', default = 5)
args = parser.parse_args()

#-----INPUT-----------------------------#

input  = args.i
output = args.o
factor = float(args.f)
hslength = float(args.s)

#-----PRINTING-INFORMATION--------------#

class bcolors:
    green='\033[0;32m'
    nc='\033[0m' # No Color

print('\n'+bcolors.green+'
    ##############################################################'+'\n##
    ')
print('##              Input file: '+args.i)
print('##             Output file: '+args.o)
print('##                  factor: '+str(args.f))
```

```
print('##      Number of BB per HS: '+str(args.s)+'\n##')
print('#################################################################'+
    bcolors.nc+'\n')

#-----PARSING-TEXT----------------------#

HS = {}
box = {}
y = 1
z = 1
oldline = 'x'
counter = 0
newhs = []
newhs = np.zeros(2)
gro = open(input, 'r')
lines = gro.readlines()
temp = []
temp = np.append(temp,1)
x = 0
for line in lines:
    if 'HS' in line:
        x = 1
    else:
        box[z] = line.split()
        z += 1
    if x == 1:
        HS[y] = line.split()
        y += 1
        if not oldline in line.split():
            if not counter == hslength:
                counter +=1
            else:
                temp = np.append(temp,oldatom)
                newhs = np.vstack((newhs,temp))
                counter = 1
                temp = [float(line.split()[2])]
        oldline = line.split()[0]
        oldatom = float(line.split()[2])
gro.close()
newhs = newhs[1:]
#newhs is a list of the start and end atoms of every molecule

#-----PARSING-COORDS--------------------#

coords = np.loadtxt(input, skiprows = 2, usecols = (-3,-2,-1))[:-1]
boxcoords = np.loadtxt(input, skiprows = 2, usecols = (-3,-2,-1))[-1]
#print(coords) # this is the numpy array of the coordinates of the
    molecules
#print(boxcoords) #this is the numpy array of the box size

#-----DEFINING-CENTER-OF-BOX------------#

xlim = boxcoords[0]/2
ylim = boxcoords[1]/2
```

```
zlim = boxcoords[2]/2
#these are the limits of the quadrants

#-----CALCULATING-GEOMETRIC-AVGS--------#

avg = {}
mol = 0
for i in newhs:
    #print(int(i[0]), int(i[1]))
    minatm = int(i[0])-1
    maxatm = int(i[1])-1
    tempavg = []
    tempavg = np.append(tempavg,np.mean(coords[minatm:maxatm,0]))
    tempavg = np.append(tempavg,np.mean(coords[minatm:maxatm,1]))
    tempavg = np.append(tempavg,np.mean(coords[minatm:maxatm,2]))
    avg[mol] = tempavg
    mol += 1

#print(avg) #avg is the dictionary of the geometic mean of all humic
    substances

#-----LIST-OF-MOLS-2-INFLATE------------#

inflatex = []
inflatey = []
inflatez = []

for mol in avg:
    #print(avg[mol])
    if avg[mol][0] > xlim:
        inflatex = np.append(inflatex,mol)
    if avg[mol][1] > ylim:
        inflatey = np.append(inflatey,mol)
    if avg[mol][2] > zlim:
        inflatez = np.append(inflatez,mol)
print('Molecules to inflate in x:')
print(inflatex)
print('Molecules to inflate in y:')
print(inflatey)
print('Molecules to inflate in z:')
print(inflatez)
#these list are the mol number of molecules whose coordinates are
    going to be inflated

#-----CHANGE-COORDS--------------------#

for mol in inflatex:
    mol = int(mol)
    minatm = int(newhs[mol][0])-1
    maxatm = int(newhs[mol][1])
    coords[minatm:maxatm,0] = coords[minatm:maxatm,0] + factor

for mol in inflatey:
    mol = int(mol)
```

```
    minatm = int(newhs[mol][0])-1
    maxatm = int(newhs[mol][1])
    #print(minatm,maxatm)
    coords[minatm:maxatm,1] = coords[minatm:maxatm,1]+factor

for mol in inflatez:
    mol = int(mol)
    minatm = int(newhs[mol][0])-1
    maxatm = int(newhs[mol][1])
    coords[minatm:maxatm,2] = coords[minatm:maxatm,2]+factor

# this corrects the coords for the factor

#-----REMOVING-LAST-ENTRY-OF-HS---------#

i = len(HS)
del HS[i]

#-----WRITEOUT-FILE---------------------#

file = open(output, 'w')
file.write('\t'+box[1][0]+'\n'+box[2][0]+'\n')
space = '          '
counter = 0
for i in HS:
    first = space + HS[i][0]
    second = space + HS[i][1]
    third = space + HS[i][2]
    forth = str(round(coords[counter,0],3))
    if len(forth) == 4:
        forth = forth + '0'
    elif len(forth) == 3:
        forth = forth + '00'
    forth = space + forth
    fifth = str(round(coords[counter,1],3))
    if len(fifth) == 4:
        fifth = fifth + '0'
    elif len(fifth) == 3:
        fifth = fifth + '00'
    fifth = space + fifth
    sixth = str(round(coords[counter,2],3))
    if len(sixth) == 4:
        sixth = sixth + '0'
    elif len(sixth) == 3:
        sixth = sixth + '00'
    sixth = space + sixth
    file.write(first[-9:]+second[-6:]+third[-5:]+forth[-8:]+fifth
   [-8:]+sixth[-8:]+'\n')
    counter += 1
x = str(round(boxcoords[0] + factor,3))
print('New box size:')
print(x)
if len(x) == 4:
    x = x + '0000'
```

```
196 if len(x) == 3:
197     x = x + '0000'
198 x = space + x
199 file.write(x[-10:]+x[-10:]+x[-10:]+'\n')
200 file.close()
201
202 print('\n'+bcolors.green+'... done'+bcolors.nc+'\n')
```

Listing 8.5: Script for inflation of VSOMM2 systems. The script uses four input parameters; an input GROMACS coordinate file (.gro suffix) with coordinates of the VSOMM2 molecules, a float that is used as a summand to either inflate or deflate the system in nm, an integer which specifies the number of building blocks per molecule and an output filename where the inflated GROMACS coordinate output file is stored.

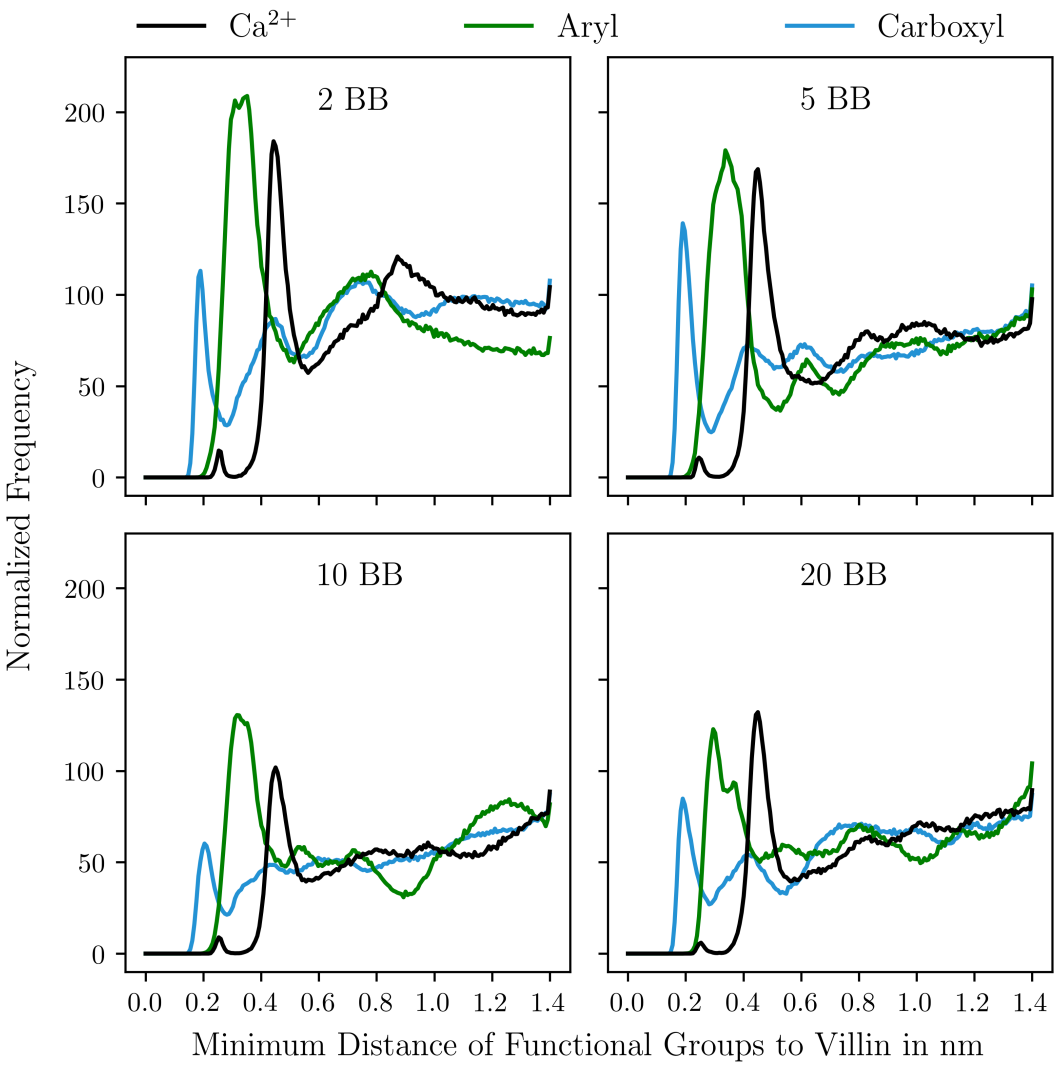

Figure 8.4: Frequency of the minimum distance functions of Ca^{2+} and selected functional groups to villin. The frequency is depicted in the number of snapshots sampled over all replicates of the simulation. Normalization was done over the number of ions/ functional groups present in the system.

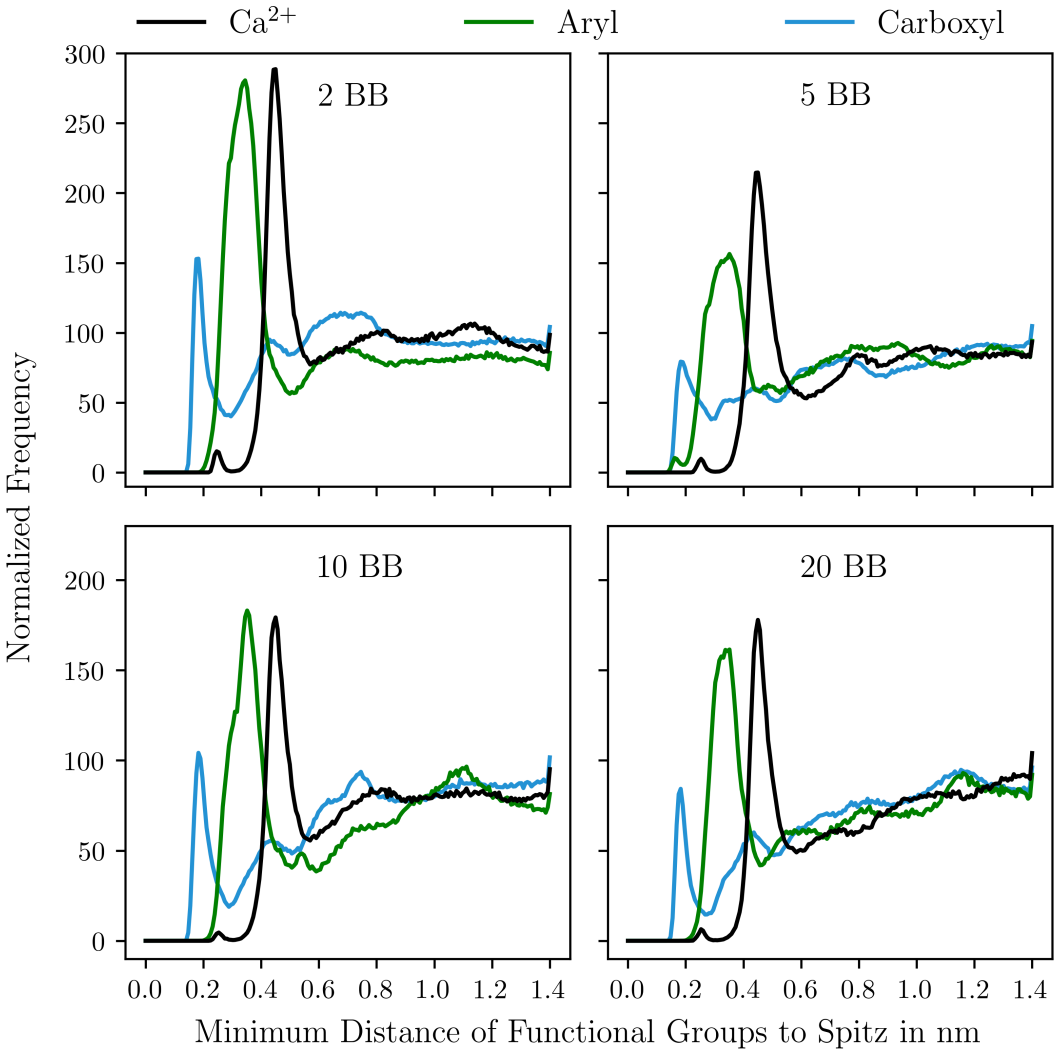

Figure 8.5: Frequency of the minimum distance functions of Ca^{2+} and selected functional groups to spitz. The frequency is depicted in the number of snapshots sampled over all replicates of the simulation. Normalization was done over the number of ions/ functional groups present in the system.

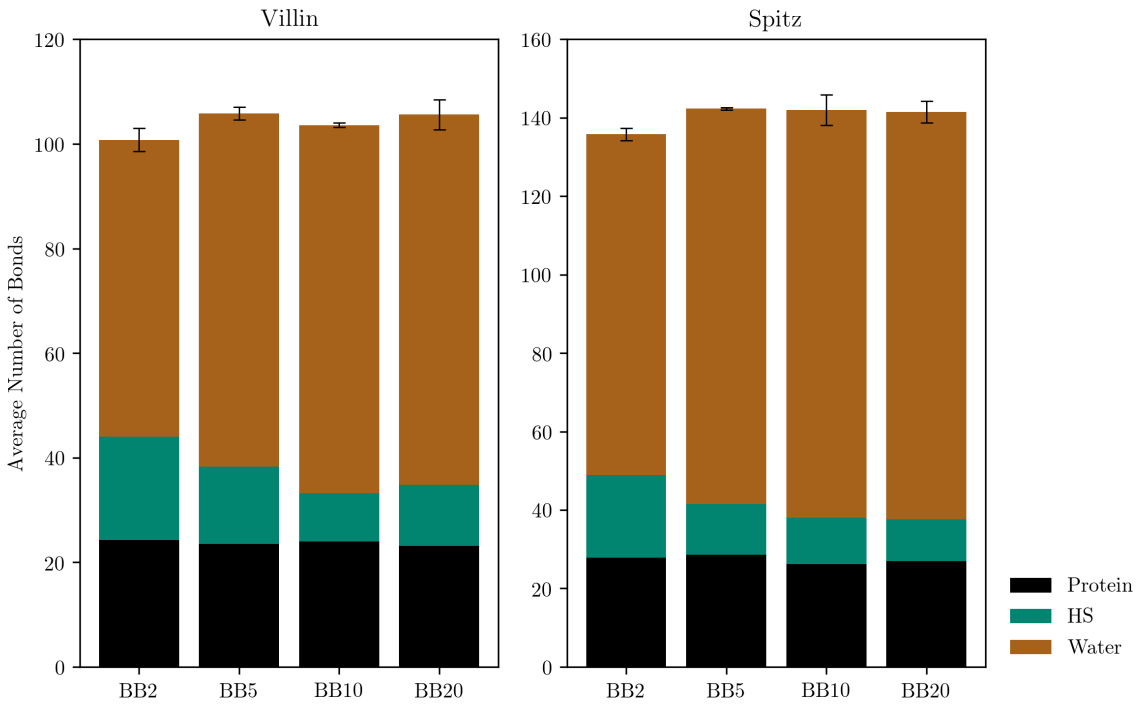

Figure 8.6: Average number of hydrogen bonds formed by the reference proteins over the simulation time. The errorbars represent the standard deviation of total numbers between replicates. The color code symbolizes different hydrogen bond partners. No differentiation was made between hydrogen donors and acceptors.

Table 8.4: Different ΔG values calculated for all amino acid residues in water. The % column shows the % of snapshots that contribute to the free energy calculation. For all values that were marked with * the $\Delta G^{TPF+OSP}_{Q\to R4}$ were used and then corrected by a linear term to approximate $\Delta G^{TPF+OSP}_{Q\to R5}$ (Figure 5.12). Abbreviations: AA = amino acids.

AA	$\Delta G^{TPF}_{Q\to N}$	$\Delta G^{OSP}_{R4\to N}$	R4 %	$\Delta G^{OSP}_{R5\to N}$	R5 %	$\Delta G^{TPF+OSP}_{Q\to R5}$
Ala	-	-10.26	65.0	-12.30	90.2	12.30
Arg	298.26	-4.16	0.8	-0.17	0.0	302.82*
Asn	254.55	-10.12	31.1	-8.75	18.9	263.29
Asp	351.37	-9.45	29.0	-8.70	17.8	360.07
Cys	18.29	-3.47	1.7	-1.42	0.9	22.34*
Gln	254.78	-10.81	16.8	-9.58	2.6	264.36
Glu	340.94	-9.32	16.0	-7.13	2.7	348.07
Hisa	31.67	-13.80	12.9	-11.48	1.3	43.15
Hisb	64.09	-12.25	10.9	-11.70	1.3	75.79
Ile	-	-9.55	3.7	-9.56	2.6	9.56
Leu	-	-12.84	14.7	-7.25	4.6	7.25
Lysh	336.35	-7.46	7.3	-7.38	0.4	343.39*
Met	12.73	-17.29	13.8	-13.97	2.4	26.71
Phe	0.80	-13.75	3.3	-9.93	0.1	12.56*
Ser	46.69	-5.81	43.0	-8.34	65.2	55.03
Thr	45.73	-4.71	17.6	-7.88	26.2	53.61
Trp	53.12	-15.93	0.2	-4.87	0.0	66.51*
Tyr	100.77	-14.72	1.9	-14.28	0.0	113.25*
Val	-	-6.78	8.5	-6.59	13.4	6.59

Table 8.5: Different ΔG values calculated for all amino acid residues in LHA systems. The % column shows the % of snapshots that contribute to the free energy calculation. For all values that were marked with * the $\Delta G^{TPF+OSP}_{Q\to R4}$ were used and then corrected by a linear term to approximate $\Delta G^{TPF+OSP}_{Q\to R5}$ (Figure 5.12). Abbreviations: AA = amino acids.

AA	$\Delta G^{TPF}_{Q\to N}$	$\Delta G^{OSP}_{R4\to N}$	R4 %	$\Delta G^{OSP}_{R5\to N}$	R5 %	$\Delta G^{TPF+OSP}_{Q\to R5}$
Ala	-	-11.21	54.0	-10.25	87.6	10.25
Arg	307.32	-8.30	1.4	-13.25	0.6	314.49*
Asn	251.97	-11.30	24.4	-9.33	21.1	261.30
Asp	408.59	-10.96	23.6	-10.25	21.0	351.16
Cys	18.70	8.59	1.7	-1.00	4.1	19.70
Gln	254.46	-14.94	9.9	-10.15	4.0	264.61
Glu	401.96	-14.82	10.4	-7.90	4.6	340.33
Hisa	29.41	-17.34	8.0	-12.94	3.5	42.35
Hisb	61.29	-14.50	6.3	-12.11	2.7	73.40
Ile	-	-6.70	5.4	-9.81	6.6	9.81
Leu	-	-10.31	13.0	-8.30	6.9	8.30
Lysh	269.05	-13.63	4.3	-10.34	1.9	347.40
Met	11.91	-18.99	9.5	-15.73	4.5	27.64
Phe	0.08	-12.84	2.9	-13.39	1.1	13.48
Ser	51.43	-7.91	43.2	-6.27	63.5	57.69
Thr	45.05	-1.40	20.5	-6.94	30.8	51.98
Trp	52.27	-15.17	1.2	-24.38	0.6	64.51*
Tyr	100.07	-14.47	2.3	-14.83	0.9	112.01*
Val	-	-5.83	14.5	-7.85	18.6	7.85

Table 8.6: $\Delta\Delta G$ and SASA values for each amino acid residue used in the SOMscore.

Residue	$\Delta\Delta G$ in kJ/mol	average SASA in nm^2
Ala	0.00	0.68
Arg	-13.71	2.09
Asn	-0.06	1.25
Asp	6.87	1.19
Cys	0.58	1.01
Gln	-2.29	1.54
Glu	5.69	1.47
His	-0.70	1.59
Ile	-2.30	1.50
Leu	-3.10	1.56
Lys	-6.05	1.75
Met	-2.98	1.62
Phe	-2.96	1.82
Ser	-4.71	0.82
Thr	-0.42	1.16
Trp	-0.05	2.20
Tyr	-0.81	1.95
Val	-3.31	1.30
Gly	-	0.00
Pro	-	1.16

```
#!/usr/bin/python3

#---------------------------------------#

import freesasa as fs
import os
import numpy as np
import argparse

#---------------------------------------INFORMATION

#Version 10
    #Introduced multiple chain compatibility
    #Works also for empty chain in pdb file
    #Gives error when interrupted protein (also when interruption by
   selenomethionine, etc.)
#To Do
    #Negative residue names give incorrect SASA values


#---------------------------------------Argument parsing
parser = argparse.ArgumentParser()
parser.add_argument('-pdb',help = 'Single pdb coordinate file of the
   protein', default = '/pool/gotsmymathias/saturation_mutagenesis/
   score/comparison/villin.pdb') #PDB file for the protein of interest
parser.add_argument('-sasa',help = 'Table with SASA values for
   residues', default = '/pool/gotsmymathias/saturation_mutagenesis/
   score/comparison/pdb_sasa_avg.lib') #List of maximum SASA values
   for each amino acid
parser.add_argument('-dG',help = 'Table with dG values for residues',
   default = '/pool/gotsmymathias/saturation_mutagenesis/score/
   comparison/pdb_dG.txt') #List of ddG value for each amino acid
parser.add_argument('-log',help = 'Boolean. Write out log file.',
   default=True)
args = parser.parse_args()

pdbfile = args.pdb
sasafile = args.sasa
dgfile = args.dG
log = args.log

#---------------------------------------Calculate SASA
def dosasa(id,o1,o2):
    x = fs.Parameters()
    x.setProbeRadius(1.4)
```

```
    x.setNSlices(10000)
    x.setAlgorithm(fs.LeeRichards)
    cls = fs.Classifier.getStandardClassifier('protor')
    structure = fs.Structure(pdbfile,cls)
    result = fs.calc(structure,x)
    area_classes = fs.classifyResults(result, structure)
    total = result.totalArea()/100 #to get from A2 to nm2
    sasa = {}
    #The backbone atoms are deselected
    sele = 'CA+C+O+N'
    if o1 == True:
        sele += '+O1'
    if o2 == True:
        sele += '+O2'
    for i in id:
        if not i[2]:
            chain = ''
        else:
            chain = ' and chain '+i[2]
        selections = fs.selectArea(('tot, resi '+i[0],'resi, resi '+i
    [0]+' and not name '+sele+chain),structure, result)
        sasa[i[0]+i[2]] = selections['resi']/100
    return sasa, total

#--------------------------------------Read the PDB file
def readpdb(dir):
    file = open(dir,'r')
    lines = file.readlines()
    file.close()
    res_start_end = []
    old = 'x'
    for line in lines:
        if 'ATOM' in line[:4]:
            temp = []
            temp.append(line[22:26].replace(' ',''))
            temp.append(line[17:20].replace(' ',''))
            temp.append(line[20:22].replace(' ',''))
            if temp == old:
                pass
            else:
                res_start_end.append(temp)
                old = temp
        if 'ENDMDL' in line[:6]:
            break
    #Check if O1 or O2 exists:
    o1 = False
    o2 = False
    for line in lines:
        if 'ATOM' in line[:4]:
            if line[13:15].replace(' ','') == 'O1':
                o1 = True
            if line[13:15].replace(' ','') == 'O2':
                o2 = True
    return res_start_end, o1, o2
```

```
90
91 #--------------------------------------Read list of ddG/SASA
92 def loaddG(dir):
93     dic = {}
94     file = open(dir,'r')
95     lines = file.readlines()
96     file.close()
97     for line in lines:
98         if '#' in line:
99             pass
100        else:
101            dic[line.split()[0]]=float(line.split()[1])
102    return dic
103
104 #--------------------------------------Back up if an old log.txt file
       exists, so nothing is overwriten
105 def backuplog():
106    count = 1
107    filename = str(count)+'log.txt'
108    if os.path.isfile('log.txt') == False:
109        filename = 'log.txt'
110    else:
111        while os.path.isfile(filename) == True:
112            count += 1
113            filename = str(count)+'log.txt'
114        print('Backed up old log file to '+filename+'.')
115        command = 'mv log.txt '+filename
116        os.system(command)
117    return
118
119 #--------------------------------------Main
120 print('#########################################################
       ##########################################################')
121 print('##
                                                    ##')
122 print('##    This program scores proteins based on free energy
       differences of sidechains in water and SOM.    ##')
123 print('##
                                                    ##')
124 print('#########################################################
       ##########################################################')
125 print('pdb file :\t',pdbfile)
126
127 if log == True:
128     backuplog()
129     logfile = open('log.txt','w')
130     logfile.write('# pdb_SOMscore_v10.py\n')
131     logfile.write('# pdb file :\t'+pdbfile+'\n')
132     logfile.write('# SASA     :\t'+sasafile+'\n')
133     logfile.write('# dG       :\t'+dgfile+'\n')
134     logfile.write('# \tAA\tchain\tSASA\tRel.SASA Score\n#\n')
135
136 dG = loaddG(dgfile)
137 max_sasa = loaddG(sasafile)
```

```
138 res_start_end,o1,o2 = readpdb(pdbfile)
139
140 #--------------------------------------Check if protein is
    interrupted
141 oldnr = 0
142 oldchain = 0
143 err = False
144 for i in res_start_end:
145     oldnr += 1
146     if int(i[0]) == oldnr:
147         oldnr = int(i[0])
148     elif i[2] != oldchain:
149         oldnr = int(i[0])
150         oldchain = i[2]
151     else:
152         print('Interrupted protein.')
153         oldnr = int(i[0])
154         oldchain = i[2]
155         err = True
156         break
157
158 #--------------------------------------Calculate Score and write
    output file
159 if err == True:
160     if log == True:
161         logfile.write('# The protein is interrupted.\n')
162         logfile.write('# Interruption ends at residue '+str(oldnr)+'
    chain '+oldchain)
163         logfile.close()
164 else:
165     sasa, totprotsasa = dosasa(res_start_end,o1,o2)
166     sum = 0
167     used_surface = 0
168     rel_surface = 0
169     for i in res_start_end:
170         try:
171             temp_dg = dG[i[1].lower()]
172             temp_msasa = max_sasa[i[1]]
173             temp_sasa = sasa[i[0]+i[2]]
174             sum += dG[i[1].lower()] * temp_sasa / temp_msasa
175             used_surface += sasa[i[0]+i[2]]
176             rel_surface += temp_sasa / max_sasa[i[1]]
177             if log == True:
178                 logfile.write('  '+i[0]+'\t'+i[1].lower()+'\t'+i[2]+'\
    t'+str(round(temp_sasa,3))+'\t'+str(round(temp_sasa / temp_msasa,3)
    )+'\t'+str(round(dG[i[1].lower()] * temp_sasa / temp_msasa,3))+'\n'
    )
179         except:
180             temp_sasa = sasa[i[0]+i[2]]
181             if log == True:
182                 logfile.write('# '+str(i[0])+'\t'+i[1].lower()+'\t'+i
    [2]+'\t'+str(round(temp_sasa,3))+'\n')
183             sum += 0
184     print('#--------------------------------------#')
```

```
185     fitness = used_surface/totprotsasa
186     print('Score   :',round(sum/rel_surface,3))
187     print('Fitness :',round(fitness,3))
188     if log == True:
189         logfile.write('#\n# total protein '+str(round(totprotsasa,4))+
    '\n')
190         logfile.write('#\n# Score   : '+str(round(sum/rel_surface,4))+
    '\n')
191         logfile.write('# Fitness : '+str(round(fitness,3))+'\n')
192         logfile.close()
```

Listing 8.6: Script for SOMscore.

9. List of Figures

10. List of Tables