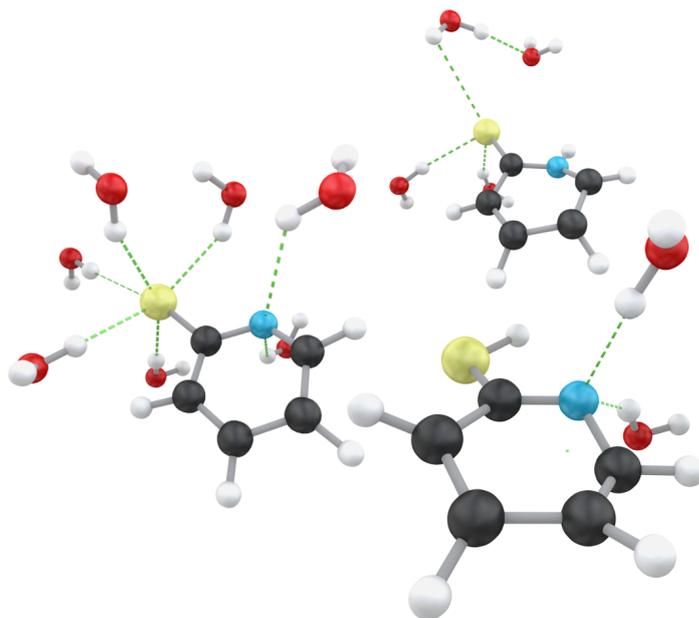


BACHELOR'S THESIS IN PHYSICS

Hydration models for thione-, thiol- and deprotonised 2-Mercaptopyridine, from molecular dynamics simulations.



Author:
Anton Ljungdahl,
anton.ljungdahl@gmail.com

Supervisors:
Michael Odelius,
Jesper Norell



Stockholm University,
Department of Physics
November 8, 2016

Abstract

The water solvation near nitrogen- and sulphur-sites, in the thione- and thiol-forms of 2-Mercaptopyridine (labeled 2TP and 2MP respectively), and in the deprotonised, negatively charged version (labeled 2AP), has been investigated. The aim of the project was to provide insights into the solvation, which can form the basis for future investigations of ground state, and photo-excited state, proton transfer processes.

Molecular dynamics simulations were performed of 2TP, 2MP and 2AP, in liquid water solution, under both constant pressure and constant volume, at ambient conditions, using a time step of 1 fs for a 90 ns production simulation time.

The first solvation shell water structures, near the molecules' nitrogen- and sulphur-sites, were obtained from analysis of both radial and spatial distributions functions. Based on this analysis, a solvation model for each organic molecule is presented.

The deprotonised 2AP had the most well defined water structure, and its N- and S-site oxygen- and hydrogen-distribution functions were, qualitatively, replicated in corresponding unprotonated sites in 2TP and 2MP. There were on average 5 water molecules within the 2AP S-site first solvation shell, and 4 within the 2TP. The 2AP had on average 2 water molecules within the N-site first solvation shell, which was also the case for 2MP. The distribution functions showed no clear indication of hydrogen-bonding between the protonated sites and water oxygen.

Sammanfattning

Förflyttning och överföring av protoner är en viktig kemisk process i biologiska system. Thione-thiol tautomerer, där en proton förflyttas mellan aktiva kväve- och svavel-grupper, är en prototyp av dessa överföringsprocesser. Vattensolvatiseringen vid kväve- och svavel-grupperna hos thione- och thiol-formerna av 2-Mercaptopyridine (kallade 2TP och 2MP), samt den deprotonerade versionen (kallad 2AP), har studerats. Detta i syfte att ge underlag för förståelse av protonöverföring i grund- och fotoexciterade tillstånd.

Molekyldynamiksimuleringar av 2TP, 2MP och 2AP utfördes, i vattenlösning, under konstant tryck och konstant volym, vid ambienta förhållanden. Tidssteget var 1 fs, och längden av simuleringen var 90 ns.

Från radiella och spatials fördelningsfunktioner, räknade från molekylernas kväve- och svavel-grupper till vattnets syre- respektive väte-atomer, kunde solvatiseringen i det första skalet bestämmas. Baserat på denna analys så presenteras tre solvatiseringsmodeller, en för varje organisk molekyl, för framtida användning i kvantkemiska beräkningar.

Acknowledgements

I want to thank my wonderful family, Anna, Tor, Vide and my children's grandparents. Without you my studies would not be possible.

Thanks to Michael and Jesper, for giving me the chance to work on this project, and for all the great support and feedback. I feel privileged to have had you as supervisors.

Thanks to my fellow thesis project workers in the C4:aquarium, Johan, Mihae and Mathias, for the good company and interesting discussions.

And special thanks to Jesper and Johan, for all the bash, text editor and general Linux tips.

Contents

1	Introduction	4
1.1	The thione-thiol tautomers	4
1.2	Purpose and motivation	5
1.3	Overview	6
2	Theory	7
2.1	Molecular dynamics	7
2.1.1	Basic principles	7
2.1.2	Molecular dynamics in statistical ensembles	9
2.2	Force field methods	10
2.3	Quantum chemistry	11
2.4	Structure and the distribution functions	13
2.4.1	Radial distribution function (RDF)	14
2.4.2	Spatial distribution function (SDF)	15
3	Simulation	16
3.1	Method	16
3.2	Production	18
4	Results	20
4.1	Performance	20
4.2	The liquid structure	22
5	Discussion and conclusion	29
5.1	Suggestions for the quantum chemistry setup	29
5.2	General discussion of the results	33
5.3	Molecular flexibility	34
5.4	Uncertainties	35
5.5	Conclusions	36
A	Appendix	38
A.1	The Born-Oppenheimer approximation	38
A.2	The leap-frog algorithm	39
A.3	Molecular dynamics using GROMACS	40
A.4	Input parameters	42

1 Introduction

The ambition when writing this thesis, was to present the project in a manner such that a hypothetical reader, with a background corresponding to an undergraduate physics student, would be able to understand and judge the method and results. This means, arguably, that the work should be somewhat self-contained, and as a consequence the background is rather extensive. A reader only interested in the scientific content could probably skip the entirety of Chapter 2, and the main part of Chapter 3 (except Section 3.2). To guide the reader through the text, a small summary of each chapter's content is presented at its beginning.

To introduce some necessary terminology and concepts, this chapter starts off with presenting the studied chemical systems, in Section 1.1. The purpose of, and motivation behind, the project is then given in Section 1.2. And last, Section 1.3 summarises the work done, and gives an overview of the thesis content.

1.1 The thione-thiol tautomers

Three very similar, small, aromatic organic molecules, were studied. Two of them, 2-Thiopyridone (2TP) and 2-Mercaptopyridine (2MP), are structurally identical, up to the placement of a proton. The third is called 2AP, an anionic (negatively charged) variant, in which a proton is lost. Figure 1.1 shows structural drawings of the molecules, which illustrates how 2TP and 2MP differ from each other only by the placement of a proton, present either on the nitrogen (N-site) in the case of 2TP, or on the sulphur (S-site) in the case of 2MP.

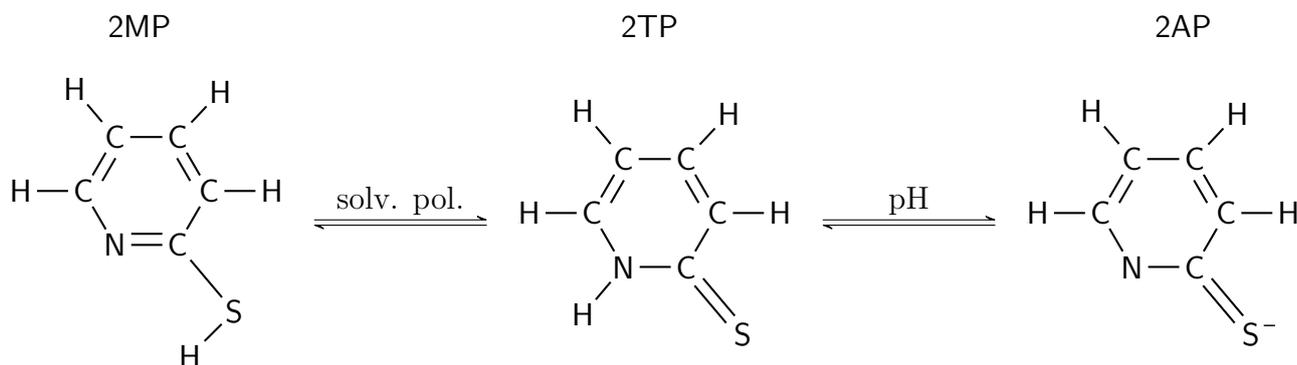


FIGURE 1.1: Schematic drawings of the two tautomers 2-Mercaptopyridine (2MP) and 2-Thiopyridone (2TP), with the deprotonised version, 2AP, to the far right. The 2AP negative charge is distributed over the whole molecule, and not only on the sulphur. The back-and-forth arrows illustrates how the equilibrium concentrations, of each form, depend on the solution environment (solvent polarity, pH, thermal equilibrium, optical excitation [8, 7]).

The organic chemistry nomenclature for such molecules is thione-thiol tautomers. Tautomers are defined as structural isomers¹, that interconvert with each other, and the equilibrium can (as in this case), be controlled by solvent-solute interactions, e.g. polarity. This interconversion, or tautomerisation process, consists of transferring a proton. It is this tautomerisation proton which is lost in 2AP. The thione-thiol prefix means that the proton transfer is between the nitrogen and sulphur groups, and 2TP is called the thione-form, while 2MP is the thiol-form.

1.2 Purpose and motivation

The scientific goal was to obtain information about the tautomer-water structure, by performing molecular dynamics simulations of the three organic molecules in aqueous solution. Based on analysis of the structural data, a solvation model for quantum chemical calculations, applied to soft X-ray spectroscopy, was to be developed. This in the context of a research project of Jesper Norell and Michael Odelius (of the Quantum Chemistry and Molecular Dynamics-group (QCMD) at the Stockholm University Physics department), in collaboration with Sebastian Eckert and Alexander Föhlich (Helmholtz-Zentrum Berlin) and coworkers. They study excited state proton transfer processes, with the thione-thiol tautomers studied herein, as model systems.

Proton-transfer processes are important chemical reactions in chemistry and biochemistry. They represent the smallest, but still everywhere present, rearrangement in the chemistry that is vital to life [8, 7]. Thione-thiol (and keto-enol, a transfer between carbon and oxygen groups) tautomerisms constitute prototypes of these processes, and has been studied extensively [8]. The amino acid cysteine, for example, has a sulphur-proton group that can be activated by thione-thiol interconversion, and this may play an important role in how a protein's biological function is defined [8]. The tautomers studied in this thesis are smaller than cysteine, however, but contains the same functional nitrogen and sulphur groups, thus the study of these smaller systems may aid in understanding the more complex biochemical processes.

The proton-transfer process was not directly studied in this project, but rather how the solvent (liquid water) reacts to the transfer. The idea was to investigate the hydrogen-bonding capabilities of the protonated and unprotonated S- and N-sites, in the three different molecules.

The primary simulation method was molecular dynamics (MD), and this method is presented in more detail in Section 2.1. The open-source simulation package GROMACS (version 5.1.2) was used for performing the simulations, and also for (most of) the analysis. Now, there is a big difference between learning how to get already available software to perform simulations corresponding to some pre-determined settings, and actually understanding *why* those settings were decided on, and *how* the results were calculated and analysed. Another purpose of this thesis, or secondary goal, was to demonstrate this understanding. This is accomplished by providing enough background on the methods used, both in the general context of computer simulations (with a focus on molecular systems), and in the specific

¹Structural isomers are molecules that share the same structural formula, in this case C_5H_5NS (for 2TP and 2MP), but have different bonds.

context of this project.

1.3 Overview

MD simulations of the three organic molecules solvated in water, in *NVT* and *NPT* ensembles, with a simulation time of 100 ns (allowing for 10 ns of equilibration time) and a time step of 1 fs, were performed (using GROMACS [3]) and analysed for structural information. The potential energy surface was obtained by force field methods, with OPLS/AA [16] as the chosen force field. The TIP4P/2005 [1] water model was used for water representation. Radial distribution functions from the N- and S-sites, to the water oxygen and hydrogen atoms, were calculated using GROMACS' analysis tools. Corresponding spatial distributions were calculated using the free software TRAVIS [5]. To obtain starting configurations of the organic molecules, as well as distributed charges, preparatory quantum chemical calculations were performed, using the software Gaussian [10] (with HF/6-31G* level of theory).

Now the above paragraph is rather technical, and probably not very transparent to someone not familiar with the methods. The entireties of Chapter 2 and Section 3.1 are devoted to its explanation. The primary concern is introducing the MD method, and understanding of the distribution functions, but a very brief overview of quantum chemistry is also included. A great deal can be said about all these subjects of course, but the focus here is on basic intuition and the underlying approximations, without a lot of time spent on technicalities.

In Chapter 4, some simulation benchmarks are discussed, before the liquid structure results are presented. It is important to have some notion of how well the simulations represent the chosen target conditions and model.

Chapter 5 begins with a presentation of the solvation models, and then continues on to discuss the structure results and uncertainties.

A slightly more formal (compared to Chapter 2) look on the Born-Oppenheimer approximation and on the leap-frog variant of the Verlet algorithm are given in Appendices A.1 and A.2. The Appendix A.3 explains the GROMACS specific workflow, to further help in understanding the molecular dynamics method. To make the performed simulations more replicable, the relevant GROMACS input parameters are given in Appendix A.4.

An observation from the early simulations was that the organic molecules showed a considerable degree of flexibility, i.e breaking of the molecular plane, and this was hypothesized to affect the distribution functions. To investigate, rigid simulations were attempted, using the SHAKE algorithm (as available in GROMACS) to constrain the atoms to the molecular plane. These attempts largely failed, since the systems “exploded” under the constraints, one could say that the simulation engine broke down. The only successful rigid simulation was of 2TP, and with a 0.1 fs time step, for 10 ns. The full investigation of the rigid versus flexible simulations was abandoned, due to time constraints, but the 2TP results are discussed in Section 5.3.

2 Theory

The point of this chapter is to present a brief overview of the ideas and techniques underlying the methods used for this thesis project. Technical details are avoided, and the emphasis is put on the basic principles and ideas. This is at the cost of completeness, and many things have been left out. The important approximations are discussed however, and examples are given in an attempt to illustrate the ideas in relation to this thesis project.

Most central is perhaps Section 2.4, on the liquid structure and distribution functions, since the results are related to this. But the chapter starts with a discussion of molecular dynamics, in Section 2.1, and then Section 2.2 further explains the concept of simulating a molecular system, without explicitly solving the Schrödinger equation, via force field methods. Preparatory quantum chemical calculations were made, to optimise geometry and determine force field charge parameters, and so a very short presentation on quantum chemistry is given in Section 2.3. Its focus is on the important approximations, and basically just mentions the two main computational methods of Hartree-Fock theory and Density Functional Theory.

2.1 Molecular dynamics

The idea of the molecular dynamics (MD) simulation technique is to follow the classical dynamics of a many-particle system and efficiently sample the configurational space, and evolve the system in time. Classical in this context means that the simulation is modeling the dynamics of the atomic nuclei by using classical mechanics, and that any quantum mechanical behavior of the nuclei is ignored. The strong coupling of electronic and nuclear degrees of freedom makes it difficult to use molecular dynamics to study reaction processes, which involve the transitions of electrons.

2.1.1 Basic principles

The core of a molecular dynamics simulation is solving (numerically) Newton's equations of motion, for all particles in the system. Since these are second order differential equations of the form

$$\mathbf{F}_i = m_i \ddot{\mathbf{r}}_i, \quad i = 1, 2, \dots, N, \quad (2.1)$$

for an N -particle system, a dynamics simulation needs a set of initial positions and velocities, or phase space coordinates, at the initial time t_0 . It also needs a potential energy function $V(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$ for the whole system, so that the force acting on each particle i can be

computed through

$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i}. \quad (2.2)$$

The initial conditions, and the potential, can then be used to estimate the position and velocities, at a later time t_1 , provided that the time step Δt is small enough, so that the interaction between the particles can be considered constant in that time step. Then the positions and velocities at t_1 can be used to obtain a new set of phase space coordinates at time $t_2 = t_1 + \Delta t$, and so forth. The principle is, that by integrating the equations of motion, for a large number of small time steps, the time evolution of the system can be obtained.

What Δt is a small enough time step? Typical dynamical properties in molecular systems, such as rotational and vibrational motion, occur with frequencies between 10^{11} - 10^{14} s^{-1} . A time step on the order of a femtosecond (fs), 10^{-15} s, is necessary to describe such motion accurately enough [15].

There are various ways of numerically integrating the equations of motion, common examples are Verlet type algorithm (like the leap-frog algorithm) or Runge-Kutta methods [15]. The leap-frog method was used for the simulations in this thesis, and this algorithm is discussed in more detail in the Appendix A.2. The algorithm illustrates how the previous positions, velocities and potential energy are used to obtain the next configuration.

Clearly, then, it is necessary to have an explicit expression for the potential energy at each time step in order to perform the integration. This potential has to take into account all the interactions between the particles of the system, which include the bonded interactions between the atoms in a molecule, as well as the intermolecular interactions. These interactions are of electronic nature [22]. It is easily imagined that, for even a moderately complex system, finding such an explicit potential is a non-trivial task. So how can this be accomplished?

The potential can in principle be calculated quantum mechanically by separating the nuclear and electronic parts of the Schrödinger equation (SE)

$$\hat{H} |\Psi(X, x)\rangle = E |\Psi(X, x)\rangle, \quad (2.3)$$

where X and x denote the positions of the atomic nuclei and electrons respectively. If we make the approximation that the nuclei are stationary, the electronic wave-function depends parametrically only on the positions of the nuclei (and not on the velocities). This is called the Born-Oppenheimer (BO) approximation, and is often very good, even for light atoms and molecules (such as H_2) [15]. See the Appendix A.1 for a more detailed treatment of the BO approximation.

In this approximation the electronic wave function gives rise to what is called a *potential energy surface* (PES), denoted by $V(X)$, which is defined by the solutions to the electronic SE for all nuclear positions $X = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N\}$ [15, 9, 22]. The PES can be written $V(X) = E_e + V_{nn}$ to illustrate that it contains both the electronic energy and the nuclear-nuclear repulsion potential.

But solving the Schrödinger equation (numerically) for the whole system in every time step is computationally very heavy. Even though such methods exist, called *ab initio* MD, a faster (but less accurate) method is to approximate $V(X)$ by a parametrised function called a *force field* [15]. The force field method was used in this thesis project, and is further explained in Section 2.2.

Another concern is that the particle counts in real world systems are in the order of Avogadro's number ($\sim 10^{24}$), while MD simulations typically have particle numbers on

the order of $10^2 - 10^5$ [9, 22]. A common method [22], to closer approximate physical systems, is the introduction of periodic boundary conditions (PBC). The idea is that the MD simulation take place in some box (not necessarily cubic), and that if a particle in the simulation passes through any side of that box it is put back inside the box on the opposite side. This simulation box is then replicated, in principle infinitely, in all directions, so that the simulated particles interact not only with the other particles in the box, but also with the replicated particles. The replicated boxes contain precisely the same configurations as the “parent” simulation box, at any given instant in the simulation. It is assumed that such PBC help in modeling a system, of the same (relatively small) size, that is embedded in a real, macroscopic system [22].

But the force calculated on a particle results from the particle’s interaction with all other particles in the system, and if there are infinitely many copies of the system, infinitely many interactions are implied. The interaction forces typically decay rapidly with distance [22, 15], however, and the particles in the replicate boxes will not contribute much to the force (an example is the Lennard-Jones interaction, proportional to r^{-6} , which will be further discussed in Section 2.2). In particular, if the interactions between particles can be neglected for separations longer than half the linear system size, the interaction force (for each particle) can be calculated efficiently by only considering the closest neighbour of the other particles [22]. This is called the minimum image convention. But electrostatic interactions can not be truncated in this way; it can be shown that the interaction potential cannot be cut off unless it decays faster than r^{-3} [9]. To take the long-range interactions into account, in a computationally feasible way, special methods, such as Ewald sums, are required [6, 9].

2.1.2 Molecular dynamics in statistical ensembles

The discussion above concerns the study of a system of N particles, in some volume V , whose time evolution is constrained to a phase space surface defined by the constant energy E . The type of statistics that can be extracted from MD simulations are *time* averages. But in a statistical ensemble, the *ensemble* averages are functions of the distribution of microstates. If we assume these averages to be equal, however, we can view the MD method above as a representation of the microcanonical (constant NVE) ensemble [9]. This assumption of equality between time and ensemble averages is called the ergodic hypothesis. We also need to assume that, for long enough simulation times, the time averages does not depend on the initial conditions. If not, the extracted data would only apply to the specific initial conditions.

But the simulations performed in this thesis were in constant NVT and constant NPT ensembles, and not the microcanonical, so how does that work? The way that the NVT , or canonical ensemble, is introduced in statistical mechanics, is by considering the system of interest as being connected to some large reservoir, of constant temperature T , in such a way that energy can be exchanged between the system and the reservoir [18]. In constant NVT MD simulations then, the idea is to simulate in such a way that a canonical distribution is obtained. That is obtaining a well defined time averaged temperature, while still allowing for temperature fluctuations typical of the canonical distribution. This is, in principle, achieved by coupling the system to a reservoir, or heat bath, but there are different coupling methods. This thesis utilised the Nosé-Hoover thermostat, which relies on a Lagrangian reformulation of the equations of motion. This Lagrangian, however, contains additional artificial coordinates and velocities [9], to couple the system to the reservoir. Similar ideas

are used in the Parrinello-Rahman scheme, which was used for the *NPT* simulations, and in order to obtain a constant (on average) pressure the box volume needs to be varied [2].

2.2 Force field methods

The point of force field methods, also called molecular mechanics (MM), is to bypass the electronic energy calculation when obtaining a potential energy surface (PES) for the system. Instead, the electronic energy is treated as a parametric function of the nuclear coordinates, and its parameters are fitted to experimental and/or computational data. For example, the OPLS/AA force field, used for the organic molecules in this thesis, has its parameters obtained partly from experiment and from quantum chemical calculations, as well as from earlier force fields [16]. Not solving the Schrödinger equation (SE) also means that information about the chemical bonds must be explicitly provided, instead of as a result of solving the SE.

The view of molecules as a composite of different units, or functional groups, is an idea that is central to organic chemistry [15]. These functional groups share their properties between different molecules. All C–H bond lengths, for example, are approximately constant in all molecules, or each CH₂ group in linear alkanes have the same energy contribution [15]. This idea of functional groups also form the basis of force field methods, where the identification and parametrisation of functional groups allow for a more general use of the force field. It is implemented by means of atom *types*. As an example, consider the nitrogen atom in the “functional group” C–N=C, of 2MP (see Figure 1.1). For the purpose of force field methods, it has some set of parameters attached to it, that reflects that it is a nitrogen atom in an aromatic ring, and is treated as some type. The protonated nitrogen in 2TP has different bonds and properties, and is of another type, even though they are both nitrogen atoms. Force fields can be viewed as a database of sorts, telling the simulation engine what functional forms, and with which parameters, should be used for calculating interactions.

Despite the idea that atom types can be used in different molecules, the amount of parameters that needs to be determined is huge, even for a relatively small set of atom types. To understand this, and what kind of parameters there are, we will first look at the force field energy. There are two main components, bonded and non-bonded energy terms, and we write the total force field energy as

$$E_{FF} = E_{stretch} + E_{bend} + E_{tors} + E_{cross} + E_{vdW} + E_{el}. \quad (2.4)$$

The first three terms represent the bonded energy, the stretching (of a bond), bending (of a bond) and torsional (rotation around a bond) energy functions. The E_{cross} term describes coupling between the bonded terms. The last two terms represent non-bonded atom-to-atom interactions. The van der Waals-term E_{vdW} describes non-polar repulsion and attraction, the non-bonded interaction that is not due to electrostatics, which are the domain of the E_{el} term.

Several different functional forms, for the terms in Equation (2.4), are possible [15], and they cannot all be discussed here. We shall discuss the example of representing the van der Waals-term by a Lennard-Jones potential

$$V_{LJ}(r) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] = \epsilon \left[\left(\frac{r_m}{r} \right)^{12} - 2 \left(\frac{r_m}{r} \right)^6 \right], \quad (2.5)$$

whose graph is illustrated in Figure 2.1. It is a pair-potential, modeling the interaction

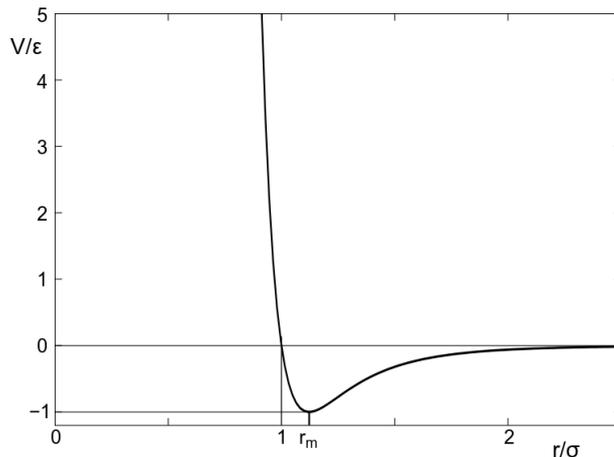


FIGURE 2.1: Graph of the Lennard-Jones potential (Equation (2.5)), distances are related by $r_m = 2^{1/6}\sigma$. From Wikimedia Commons, used under CC BY-SA 3.0 license.

between two atoms. So r denotes the distance between the atoms, ϵ is a parameter that corresponds to the depth of the potential well, σ is the finite distance at which the potential is zero, and r_m is the distance at which the potential is at a minimum. The r^{-12} term represents the strong repulsion for small separation distances, due to the Pauli exclusion principle, and the r^{-6} term is attraction due to resonant charge polarisations (induced dipole-induced dipole interactions) [15].

But the Lennard-Jones potential only need two parameters, σ and ϵ , determined. For a set of 100 atom types, 200 parameters would need to be determined. This is perhaps not too bad. Consider now the torsional energy term E_{tors} , which describes rotation around a bond, involving four atoms for well defined motion. This term will not be treated in detail, the point is to illustrate the number of parameters that needs to be decided. All atom types cannot form bonds to every other atom type, probably, but let's say that 30 atom types can. Then, since we have four atoms participating in the torsional energy, there are $\frac{1}{2} \times 30^4 = 405000$ possible different torsional energy terms. A common functional form of E_{tors} employs three parameters [15], and so $3 \times 405000 = 1215000$ parameters would need to be determined.

The time it would take to determine all parameters by experiment would not be worth it, compared to just performing the calculations *ab initio*, and solving the Schrödinger equation. This is one reason for why parametrisation is a combined effort between experiment, quantum computational chemistry and theoretical derivation [15].

2.3 Quantum chemistry

Employing the Born-Oppenheimer approximation (see Appendix A.1), the main problem [15] of quantum chemistry is solving the (time-independent and non-relativistic) molecular electronic Schrödinger equation

$$\hat{H}_e |\Phi\rangle = E_e |\Phi\rangle. \quad (2.6)$$

The state $|\Phi\rangle$ is represented by the *molecular* wave function, which concerns, in general, a system of multiple electrons and nuclei. Only the H_2^+ system, and systems similar to it, can

be solved analytically, and so numerical methods are generally required [15, 11]. Various approximations need to be made in order to obtain both analytical and numerical solutions.

An approximation that simplifies the problem, conceptually and computationally, is the *independent particle* approximation. The idea is that, dynamically, any electron in the system is independent of all the other electrons, and the electron-electron interaction is instead taken care of by a *mean field* approximation. This means that the molecular wave function, for the n -electron system, can be written as a product of molecular, one-electron, spin-orbitals $|\phi_i\rangle = |\psi_i\rangle|\sigma\rangle$,

$$|\Phi\rangle = |\phi_1\rangle|\phi_2\rangle\dots|\phi_n\rangle = |\phi_1\phi_2\dots\phi_n\rangle. \quad (2.7)$$

In the spin-orbital, $|\psi_i\rangle$ denotes the spatial part of the state of electron i , and $|\sigma\rangle$ the spin part (either spin up or spin down). This works for atomic orbitals as well, and the periodic system is based on this approximation [11], where, for example, the electronic configuration of carbon,

$$\text{C} : 1s^2 2s^2 2p^2, \quad (2.8)$$

implies the separation of each electronic state. But Equation (2.7) does not take the Pauli principle into account! Electrons are fermions, and so $|\Phi\rangle$ needs to be antisymmetric under exchange of any two electrons. This is taken into account by writing the molecular wave function as a Slater determinant,

$$|\Phi\rangle = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) & \dots & \phi_n(1) \\ \phi_1(2) & \phi_2(2) & \dots & \phi_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(n) & \phi_2(n) & \dots & \phi_n(n) \end{vmatrix}, \quad (2.9)$$

where $\phi_j(i)$ denotes the state $|\phi_j\rangle$ occupied by electron i . In the two-electron case we get

$$|\Phi\rangle = \frac{1}{\sqrt{2}} (|\phi_1(1)\phi_2(2)\rangle - |\phi_1(2)\phi_2(1)\rangle), \quad (2.10)$$

which, for example, can represent the quantum state of a $1s^2$ atomic orbital, where ϕ_1 and ϕ_2 only differ by their spin states.

One of the two most common methods, for solving the molecular electronic problem, is the Hartree-Fock (HF) method. It employs variational principles to find the best approximation to the energy eigenfunction $|\Phi_0\rangle$ of the molecule's electronic ground state, by considering the eigenfunction as a Slater determinant [15].

The other dominant method is Density Functional Theory (DFT). The concept underlying DFT, is that the ground state electronic energy is completely determined by the electron density, proven by Hohenberg and Kohn in 1964 [14]. The ground state energy can be considered as a functional $E[\rho]$ of the electron density function $\rho = \rho(x, y, z)$. This problem only concerns three variables, in stark contrast to the Hartree-Fock method. But the preferred method, for computational chemistry, is Kohn-Sham theory, which reintroduces orbitals into DFT, to increase the accuracy in the representation of electronic kinetic energy, at the cost of now dealing with $3N$ variables.

A key idea in these methods is to express a molecular orbital $|\phi_i\rangle$ as a linear combination of *known* basis states $|\chi_\alpha\rangle$, from some basis set, which is formally written

$$|\phi_i\rangle = \sum_{\alpha} c_{i\alpha} |\chi_\alpha\rangle, \quad (2.11)$$

where $c_{i\alpha}$ denotes the expansion coefficients. This is called Linear Combination of Atomic Orbitals (LCAO), even though the basis states does not have to be solutions to the atomic problem (i. e the the combination of spherical harmonics), but can also be, for example, plane waves or gaussian functions [15]. The geometry optimizations performed in this thesis used the Hartree-Fock method and the 6-31G* basis set, or summarised as HF/6-31G* *level of theory*, which is an often occurring terminology in quantum chemistry.

The above exposition is to be taken as the briefest of overviews of the subject, and no claim of completeness is made. Nothing has been said about how the methods consider the electronic correlation energies [15, 11], for example. Since the methods used for this project were not, primarily, of quantum mechanical nature, a more detailed account of quantum chemistry is considered to be outside the scope of this thesis.

2.4 Structure and the distribution functions

The word Structure, in the context of this discussion, refers to the relative positions and orientations of the atoms of a molecular system, its spatial configuration. For these kinds of system it is most meaningful to discuss the average (most probable) spatial configuration over any given time interval. Structure then becomes a measure of how correlated the spatial configurations are over time. No correlation (over long periods of time) corresponds to no structure, and a system whose configurations show high correlation is considered to be highly structured.

The particles in an ideal gas are approximately assumed not to interact, the separation between its molecules are fairly large, and the gas molecules move fast, compared to the solid phase. It is not common to talk about any local structure for substances in the gas phase, the spatial configurations at different times show no correlation with each other. On the other end of the spectrum; a crystalline solid at 0 K (with minimal lattice vibrations) has its structure completely determined by the crystal lattice unit cell. In other words, the unit cell is the minimum unit of volume that contains all the information about the crystal's structure and symmetry. This is how the solid phase is considered to be very structured.

In an liquid however, the molecules are closely packed (at least compared to a gas), on the one hand. On the other hand, the molecules are also constantly moving and shuffling around. In this sense the liquid phase can be viewed as intermediate to the gas and solid phases. It should be noted that liquid refers to the isotropic kind here, and not crystalline.

The structural information presented and discussed in this thesis was obtained from radial distribution functions (RDF) and spatial distribution functions (SDF). These are common methods [9, 15, 22, 17, 21] for investigating liquid structure averages. The SDF is only directly obtainable from simulations, since it is a function of the actual spatial coordinates at all time steps, for any given particle in the system, which is not feasible to actually measure for a fluid. The RDF, however, is related to the *structure factor*, which is a measurable quantity (by means of diffraction) [12]. So the RDF can be determined indirectly via experiment, which makes it a useful instrument for comparing computer simulations and experimental results. The following two sections expand a bit on the theory of radial and spatial distribution functions, while further discussion of the structure factor is beyond the scope of this thesis.

2.4.1 Radial distribution function (RDF)

The intuitive presentation of the radial distribution function in this section follows the one given in [13].

Consider a given atom in a molecular fluid (gas or liquid). A question we might ask is how many nearest neighbours this atom has on average, or how many second nearest neighbours it has on average, and so forth. The general question is how many neighbours there are, on average, within some radial distance r from the given atom.

In order to give some intuition on, and to define, the RDF let us consider a homogeneous, uniform and isotropic fluid. Let there be N fluid particles occupying a total volume V , at some temperature T , and let the bulk number density of the fluid be $\rho = N/V$.

Now imagine following the central atom as it moves through the fluid over some interval of time. Since the system is a fluid, the average distribution of other particles (other atoms or, in the case of isotropic systems, molecules) around the central atom will display spherical symmetry. In the case of an ideal gas, the mean number of particles in an infinitesimal volume element $d\tau^1$, at distance r from the central atom, is given by $\rho d\tau$. In the general case however, the particles in the fluid are interacting, and so the central atom is perturbing its environment as it moves, and the mean number of particles in a volume element around the central atom will *not* equal $\rho d\tau$. Due to the exclusion principle, for example, there can be no particles neighbouring the central atom for sufficiently small r . Depending on the interaction potential between the central atom and the surrounding particles, the general mean number of particles in a volume element $d\tau$ can thus be written as $\langle \rho_N(r) \rangle d\tau = \rho g(r) d\tau$. This leads to the definition of the radial distribution function

$$g(r) = \frac{\langle \rho_N(r) \rangle}{\rho}, \quad (2.12)$$

where $\langle \rho_N(r) \rangle$ denotes the average number density at distance r from the central atom and ρ the bulk number density of the fluid.

For typical systems the interparticle potential fall-off is inversely proportional to some power of r [4, 22], so for large enough distances we have that $g(r) \rightarrow 1$, which gives back the ideal gas limit of non-interacting particles. It is thus motivated to think of $g(r)$ as a factor that determines how much the mean *local* density $\langle \rho_N(r) \rangle$ deviates from the bulk density ρ . This provides a way of quantifying how structured the system is; a greater deviation from the complete randomness of an ideal gas implies more structure.

It is important to point out that the average $\langle \rho_N(r) \rangle$ is taken over *all particles* of interest, as well as over time. As a typical example we can consider the radial distribution function $g_{OO}(r)$ between all oxygen atoms in liquid water. An RDF of this kind is plotted in figure 2.2, using data from the 2TP *NPT*-simulation performed for this thesis.

¹We take this volume element to be a spherical shell of radius r and thickness dr around the central atom, corresponding to a shell volume of $d\tau = 4\pi r^2 dr$.

The O to O radial distribution $g_{OO}(r)$.

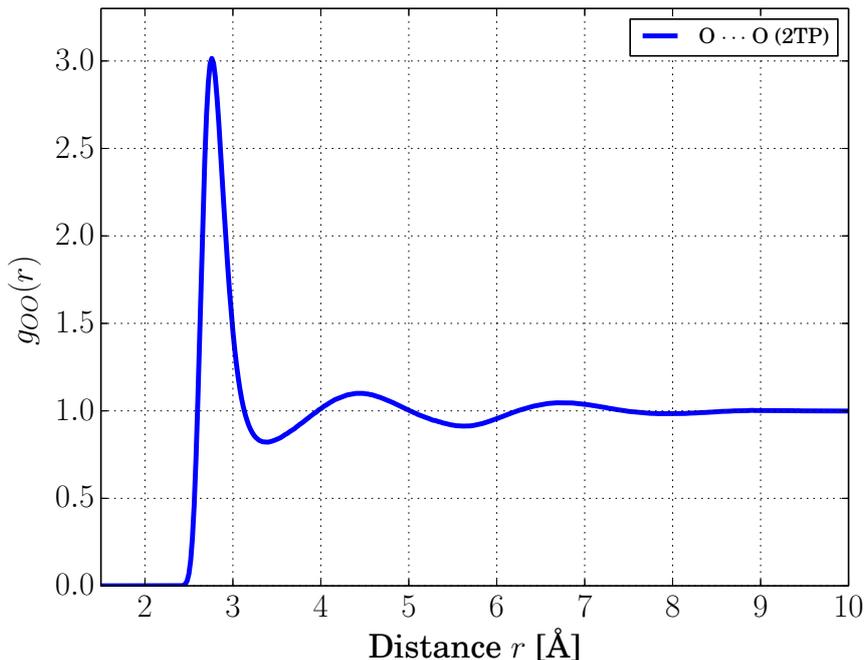


FIGURE 2.2: The oxygen-oxygen RDF calculated from the 2TP NPT-simulation. The vanishing behaviour as $r \rightarrow 0$ as well as the convergence to 1 for increasing r is demonstrated.

If we want to know the average number of particles, or the coordination number, within the first solvation shell, we calculate

$$n_1 = \int_{r_0}^{r_1} \langle \rho_N(r) \rangle d\tau = \int_{r_0}^{r_1} g(r) \rho d\tau = 4\pi \int_{r_0}^{r_1} r^2 g(r) \rho dr, \quad (2.13)$$

where we, in the example of Figure 2.2, take $r_0 \approx 2.5 \text{ \AA}$ and $r_1 \approx 3.3 \text{ \AA}$. Thus integrating up to the first minima in the distribution function. This gives the average number of oxygen atoms within the first solvation shell.

The way $g(r)$ is calculated from simulations is usually by measuring the distances between atoms, sorting the distances into discrete bins, and normalising properly [4].

2.4.2 Spatial distribution function (SDF)

Since the RDF is averaging over all spatial coordinates for some given r , it is a one-dimensional representation of a three-dimensional structure. If the interaction potential in the liquid is not isotropic, which is the case for many molecular liquids [21], the actual local structure is not spherically symmetric. The *spatial* distribution function (SDF) is a useful tool for visualising the particle densities as a function of spatial coordinates relative to the central atom or molecule. Instead of $g_{OO}(r)$, the oxygen-oxygen spatial distribution function in liquid water would be $g_{OO}(x, y, z)$, or $g_{OO}(r, \theta, \varphi)$ in spherical coordinates, which explicitly specifies where in space the neighboring particle is likely to be found.

The SDF defines a three-dimensional map and, by enclosing regions of probability density over some threshold, the three-dimensional structure of the liquid can be visualised. This is exemplified and elaborated upon, in the case of liquid water, in [17, 21].

3 Simulation

This chapter concerns the simulation work performed for this thesis, the method and the production simulation setup. First, in Section 3.1, the decision-making process and simulation procedure is explained, by tracing a path from initial idea to final trajectory. A flowchart of the process is presented and the following text explains the different steps (roughly represented by the indented sections). A summary of the production simulation settings is given in Section 3.2.

3.1 Method

Before an MD-simulation can be performed, various decisions need to be made. Will the potential energy surface come from *ab initio* calculations or are force field methods to be used? If so, what force field? How many (solute) molecules are going to be simulated, and in what solution, if any? If simulating in water solution, what water model should be used? How many water molecules, i. e how big of a simulation box? What will the equilibrium conditions be, and which ensemble should be used? These are some examples of the questions that need answers, before the setup of the simulation can begin. The open-source software GROMACS (version 5.1.2) was used for all molecular dynamics simulations performed for this thesis, and as the thesis title indicates the simulations were performed in liquid water solution. GROMACS is based on force field methods (as opposed to *ab initio*¹), so the quantum chemistry calculations performed were only preparatory, i.e for optimising the molecular structure. A flowchart of the simulation procedure used is shown in Figure 3.1.

¹The thiol-form, 2MP, is almost non-existent in water solution [8], and the simulation of this molecule (in water) is in some sense unrealistic. It is still interesting to perform this simulation to understand the differences in solvation. But it means that *ab initio* MD (without constraints) was not an alternative.

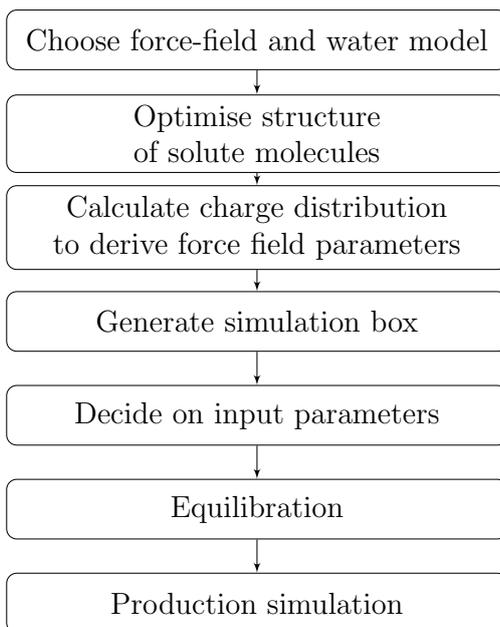


FIGURE 3.1: *Flowchart of the simulation procedure.*

The first step in Figure 3.1, the choice of water model and force field, was the only step that was performed once and for all. All other steps were performed three times, one time for each molecule (2TP, 2AP and 2MP). In GROMACS, the force field defines the inter- and intramolecular interactions of the organic molecules, while the water model defines those interactions for the water molecules [2]. The chosen force field was OPLS/AA, which stands for “*Optimised Potential for Liquid Simulations (All Atom)*”, and is included in GROMACS. It was chosen for its availability and success in modeling organic molecules in solution [16]. The rigid, planar, four-point TIP4P/2005 model was chosen as water model (also readily available in GROMACS), for its ability to reproduce observed liquid water properties at ambient (room temperature and atmospheric pressure) conditions [1]. The word “four-point” refers to how TIP4P/2005 represents a water molecule by four centers: three point charge centers (the atoms), as well as a center for Lennard-Jones interaction, placed in the molecular plane (see Figure 3.2).

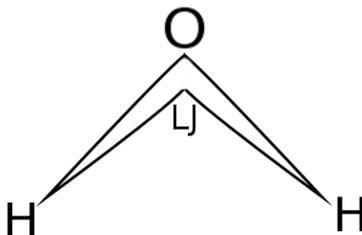


FIGURE 3.2: *Schematic of the TIP4P/2005 four-point water model. In addition to the three atoms there is a Lennard-Jones center, in the molecular plane, bonded to the hydrogens. The vertices of the graph represents the position of the four centers. For further reference see [1].*

The next two steps in the flowchart is the optimisation of the molecular structure and the calculation of the charge distribution. To simulate a molecule, the simulation software needs

a representation of the molecule’s structure; a set of space coordinates for the atoms. To obtain such a set of coordinates, the basic relative positions of the atoms (as sketched in Figure 1.1) were set up using the software MOLDEN. This provided a starting point geometry for optimisation. The software Gaussian was used for the structure optimisation, which employs quantum chemical methods [10]. The basic idea of optimisation is to find the configuration corresponding to the minimum of the potential energy surface (PES), defined by the electronic configuration. Since some parameters of the OPLS/AA force field are derived from quantum chemical calculations, using HF/6-31G* level of theory [16], the same level of theory was used for optimisation. If the optimisation has converged properly on a minimum on the PES, all of the molecular vibrational modes should exhibit positive frequencies. This molecular vibrational data was also provided from the optimisation, and it was checked that these frequencies were indeed positive. Another calculation performed on the quantum level was the charge distribution, also computed by Gaussian. This calculation gave the partial charge of each atom, and accounts for the molecule’s polarisation. The sum of the partial charges should add up to the total molecular charge. This partial charge distribution is a property of the specific molecule, while the force field parameters are more general - applicable to a large set of molecules (see Section 2.2). Charge parameters are difficult to parametrise for general use, which is why the charge calculation was necessary.

Next comes the generation of the simulation box, and thereafter the decision of input parameters. The order here does not really matter, the point is that, in order to run any kind of simulation, the basic simulation settings need to be decided, for example total simulation length, target ensemble and so forth. Generation of the simulation box consists of setting up the size, and type, of the simulation box, as well as putting the solute and solvent molecules in it. More specific to GROMACS, this step also includes the setup of the simulation topology (how GROMACS handles the connections between the atoms). An example of the box generation procedure in GROMACS is given in Section A.3, and the settings used for this thesis is presented in Section 3.2. The input parameters are explained in more detail in Section A.4.

Now the production simulation can start. But the system probably won’t start out exactly at the target equilibrium conditions. So it is natural to divide the parts of the simulation that are not representing the desired conditions (in the time average sense), from the part that is. This first part is called equilibration, and it simply consists of letting the simulation run for a desired number of time steps, until the simulation seem to correctly represent the conditions sought after. The production simulation then consists of all the time steps in the second part, and it is only from this part that any statistics are gathered.

3.2 Production

Three main cases of simulations were performed, one each for 2TP, 2AP and 2MP. Production simulations over 100 ns, in both *NVT* and *NPT* ensembles, were performed, giving a total of six production simulations. The simulations were performed with a time step of 1 fs (10^{-15} s) and thus 100 million steps were calculated for each simulation. Frames were collected every 500th step corresponding to two hundred thousand frames of data for each simulation. From the obtained simulation data the first 10 ns was taken as equilibration time, and so 90 ns of simulation data were used as the basis for analysis.

The simulations were performed in a cubic box with sides of 2.2 nm and periodic boundary conditions. The total number of molecules in each case were 350, where for 2TP and 2MP this meant 349 water molecules. For 2AP, the simulation box contained (except for 2AP itself), 348 water molecules and one Na^+ ion, so that the total system charge would be neutral. The box size and molecule numbers corresponded to a pre-simulation density of $\rho = 997.95 \text{ kg/m}^3$, which is in very good agreement with the reference ambient condition TIP4P/2005 water density of $\rho = 997.9 \text{ kg/m}^3$, given in [1].

4 Results

Production simulations of both constant number, volume and temperature (NVT), and constant number, pressure and temperature (NPT), ensembles were performed. This section only presents and discusses the results from the NPT -simulations, since the structural results from the NVT -simulations were practically identical. The NPT ensemble is what most closely resembles experimental settings, which is why these results form the basis for the following discussion. The benchmark of the simulations is treated first, presenting statistics and structural results for pure water. These performance data are discussed as they are presented. Following this, the radial distribution results, as well as the charge distribution quantum calculation, are presented. These results are discussed in more detail in Chapter 5.

4.1 Performance

When a simulation run is completed the first question one might ask is often “*how well did the simulation go?*”. The answer is not always straightforward, especially if there are no experimental results available for comparison. There exist some natural benchmarks, however. For example, since simulations were performed with a target temperature $T = 300$ K and target pressure $P = 1$ bar, the statistics of these quantities (and others) may be extracted from the simulation (in practice by the GROMACS module `gmx energy`). Ideally the quantity averages should not only approach the target values, they should also remain relatively constant over time. This provides a way of judging how well the performed simulations represent the target ensembles. Due to the statistical nature of the quantities, and the finite simulation time (and to a lesser extent the finite machine precision), the averages will not be precisely constant [9, 2].

The GROMACS output of the mean total energy, temperature, pressure, and system density is shown in Table 4.1. Compared to the average values, the total drift is small (except perhaps in the case of the 2MP pressure), which indicates that the system has converged. GROMACS calculates drift by fitting a straight line to the quantity data, and then simply taking the difference, between the first and last points of that linear fit, as the total drift value [2].

<i>NPT</i> production simulation statistics.					
Quantity	System	Average	Err. Est.	RMSD	Total drift drift
Total Energy [kJ/mol]	2TP	-11786.40	0.34	154.17	-0.05
	2AP	-12803.3	0.2	152.7	-0.8
	2MP	-12056.3	0.3	153.9	-0.8
Temperature T [K]	2TP	299.99	0.01	9.42	-0.03
	2AP	299.995	0.007	9.395	-0.001
	2MP	299.99	0.01	9.44	0.06
Pressure P [bar]	2TP	2.7	0.2	722.2	0.5
	2AP	2.4	0.2	732.3	0.3
	2MP	2.28	0.35	721.45	-1.04
Density ρ [kg/m ³]	2TP	995.7	0.1	14.2	0.2
	2AP	1003.48	0.07	14.11	-0.07
	2MP	995.82	0.05	14.03	-0.12

TABLE 4.1: Table of the simulation statistics from the *NPT* runs.

Looking at the temperature and pressure, the target temperature of 300 K lies within the temperature error limits, but the obtained pressure is closer to 3 bar than to 1 bar. In addition, the RMSD column shows how the pressure fluctuates wildly between datapoints. This fluctuating behavior of the pressure is expected, however, since pressure is a macroscopic, time-averaged, property [9, 2]. Even though the target pressure of 1 bar was not reached, this effect on the structural result seemed small, as is demonstrated in the oxygen-oxygen example below. How much the pressure oscillates depends on both the type of pressure coupling used (in this case the Parrinello-Rahman scheme), and that coupling’s constants, as well as the size of the system. According to the GROMACS manual, the expected fluctuations of a system of 216 water molecules are in the order of 500-600 bar.

Another case for the validity of the simulations can be seen in figure 4.1, which shows the oxygen-oxygen RDFs from the *NPT*-simulations. This structure data is in agreement with the TIP4P/2005 reference [1], which also compares the simulated RDFs with experimental data. Additionally the RDFs in figure 4.1 compare well with experimental results in [19] and [20]. This argues that the simulations were, at least, representative of the chosen model (OPLS/AA and TIP4P/2005).

The O to O radial distribution $g_{OO}(r)$.

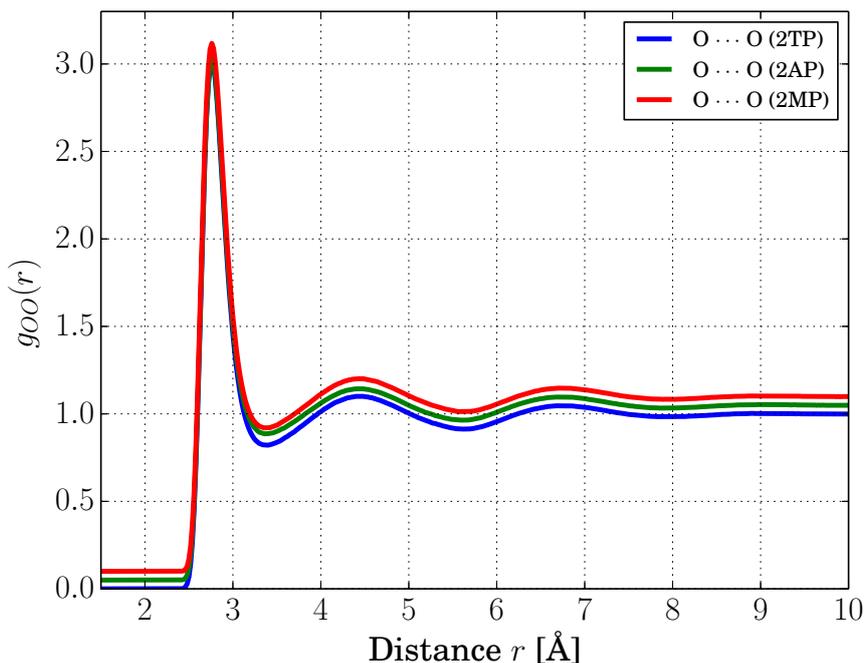


FIGURE 4.1: The oxygen-oxygen RDFs calculated from the *NPT*-simulations, with (2AP) and (2MP) shifted in y by 0.05, 0.1 respectively, for clarity. They are in qualitative agreement with the TIP4P/2005 oxygen-oxygen structure under ambient conditions as shown in [1], as well as experimental results in [19] and [20].

4.2 The liquid structure

In order to compare the water structure, around the solute, the RDFs were computed from the solute molecules' S and N atoms, to all water oxygens as well as to all water hydrogens. Figure 4.2 compares the different radial distributions, while the plots in Figures 4.3, 4.4, 4.5 and 4.6 show the radial distribution functions and cumulative number functions for each RDF calculation (for all three simulations). Figure 4.7 show the distribution calculated from the tautomer proton to water oxygens, for both 2TP and 2MP.

The atom charge distributions, obtained by quantum chemistry calculations, are shown in Table 4.2.

The liquid structure results are discussed in more detail in Chapter 5.

S-H, S-O, N-H and N-O radial distributions.

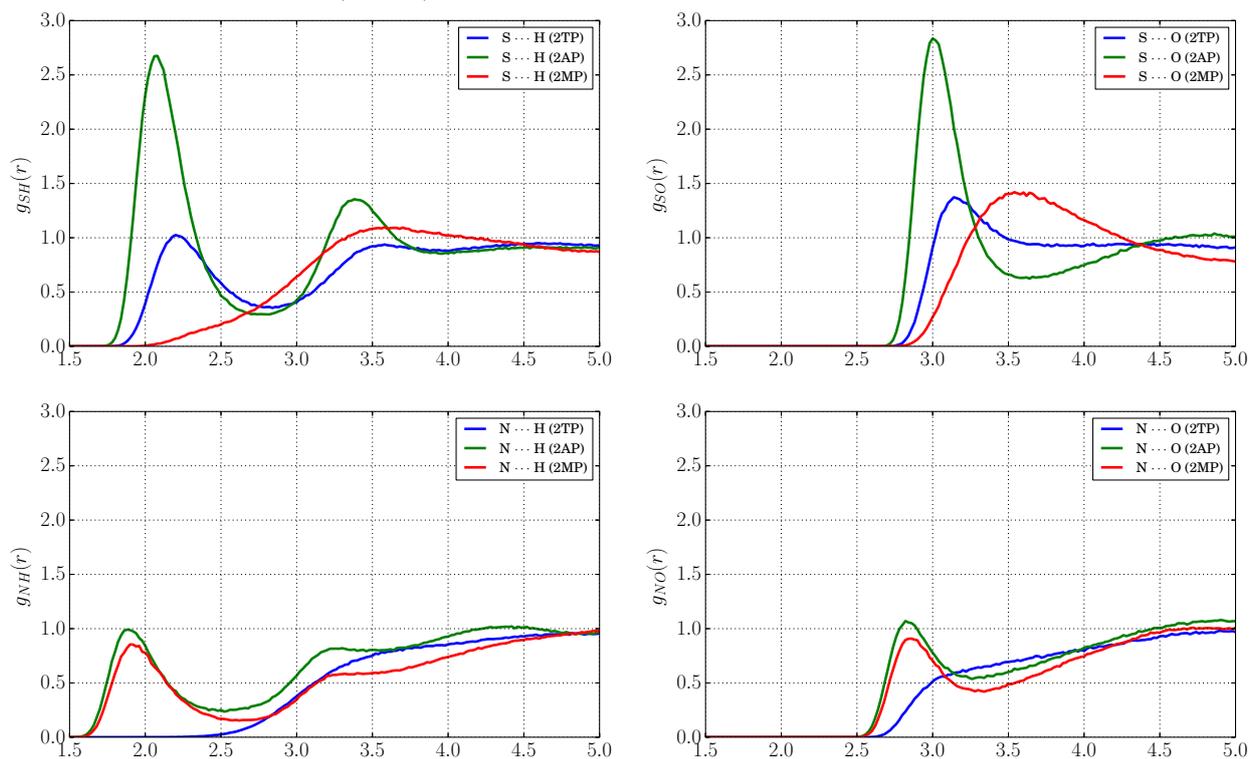


FIGURE 4.2: The sulphur-hydrogen, sulphur-oxygen, nitrogen-hydrogen and nitrogen-oxygen radial distribution functions, from the three main NPT-simulations.

2TP charge distribution		2AP charge distribution		2MP charge distribution	
C1	0.20536	C1	-0.40634	C1	-0.41068
C2	-0.07644	C2	0.74631	C2	0.56624
N1	-0.07006	N1	-0.76523	N1	-0.68527
C3	-0.16707	C3	0.48764	C3	0.45929
C4	-0.00618	C4	-0.56088	C4	-0.41416
C5	-0.07625	C5	0.17629	C5	0.17360
S1	-0.52134	S1	-0.89809	S1	-0.35752
H1	0.17155	H1	-0.02979	H1	0.01916
H2	0.12150	H2	0.12890	H2	0.14865
H3	0.11610	H3	0.04242	H3	0.07730
H4	0.08262	H4	0.07876	H4	0.19251
H5	0.22021			H5	0.23087

TABLE 4.2: The computed charge distribution among the atoms in the three molecules. The 2TP and 2MP net charge is zero, while 2AP has a net charge of -1 . Worth noting is the difference of around 0.4 between the 2AP and 2TP S-site charge distribution, which is greater than the difference of less than 0.1 between the 2AP and 2MP N-site distributions. This is discussed further in Section 5.2.

The S to H radial distribution $g_{SH}(r)$ and cumulative number $n_{SH}(r)$.

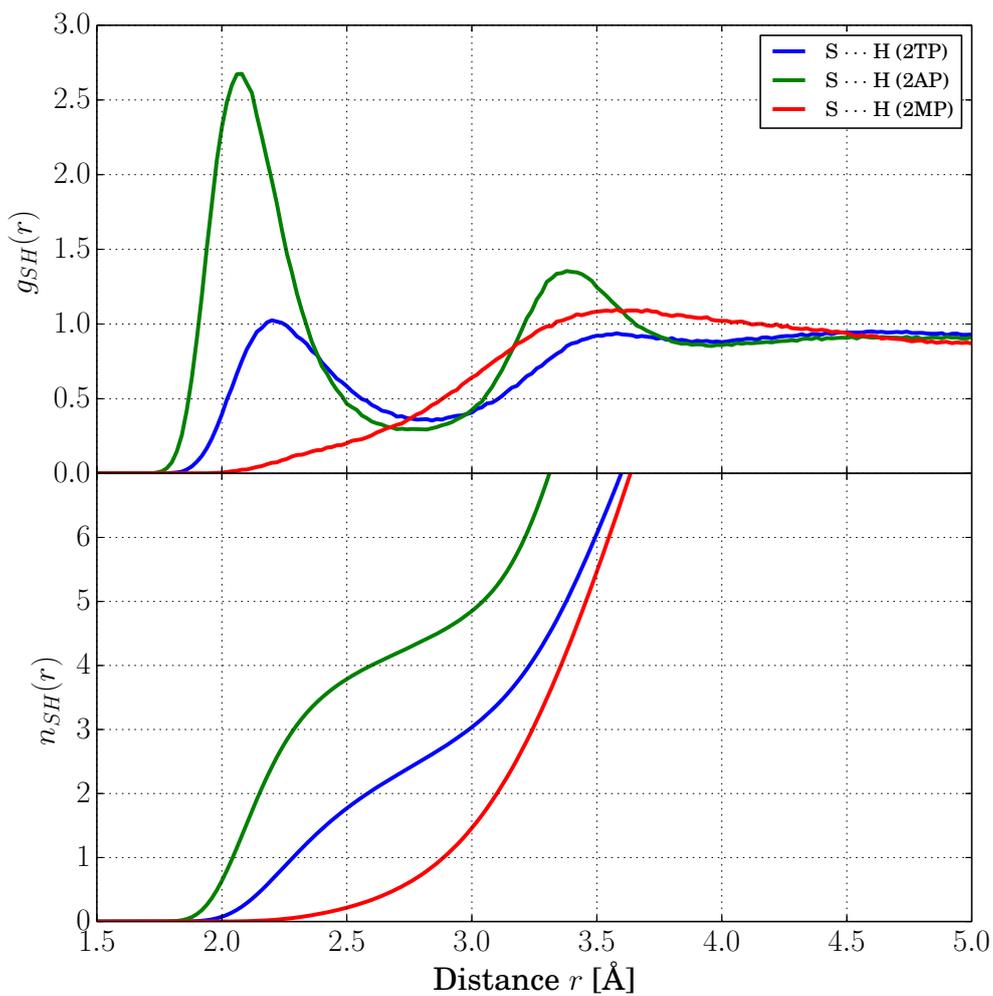


FIGURE 4.3: A comparison between the sulphur-hydrogen(water) RDFs and CNFs from the three main NPT-simulations.

The S to O radial distribution $g_{SO}(r)$ and cumulative number $n_{SO}(r)$.

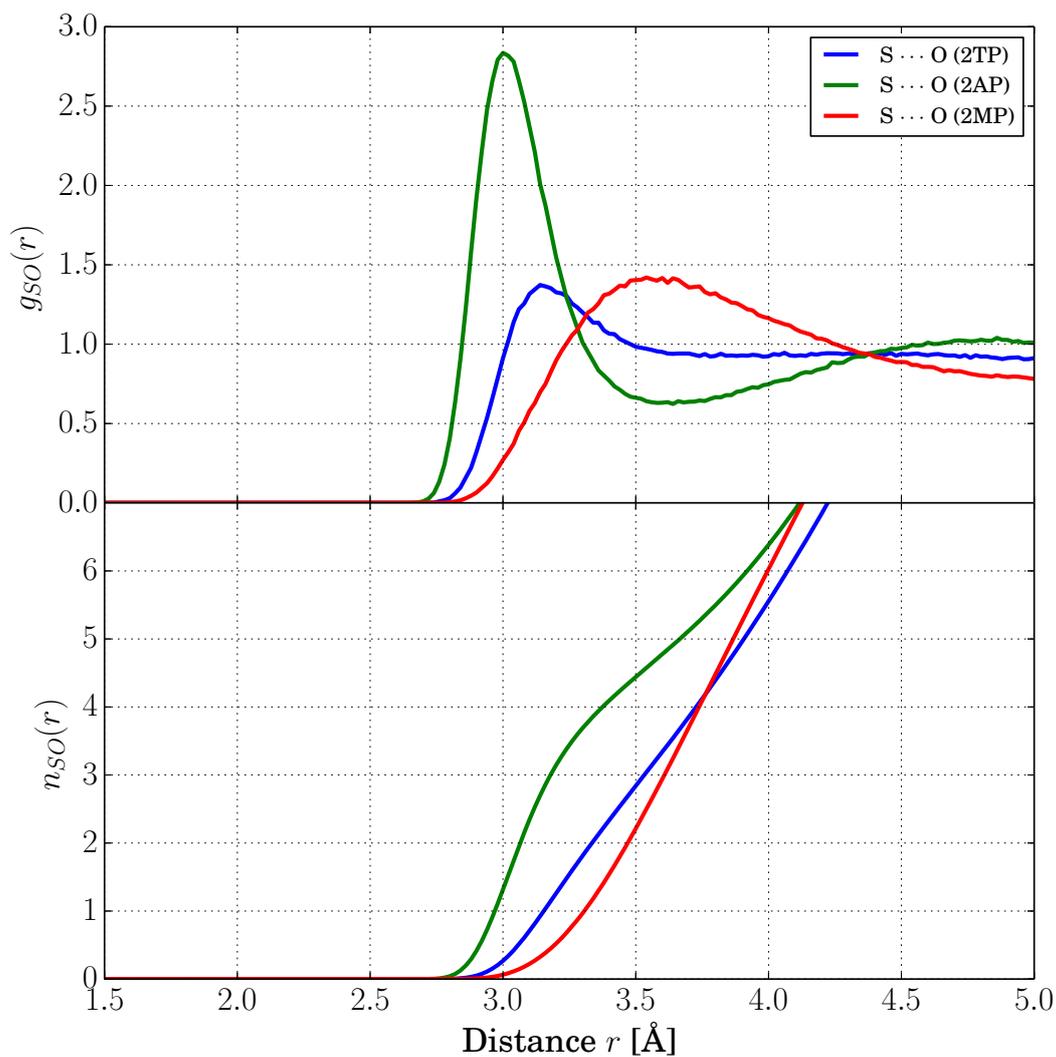


FIGURE 4.4: A comparison between the sulphur-oxygen RDFs and CNFs from the three main NPT-simulations.

The N to O radial distribution $g_{NO}(r)$ and cumulative number $n_{NO}(r)$.

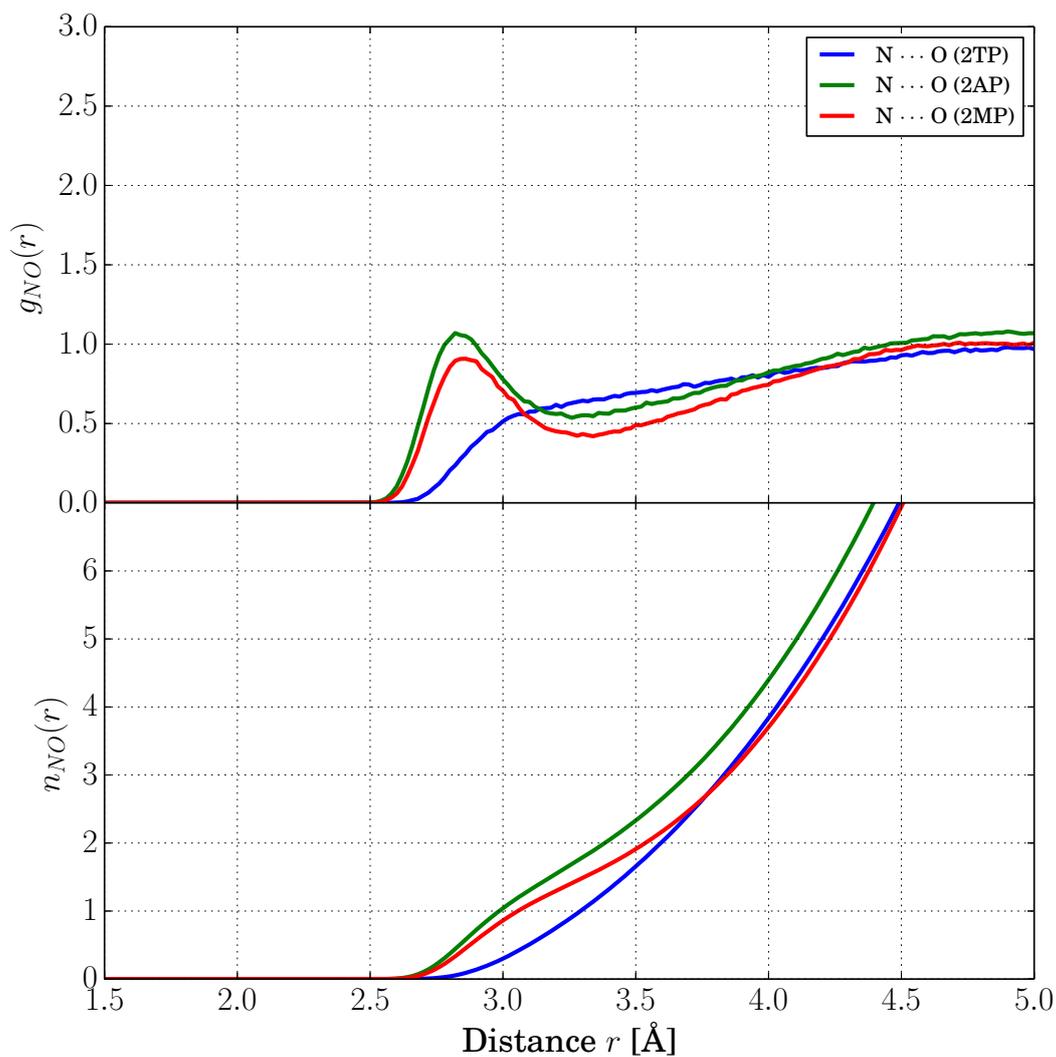


FIGURE 4.5: A comparison between the nitrogen-oxygen RDFs and CNFs from the three main NPT-simulations.

The N to H radial distribution $g_{NH}(r)$ and cumulative number $n_{NH}(r)$.

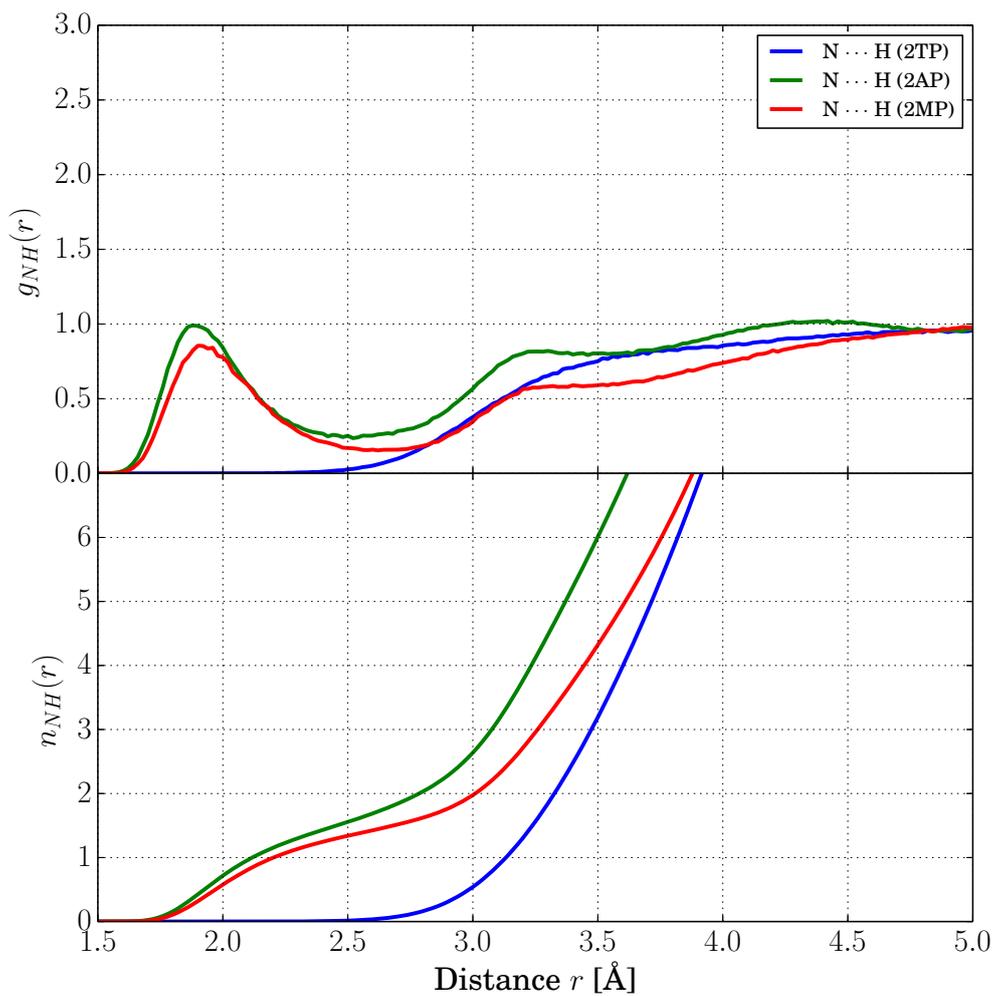


FIGURE 4.6: A comparison between the nitrogen-hydrogen(water) RDFs and CNFs from the three main NPT-simulations.

The tautomer proton to O radial distribution $g_{\text{proton-O}}(r)$ and cumulative number $n_{\text{proton-O}}(r)$.

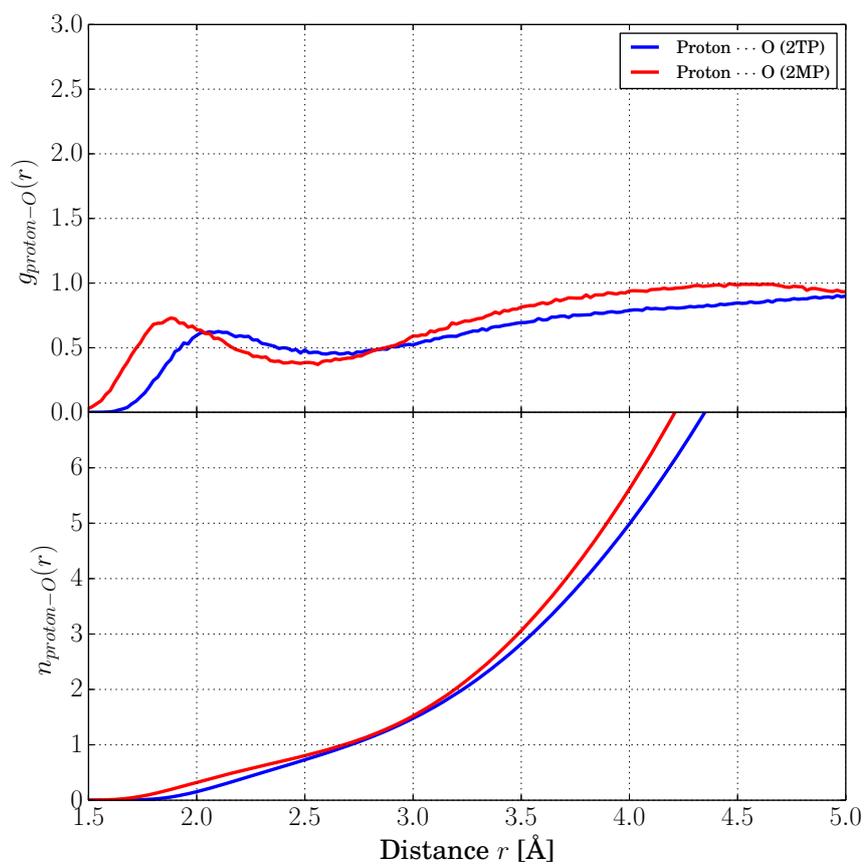


FIGURE 4.7: *RDF and CNF calculated from tautomer proton to water oxygen. In 2TP (blue) the proton is on the N-site, and in 2MP (red) the proton is on the S-site*

5 Discussion and conclusion

This chapter begins by a summary of the hydration structure (derived from the distribution functions), including a suggestion of a specific quantum chemical model for each of the three organic molecules, presented on separate pages. After this, the radial distribution results from Section 4.2 are discussed and compared between the molecules. The flexibility of the studied molecules is discussed, in Section 5.3, where the 2TP S-H radial distribution data is compared between flexible and rigid simulations. A treatment of errors and uncertainties follows, in Section 5.4, before the conclusion is made (Section 5.5).

5.1 Suggestions for the quantum chemistry setup

Based on the analysis of the distribution functions a suggestion for a first solvation shell water configuration is presented, for the purpose of providing a starting point for quantum chemical investigations of the thione-thiol proton transfer process, as outlined in Section 1.2. In addition to coordination numbers and spatial distributions, which indicate the possible configurations, explicit configuration examples, consistent with the distributions, are also given. The presentation first gives the suggestion in text, then illustrates the configuration example and the SDF in a figure, and motivates the suggestions based on the radial distribution results.

It should be kept in mind that the distribution data only gives averages. A look inside, for example, a 4 Å radius of either the S- or N-site, in any given time step of the simulation, will not necessarily show a configuration compatible with the following suggestions. So the configuration examples are constructed to be consistent with the distribution data, and are not based on any particular simulation time step.

2AP

The suggested near N-site water configuration is two water molecules, such that their oxygens are within a 3.3 Å radius.

For the S-site there are on average 5 water molecules within the first solvation shell, in this case with oxygens within an 3.6 Å radius, and this is the suggested configuration. The spatial distribution of water hydrogen and oxygen, as well as an example of the suggested configuration, is shown in Figure 5.1.

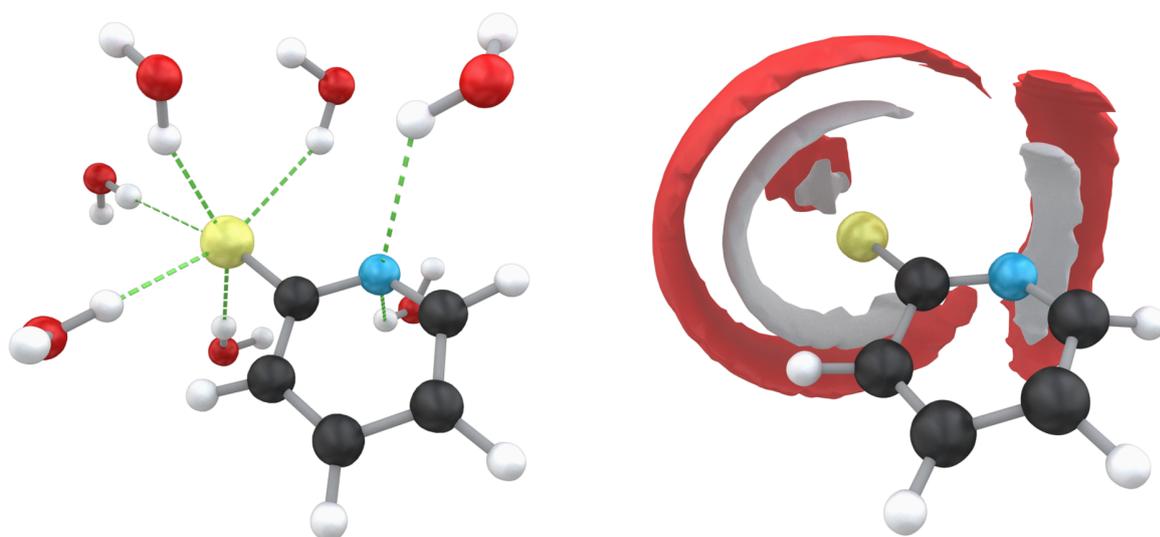


FIGURE 5.1: An example snapshot of the suggested 2AP water configuration (left), green dotted lines added to clarify placements. To the right is the 2AP SDF, calculated from the 90 ns NPT-simulation.

The first minimum of the nitrogen-oxygen RDF, in Figure 4.5, occurs at around 3.3 Å, and integrating up to this point yields a bit more than 1.5 oxygens on average, within this shell. The N-H first minimum is at approximately 2.5 Å, as shown in Figure 4.6, and integration gives 1.5 hydrogens on average. This motivates the suggested water orientation (one water hydrogen toward the N-site and the other toward the bulk liquid), and it is supported by the spatial distribution. Rounding up the integrated atom counts suggests the total of two water molecules near the N-site.

Integrating to the first S-O minimum (at 3.6 Å in Figure 4.4) gives about 4.7 oxygens, while the S-H cumulative number up to the first minimum (at 2.8 Å in Figure 4.3) gives about 4.4 hydrogens, which, together with the SDF, makes the case for the example in Figure 5.1.

2TP

Suggested 2TP near S-site water configuration is 4 water molecules within the first solvation shell, that is with oxygens within the 3.7 Å radius. An example configuration is shown in Figure 5.2.

The 2TP N-site RDFs (Figures 4.6 and 4.5) show no apparent structure, there is no indication of hydrogen bonding between the bulk liquid and the N-site proton. But the spatial distribution (Figure 5.2) show overlapping oxygen and water hydrogen densities near the N-site hydrogen. A possible suggestion, if any, could be to treat this area as a polarisable medium.

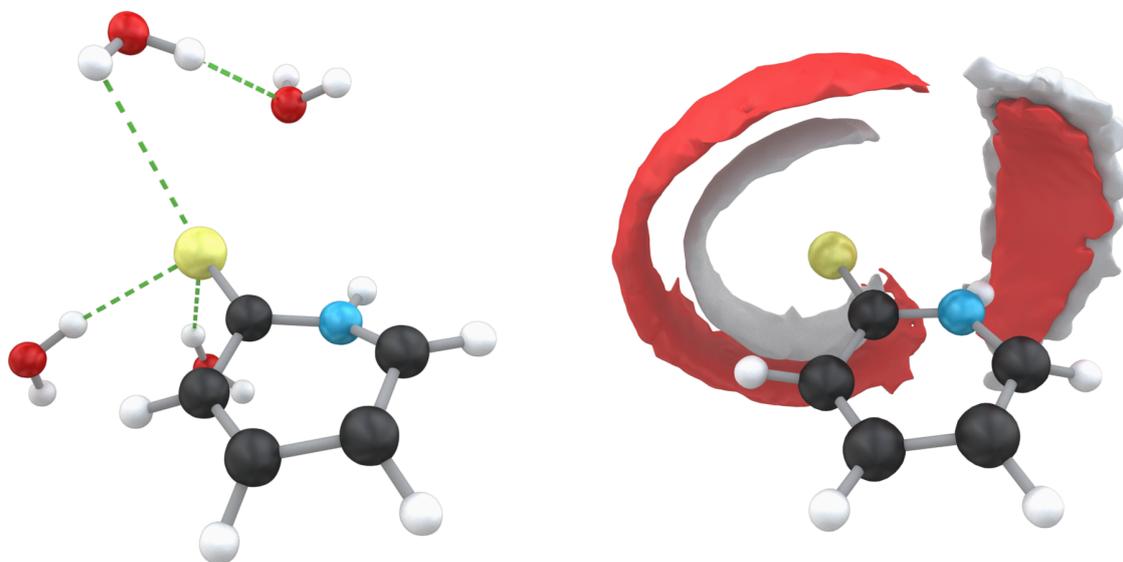


FIGURE 5.2: An example snapshot of the suggested 2TP water configuration (left), green dotted lines added to clarify placements. To the right is the 2TP SDF, calculated from the 90 ns NPT-simulation.

Looking at Figure 4.4, the first S-O minimum occurs at 3.7 Å, and integrates to 4 oxygens within that shell. The S-H RDF for 2TP, in Figure 4.3, integrates to a bit more than 2.5 hydrogens up to the minimum at 2.8 Å. Rounding up, three hydrogens can then be connected with the four oxygens, which indicates that one of the water molecules within this shell is oriented (on average) with its hydrogens outside the first solvation shell.

2MP

The water configuration near the 2MP N-site is suggested to consist of 2 water molecules within the first solvation shell, defined by the oxygens within an 3.3 \AA radius. An example configuration is shown in Figure 5.3.

No apparent structure can be seen in the 2MP S-site RDFs. There are, however, broad peaks in both the S-H (Figure 4.3) and S-O (Figure 4.4) distributions, at 3.5 and 3.6 \AA respectively. The SDF in Figure 5.3 also show an increased density distribution around the S-site, with overlapping oxygen and hydrogen densities. This region could be treated as a polarisable medium in the quantum chemical calculation.

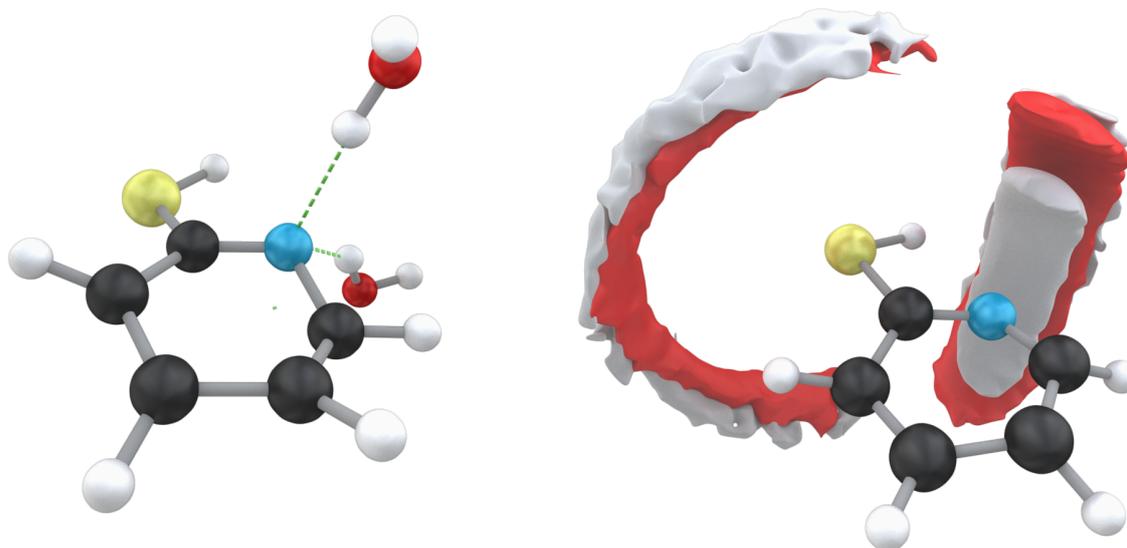


FIGURE 5.3: An example snapshot of the suggested 2MP water configuration (left), green dotted lines added to clarify placements. To the right is the 2MP SDF, calculated from the 90 ns NPT-simulation.

Figure 4.5, presenting the N-O RDF, shows a peak for the 2MP distribution at 2.8 \AA , and the first minimum occurring at $r \approx 3.3 \text{ \AA}$. Integration gives 1.5 oxygens within the 3.3 \AA radius. The first peak in the N-H 2MP RDF (Figure 4.6) falls off to the first minimum at about 2.7 \AA , and integrates to 1.5 hydrogens. Rounding up then gives the total of two water molecules within the first solvation shell, and motivates the example orientation in Figure 5.3.

5.2 General discussion of the results

From the RDF data (Figure 4.2) it can be seen that the N-site structure present in the 2AP plot also seem to be present in 2MP, while the S-O and S-H RDFs show similarities between the 2AP and 2TP structures. The Figures 5.1, 5.2 and 5.3 also show that this behaviour is spatially present as well. Considering that all three molecules are structurally the same, except for that proton on the S-site (2MP), N-site (2TP) or missing (2AP), this characteristic of the distribution functions seem reasonable.

In general the 2AP RDFs have more pronounced maxima and minima, and they are shifted towards lower r , relative to the 2TP and 2MP plots. This difference is more exaggerated in the S-site plots, comparing 2AP and 2TP, than it is comparing the 2AP and 2MP N-site RDFs. The general shift in, and stronger shape of, the 2AP peaks can be explained by the charge distribution, shown in Table 4.2. 2AP has a net charge of minus one, so the interaction with the water dipoles should be stronger than the neutral 2MP and 2TP.

The charge distribution can also serve to explain the stronger difference between the 2AP and 2TP S-site structures, compared to the 2AP and 2MP N-site structure. As seen in Table 4.2, the charge difference between the 2AP and 2TP sulphur is 0.4 e, while the 2AP and 2MP nitrogen difference is 0.1 e. In the electronegative sense, the difference between the 2AP sulphur, and the 2TP sulphur, is greater than between the 2AP nitrogen and the 2MP nitrogen, and this is also reflected in the RDFs.

The presence of the proton on the N- and S-sites seem to block the solvation structure. The RDFs show no indication of strong hydrogen-bonding between water oxygens and the tautomer proton. This is supported by the tautomer proton-to-oxygen distribution in Figure 4.7.

In the context of ultra-fast deprotonation, substantial changes in the solvation is noticed, which will influence the proton dynamics. The proton transfer process is likely coupled to changes in the solvation structure, which can be investigated by further study of excited state dynamics.

5.3 Molecular flexibility

Something not illustrated (at all) by the RDFs, or (very well) by the SDFs, is how flexible the organic molecules were during simulations. That is, how the molecule was flexing between time steps. This is better seen in Figure 5.4, where three successive frames (corresponding to a time difference of 0.5 ps between each frame) from the 2TP simulation is compared. This flexing behaviour was observed in all simulations.

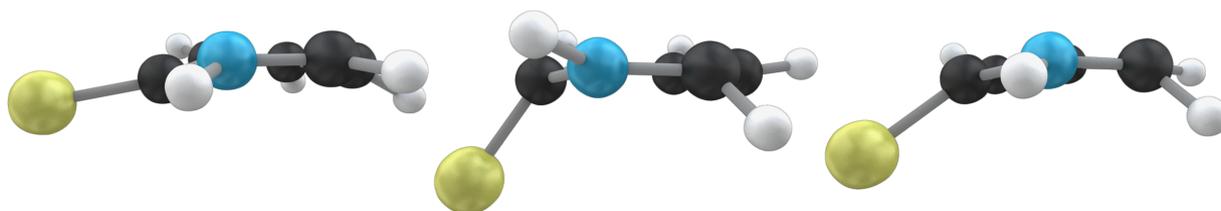


FIGURE 5.4: *Three snapshots of 2TP, taken from successive frames (a time difference of 0.5 ps between each) in the production simulation, to illustrate the flexing; relative atom positions change between frames. Similar behaviour was also observed in the 2MP and 2AP simulations.*

Some flexing is expected, but the amount of flexibility was not quantified (ultimately due to time constraints). However, as mentioned in Section 3.1, the optimisation of a molecule also provided the molecular vibrations. For all three molecules, the motions of the first few vibrational levels were such that the molecular plane was broken, in a similar (but not as extreme) manner as shown in Figure 5.4. Interactions between organic molecule and water could also emphasise these vibrations, which argues for expecting the flexing behaviour. Still, it would be interesting to see how the flexibility of the organic molecules effects the distribution functions. To accomplish this, the production simulations could be performed once more, but with a completely (or as close as possible) rigid molecule; its atoms constrained to its plane, bond stretching and rotations strongly dampened. This was attempted, as mentioned in Section 1.3, but only the rigid 2TP could be simulated properly, and only for a simulation time of 10 ns (with a 0.1 fs timestep, thus with the same number of datapoints as the 90 ns simulations). An example of the RDF data from this 10 ns, rigid, simulation is shown in Figure 5.5, plotted together with the non-rigid 2TP production distribution data.

The S to H radial distribution $g_{SH}(r)$ and cumulative number $n_{SH}(r)$.

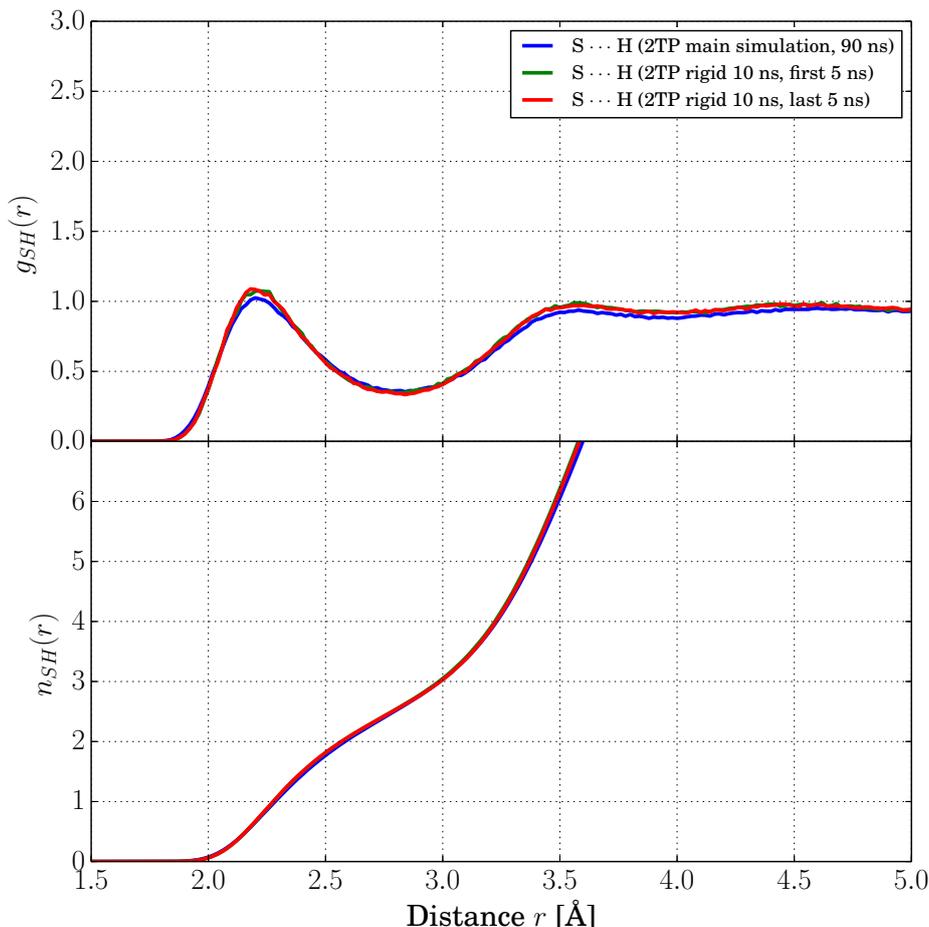


FIGURE 5.5: The sulphur-hydrogen radial distribution and cumulative number from the rigid 2TP simulation. The 100 ns (non-rigid) production data is also plotted for comparison. No major difference between the rigid and non-rigid radial distribution can be seen, and the same was true for the S-O, N-H and N-O data.

It can be seen, in Figure 5.5, that the radial distribution is essentially the same (especially looking at $n(r)$) in both the rigid and non-rigid cases. The distribution data, from the first half of the simulation, aligns well with the data from the second half, which indicates that the rigid simulation has equilibrated. Thus it seems like the rigidity of 2TP does not effect the water structure to any significant degree. Rigid simulations of 2AP and 2MP need to be performed, as well, to determine the flexibility dependence of the water structure in those cases.

5.4 Uncertainties

It does not seem common to provide error bars, for simulated radial distribution data, when the amount of statistics is large (in this case 200 000 datapoints). The statistical errors are not significant for the type of analysis that the radial distributions were used for in this thesis. To provide an example, the standard error of the 2AP sulphur-hydrogen coordination

number (Figure 4.3), up to $r_{2.8} = 2.8 \text{ \AA}$, was calculated, by cutting up the distribution data into five blocks. The blocks consisted of 18 ns of simulation time (corresponding to 36 000 datapoints each), which gave a set of five 2.8 \AA coordination numbers $n_i = n_{SH}(r_{2.8})$, one for each blocks. This is the way GROMACS calculates the statistical errors in the average quantities in Table 4.1, in order to account for any time-correlation effects [2]. The standard deviation was calculated by

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{i=5} (\bar{n} - n_i)^2}, \quad (5.1)$$

and the standard error by

$$\sigma_{err} = \sigma/\sqrt{N}. \quad (5.2)$$

With $N = 5$, and with coordination numbers $n_1 = 4.334$, $n_2 = 4.515$, $n_3 = 4.363$, $n_4 = 4.349$ and $n_5 = 4.515$, this gave an average of

$$n_{SH}(r_{2.8}) = 4.42 \pm 0.04, \quad (5.3)$$

for the 2AP sulphur-hydrogen coordination number, corresponding to a relative error of 1%. This is not big enough of an error to affect the rounding of the coordination numbers (as in the suggestions), nor is it big enough to yield any readable error bars in the plots, given the presented ranges. All radial distributions were based on the same amount of datapoints, and so the error of the above example can be considered representative for the size of the error in all RDFs.

It is worth to reiterate that the reliability of the results should be considered in the context of the chosen model (that is the OPLS/AA force-field and the TIP4P/2005 water model). For example, as shown in [17], the spatial distributions change considerably depending on water model. But as discussed in Section 4.1, the performed simulations pretty closely reproduces water oxygen-oxygen radial distributions, from experiments and TIP4P/2005 simulations both. This makes the case for the simulations being an accurate representation of the chosen model (i. e the method is working as intended), it does *not*, however, say anything about how well the organic molecule-water structural results represents the actual physical system. The judge of that representation is ultimately comparison with experiment.

5.5 Conclusions

The performed 90 ns production simulations, of the tautomers 2AP, 2MP and 2TP, in constant number, pressure and temperature ensembles at near ambient (300 K, 2-3 bar) conditions, reproduced the water oxygen-oxygen structure from both experiment [20, 19] and simulations (using the same water model) [1]. This argued that the simulations were correctly performed, within the chosen model.

The first solvation shell structure, near the N- and S-sites of 2AP, was found to be similar to the structure near respective unprotonated sites in 2MP (the N-site) and 2TP (the S-site). The total negative charge of 2AP could serve to explain its tighter, and more defined, organic molecule-water structure (compared to 2TP and 2MP).

Simulations over 10 ns, of a rigid 2TP molecule (in contrast to the flexible molecules of the 100 ns simulations) showed that the 2TP radial distributions did not significantly depend on the flexibility of the organic molecule. To include 2AP and 2MP in that statement, rigid simulations of those molecules would need to be performed as well.

To see how well the model represents the physical systems, comparison with experiments should be made. The results did, however, provide some substantiated starting points for quantum chemical calculations.

A Appendix

A.1 The Born-Oppenheimer approximation

This section gives a more detailed presentation of the Born-Oppenheimer (BO) approximation, which is central to both molecular dynamics and quantum chemistry. The idea is that the coupling between nuclear and electronic motion is neglected, motivated by the large time scale separation between the respective motions, i. e. the electrons adjust very fast to change in nuclear positions. Under this approximation the electronic part of the Schrödinger equation can be solved separately, using the nuclear positions only parametrically. The resulting potential energy surface then forms the basis for solving the nuclear motion.

The non-relativistic, time-independent Schrödinger equation (SE) is

$$\hat{H}\Psi(X, x) = E\Psi(X, x), \quad (\text{A.1})$$

where X denotes the nuclear coordinates and x the electron coordinates. We expand the Hamiltonian as

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} + \hat{V}_{nn}, \quad (\text{A.2})$$

where the subscripts indicate nuclear and electronic kinetic energies, and nuclear-electronic, electronic-electronic and nuclear-nuclear potential energies. Employing the BO approximation, we can separate the nuclear and electronic part of the wavefunction,

$$\Psi(X, x) = \Psi_n(X)\Psi_e(X, x). \quad (\text{A.3})$$

Now let the electron part of the Hamiltonian,

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} + \hat{V}_{nn}, \quad (\text{A.4})$$

act on the electronic wavefunction,

$$\hat{H}_e\Psi_e(X, x) = E_e(X)\Psi_e(X, x). \quad (\text{A.5})$$

The resulting energy $E_e(X) = \text{nuclear repulsion} + E_{electron}$ depends parametrically on the nuclear coordinates. Combined with the nuclear kinetic energy operator, $E_e(X)$ can be used to solve the nuclear Schrödinger equation

$$\left(\hat{T}_n + E_e(X)\right)\Psi_n(X) = E_{tot}\Psi_n(X). \quad (\text{A.6})$$

This illustrates how, in the BO approximation, the electronic solution to the SE provides a potential energy surface $E_e(x)$, on which the nuclei move.

When the energies of different solutions to the SE come too close to each other, however, the BO approximation breaks down [15].

A.2 The leap-frog algorithm

The leap-frog algorithm is a Verlet-type algorithm, and is the numerical method by which the equations of motions were solved, in the performed simulations. Let us first consider the Verlet algorithm, before showing in what way the leap-frog algorithm is different.

The problem is to solve the equations of motion, for a system of N particles, and given a potential energy surface $V(\mathbf{r})$, where the vector

$$\mathbf{r} = \mathbf{r}(x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_N, y_N, z_N; t), \quad (\text{A.7})$$

denotes a point in the $3N$ -dimensional configuration space (and each configurational coordinate depends implicitly on time). We write the equations of motion for the entire system as

$$-\frac{dV}{d\mathbf{r}} = m \frac{d^2\mathbf{r}}{dt^2}, \quad (\text{A.8})$$

where it is understood that the left hand side represents the negative energy gradient, which equals the configurational space force \mathbf{F} on the system. If we start out at some configuration \mathbf{r}_i (at time t_i), the configuration \mathbf{r}_{i+1} , after the elapse of some small time step Δt , can be obtained by a Taylor series expansion

$$\mathbf{r}_{i+1} = \mathbf{r}_i + (\Delta t) \mathbf{v}_i + \frac{1}{2} (\Delta t)^2 \mathbf{a}_i + \frac{1}{6} (\Delta t)^3 \mathbf{b}_i + \dots, \quad (\text{A.9})$$

where $\mathbf{v}_i = d\mathbf{r}/dt$ is the first time derivative of the configurational position, evaluated at $t = t_i$, and \mathbf{a}_i is the second time derivative (evaluated at $t = t_i$), and so on. By letting $\Delta t \rightarrow -\Delta t$, the configuration at a Δt time *earlier* is obtained:

$$\mathbf{r}_{i-1} = \mathbf{r}_i - (\Delta t) \mathbf{v}_i + \frac{1}{2} (\Delta t)^2 \mathbf{a}_i - \frac{1}{6} (\Delta t)^3 \mathbf{b}_i + \dots \quad (\text{A.10})$$

Adding Equations (A.9) and (A.10) yields

$$\mathbf{r}_{i+1} + \mathbf{r}_{i-1} = 2\mathbf{r}_i + (\Delta t)^2 \mathbf{a}_i + \dots, \quad (\text{A.11})$$

showing that the algorithm is correct to third order in Δt . We can solve for \mathbf{r}_{i+1} to get a recipe for obtaining the configuration at time $t = t_{i+1}$,

$$\mathbf{r}_{i+1} = (2\mathbf{r}_i - \mathbf{r}_{i-1}) + (\Delta t)^2 \mathbf{a}_i, \quad (\text{A.12})$$

which is preceded by obtaining the acceleration through Equation (A.8),

$$\mathbf{a}_i = -\frac{1}{m} \frac{dV}{d\mathbf{r}_i}. \quad (\text{A.13})$$

The notation here hides the fact that the mass m is not, in general, the same for all particles in the system, but the point of the equations is to illustrate the concept.

The recipe of Equations (A.12) and (A.13) is based on knowing the previous configuration \mathbf{r}_{i-1} , which we don't know if we start at configuration \mathbf{r}_0 . But it can be estimated by approximating Equation (A.10) to first order,

$$\mathbf{r}_{-1} = \mathbf{r}_0 - \mathbf{v}_0 \Delta t. \quad (\text{A.14})$$

A smaller time step increases the accuracy of the algorithm, but also the computational effort, and the finite numerical precision of computers are a limiting factor on how small Δt can be taken, as always.

The Verlet algorithm has some disadvantages, specifically a numerical disadvantage in the scheme for obtaining new positions (Equation (A.12)), and the fact that velocities are not explicitly apparent. The acceleration term in Equation (A.12) is proportional to $(\Delta t)^2$, which is a small number, while the first term $(2\mathbf{r}_i - \mathbf{r}_{i-1})$ is a difference between two large numbers. This can lead to truncation errors in the addition of these terms, due to finite machine precision. If a constant temperature ensemble is to be simulated, velocities corresponding to the correct distribution needs to be generated, which is problematic when the velocities are not explicit in the integration.

The purpose of the leap-frog variant of the Verlet algorithm is to come to terms with these disadvantages. In order to introduce velocities explicitly into the algorithm, first consider the Taylor expansions of the velocity, at half of the time step Δt ,

$$\begin{aligned}\mathbf{v}_{i+1/2} &= \mathbf{v}_i + \frac{1}{2}\Delta t\mathbf{a}_i + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2\mathbf{b}_i + \dots, \\ \mathbf{v}_{i-1/2} &= \mathbf{v}_i - \frac{1}{2}\Delta t\mathbf{a}_i + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2\mathbf{b}_i + \dots\end{aligned}\tag{A.15}$$

To second order, we get that

$$\mathbf{v}_{i+1/2} - \mathbf{v}_{i-1/2} = \Delta t\mathbf{a}_i,\tag{A.16}$$

or, solving for the forward step,

$$\mathbf{v}_{i+1/2} = \mathbf{v}_{i-1/2} + \Delta t\mathbf{a}_i.\tag{A.17}$$

Now we can define the velocity at $i + 1/2$ as the finite difference

$$\mathbf{v}_{i+1/2} \equiv \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{\Delta t},\tag{A.18}$$

which immediately, together with Equations (A.17) and (A.13), gives another recipe for acquiring \mathbf{r}_{i+1} ,

$$\mathbf{r}_{i+1} = \mathbf{r}_i + \mathbf{v}_{i+1/2}\Delta t.\tag{A.19}$$

The cause of the leap-frog name now becomes evident, since the velocities and positions are always out of phase by $\frac{1}{2}\Delta t$.

These Verlet type integrators are so called *symplectic* integrators, and are preferable to for example Runge-Kutta methods, since they conserve the total energy much better, over long simulation times [15].

A.3 Molecular dynamics using GROMACS

The GROMACS simulation engine is based on the principles outlined in Section 2.1 [2], but it is not obvious how the theory applies in actual software. This section serves to explain the general simulation workflow in GROMACS.

Terminal-type text in this section, like `this`, either refers to a terminal command, or to software concepts like file types and GROMACS attributes, and what is meant should be

obvious from the context.

GROMACS is a command-line software, built as a collection of modules (as of version 5), each performing different tasks in the simulation pipeline. The modules handles various input and output files, which often are human-readable text files. The modules are called via the command-line by `gmx <module> <I/O, flags, ...>`, so for example `gmx rdf` calls the module for computing radial distribution functions.

There are three basic input files that GROMACS needs in order to run an MD-simulation. There needs to be a *configuration file*, specifying all atoms in the system (including solvent atoms) by their element, identifier, spatial coordinates (relative to some arbitrary origin) and (if applicable) velocities, as well as to which molecule each atom belong.

The software also needs a *topology file* (file extension `.top`). This file includes the *constant* attributes of each atom, in contrast to the coordinates and velocities, which obviously change during simulation. An example of a constant attribute is `atomtype` which tells GROMACS what force field parameters should be used for the atom in question. The parameters are then pulled from the force field database during simulation. Another constant attribute is the partial *charge* on each atom in a given molecule.

The topology file also lists the bonds and interactions between the atoms in a molecule, which motivates the name *topology*, since the file defines how the atoms in the system are connected. The attributes, bonds and interactions specified in the topology file only concern the solute molecule (or molecules). The solvent molecules, usually H_2O , are simply grouped together under the default label `SOL`. The solvent parameters and attributes for the atoms in each molecule belonging to `SOL` are then pulled from the chosen solvent model database (in this case `TIP4P/2005`). Both force field (in this case `OPLS/AA`) and water model are specified at the top of the topology file.

The configuration file and the topology file describes the microscopic chemical system (i. e atoms, molecules and the parameters for their interaction), whereas the *simulation input parameter file* (extension `.mdp`) specify the simulation settings, such as the number of simulation steps, time step size, target temperature and canonical ensemble. The `.mdp`-file for this project's production simulations is explained in more detail in Section A.4.

Below is an example of the simulation box generation, which is performed by the `editconf` module in GROMACS:

```
gmx editconf -f 2MP.pdb -o 2MP_box.gro -bt cubic -box 2.2
```

This command generates a cubic box with sides of 2.2 nm, and puts the (in this case) 2MP-structure inside the box, providing an output file with this information. This output file is then used when the box is filled with solvent, exemplified by the command

```
gmx solvate -cs tip4p.gro -cp 2MP_box.gro -p topol.top -o 2MP_solv_box.gro  
-maxsol 349.
```

The `solvate` command takes the 2MP box configuration and fills it with the selected solvent configuration, `TIP4P/2005` (in this case with a maximum of 349 water molecules), and the topology file is also updated to reflect that there now is solvent in the system. Out comes a new configuration file specifying all the initial data for the water molecules and 2MP. This

file is then used as basis for energy minimisation.

Now assume the relaxation of the system has been performed; a relaxed configuration file has been obtained. This configuration file is used as the starting point for the production simulation. The command

```
gmx grompp -f input_params.mdp -c relaxed_config.gro -p topol.top
-o mdready.tpr
```

calls the `grompp` module, which takes input parameters, configuration file and topology and produces a binary run `.tpr`-file. This binary run file is then supplied to the module `mdrun`, which is the main simulation engine. After the simulation is done (approximately 25 hours of simulation time for each production simulation for this thesis), the `mdrun` module has produced a number of files, which contain the simulation data. Of particular interest are the `.trr`-files; the trajectories. These contain the configurations and velocities in every saved time step, and are the basis for analysis.

A.4 Input parameters

The GROMACS input parameters for the *NPT* simulations are listed in Table A.1. The *NVT* input parameters were identical except that all barostat parameters were removed. They are listed as they appear in the `.mdp`-file, reformatted for readability. Some are not very interesting (like the seed for velocity generation), and to consider them all, in detail, is outside the scope of this thesis. The point is to give further insight on the simulation process, and for simulation replicability. The integrator, the thermo/barostat, and Ewald summation, have been discussed (or briefly mentioned) in Section 2.1. But Particle Mesh Ewald is an $N \log(N)$ -method for computing long-range electrostatics [6], which gives much shorter simulation times.

GROMACS .mdp-file for the <i>NPT</i> simulations	
Integrator:	md (leap-frog)
Number of steps:	100 000 000
Timestep:	0.001 ps
Cut-off scheme:	Verlet
r_{coulomb} :	1.0 nm
r_{vdW} :	1.0 nm
Coulomb-type	PME (Particle Mesh Ewald)
PME order:	4
FFT grid spacing:	0.12
Thermostat:	Nosé-Hoover
Time constant τ_T :	0.2 ps
Reference temperature T :	300 K
Barostat:	Parrinello-Rahman
Type:	Isotropic (uniform scaling of box vectors)
Time constant τ_P :	1.0 ps
Reference pressure P :	1.0 bar
Isothermal compressibility of water:	$4.5 \times 10^{-5} \text{ bar}^{-1}$
Periodic boundary conditions:	<i>xyz</i>
Generate velocities:	yes
Temperature for <i>v</i> -generation:	300 K
Generation seed:	3653

TABLE A.1: *Input parameters for the constant pressure simulations. The parameters, as presented here, are re-written to enhance readability.*

Bibliography

- [1] J. L. Abascal and C. Vega. A general purpose model for the condensed phases of water: TIP4P/2005. *The Journal of chemical physics*, 123(23):234505, 2005.
- [2] M. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team. *GROMACS users manual version 5.1.2 (2016)*.
- [3] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19 – 25, 2015.
- [4] M. P. Allen and D. J. Tildesley. *Computer Simulations of Liquids*. Oxford University Press, 1st edition, 1991.
- [5] M. Brehm and B. Kirchner. Travis-a free analyzer and visualizer for Monte Carlo and molecular dynamics trajectories. *Journal of chemical information and modeling*, 51(8):2007–2023, 2011.
- [6] T. Darden, D. York, and L. Pedersen. Particle Mesh Ewald: An NlogN method for Ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.
- [7] R. Du, C. Liu, Y. Zhao, K.-M. Pei, H.-G. Wang, X. Zheng, M. Li, J.-D. Xue, and D. L. Phillips. Resonance raman spectroscopic and theoretical investigation of the excited state proton transfer reaction dynamics of 2-thiopyridone. *The Journal of Physical Chemistry B*, 115(25):8266–8277, 2011.
- [8] S. Eckert, P. Miedema, W. Quevedo, B. O’Cinneide, M. Fondell, M. Beye, A. Pietzsch, M. Ross, M. Khalil, and A. Föhlisch. Molecular structures and protonation state of 2-mercaptopyridine in aqueous solution. *Chemical Physics Letters*, 647:103 – 106, 2016.
- [9] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to applications*. Academic Press, 2nd edition, 2012.
- [10] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. nox, J. B. Cross,

- V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision E.01. Gaussian Inc. Wallingford CT 2009.
- [11] G. H. Grant and W. G. Richards. *Computational Chemistry*. Oxford University Press, 1st edition, 1996.
- [12] J.-P. Hansen and I. R. McDonald. *Theory of Simple Liquids*. Academic Press, 3rd edition, 2006.
- [13] T. L. Hill. *Introduction to Statistical Mechanics*. Addison-Wesley, 1st edition, 1962.
- [14] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.
- [15] F. Jensen. *Introduction to Computational Chemistry*. Wiley, 2nd edition, 2009.
- [16] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.
- [17] P. G. Kusalik, I. M. Svishchev, et al. The spatial structure in liquid water. *Science(Washington)*, 265(5176):1219–1221, 1994.
- [18] D. V. Schroeder. *An Introduction to Thermal Physics*. Addison-Wesley, 1st edition, 2000.
- [19] L. B. Skinner, C. Huang, D. Schlesinger, L. G. M. Pettersson, A. Nilsson, and C. J. Benmore. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide Q-range. *The Journal of chemical physics*, 138(7):074506, 2013.
- [20] A. Soper. The radial distribution functions of water and ice from 220 to 673 K and at pressures up to 400 MPa. *Chemical Physics*, 258(2):121–137, 2000.
- [21] I. M. Svishchev, A. Y. Zassetsky, and P. G. Kusalik. Solvation structures in three dimensions. *Chemical Physics*, 258(2–3):181 – 186, 2000.
- [22] J. M. Thijssen. *Computational physics*. Cambridge University Press, 2nd edition, 2010.