HOVEDOPPGAVE
2004

Polarizable force fields for flexible molecules

Kandidatens navn
Asbjørn Holt

Faglærer:
Professor Per-Olof Åstrand, NTNU
Professor Gunnar Karlström, Lunds Universitet

NTNU
Fakultet for naturvitenskap og teknologi
Institutt for kjemi

Erklæring/Declaration

Jeg erklærer herved at dette er et uavhengig arbeid utført i henhold til de retningslinjer som er fastlagt av Norges Teknisk-Naturvitenskaplige Universitet (NTNU).

I hereby declear that this is an independent work according to the exam regulations of the Norwegian University of Science and Technology.

Trondheim, 12th July 2004

....
Asbjørn Holt

Preface

This master thesis concludes the Sivilingeniør (MSc) education in chemistry at the Norwegian University of Science and Technology (NTNU). The work has been performed under the supervision of Professor Per-Olof Åstrand (NTNU) and Professor Gunnar Karlström (Lund University), and has mainly been conducted at the Department of Theoretical Chemistry at Lund University. The total weight of the thesis is 30 STP, equivalent to 20 weeks work. During this study the following things have been done;

- Quantum chemical computations at HFSCF, CASSCF and MP2 level of theory
- Comparison of two different methods for calculation of atomic charges
- Parameterization of atomic charges, dipoles and polarizabilities
- Parameterization of the intramolecular energy
- Implementation of two simulated annealing algorithms
- Parameterization of a non-linear repulsion energy expression using three different algorithms
- Parameterization of the dampening term in the dispersion energy
- Implementation of a new Monte Carlo integrator in Protomol
- Implementation of algorithms for;
  - repulsion energy
  - dispersion energy
  - induction energy
  - computation of the cross term energy
  - charge computation
  - computation of the dipole moment
  - computation of the polarizability tensor
  - perturbation of the system, including update of the maximum displacement for the various degrees of freedom
  - computation of the zero-point vibrational energy
  - Jacobi routine for computation of eigenvalues

- Samplings routines for;
  - the radial distribution function
  - bond and angle distribution
  - zero-point vibrational spectra
  - combinations of bonds and angles
- Modification of;
  - the computation of the angle energy
  - the computation of the bond energy
  - electrostatic energy
- Monte Carlo simulations of various scenarios
- Report writing

I want to thank Professor Gunnar Karlström for all his help during this project, and for welcoming me to Lund. It has been 20 fantastic weeks. Furthermore, I would like to thank Professor Per-Olof Åstrand for all his help and for suggesting to travel Lund. Last I would like to thank Dr. Thierry Matthey for his help during the implementation phase of the project.

According to the regulations of NTNU every source which has contributed to the work in a masters thesis must be quoted. Due to this some unorthodocs quoting may occur, however the author will do his best to provide these sources to anyone interested[1].

The modified version of Protomol used in this work and a Portable Document File (pdf) version of this report is included on the CD attached to this report.

[1] e-mail; asbjorh@phys.chem.ntnu.no

Abstract

Simulations provides the possibility to study large molecular systems. In order to increase the accuracy of results from such simulations, it is important to improve the physical description of the molecular interactions. In this study a force field for flexible water molecules is constructed. To describe the electrostatic properties of the water molecules in this force fields, atomic charges and dipole moments are used. These charges and dipole moments are formulated as functions of the molecular geometry. The intramolecular energy of the water molecules are formulated using Simon-Parr-Finlan terms for the bond stretches and the angle bending is described by a Urey-Bradley and cosine term. In addition to these a cross term between the two bond lengths is included. Polarization is also included in the model in the form of induced point dipoles. To describe this polarization, atomic polarizability tensors are used. The atomic polarizability tenors are described as a function of the molecular geometry, in a manner similar to that of the electrostatic properties. The Van der Waals interactions between the molecules are modeled using a repulsion energy term similar to the one used in the NEMO potential, and the dispersion energy is modeled using a London expression with a dampening term. In addition to these energy terms, the zero-point vibrational energy is included in the force field.

The force field is parameterized using data from ab inito calculations. The electrostatic properties, the polarizability tenors and the intramolecular energy are modeled from water monomer calculations at the CASSCF level of theory, using the aug-cc-pVTZ basis set. To describe the atomic properties the LoProp method is used. The parameters in the expression for the repulsion energy is parameterized from a series of water dimer calculations at the HFSCF level of theory using the same basis set. The dampening term in the dispersion energy is modeled from dimer calculations at the MP2 level of theory, using the aug-cc-pVTZ basis set. To correct for BSSE, counter poise correction are used in all dimer calculations.

The force field is implemented in the Protomol framework, and Metropolis Monte Carlo simulations are used to study its performance. These simulations are performed on a system containing 216 water molecules, using periodic boundary conditions. Simulations are done for rigid molecules, and flexible molecules with and without the zero-point vibrational energy. The vibrational spectra, distributions of the internal degrees of freedom and radial distribution functions are computed. From these simulations it is found that the zero-point vibrational energy has the effect of increasing the bond lengths and angle. The water geometry found are in agreement with other theoretical and experimental values. Furthermore it is found that the force field gives a shift in the bond length comparable to that of experimentally measured, with a slightly better value when the zero-point vibrational energy is included. The vibrational spectra calculated were in the same area as the experimental values, however the results were not fully satisfactory. The radial distribution functions computed shows an agreement with the experimental values, however it is found that further improvements of the repulsion energy are necessary.

Contents

List of Figures

List of Tables

List of symbols

Symbol	*Definition/description*	*Dimension*
i	Imaginary number	Scalar
h	Planck's constant	Scalar
$\hbar$	$h/2\pi$	Scalar
$\frac{\partial}{\partial q}$	Partial derivative with respect to q	Scalar
$\hat{H}$	Hamilton operator	Operator
ϕ	Wave function	
E	Eigenvalue	Scalar
M	Mass of nucleus	Scalar
∇_i^2	Momentum operator	Operator
m_0	Mass of electron	Scalar
V	Potential energy	Scalar
Z_i	Number of elementary charges	Scalar
ϵ_0	Dielectric constant *in vacuo*	Scalar
e	Elementary charge	Scalar
ψ_i	Eigenfunction	
$E_{Correlation}$	Correlation energy	Scalar
E_{Exact}	Exact solution to the Schödinger equation	Scalar
$E_{Hartee\text{-}Fock}$	Energy at the Hartree-Fock level of theory	Scalar
$E_{Relativistic}$	Relativistic energy	Scalar
C_i	Coefficient	Scalar
λ_i	Expansion parameter	Scalar
λ_i	Eigenvalue	Scalar
$a_{p,i}^{(n)}$	Coefficient	Scalar
$E_i^{(n)}$	nth order energy correction	Scalar
$\mathcal{V}$	Perturbation in MP2	Operator
$\mathcal{J}_j$	Coulomb operator	Operator
$\mathcal{K}_j$	Exchange operator	Operator
$< O >$	Expectation value of operator $\hat{O}$	Scalar
$\mathbf{D}$	1-electron density matrix	Matrix
$\mathbf{O}$	Integrands of operator $\hat{O}$	Matrix
$\mathbf{Y}$	Dependent variable	Vector/Matrix
$\hat{\mathbf{Y}}$	Predicted values of dependent variable	Matrix
$\mathbf{B}$	Regression coefficients	Vector/Matrix
$\mathbf{X}$	Independent variable	Matrix
ϵ	Root mean square error	Scalar

B_0	Parameter in Simon-Parr-Finlan potential	Scalar
b_n	Parameter in Simon-Parr-Finlan potential	Scalar
r	Bond length	Scalar
r_e	Equilibrium bond length	Scalar
V_{angle}	Energy contribution from angle bending	Scalar
k_{UB}	Urey-Bradley force constant	Scalar
$r_{1,2}$	Distance between hydrogen atoms	Scalar
r_{UB}	Rest length between hydrogen atoms	Scalar
k_θ	Angle bending force constant	Scalar
θ	Angle	Scalar
θ_e	Angle rest size	Scalar
V_{cross}	Cross term energy contribution	Scalar
C_1	Force constant in cross term	Scalar
V_{elec}^{AB}	Electrostatic energy between two atoms	Scalar
$\vec{R}$	Vector between two atoms	Vector
$\lvert\vec{R}\rvert$	Norm of vector between two atoms	Scalar
$\vec{\mu}_B$	Atomic dipole moment of atom B	Vector
$V_{induced}$	Polarization energy contribution	Scalar
$\mathbf{T}_{IJ}$	Interaction tensor	Matrix
T_{ik}	Component ik of interaction tensor	Scalar
$\vec{\mu}_I$	Induced dipole moment	Vector
$\vec{E}_I$	External electric field	Vector
α_I	Atomic polarizability	Matrix
$\vec{\mu}_J^{perm}$	Permanent atomic dipole moment	Vector
σ	Lennard-Jones parameter	Scalar
ϵ	Lennard-Jones parameter	Scalar
ΔV_{HFSCF}	Energy difference between dimer and monomer at HFSCF level	Scalar
V_{rep}	Exchange repulsion energy contribution	Scalar
κ_{ij}	Atom pair exchange repulsion parameter	Scalar
$\alpha_{i,j}$	Atom pair exchange repulsion parameter	Scalar
V_{disp}	Dispersion energy contribution	Scalar
ΔV_{MP2}	Energy difference between dimer and monomer at MP2 level	Scalar
B_{mn}	Dispersion energy parameter	Scalar
b_{mn}	Dispersion energy parameter	Scalar
V_{zpv}	Zero-point vibrational energy contribution	Scalar
$\mathbf{V}''_{pot}$	Hessian of the potential energy	Matrix
$\mathbf{M}$	Matrix for mass-weighting of coordinates	Matrix
ν_i	Vibrational frequency	Scalar
Γ	Phase space	
$d\Gamma$	Sub-volume of phase space	
H	Hamiltonian	
k	Boltzmanns constant	Scalar
$\mathcal{P}$	Probability of configuration in phase space	Scalar

$\vec{p}$	Momentum of particle	Vector
$\vec{q}$	Position of particle	Vector
Z	Configurational partition function	
$n^{(N)}$	N-particle distribution function	
$g(r)$	Radial distribution function	
ρ	Particle density	Scalar
V	Volume	Scalar
ρ_m	Probability of configuration	Scalar
$\pi_{m,n}$	Probability of going from configuration m to n	Scalar
$a.u$	Atomic unit	Scalar
E_h	Hartree (unit of energy in the atomic unit system	Scalar
e	Charge of an electron ($1.6022 \cdot 10^{-19}$ C)	Scalar
ea_0	Unit of dipole in atomic unit system ($8.47384 \cdot 10^{-30} Cm$)	Scalar
$\mathring{A}$	Ångström (10^{-10} m)	Scalar
D	Debay ($3.33564 \cdot 10^{-30}$ Cm)	Scalar
Deg	Degrees	Scalar
a_0	Bohr radius ($5.29177 \cdot 10^{-11}$ m)	Scalar
$\vec{v}_B$	Coordinate vector	Vector
$\mathbf{B}$	Matrix with basis vectors	Matrix

Abbreviations used in this work

Abbr.	*Abbreviation(s)*
HFSCF	Hartree-Fock self consistent field
MP2	Many-body perturbation theory
CI	Configuration interaction
MCSCF	Multi configurational self consistent field
CASSCF	Complete active space self consistent field
LoProp	Localized Properties
BSSE	Basis set superposition error
cp	Counterpoise (correction)
DH	Dinur-Hagler
ZPV	Zero-point vibrational (energy contribution)
MC	Monte Carlo
MD	Molecular Dynamics
GATP	Generalized atomic polar tensor
OLS	Ordinary Least Squares
PLS	Partial Least Squares
SA	Simulated annealing
SA-M	Simulated annealing using Metropolis algorithm
SA-NR	Simulated annealing using Numerical Recipes algorithm
GA	Genetic algorithm
RMSE	Root mean square error
SPF	Simon-Parr-Finlan (potential)
QM	Quantum mechanics
QM/MM	Quantum mechanics/molecular mechanics
RDF	Radial distribution function

For most of the numerical results the following notation are used;

$$1.0(3) = 1.0 \cdot 10^3 = 1000$$

"All is water"

- Thales from Milet, 500 B.C.

Chapter 1

Introduction

According to Allen and Tildesley [2] the early theoretical studies of liquids where done studying the packing of large numbers of gelatin balls. This method of studying liquids was of course quite crude, but managed to give a relatively good picture of the structure of a simple liquid. As the development of electronics gave increasingly more powerful computers, it became feasible to solve complex mathematical models of liquids through the use of simulations. The first simulation of a liquid was done in the early fifties by Metropolis et al. [3]. The first systems to be studied with computer simulations where atomic systems, which where modeled using hard spheres or Lennard-Jones potentials. In the beginning of the seventies one started to simulate molecular systems, and since then the rapid increase in computational power has enabled studies of larger systems and molecules, using more and more advanced mathematical models.

In the earliest simulations, the forces acting between the different particles in the liquids where adjusted to functional forms using empirical data. However, the increase in computational power also gave rise to the possibility of performing quantum mechanical computations on increasingly larger systems. This made it possible to use data from such calculations to model the interacting forces between atoms and molecules. From this the mathematical models may be branched into three categories; empirical-, semi-empirical- and ab inito models[1].

One of the liquids which attracted attention early on, and which still is one of the most interesting liquids to study, is water. The importance of water cannot be overestimated. It is one of the basics organisms need to survive, and a good illustration of its importance is the resent Mars expeditions by ESA (Beagle and Mars Express) and NASA (Spirit and Opportunity) searching for water on the planet [2]. Water is not a simple liquid to study due to the strong interactions between the molecules. In addition water has some unusual properties, such as the larger volume of the solid phase compared to liquid phase [4]. Several

[1] by ab inito models, it is not meant Car-Parrinello models, but models based on results from ab inito calculations

[2] for those able to read Norwegian see; http://www.forskning.no/Temaer/1074176593.12

different water models has been proposed, and these can be divided into three categories; rigid models, flexible models and models which includes many-body effects [1].

One such many-body effect is polarization. Polarization is electrostatic interactions which arises from changes in the charge distribution around a molecule [1]. Several different models which includes polarization have been reported in the literature, and it is found that such models provide a far better physical model when compared to non-polarizable models [5].

Even at 0 K there exists atomic motion, in terms of vibrations, as a consequence of the Heisenberg uncertainty relation. This gives rise to an energy contribution which can be quite substantial [1]. This is usually not included in in simulation models. Inclusion of this would thus refine the models.

The aim of this study will be to construct a flexible, polarizable model for water which includes fluctuations of the atomic charges, dipoles, and polarizabilities and zero point vibrational corrections. The model will be an ab initio model, and will be verified using simulations.

The report has the following structure: Chapter 2 contains a description of the theoretical background for the method, of the methods used to construct the model and of the various terms in the model and the theoretical background for the simulations. In chapter 3 the results from the model constructions are reported. Chapter 4 contains a description of the computational details of the simulation procedure. The results from the simulations study is presented in chapter 5. A discussion of these results are given in chapter 6. Conclusive remarks is given in chapter 7.

Chapter 2

Theoretical background

The fundamental equation describing the properties of a chemical system is the Schödinger equation, which for a given system can be written as [6];

$$i\hbar\frac{\partial\phi(x,y,z,t)}{\partial t} = \hat{H}\phi(x,y,z,t) \tag{2.1}$$

where $\hat{H}$ is the Hamiltonian operator and ϕ is the wave function (postulated to contain all information about the system). One can separate the time dependent and the time independent part of the Schödinger equation, and write the time independent part as;

$$\hat{H}\phi(x,y,z) = E\phi(x,y,z) \tag{2.2}$$

where E is the energy eigenvalue. This equation is called the time independent Schödinger equation. The Hamiltonian operator for a molecular system can generally be written as;

$$\hat{H} = -\sum_{j=1}^{r}\frac{\hbar^2}{2M}\nabla_j^2 - \frac{\hbar^2}{2m_0}\sum_{i=1}^{s}\nabla_i^2 + V \tag{2.3}$$

here V is the potential between the particles in the system, given by;

$$V = \sum_{i,i'}\frac{e^2}{4\pi\epsilon_0 r_{ii'}} + \sum_{j,j'}\frac{Z_j Z_{j'} e^2}{4\pi\epsilon_0 r_{jj'}} - \sum_{i,j}\frac{Z_j e^2}{4\pi\epsilon_0 r_{ij}}. \tag{2.4}$$

Where the first part and second part is the coulombic repulsion between the electrons and nuclei respectively, and the last part is the coulombic attraction between the nuclei and electrons. To simplify things further the Born-Oppenheimer approximation[1] is invoked. The Born-Oppenheimer approximation simply states that due to the significantly smaller mass of the electrons compared to the nuclei, the electrons will instantly adjust themself

[1] this approximation is also called the adiabatic approximation [7]

to the motions of the nuclei[2]. The time independent Schödinger equation can therefore be divided into two parts; the electronic Schödinger equation dealing with the electrons for given nuclei coordinates, and the nuclear Schödinger equation where the solutions of the electronic Schödinger equations enters as the potential in which the nuclei move.

The Born-Oppenheimer approximation thus introduces us to one of the most central topics of theoretical chemistry, namely the potential energy surfaces. A potential energy surface is the energy from the electronic Schödinger equation given as a function of the internal coordinates of the nuclei. An illustration of such a potential energy surface is given in figure 2.1.

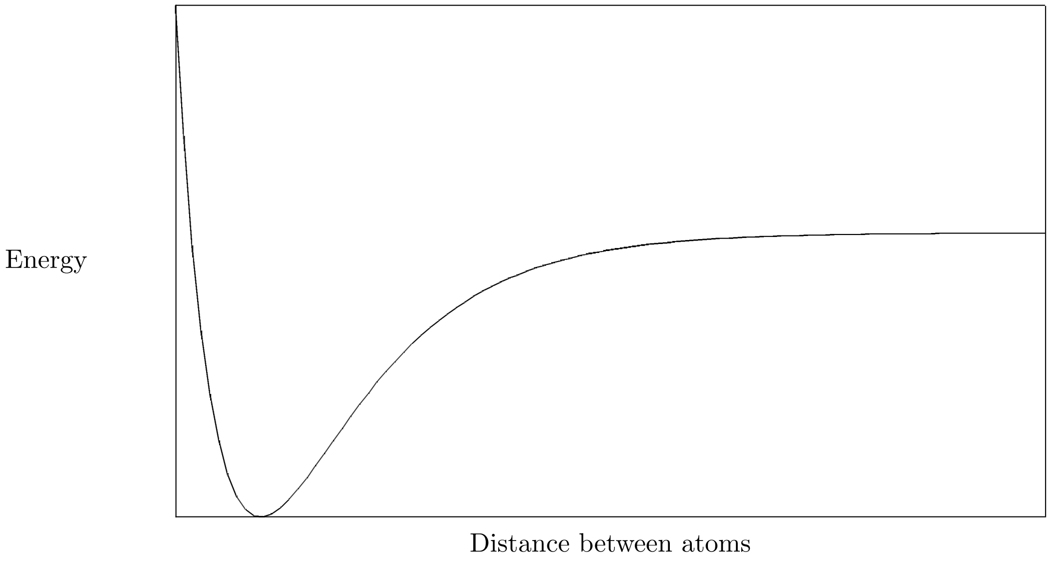


Figure 2.1: Illustration of a potential energy surface for a diatomic molecule

For small molecules (and collections of small molecules) it is possible to solve the equation above in an approximative manner. However when we turn to larger systems such as proteins or liquids such a scheme is no longer possible due to the incredibly large number of degrees of freedom in the system. The solution to this problem is to turn to classical mechanics, as classical mechanics is a special case of quantum mechanics. The concept of using classical mechanics on molecular system is often referred to as molecular mechanics and the potential energy surface is often called a molecular force field (or simply force field). This definition will be used further on in this report.

[2] this was shown by Max Born and J. Robert Oppenheimer in 1927 through the use of perturbation theory [8]

2.1 About force fields

Force fields are as described in the introduction of this report, often divided into three different categories depending on the data used in the parameterization. These three categories are theoretical (or ab inito) force field, semi-empirical force field and empirical force field. Another point which must be made about force fields in general is that they are often parameterized for a specific group of molecules (e.g. CHARMM [9] and AMBER[10] which both are modeled for proteins), and does therefore not necessarily perform well for other kinds of molecules.

In molecular force fields the energy is divided into various contribution. These contributions can be intramolecular interactions between bonded atoms and atoms defining an angle or dihedral angle. It can also be long-range non-bonded interactions such as the electrostatic (coulomb) interaction and Van der Waals interactions. There are several different ways of describing these interactions, all based on different physical approximation. The partition used in this study is described in section 2.2.4, together with other methods usually used. The various ways to partition the energy has given rise to a classification of force fields which depends on the incorporation of cross terms between the degrees of freedom in the system [1]. In this classification a force field with no cross terms and harmonic terms is a class I force field, a force field with explicit cross terms and anharmonic terms is a class II force field. Some further classification has also been done, defining a class III force field as a model which incorporates chemical and other effects. In this respect the force field constructed in this study is a class III force field as it incorporates polarizability and a coupling between the intra- and intermolecular degrees of freedom.

In the resent years there has been an increasing amount of attention directed construction of force fields which incorporates polarization. Incorporation of polarizability in force fields is not a new idea, but have been done for at least 20 years [5]. In general there are two ways to incorporate polarizability in a force field; the induced point dipole model and the fluctuating charge model. In the induced point dipole model, a point dipole is induced on an atom (or another contributing center on the molecule), as a response to the electric field on the atom [5]. In the fluctuating charge model, the atomic charges in the molecule changes as the environment of the atom changes. This is done according to the principles of electronegative equalization. For more information about polarizable force field, see the review article by Halgren and Damm [5].

As mentioned in chapter 1, water is one of the most studied and modeled liquids. The simplest force fields constructed for water are rigid models with three to five interaction cites. To these belong the TIP3P model of Jorgensen et al. [11] and the SPC model of Berendsen et al. [12] and modifications such as TIP4P [11]. In these models the interaction between the molecules are modeled using a positive charge on the hydrogen atoms, balanced by a negative charge on the oxygen atom or in its vicinity. Furthermore dispersion and repulsion energy between the molecules are modeled using a single cite on the molecule placed on the oxygen atom. Some refinement of these exists, such as the ST2 model of

Stillinger and Rahman [13], where the negative charges have been divided into two and placed on the "lone pairs" cites of the oxygen atom. The advantage with these simple models is the fact that they are computationally inexpensive.

The models described above has also been extended into flexible water models. This was for instance done by Wallqvist and Teleman [14], who used the SPC as basis for their flexible model. This model was then used to calculate the vibrational frequencies of water, however they did not manage to reproduce the experimental values (this might be due to the fact that they, at least for their harmonic model, did not include any cross terms, see section 2.2.4).

More advanced models have also been developed. These include refinements in both electrostatic and other long range, non-bonded forces. Polarizable water models have been formulated using both fluctuating charge models and atomic centered polarizabilities. In a study Wallqvist [15] included a molecular polarizability and atomic charges which depended on the molecular geometry. The conclusion of the study was that the coupling between the internal and external (intermolecular) degrees of freedom has a *"... large effect on liquid water properties such as structure and energetics ..."* (pp.448). Saint-Martin et al. [16] formulated a similar model, where they allowed for a fluctuation of the charge distribution through the use of a mobile negative charge coupled to the oxygen atom. Using this model they found a good description of the structure, energetics and dipoles in small water clusters, along with reproducing the second virial coefficient of steam. In a series of articles, the NEMO potential has been used to simulate water and systems containing water [17–21]. In the NEMO potential a multicenter multipole expansion of the electric field is used and atomic charges, dipole moments, quadrupole moments and polarizabilities are therefore included. These properties can be used to improve both the electrostatic part of the water potential and the other long range contributions (see section 2.2.4). The NEMO models show good agreement with both high level ab inito calculations and experimental data, and is one of the force fields which has inspired this study. In an article by Millot and Soetens [22] atomic charges, dipoles and polarizabilities as described by Stones Distributed Multipole Analysis (DMA) [23] and Baders Atoms in Molecules [24] (AiM, see section 2.2.2) were used. The conclusion of this study was that distribution of the polarizability from a single molecular cite to the atoms has an important influence on the properties of water, and that the liquid properties were sensible to the values of the various charges and dipoles. An extension of the TIP4P model to include fluctuating charges has been done by Rick et al. [25]. Inclusion of fluctuating charges, along with polarizable dipole has also been done by Stern et al. [26] who reported a good model performance in the whole temperature region of liquid water. This model has also been applied to other molecules [27]. A comparison study between two non-polarizable and three polarizable models, done by Jedlovszky and Richardi [28], reported that the polarizable models gave a better reproduction of the experimental structure at high temperature, but were unable to give correct temperature dependence.

2.2 Force field construction

The force field used in this study was constructed using quantum chemical data. A short presentation of the basis of the quantum chemical methods used, and a description of the various parts entering the force field is given below. A description of the parameterization scheme used is also presented.

2.2.1 Quantum Chemical methods

Hartree-Fock Self Consistent Field (HFSCF)

The most basic way to solve the electronic Schödinger equation for a molecular system is the HFSCF method. In Hartree-Fock theory the electronic Hamiltonian operator is modified so that each electron moves in a "mean field" set up by all the other electrons in the system. The wave function is then spanned out in a subspace of the Hilbert space through the use of so called basis functions, and the expansion coefficients which give the lowest energy of the system is found through an iterative process (uses the Variational theorem). For more information the reader is referred to Leach [1], Atkins and Friedman [29] or similar textbooks.

Complete Active Space Self Consistent Field (CASSCF)

One major drawback with Hartree-Fock level theory is the fact that it does not account for the so called correlation energy. The correlation energy is the energy which accounts for the fact that the electron does not move in an average field set up by all the other electrons, but "feels" the other electrons explicitly [1]. One definition of the correlation energy is the difference between the exact solution of the Schödinger equation and the sum of the Hartree-Fock and relativistic contribution [30];

$$E_{correlation} = E_{exact} - E_{Hartree-Fock} - E_{relativistic} \tag{2.5}$$

There are several ways to correct this error. Configurational Interaction (CI) is one such method which utilizes a theorem stating that, within the limits of the Born-Oppenheimer approximation and classical quantum mechanics, inclusion of all possible excitation in the wave function will, at the limit of an infinite basis set, yield the exact energy [31]. In CI this is done by using a wave function which may be written as;

$$\psi = c_0\psi_0 + c_1\psi_1 + c_2\psi_2 + \dots \tag{2.6}$$

where ψ_0 is the wave function from HFSCF and $\psi_{i>0}$ is the wave functions representing configurations which includes excitations (all ψs are expressed as Slater determinants). The coefficients c_0,c_1 etc are then found using the variational principle. In regular CI only these coefficients are varied, not the coefficients before the basis functions determinants. A refinement of CI which includes variation of both the CI coefficients and the coefficients before the basis functions is the Multi Configurational Self Consistent Field (MCSCF) techniques [1]. One popular version of MCSCF is the Complete Active Space Self Consistent Field method of Roos et al. [32]. In CASSCF the spin orbitals of the molecule is partitioned into three different contributions;

- Inactive orbitals
- Virtual orbitals
- Active orbitals

The inactive orbitals are orbitals which are always doubly occupied. The virtual orbitals are orbitals for which the energies are so large that their contributions are neglectable. Active orbitals are orbitals which are between the two others in energy. It is these orbitals we would like to vary the coefficients of the basis for. The configurations are then given by all possible arrangements of the electrons in the active space.

Many-body Perturbation theory (MP)

Many-body Perturbation theory provides another way to incorporate the correlation energy into the wave function. This approach was first formulated by Møller and Plesset [33] uses Rayleigh-Schödinger perturbation theory (RSPT). In RSPT the Hamiltonian, wave function and eigenvalues of the Hamiltonian are expanded in series, such that;

$$H = \sum_{n=0}^{\infty} \lambda^n H^{(n)}, \tag{2.7}$$

$$\Psi_i = \sum_{n=0}^{\infty} \lambda^n \Psi_i^{(n)}, \tag{2.8}$$

$$E_i = \sum_{n=0}^{\infty} \lambda^n E_i^{(n)}, \tag{2.9}$$

where n denotes the correction order, i the vibrational state and λ is an expansion parameter. The wave function is spanned out by the eigenfunctions of the zero order contribution;

$$\Psi_i^{(n)} = \sum_{r=0, r\neq i}^{\infty} a_{r,i}^{(n)} \Psi_r^{(0)}, \tag{2.10}$$

for which the exact solution is known (thus giving a complete set of basis functions). This gives the following recursive expressions for the energy corrections and the expansion coefficients [34];

$$E_i^{(n)} = H_{ii}^{(n)} + \sum_{m=1}^{n-1} \sum_{r=0,r\neq i}^{\infty} H_{ir}^{(m)} a_{r,i}^{(n-m)}, \tag{2.11}$$

$$a_{p,i}^{(n)} = \frac{-H_{pi}^{(n)} + \sum_{m=1}^{n-1}(E_i^{(m)} a_{p,i}^{(n-m)} - \sum_{r=0,r\neq i}^{\infty} H_{pr}^{(m)} a_{r,i}^{(n-m)})}{E_p^{(0)} - E_i^{(0)}}, p \neq i, \tag{2.12}$$

where the the short-hand notation $< p|H^{(n)}|i >= H_{pi}^{(n)}$ has been introduced. In MP the zero order contribution used is the HFSCF solution, and the following expression for the higher order terms of the Hamiltonian operator [1];

$$\mathcal{V} = \sum_{n=1}^{N} \lambda^n H^{(n)} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{1}{r_{ij}} - \sum_{j=1}^{N}(\mathcal{J}_j + \mathcal{K}_j) \tag{2.13}$$

where $\mathcal{J}_j$ is the Coulomb operator and $\mathcal{K}_j$ is the exchange operator. The length of the expansion does then give name to the method, such that MP2 is an expansion to the second order and so on.

2.2.2 Atomic properties

A common way to represent the physical description of a molecule is through localized atomic properties. Such atomic properties are charges, dipoles etc. One problem with using atomic properties is that they are not observables, and thus they cannot be measured experimentally. Another problem arising from this is that there is no correct[3] mathematical way of expressing an operator in quantum mechanics. The reason for using such properties is to reproduce the electronic field of the molecule, and give a more correct description of the intramolecular interactions.

There are several different ways to derive these properties, all depending on the way in which the rest of the force field is constructed. For empirical force field the atomic charges are often adjusted in such a way that they reproduce the molecular dipole moments, quadrupole moments etc. One significant drawback with this method is that one often encounters unphysical solutions such as a negative and positive charge close to each other, mimicking a point dipole [35]. The classic quantum mechanical method to produce atomic charges is the Mulliken scheme [1, 36]. One significant drawback of the Mulliken scheme is the fact that it has very poor basis set convergence. Another method to calculate atomic multipoles is Baders Atoms in Molecules (AiM) [22, 24, 36]. This method uses the topology of the one-electron density to give a unique definition of the molecular atom which enables calculation of various atomic property. Two drawbacks of this method are its computational expense and that it may give charges which are not localized to any atoms [35].

[3] correct in the sense that its the only right way

The Generalized Atomic Polar Tenor (GAPT) [36, 37] and Dinur-Hagler method (DH) [38] provides yet another way to define localized charges and higher order moments. Here the cartesian derivatives of the molecular dipole moments is used to define the atomic charges. The Atomic Polar Tensor (APT) [39] uses the mean of the sum of the cartesian derivatives, while DH uses the off plane component. In a recent article Solheim et al. [40] extended the GAPT and DH scheme to include atomic dipoles. In their analysis they found that the GAPT did not reproduce the molecular quadrupole moments, while DH methods (though restricted to plane molecules) did this. The method used in this study is the LoProp method. A description of this method is given below.

Localized Properties method (LoProp) The LoProp method [41] is newly developed method which partitions the molecular properties (such as multipoles) into local (atomic and inter-atomic) contribution through the use of a series of orthogonalizations of the basis set. These orthogonalization steps are;

1. Gram-Schmidt orthonormalization of the atomic basis functions
2. Two Löwdin orthonormalizations, one in the occupied space and one in the virtual space
3. Gram-Schmidt orthonormalization projecting out any components of the occupied subspace out of the virtual space
4. Löwdin orthonormalization in the virtual space

This gives an orthonormal atomic basis set from which the atomic properties can be calculated. In general the expectation value of a property is;

$$< O >= Tr(\mathbf{DO}) = \sum_{\mu\nu} D_{\mu\nu} < \mu|\hat{O}|\nu >, \tag{2.14}$$

where $\mathbf{D}$ is the 1-electron density matrix (and $D_{\mu\nu}$ is an element of this matrix) and $\mathbf{O}$ is the integrands of $\hat{O}$ which is the operator of the property. The localized property can then be calculated by transforming the integrals of the property and the 1-electron density matrix into the new basis and taking the trace over the subspace partitioned to a single center. This is used to calculate the atomic charges and dipole moments. The atomic polarizabilities are then calculated by using a second order central differentiation, and introducing a penalty function to avoid charge transfer over larger distances (as the method only assumes charge transfer through the bonds).

The LoProp method is origo independent, shows a good basis set convergence, is transferable and computationally inexpensive [41].

2.2.3 Regressional methods

There exists several methods to find the parameters used in a force field. Among the methods used are both manual methods and automated procedures [42, 43]. To attain the parameters needed in this force field two methods were used; linear regression and parameter optimization. The basis of the two methods are described below.

Partial Least Squares regression (PLS)

In Ordinary Least Squares (OLS) regression the dependent variables $\mathbf{Y}$ are connected to the independent variables $\mathbf{X}$ though a coefficient matrix $\mathbf{B}$ such that;

$$\hat{\mathbf{Y}} = \mathbf{XB}. \tag{2.15}$$

Furthermore $\mathbf{B}$ is given as;

$$\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2.16}$$

by minimizing the sum of difference between the measured values $\mathbf{Y}$ and the modeled values $\hat{\mathbf{Y}}$ squared. The OLS method is commonly used and well documented in literature, however it has some serious drawbacks. It does not handle collinearity in the independent variables, does not handle interference between the independent variables and it is unable to handle cases where there are more variables than samples. Even though one in general should try to avoid these problems, there is still a need for more robust methods to handle such cases. One such method is the Partial Least Squares (PLS) regression, in which one finds the latent variables in $\mathbf{X}$ which have relevance for prediction of the dependent variable. These latent variables are chosen to maximize the covariance between the two matrices $\mathbf{X}$ and $\mathbf{Y}$. For more information about regressional methods the reader is referred to Esbensen [44] and Walpole et al. [45].

Parameter optimization

It is not possible to linearize all equations so a linear regression method might be used. In such cases minimization techniques can be utilized to give the optimal choice of parameters.

Two such methods, both inspired by nature, are Simulated Annealing (SA) and Genetic Algorithms (GA). The SA algorithm is inspired by the process of crystallization [46]. Given ample time and a slow cooling, nature always builds completely ordered and pure crystal, representing the minimum energy conformation of the atoms. In SA one attempts to reproduce this process, by allowing some random, “uphill” walks on the hyper surface of the function one attempts to minimize. The probability of such a “uphill” walk is governed by a “temperature”, which is gradually reduced making the algorithm more and more like a regular minimization algorithm. By carefully regulating the “temperature”, the algorithm hopefully ends up in an equilibrium. SA has several analogies to statistical mechanics which might be utilized to increase the efficiency of the algorithm [47]. The inspiration of

GA is Darwin's theory of evolution, and is a part of a larger group of algorithms called Evolutionary Algorithms [1]. In a GA one starts with a set of possible solution to the optimization problem (a "population"), called "parents". Each member of the population has a certain "fitness" determined by a fitness function. The properties of the various members are stored in bits (0s and 1s). These are used to generate a new population called "children" through crossover (interchanging some bits between a pair of parents) or mutation (small random bit change on individuals). The pair of parents are chosen random, however the choice is biased towards the most "fit" in the population. In this way the most optimal choice is hopefully encountered [4]. Both SA and GA has previously been used to derive force field parameters [48–50]. There has also been used other optimization schemes to parameterize force fields, such as the one by Norrby and Liljefors [51] and the neural network of Cho et al. [52], however it is not the object of this study to compare such methods nor does the complexities of the problems demand more methods, and these have therefore not been studied.

In order to use the techniques described above to find the best fit to a set of data, one must define a function to minimize. One such function is the root-mean-square error (RMSE) defined as;

$$\epsilon = \sqrt{\frac{1}{N} \sum_{i}^{\textit{All data points}} (V_{exact}(\vec{q}_i) - V_{appr.}(\vec{q}_i; \vec{b})^2)} \tag{2.17}$$

and then search for the smallest value of ϵ. A problem one might encounter using the expression above is large errors between the exact values and the approximated solutions in regions where the exact solution has very large values (compared to the rest of the data). When fitting energy data to a function, one way to prevent these regions from having too large influence is by weighting their contribution to the error using a Boltzmann distribution. In figure 2.2 the effect of the temperature on the Boltzmann weight is illustrated. Here we see that using a temperature of 1000 K the influence on points with a energy higher than 10 kcal/mol is almost zero, weighting with 6000 K gives a reduction of the error of about 40% for a point with 10 kcal/mol. It is obvious that such weighting must be used with care and adjusted to the given problem, however it does provide a good way of enhancing the influence on the most important regions.

2.2.4 The energy contributions

Intramolecular energy

The intramolecular energy is the part of the energy arising from interactions between atoms in the same molecule. These interactions are partitioned into contributions from bonds, angles and for larger molecules dihedral angle and improper rotation. As the force field

[4] For more information about genetic algorithms see; http://cs.felk.cvut.cz/ xobitko/ga/

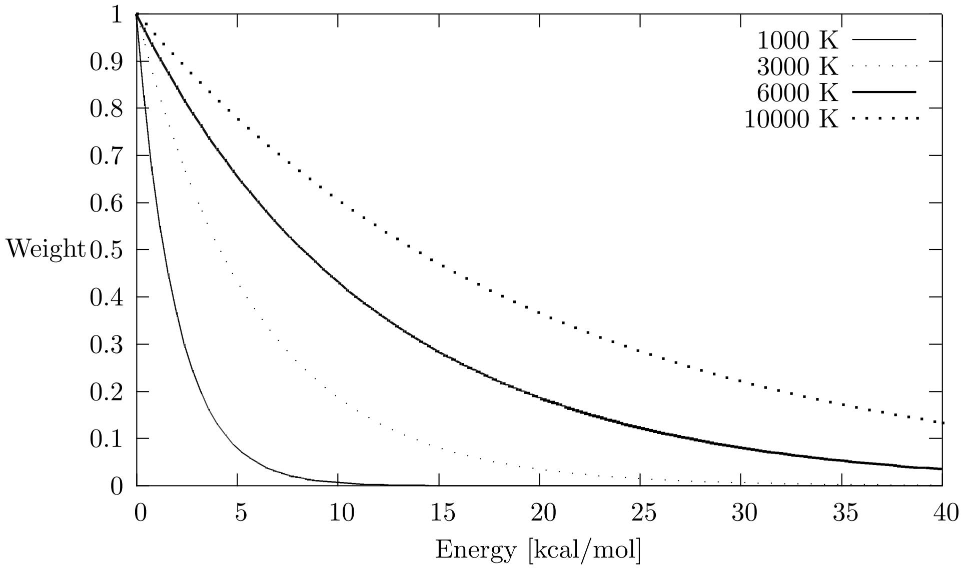


Figure 2.2: Effect of Boltzmann weighting as function of temperature

constructed here is for water, there are no dihedrals or improper rotation in the molecule, and hence the intramolecular energy is partitioned into bond stretches and angle bending.

There exists several ways to fit the intramolecular energy to an expression. One such way proposed by Lifson and Warshel [53] is to fit the potential to a Taylor-expansion (the so-called consistent force field). The advantage with this is that a Taylor-expansion is the most correct way to represent a potential around an equilibrium structure. Another way to fit the potential to a theoretically based expression of the potential energy surface. This can be done using multiple linear regression, and this is the approach chosen in this study.

Bond stretching The bond stretching term describes the forces between two neighboring atoms in a molecule. There are several ways to describe these forces. One of the most used ways to do this is using a harmonic potential, in which the atoms are connected with a ideal spring (such as described by Hooks law). Although such a model is simple and correct for energies near the equilibrium energy, it does not account for larger deviations from the equilibrium positions. Two potentials that does this is the well known Morse potential [54] and the Simon-Parr-Finlan (SPF) potential [55]. The model used in this force field is the latter one.

The SPF potential is a potential based on a perturbation approach, and was originally introduced by Fougere and Nesbet [56]. The SPF potential is given by;

$$V(r) = B_0 \left[\frac{r - r_e}{r}\right]^2 \left(1 + \sum_{n=1}^{\infty} b_n \left[\frac{r - r_e}{r}\right]^n\right), \tag{2.18}$$

and is thus a power series in $\frac{r-r_e}{r}$, and a special case of the Thakkard potential [57]. One of the attractive features with this potential is the fact that it dissolves correctly as $r \to \infty$, another that the potential is well behaved.

Angle bending Angle bending is the contribution to the intramolecular energy which occurs when a angle is perturbed from it equilibrium position. There are several ways to describe the angle bending, such as the use of harmonical terms, polynomial expansion and Urey-Bradley terms[5] [1]. The model for the angle bending chosen in this study is a combination of a modificated Urey-Bradley expression and cosine term;

$$V_{angle} = k_{UB} \left(\frac{r_{1,3} - r_{UB}}{r_{1,3}} \right)^2 + k_\theta \left(\cos\theta - \cos\theta_e \right)^2 \tag{2.19}$$

Such a choice for the angle bending will also act as a cross term[6]. The expression used in eq. 2.19 has two advantages; it goes to infinity as the distance between the atoms goes towards zero, and it provides a barrier at 180^o which makes it possible, but not probable that the molecule will “flip” (pass through a linear state). It will therefore (within a certain range of $r_{1,3}$) give a physically correct description of the energy as the angle changes.

Cross terms To be able to predict molecular vibrations, on must include so called cross terms in the force field [1, 30]. As the choice for describing the angle bending term in eq. 2.19 already includes a bond-angle cross term, the cross term which left to be included in the model is the bond-bond cross term. This term can be formulated as;

$$V_{cross} = C_1 \left[\frac{r_1 - r_e}{r_1} \right]^2 \left[\frac{r_2 - r_e}{r_2} \right]^2 \tag{2.20}$$

The choice of using the same expression for the bonds in eq. 2.20, as in the SPF potential arises from the fact that it would be preferable to keep the correct dissociation of the bonds.

[5] writes the angle bending in terms of distances between the 1,3 atoms, rather than as angles

[6] as we for a general triangle have the following expression; $a^2 = b^2 + c^2 - 2bc\cos\theta$. For water b and c here is the bond lengths and θ the angle

Total The total energy contribution from the intramolecular energy of one water molecule then becomes;

$$
\begin{aligned}
V_{intra}(r_1, r_2, r_{1,3}) &= V_{SPF}(r_1) + V_{SPF}(r_2) + V_{angle} + V_{cross} \qquad (2.21)\\
&= B_0 \left[\frac{r_1 - r_e}{r_1}\right]^2 \left(1 + \sum_{n=1}^{\infty} b_n \left[\frac{r_1 - r_e}{r_1}\right]^n\right)\\
&\quad + B_0 \left[\frac{r_2 - r_e}{r_2}\right]^2 \left(1 + \sum_{n=1}^{\infty} b_n \left[\frac{r_2 - r_e}{r_2}\right]^n\right)\\
&\quad + k_\theta \left(\cos\theta - \cos\theta_e\right)^2\\
&\quad + k_{UB} \left(\frac{r_{1,3} - r_{UB}}{r_{1,3}}\right)^2\\
&\quad + C_1 \left[\frac{r_1 - r_e}{r_1}\right]^2 \left[\frac{r_2 - r_e}{r_2}\right]^2\\
&\quad + constant
\end{aligned}
$$

Electrostatic energy

Different atoms in a molecule have different electronegative potential. As the atoms with a high electronegativity have a larger affinity for electrons than those with smaller electronegativity, the molecules will be surrounded by a permanent electric charge distribution. The interaction between two such charge distribution (A and B) may be written as a multipole expansion, which gives the following expression [58];

$$
V_{elec}^{AB} = \frac{q_a q_b}{|\vec{R}|} + \frac{q_A \vec{\mu}_B \cdot \vec{R}}{|\vec{R}|^3} + \frac{(\vec{\mu}_A \cdot \vec{R}) q_B}{|\vec{R}|^3} + \frac{(\vec{\mu}_A \cdot \vec{R})(\vec{\mu}_B \cdot \vec{R})}{|\vec{R}|^5} + \ldots \qquad (2.22)
$$

here truncated at the dipole. The first term in the expression above is the charge-charge interaction (often called Coulombs law), the second and third term is the dipole-charge interactions and the last term is the dipole-dipole interaction. This is a general expression of the interaction energy of two charge distributions. In the force field used in this study a multicenter multipole expansion has been done in which the charge distributions of the molecule has been divided into atomic charges and dipole moments. The expression in eq. 2.22 is then the electrostatic interaction between two atoms on different molecules.

Induction energy

Induction energy is the energy arising from changes in the charge distribution caused by an external field. Such change in charge distribution are called polarization, and the physical property governing this process is called the polarizability.

The energy of a system contain N atoms[7] the induction energy can be written as [59];

$$V_{induced} = -\frac{1}{2}\sum_{I,J}^{N} \vec{\mu}_I^{ind}\mathbf{T}_{IJ}\vec{\mu}_J^{ind} + \sum_{I}^{N} V_{I,self} - \sum_{I}^{N} \vec{\mu}_I{}^{ind} \cdot \vec{E}_I, \tag{2.23}$$

where $\vec{\mu}_I^{ind}$ is an induced dipole, E_I the external electric field and $\mathbf{T}_{IJ}$ the second order interaction tensor given by;

$$\begin{aligned} \mathbf{T}_{IJ} &= \nabla^2 \left(\frac{1}{\vec{R}_{IJ}}\right) && (2.24) \\ &= \frac{3}{|\vec{R}_{IJ}|^5}\vec{R}_{IJ}^T\vec{R}_{IJ} - \mathbf{I}\frac{1}{|\vec{R}_{IJ}|^3}, && (2.25) \\ \mathbf{T}_{II} &= 0 \end{aligned}$$

here $\vec{R}_{IJ}$ is the vector between atom I and atom J. The first term in eq. 2.23 is the dipole-dipole interaction between two induced atomic dipoles, the second term is the work required to create the induced dipole moments and the last term is the interaction between the induced dipole and the external electric field. Each atom in the system will at any given moment have an induced dipole moment such that the induction energy is minimized, giving the following expression for the induced dipole;

$$\vec{\mu}_I^{ind} = \alpha_I(\vec{E}_I + \sum_{J}^{N} \mathbf{T}_{IJ}\vec{\mu}_J^{ind}), \tag{2.26}$$

where E_I is the electric field arising from the permanent charges and dipoles [58];

$$\vec{E}_I = \sum_{J=1,J\neq I}^{N} \left[\frac{q_J^{perm}\vec{R}_{IJ}}{|\vec{R}_{IJ}|^3} + \frac{\vec{R}_{IJ}(\vec{R}_{IJ}\cdot\vec{\mu}_J^{perm})}{|\vec{R}_{IJ}|^5}\right]. \tag{2.27}$$

Using these equations the expression for the induction energy can be rewritten as;

$$V_{induced} = -\frac{1}{2}\sum_{I}^{N} \vec{\mu}_I^{ind}\vec{E}_I \tag{2.28}$$

Repulsion- and dispersion energy

Besides electrostatic and induced electrostatic interactions, there are also Van der Waals interactions between non-bonded atoms. The Van der Waals interactions can be divided into two individual contributions; the dispersive and the repulsive interaction. The traditional

[7] using atomic polarizabilities

way to model the repulsion- and dispersion energy is through the famous Lennard-Jones potential [1]:

$$V_{disp,rep} = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{2.29}$$

However, there are other ways to model these interactions. The dispersive- and repulsive interactions used in the force field in this study is a simplification of the ones used in the NEMO potential [58].

In ab inito force fields the dispersion and repulsion energy can be found calculating the interaction energy between various dimer configurations. Due to the incompleteness of the basis functions the so called Basis Set Superposition Error (BSSE) must be adjusted for. The problem of BSSE arises as the basis functions of one molecule starts to improve the description of the electrons on the other molecule, and hence lowers the energy of the system [1]. A correction to this is the so called counterpoise (cp) method of Boys and Bernardi [60]. In this method the interaction energy between two molecules are calculated as the difference in energy between the complex and the monomers given the orbitals of the other monomer.

Repulsion energy The repulsion energy is understood as a consequence of the Pauli principle [1], and can be obtain from ab initio calculations using the following energy balance;

$$V_{rep}^{AB} = \Delta V_{HFSCF} - V_{elec} - V_{induced} \tag{2.30}$$

where ΔV_{HFSCF} is the difference between the energy of the dimer and the two BSSE corrected monomers at Hartree-Fock level.

In the NEMO potential this energy difference is fitted to the following expression [17];

$$V_{rep}^{AB} = \sum_{i}^{N_A} \sum_{j}^{N_B} \gamma_{ij} e^{-R_{ij}/\sigma_{ij}} \tag{2.31}$$

here γ_{ij} is given by $\gamma_{ij} = q_i q_j \kappa_i \kappa_j$ where q_i and q_j are the atomic charges of atom i and j, and κ_i and κ_j atomic parameters fitted to calculations. The parameter σ_{ij} is given by;

$$\sigma_{ij} = \frac{1}{|M_{ij}|^{1/2}(\alpha_i + \alpha_j)} \tag{2.32}$$

where $|M_{ij}|^{1/2}$ gives a measure of the extension of the electron density (see Hermida-Ramón et al. [17]) and α_i and α_j are atom type parameters. In this study the eq. 2.31 was simplified by using atom pair parameters, thus giving the following equation;

$$V_{rep}^{AB} = \sum_{i}^{N_A} \sum_{j}^{N_B} \kappa_{ij} e^{\alpha_{i,j} R_{ij}} \tag{2.33}$$

where $\alpha_{i,j}$ must be negative.

Dispersion The dispersion energy is an energy contribution from interactions between immediately induced dipole moments, arising from fluctuations in the electron distributions. Using an energy balance the dispersion energy can be approximated to be;

$$V_{disp} = \Delta V_{MP2} - V_{elec} - V_{ind} - V_{rep} \tag{2.34}$$

where ΔV_{MP2} is the difference between the energy of the dimer and the two BSSE corrected monomers at the MP2 level.

The starting point for the dispersion energy in the NEMO potential is the London expression for the dispersion energy [58];

$$V_{disp}^{Lon} = -\sum_{m,n}^{atoms} f_{mn} \frac{CE_{12}}{4} \sum_{i,j,k,l}^{3} \alpha_{ij}^{m} \alpha_{kl}^{n} T_{ik} T_{jl}, \tag{2.35}$$

where $\mathbf{T}$ is the second order interaction tensor and E_{12} is the average molecular excitation energy;

$$E_{12} = \frac{E_1 E_2}{E_1 + E_2}, \tag{2.36}$$

and C a constant. The term f_{mn} is a dampening term which is given by;

$$f_{mn} = 1 - e^{-b_{mn} r_{mn}} \sum_{k=0}^{6} \frac{(b_{mn} r_{mn})^k}{k!} \tag{2.37}$$

where r_{mn} is the distance between the atoms and b_{mn} is given by;

$$b_{mn} = \frac{1}{c(r_m + r_n)}. \tag{2.38}$$

where c is a constant. The r_m terms are given by the trace of the second moment (Q_m) and the valence charge q_m;

$$r_m = \sqrt{\frac{Tr(Q_m)}{q_m}}. \tag{2.39}$$

In the model used in this study a slight modification of the London expression for the dispersion energy is used. In this modification the $\frac{CE_{12}}{4}$ term is changed to an atom pair constant B_{mn} and only the first term in the sum in eq. 2.37 is used. The value of b_{mn} is no longer found by eq. 2.38, but is determined along with B_{mn} from a set of dimer calculation. The expression of the dispersion energy then becomes;

$$V_{disp} = \sum_{m,n}^{atoms} B_{mn}(1 - e^{-b_{mn} r_{mn}}) \sum_{i,j,k,l}^{3} \alpha_{ij}^{m} \alpha_{kl}^{n} T_{ik} T_{jl}. \tag{2.40}$$

Zero-point vibrational corrections (zpv correction)

The zero-point vibrational correction is found through calculations of the vibrational frequencies of the molecule. This can be done using the FG-matrix approach of Wilson et al. [61]. A short description of this method is given below, following the description of Cotton [62].

In the FG-matrix approach the vibrational frequencies are found solving the following equation;

$$|\mathbf{FG} - \mathbf{E}\lambda| = 0 \tag{2.41}$$

where **F**, **G** and **E** are matrices and λ eigenvalues of the product from the multiplication of the **F** and **G** matrices. The F-matrix are constructed using the Hessian of the potential energy, and accounting for the symmetry of the molecule such that;

$$\mathbf{F} = \mathbf{U}\mathbf{V}''_{pot}\mathbf{U}^T \tag{2.42}$$

where $\mathbf{V}''_{pot}$ is the Hessian and **U** for the special case of a water molecule is;

$$\mathbf{U} = \begin{bmatrix} 0 & 0 & 1 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix}$$

The G-matrix introduces the masses of the atoms into the equation, and is found using the following equation;

$$\mathbf{G} = \mathbf{U}\mathbf{g}\mathbf{U}^T \tag{2.43}$$

where the various components of **g** can be found tabulated in books such as Cotton [62] and Wilson et al. [61]. Using atomic units the ZPV energy becomes;

$$V_{zpv} = \sum_{i=1}^{N_{mn}} \frac{\sqrt{\lambda_i}}{2} = \sum_{i=1}^{N_{mn}} \frac{\nu_i}{2} \tag{2.44}$$

where ν_i is the vibrational frequencies.

2.3 Statistical Mechanics and Simulation Methods

Statistical mechanics[8] is the theory linking the microscopic theory of quantum mechanics to the macroscopic theory of thermodynamics. The postulatory basis for statistical mechanics is the postulate of equal a prior probability;

> ”All available states for an isolated system occur equally likely”,
> Kjellander [63], pp. 7

[8] in this report the name statistical mechanics equals equilibrium statistical mechanics

and the ergodic hypothesis, which states that;

> ”The (long) time average of any mechanical variable is equal to the ensemble average of the same variable in the limit of infinite number of member systems in the ensemble”,
> Kjellander [63], pp. 20

The last postulate demands a definition of an ensemble;

> ”a set of sampled configurations of a system, in contact with a thermal bath, where each individual sample may have different state while sharing three common macroscopic properties”,
> Seddon and Gale [64], pp. 69

These postulates and definitions gives us the basis for of how to interpret the experimental results in the light of statistical mechanics, and how to set up computer simulations to solve the equations of statistical mechanics. There exists several methods to use simulations to solve the equations of statistical mechanics. Among the most popular methods are Molecular Dynamics (MD) and Monte Carlo (MC) simulations. These two methods are fundamentally very different, as MD calculates a time average solving Newton’s equations of motion for a molecular system, while MC calculates an ensemble average through the generation of random configurations. The two methods are therefore linked together through the ergodic hypothesis. This point is of importance as simulations must be ergodic (able to, in theory, to investigate the whole of the phase space). The simulation method used in this study is MC.

This section will try to outline the properties of statical mechanics which will be studied in this report, and give a description of simulation method used. No detailed derivation of the basis for statistical mechanics will be presented in this chapter as this can readily be found in textbook in physical chemistry such as Atkins [65] and Chang [66], or introduction book on the subject such as Seddon and Gale [64] and Hill [67].

2.3.1 The radial distribution function

In studies of liquids, the structure of the liquid is of special interest, as it is related to the properties of the liquid, such as for instance the excess part of thermodynamical properties. A description of the structure of the liquid is given by the radial distribution function. In the following section the concept of the radial distribution function is derived. The derivation follows the one given in Kjellander [63].

In classical statistical mechanics the energy of a system can be expressed in a Hamiltonian which depends on all dimensions in phase space (Γ);

$$H = H(\vec{p}_1, \ldots, \vec{p}_n, \vec{q}_1, \ldots, \vec{q}_n) \tag{2.45}$$

and hence the probability for the system to be in a certain volume in phase space ($d\Gamma$) is;

$$\mathcal{P}(\Gamma)d\Gamma = \frac{e^{-H(\Gamma)/kT}d\Gamma}{\int_\Gamma e^{-H(\Gamma)/kT}d\Gamma'},\ \Gamma = \vec{p}_1, \ldots, \vec{p}_n, \vec{q}_1, \ldots, \vec{q}_n \tag{2.46}$$

where k is Boltzmanns constant and T is the temperature of the system. Integrating over all the momentums we get;

$$\mathcal{P}(\vec{q}_1, \ldots, \vec{q}_n) = \frac{e^{-V_{pot}(\vec{q}_1,\ldots,\vec{q}_n)/kT}}{\int_{\vec{q}_1\ldots\vec{q}_n} e^{-V_{pot}(\vec{q}_1',\ldots,\vec{q}_n')/kT}d\vec{q}_1' \ldots d\vec{q}_n'} \tag{2.47}$$

This is the probability of a certain point in phase space, independent of the velocities of the particles. Introducing the configurational partition function;

$$Z = \int_{\vec{q}_1\ldots\vec{q}_n} e^{-V_{pot}(\vec{q}_1',\ldots,\vec{q}_n')/kT}d\vec{q}_1' \ldots d\vec{q}_n' \tag{2.48}$$

eq. 2.47 may then be written as;

$$\mathcal{P}(\vec{q}_1, \ldots, \vec{q}_n) = \frac{e^{-V_{pot}(\vec{q}_1,\ldots,\vec{q}_n)/kT}}{Z}. \tag{2.49}$$

The equation above implies that the particles in the system are distinguishable, which is not the case. To account for the fact that the particles in the system are indistinguishable, the N-particle distribution function is introduced;

$$n^{(N)}(\vec{q}_1, \ldots, \vec{q}_n) = N!\mathcal{P}(\vec{q}_1, \ldots, \vec{q}_n) \tag{2.50}$$

and more specifically the 2-particle distribution function;

$$n^{(2)}(\vec{q}_1, \vec{q}_2) = N(N-1)\frac{\int e^{-V_{pot}(\vec{q}_3,\ldots,\vec{q}_N)/kT}d\vec{q}_3 \ldots d\vec{q}_N}{Z}. \tag{2.51}$$

The 2-particle distribution function gives the probability of finding particle 1 in volume $d\vec{q}_1$ and particle 2 in volume $d\vec{q}_2$. This may be written in terms of the conditional probability of finding particle 1 in $d\vec{q}_1$ given particle 2 in $d\vec{q}_2$;

$$n^{(2)}(\vec{q}_1, \vec{q}_2) = n(\vec{q}_1|\vec{q}_2)n(\vec{q}_2) \tag{2.52}$$

where $n(\vec{q}_2)$ is the probability of finding particle 2 in $d\vec{q}_2$ and $n(\vec{q}_1|\vec{q}_2)$ the conditional probability described above. Using this it is possible to introduce a pair distribution $g(\vec{q}_1, \vec{q}_2)$ such that;

$$\begin{aligned} n^{(2)}(\vec{q}_1, \vec{q}_2) &= n(\vec{q}_1|\vec{q}_2)n(\vec{q}_2) \\ &= n(\vec{q}_1)n(\vec{q}_2)g(\vec{q}_1, \vec{q}_2). \end{aligned} \tag{2.53}$$

Due to the symmetry[9] in bulk fluids one may introduce the radial distribution function $g(r)$, which only depends on the distance r between the particles. The radial distribution function is often defined as;

$$g(r) = \frac{V}{N^2} < \sum_i \sum_{i \neq j} \delta(\vec{r} - \vec{r}_{ij}) > \tag{2.54}$$

where the sum is over all atom pairs, N the total numbers of atoms and V the volume. In a simulation this is realized by sampling the numbers of atoms in a spherical shell between r and $r + \delta r$ into bins, and then normalize each bin giving the following;

$$g(r) = \frac{N^{bin}}{4/3\pi\left((r+\delta r)^3 - r^3\right)\rho N^{atoms}\tau_{MC\text{-}steps}}, \tag{2.55}$$

where

$$4/3\pi\left((r+\delta r)^3 - r^3\right) \tag{2.56}$$

is the volume of the shell, ρ is the particle density, N^{atoms} the numbers of atoms and $\tau_{MC\text{-}steps}$ the numbers of sampled MC steps.

2.3.2 Monte Carlo (MC)

Monte Carlo (MC) technique provides a method to solve the equations of statistical mechanics, through the ergodic hypothesis. The description below follows the one given in a previous report [68].

As described earlier a MC simulations works by generating random configurations and taking the ensemble average over these. The use of regular MC simulations of statistical mechanics ensembles is, unfortunately, impossible. The reason for this is that the regions in phase space such a simulation would sample are primary regions of low probability, and hence of little importance for the properties of a given system as the average value of a property is defined as;

$$< A >= \int_\Gamma A(\Gamma')\mathcal{P}(\Gamma')d\Gamma' \tag{2.57}$$

Due to this one must use a modification of the MC method which avoids these areas of phase space, and instead focus on the “important” regions. Doing so in a way that gives an unbiased estimate of the given property is called an “importance sampling MC” [69]. The type of importance sampling usually used in MC simulations of molecular systems are the Metropolis algorithm[10] first described by Metropolis et al. [3].

[9] no direction is preferred more than the others

[10] the Metropolis algorithm is also called the M(TR)^2 method [70]

Metropolis MC

The Metropolis algorithm provides a Markov chain. A Markov chain is a set of events (or in the case these of simulation, the configurations of the system), which fulfills two conditions [1];

1. Each generation of a new configuration depends only on the preceding configuration, and not the history of the simulation
2. Each configuration belongs to a finite step of possible outcomes

From this one can formulate the probability of a new configuration n as;

$$\rho_n = \pi_{m,n}\rho_m, \tag{2.58}$$

where ρ_m is the probability of the old configuration and $\pi_{m,n}$ is the probability of going from configuration m to n. The Metropolis algorithm utilizes this and furthermore imposes the condition of microreversiblility, which mathematically gives us;

$$\pi_{n,m}\rho_n = \pi_{m,n}\rho_m, \ \forall n, m \tag{2.59}$$

which in the end gives us the criteria used in the Metropolis algorithm;

$$\frac{\rho_n}{\rho_m} = \frac{\pi_{m,n}}{\pi_{n,m}} = e^{-(V_n - V_m)/kT}. \tag{2.60}$$

A pseudocode of the Metropolis algorithm is given in figure 2.3. The first step in the algorithm is a perturbation of one or more variables in the system (usually some or all coordinates). The perturbation is random, but limited by introduction of a maximum displacement parameter, $\vec{dr}_{max}$. This parameter controls the acceptance ratio (fraction of steps which is accepted) of the simulation, and should be regulated during the simulation so that one has an acceptance around a certain value (most usually 50%, however some has suggested a value of as low as 10% [2]).

```
Perturb the system randomly;
$\vec{q}_{new} = \vec{q}_{old} + \vec{dr}_{max} * random(-1,1)$
Calculate the new energy; $V_{new} = V(\vec{q}_{new})$
if $\frac{\rho_{new}}{\rho_{old}} = e^{(V_{new}-V_{old})/kT} > 1$:
    Accept move. Count new point
else if $\frac{\rho_{new}}{\rho_{old}} = e^{(V_{new}-V_{old})/kT} > random(0,1)$:
    Accept move. Count new point
else:
    Reject move. Old point is counted once more
```

Figure 2.3: Pseudocode for the Metropolis algorithm

The energy of the new configuration is then calculated. If this energy is lower than the that of the old configuration, or if $e^{(V_{new}-V_{old})/kT}$ is larger than a random number between 0 and 1 (chosen from a uniform distribution), the new configuration is accepted. Otherwise the configuration is rejected, and the system returned to the old configuration, which is sampled once more.

Chapter 3

The Force Field Parameters

To model the intramolecular energy and the atomic charges, dipoles and polarizability, 142 calculations (giving a total of 282 data points) were performed at various configurations of the water monomer. The software used was Molcas 6 [71]. The computations where performed at the CASSCF level of theory using one inactive orbital and eight active orbitals. The basis set used was the correlation consistent aug-cc-pVTZ basis set of Dunning [72]. The calculations were done at the CASSCF level of theory to get results consistent with experimental values [15].

The parameters for the repulsive energy and the dampening term in the dispersion energy were fitted to a series of 50 HFSCF and MP2 calculations on the water dimer. These calculations w ere performed in Molcas 5.4 [71] using the aug-cc-pVTZ basis set.

The parameter fitting to the linearized equations were done using the Minitab 14 software [73]. For the repulsion energy three different algorithms were used; the simulated annealing in Numerical Recipes [46] (SA-NR), a simulated annealing algorithm based on the metropolis algorithm (SA-M) and a genetic algorithm (GA) [74]. The SA-NR algorithm was used to parameterize the dampening term in the dispersion energy.

For the intramolecular energy and the atomic properties the reference structure was found by geometry optimizing the water monomer at the same level of theory and with the same basis set as the other monomer calculations. The optimized geometry is presented in table 3.1.

Table 3.1: Equilibrium geometry of water at the CASSCF level using aug-cc-pVTZ basis set

	Symbol	*Value*	*Units*
Angle	θ_e	104.583	Degrees
Bond length	r_e	0.963039	Å

The position of the water molecule is illustrated in figure 3.1, and in the following section bond length with index 1 will refer to the hydrogen atom which lies along the x-axis (the position relative to the axis are of importance when modeling the components of the dipole moment and the polarizability).

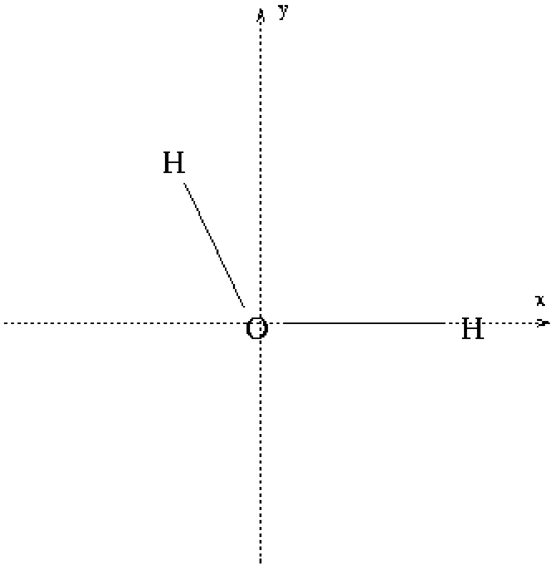


Figure 3.1: Illustration of the position of the atoms in the water molecule, with respect to the axis

3.1 Intramolecular energy

To find the parameters in the expression for the intramolecular energy in eq. 2.22, PLS regression was used. To be able to perform a linear regression on this equation one must transform the expression into a linear form.

Lets first of consider the SPF potential. The SPF potential is (as written in eq. 2.18) given by;

$$V(r) = B_0 \left[\frac{r-r_e}{r}\right]^2 \left(1 + \sum_{n=1}^{\infty} b_n \left[\frac{r-r_e}{r}\right]^n\right)$$

describing the energy between two bonded atoms. To adjust this potential to a series of data points, one must first of all truncate the sum (here truncated at 4) and then do the following changes;

$$\begin{aligned}
V_{SPF}(r) &= B_0 \left[\frac{r-r_e}{r}\right]^2 \left(1+\sum_{n=1}^{4} b_n \left[\frac{r-r_e}{r}\right]^n\right) \\
&= B_0 q^2 \left(1+\sum_{n=1}^{4} b_n q^n\right) \\
&= B_0 q^2 + b_1 q^2 B_0 q + b_2 q^2 B_0 q^2 + b_3 q^3 B_0 q^2 + b_4 q^4 B_0 q^2 \\
&= B_0 q^2 + C_1 q^3 + C_2 q^4 + C_3 q^5 + C_4 q^6 \qquad (3.1) \\
&= a x_1 + b x_2 + c x_3 + d x_4 + e x_5
\end{aligned}$$

giving a linear expression. The angle bending part must also be transformed;

$$V_{angle} = k_\theta \left(\cos\theta - \cos\theta_e\right)^2 = k_\theta x_\theta. \qquad (3.2)$$

The cross terms are transformed in a similar manner, and thus the total potential becomes;

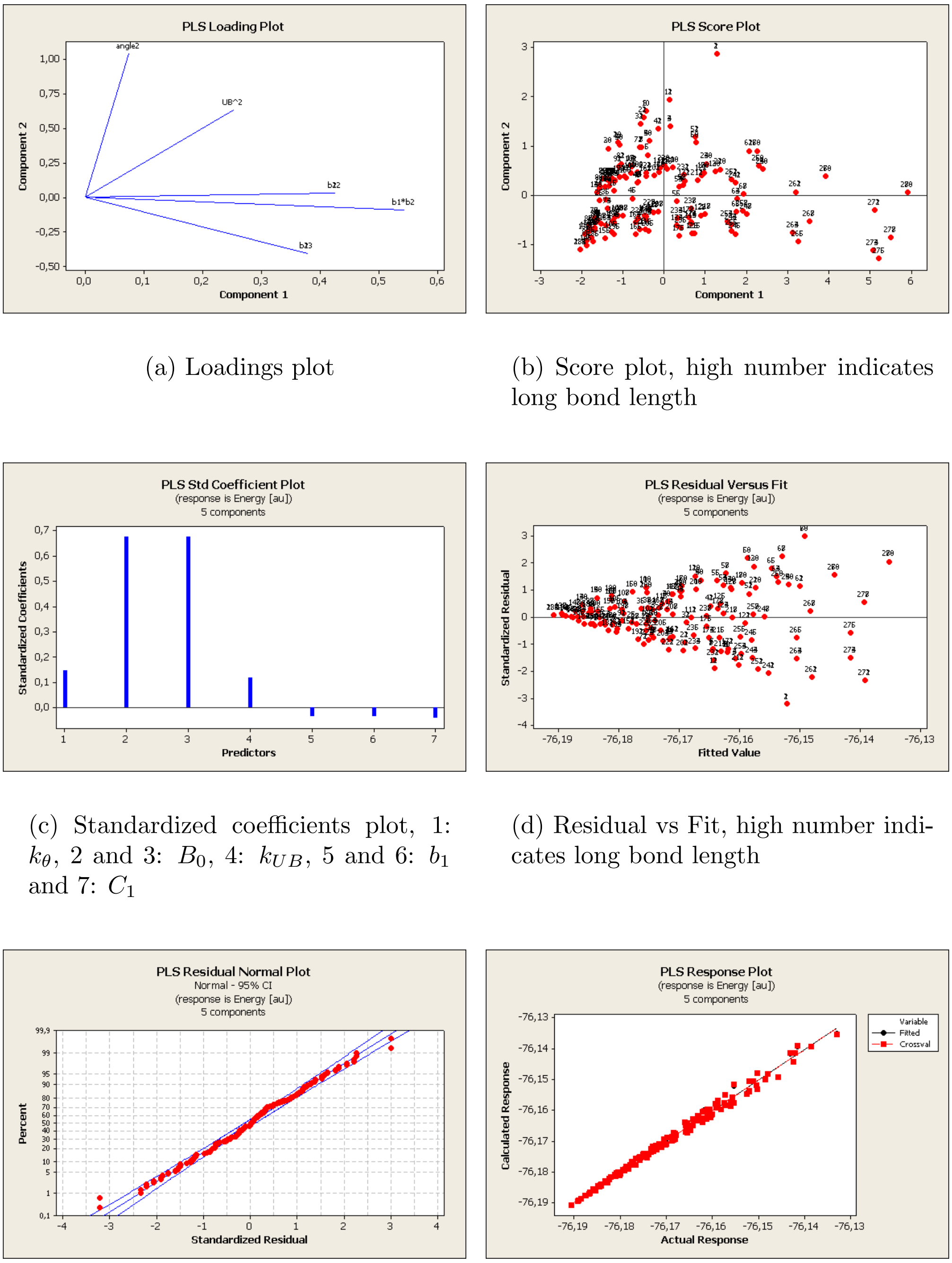
$$\begin{aligned}
V_{mol} &= a_1 x_{1,1} + b_1 x_{2,1} + c_1 x_{3,1} + d_1 x_{4,1} + e_1 x_{5,1} \qquad (3.3) \\
&\quad + a_2 x_{1,2} + b_2 x_{2,2} + c_2 x_{3,2} + d_2 x_{4,2} + e_2 x_{5,2} \\
&\quad + k_\theta x_\theta \\
&\quad + c_1 x_{r_1,r_2} \\
&\quad + k_{UB} x_{UB} + constant
\end{aligned}$$

which is a linear expression.

The partial least squares fit was done expanding the SPF potential only to the first part (b_1 part), as it was found that the model was not improved further by increasing the numbers of terms. The intramolecular energy was thus fitted to the following expression;

$$\begin{aligned}
V_{intra}(r_1, r_2, \theta) &= B_0 \left[\frac{r_1 - r_e}{r_1}\right]^2 \left(1 + b_1 \frac{r_1 - r_e}{r_1}\right) \qquad (3.4) \\
&\quad + B_0 \left[\frac{r_2 - r_e}{r_2}\right]^2 \left(1 + b_1 \frac{r_2 - r_e}{r_2}\right) \\
&\quad + k_\theta \left(\cos\theta - \cos\theta_e\right)^2 \\
&\quad + k_{UB} \left(\frac{r_{1,3} - r_{UB}}{r_{1,3}}\right)^2 \\
&\quad + C_1 \left[\frac{r_1 - r_e}{r_1}\right]^2 \left[\frac{r_2 - r_e}{r_2}\right]^2 \\
&\quad + constant
\end{aligned}$$

The results from the regression are given in figures 3.2(a) to (f) and table 3.2. The loadings plot in figure 3.2(a) tells us that the most important contributions to the intramolecular



(a) Loadings plot

(b) Score plot, high number indicates long bond length

(c) Standardized coefficients plot, 1: k_θ, 2 and 3: B_0, 4: k_{UB}, 5 and 6: b_1 and 7: C_1

(d) Residual vs Fit, high number indicates long bond length

(e) Residual normal probability plot

(f) Response plot, R^2=0.9925

Figure 3.2: Results from PLS regression on the intramolecular energy

energy arises from the bond length terms, as they span out the first PLS component, and that these terms are correlated. The main contributer to the second PLS component is the angle term, with the Urey-Bradley term somewhere between the pure angle and the bond terms. This is not surprising, as the distances and angle are related through following equation $a^2 = b^2 + c^2 - 2bc\cos\theta$. In the score plot (figure 3.2(b)) we observe that the data points with the largest importance for the first PLS component are points where the bond length are particularly large, while the second PLS component is dominated by a point with a small angle (denoted 2). It therefore seems that for the first two PLS components the extreme points are the ones given importance when adjusting the dataset to the functional form. The standardized coefficients plot in figure 3.2(c) tells us that the most important coefficient is B_0. This is not surprising as it is the equivalent of the spring constant of harmonic oscillators. The two angle terms seem to be equally important. The reason for this might be that the Urey-Bradley term describes the molecule better for smaller angles, whereas the harmonic cosines term describes the equilibrium geometry better. The anharmonic terms of the bond energy and the cross term between these seem to give small corrections to the model. There seems to be hetroscedatic noise, when looking at the residual vs. fit plot in figure 3.2(d). This hetroscedasticity arises from the cosines term in the bond energy, as this one does not go towards zero at infinite separation between the hydrogen atoms. The residuals seems to follow a normal distribution, as they in the residual normal probability plot in figure 3.2(e) lies more or less along a straight line. The response plot in figure 3.2(f) shows a good fit between the actual values and the model, with a R^2 of 0.9925.

The numerical results from the regression are presented in table 3.2 and a plot of the energy as a function of the bond lengths is presented in figure 3.3. As we can see from this figure there exists a distinct minima around the equilibrium energy, and as the one or both of the bond lengths approach zero the energy goes towards infinity. We also observe that for infinite separation, the energy converges towards a constant value which can be chosen as a reference value.

Table 3.2: Parameters determining the intramolecular potential

Parameter	*Value*	*Units*
B_0	8.737(-1)	E_h
b_1	-2.568(-1)	E_h
k_θ	6.86(-2)	E_h
k_{UB}	2.84(-2)	E_h
C_1	-2.6220	E_h
Constant	-76.1907	E_h

The energy as a function of the angle is presented in figure 3.4, and as we can see from this figure there is a local maxima at 180^o, and the energy goes towards infinity as the angle goes towards 0 or 360^o.

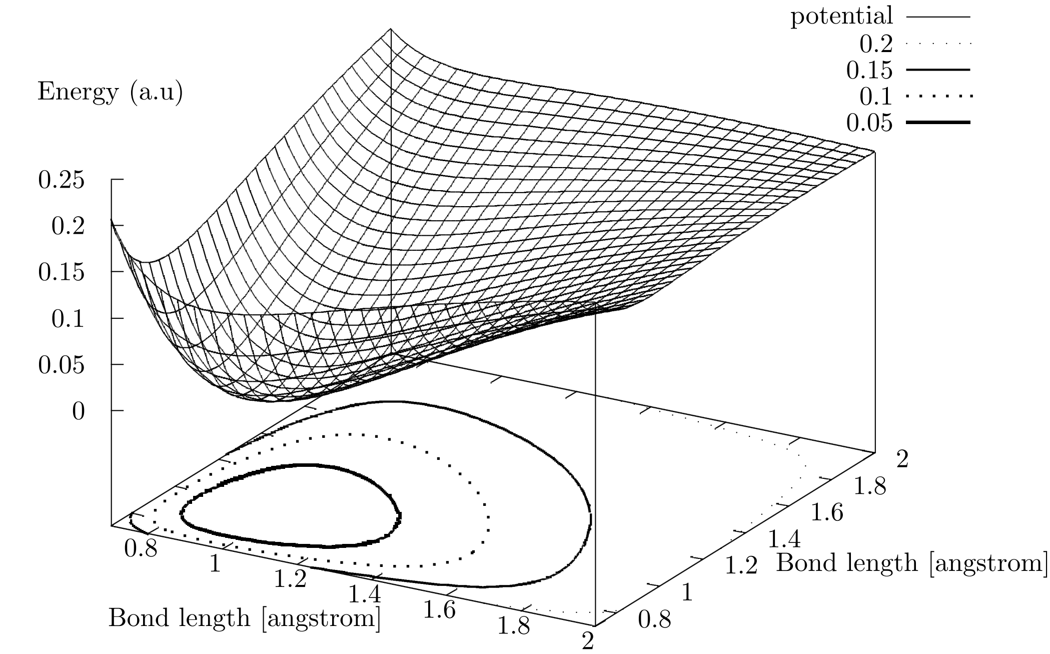


Figure 3.3: The energy as a function of the bond length (at equilibrium geometry of the angle)

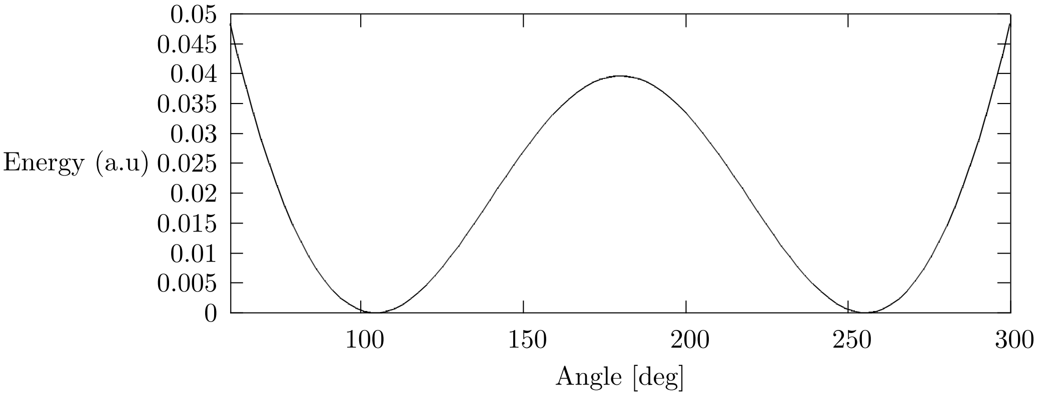


Figure 3.4: The energy as a function of the angle (at equilibrium geometry of both bonds)

3.2 Atomic charge

The atomic charge of the hydrogen atom was fitted to the following;

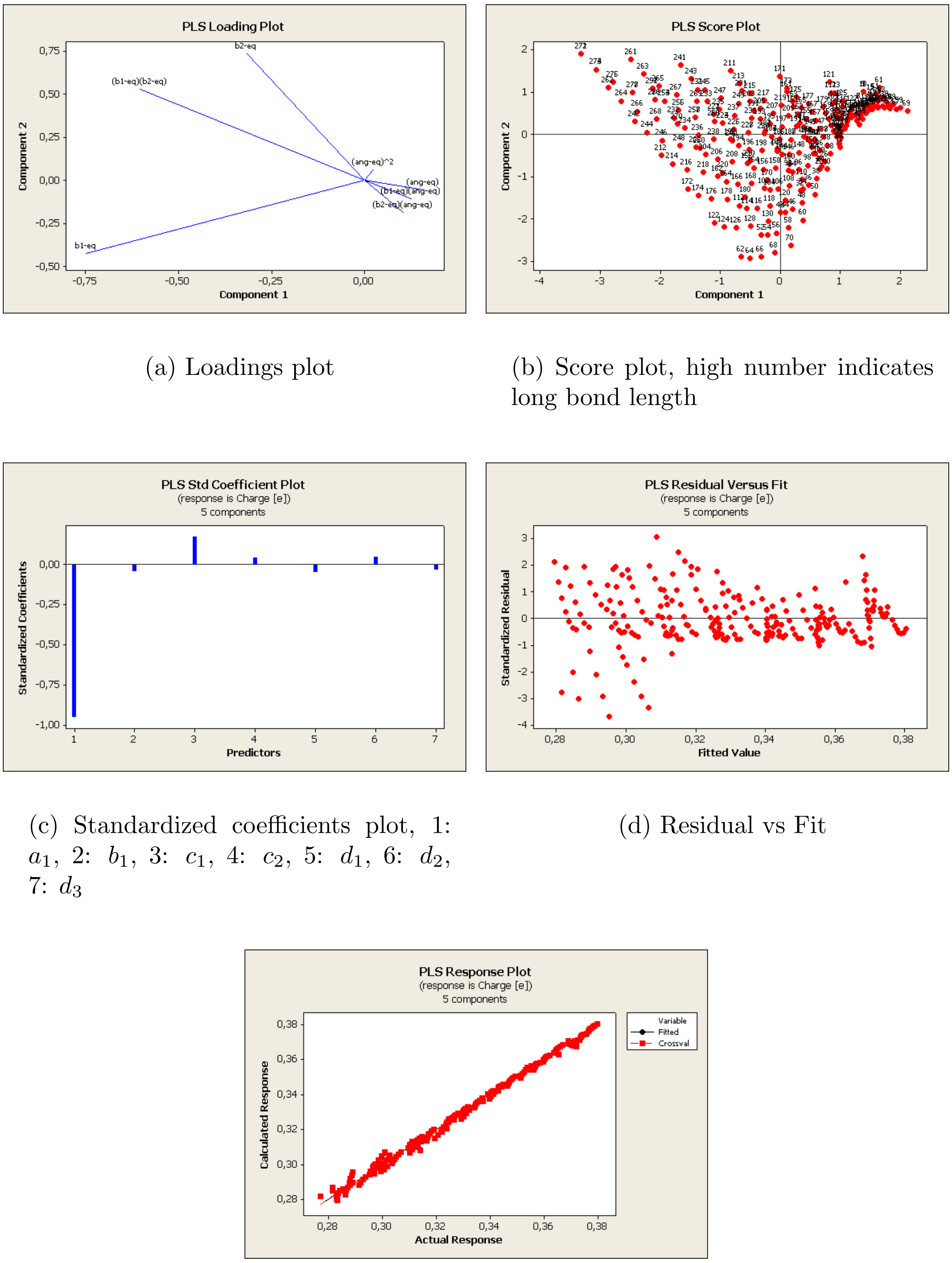
$$\begin{aligned} q_H &= a_1(r_1 - r_e) \\ &+ b_1(r_2 - r_e) \\ &+ c_1(\theta - \theta_e) + c_2(\theta - \theta_e)^2 \\ &+ d_1(r_1 - r_e)(r_2 - r_e) + d_2(r_1 - r_e)(\theta - \theta_e) + d_3(r_2 - r_e)(\theta - \theta_e) \\ &+ constant \end{aligned} \tag{3.5}$$

where r_1 is the distance between the hydrogen in question and the oxygen, r_2 the distance between the other hydrogen and the oxygen and θ the angle between the two hydrogens. This was done using PLS regression. The quadric bond terms were excluded as it was found that they did not give a significant contribution to model.

The results of the regression are presented in figures 3.5(a) to (e) and table 3.3. In figure 3.5(a) the loadings plot from the regression is presented. Here both the first and the second PLS component are span out by the bond terms, and that the main contribution of the angle terms (but not the quadric angle term) are along the first component. This is of course reflected in the score plot in figure 3.5(b). Here the data points with high bond lengths are spread widely to the left of the plot, however there seem to be a large group of data points in the second quadrant of the plot which is the area where the angle is most important. This is further reflected in the plot of the standardized coefficients (figure 3.5(c)), which describe the "importance" of the various coefficients in the model. There the bond to the oxygen atom is by far the most important part of the model (and indeed using the following model: $q_H = a_1(r_1 - r_e) + const$, gives a good fit), followed by the angle. The role of the angle seems thus to be to distinguish between the points mentioned in the score plot. The other variables seem to give small corrections to the model. The standardized residuals are plotted against the fitted value in figure 3.5(d), and seem to be evenly spread, thus indicating that there is no hetroscedastic noise in the residuals. The response plot in figure 3.5(e) tells us that the model reproduces the LoProp charges very well.

Table 3.3: Parameters determining the atomic charge of hydrogen

Parameter	*Value*	*Unit*
a_1	-2.66349(-1)	e/Å
b_1	-1.2049(-2)	e/Å
c_1	4.58(-5)	e/Deg
c_2	1.3(-5)	e/Deg^2
d_1	9.9230(-2)	e/Å^2
d_2	1.161(-3)	e/Å Deg
d_3	-7.19(-4)	e/Å Deg
constant	3.43976(-1)	e



(a) Loadings plot

(b) Score plot, high number indicates long bond length

(c) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: c_2, 5: d_1, 6: d_2, 7: d_3

(d) Residual vs Fit

(e) Response plot, R^2=0.996277

Figure 3.5: Results from PLS regression on the atomic charge of hydrogen

The parameters of eq. 3.5, are presented in table 3.3.

To obey the law of charge conservation, the atomic charge on the oxygen atom is chosen to be;

$$q_O = -(q_{H_1} + q_{H_2}) \tag{3.6}$$

as was done by Stern et al. [26] and Wallqvist [15].

3.3 Atomic dipole moments

To be able to fit the dipole moment to a function (or more correctly a set of functions), we need to transform the components of the dipole moment obtained in the ab inito calculation onto a common basis which is only dependent on the molecular geometry, and not on an arbitrarily chosen origin. According to Edward and Penny [75] such a change of basis can be written as;

$$\mathbf{M}_B\vec{v}_B = \mathbf{M}_{B'}\vec{v}_{B'} \tag{3.7}$$

where $\mathbf{M}_B$ and $\mathbf{M}_{B'}$ are matrices with the basis vectors of basis B and B' respectively as columns, and $\vec{v}_B$ and $\vec{v}_{B'}$ are the coordinate vectors in the two different basis. The coordinate vector in the new basis can then be written as;

$$\vec{v}_B = \mathbf{M}_B^{-1}\mathbf{M}_{B'}\vec{v}_{B'} \tag{3.8}$$

Details on this change of basis can be found in appendix A.2.1, where the mathematical details are described. Such a change of basis may also be seen as a rotation of the coordinate system.

Using this equations one get a well defined atomic dipole moment using a unique basis which is only dependent on the molecular geometry. During a simulation this basis must be calculated. Together with the dipole moment in this basis (found through the use of the regressional model), it must be used to calculate the dipole moment in the basis of the system, before the forces are calculated. The molecular basis used in this study was the following; x-axis was defined to be the unity vector between the oxygen and the hydrogen atom, the y-axis was defined to be the cross product between the x-axis and the (unity) normal vector of the molecular plane and the z-axis was defined to be the (unity) normal vector of the molecular plane. A rotation with an angle equal to the angle of the molecule must of course be done on one of the hydrogen atoms before the rotation, so that the geometry in figure 3.1 is attained. This also corresponds to the axis used in the ab inito calculations. The calculation of a common molecular basis is therefore not necessary for the adjusting of the dipole components, but utilizing this gives a reduced number of calculations.

The components of the dipole moment was fitted to a polynome in the same manner as the atomic charges;

$$\begin{aligned}\mu_{A_q} &= a_1(r_1-r_e)+a_2(r_1-r_e)^2 \\ &+ b_1(r_2-r_e)+b_2(r_2-r_e)^2 \\ &+ c_1(\theta-\theta_e)+c_2(\theta-\theta_e)^2 \\ &+ d_1(r_1-r_e)(r_2-r_e)+d_2(r_1-r_e)(\theta-\theta_e)+d_3(r_2-r_e)(\theta-\theta_e) \\ &+ const\end{aligned} \tag{3.9}$$

where A is H or O, and q is x, y or z. This was done using PLS regression

The parameters for the various components in eq. 3.9 are presented in table 3.4 and 3.5 for hydrogen and oxygen respectively. Due to the shear number of plots involved, the discussion of the various models for the components have been moved to appendix B.1.1. Summing up the discussions one can conclude that the model in eq. 3.9 is somewhat too complicated for hydrogen, and a simpler model might be used for this element. For oxygen, having a more complex electron configuration, the model used seems to be of the correct size. Furthermore it is found that the regressional model for both elements gave results which where very much in agreement with the results from the LoProp calculations.

Table 3.4: Parameters determining the atomic dipole moment for hydrogen

Parameter	*x-component*	*y-component*	*Unit*
a_1	2.06850(-1)	6.69797(-2)	$ea_0/\text{Å}$
a_2	-4.4398(-2)	-4.65204(-2)	$ea_0/\text{Å}^2$
b_1	1.12113(-2)	9.0968(-3)	$ea_0/\text{Å}$
b_2	-8.919(-3)	5.8527(-3)	$ea_0/\text{Å}^2$
c_1	3.43(-4)	8.618(-4)	ea_0/Deg
c_2	-1(-6)	-1.35(-5)	ea_0/Deg^2
d_1	8.9868(-2)	-4.05265(-2)	$ea_0/\text{Å}^2$
d_2	4.27(-4)	8.5(-6)	$ea_0/\text{ÅDeg}$
d_3	-5.73(-4)	-3.096(-4)	$ea_0/\text{ÅDeg}$
const	-1.42383(-1)	-4.093(-3)	ea_0

Table 3.5: Parameters determining the atomic dipole moment for oxygen

Parameter	*x-component*	*y-component*	*Unit*
a_1	1.25630(-1)	1.31550(-1)	$ea_0/\text{Å}$
a_2	-4.99838(-1)	-1.22641(-1)	$ea_0/\text{Å}^2$
b_1	1.25630(-1)	1.31550(-1)	$ea_0/\text{Å}$
b_2	-4.99838(-1)	-1.22641(-1)	$ea_0/\text{Å}^2$
c_1	-1.864(-3)	2.361(-3)	ea_0/Deg
c_2	-2.7(-5)	-2.6(-5)	ea_0/Deg^2
d_1	6.17129(-1)	-4.21063(-1)	$ea_0/\text{Å}^2$
d_2	-2.33(-4)	1.610(-3)	$ea_0/\text{ÅDeg}$
d_3	-2.33(-4)	1.610(-3)	$ea_0/\text{ÅDeg}$
const	1.97459(-1)	2.66008(-1)	ea_0

3.4 Atomic polarizabilities

The components of the atomic polarizabilities were fitted to eq. 3.10 for both hydrogen and oxygen. The geometry used (relative position along the axis) was the one illustrated in figure 3.1, and during the simulations a rotation or change of basis must be performed on the polarization tensor. The mathematical background for this is presented in appendix A.2.2. The same molecular basis was used for the atomic polarizabilities as was used for the atomic dipoles.

The non-zero components of the atomic polarizabilities were fitted to the following equation;

$$\begin{aligned}\alpha_{A_{ij}} &= a_1(r_1 - r_e) + a_2(r_1 - r_e)^2 \\ &+ b_1(r_2 - r_e) + b_2(r_2 - r_e)^2 \\ &+ c_1(\theta - \theta_e) + c_2(\theta - \theta_e)^2 \\ &+ d_1(r_1 - r_e)(r_2 - r_e) + d_2(r_1 - r_e)(\theta - \theta_e) + d_3(r_2 - r_e)(\theta - \theta_e) \\ &+ const\end{aligned} \tag{3.10}$$

where A is H or O, and $i, j \in [x, y, z]$. The method used was PLS regression. A comprehensive discussion of the results obtained can be found in appendix B.1.2, and the parameters obtained in the regression can be found in table 3.6 and 3.7 for hydrogen and oxygen respectively.

Table 3.6: Coefficients for the various components of the polarizability of the hydrogen atom

	α_{xx}	α_{xy}	α_{yy}	α_{zz}	*Unit*
Constant	9.4205(-1)	6.2803(-2)	5.2569(-1)	7.0104(-1)	a_0
a_1	1.83685	1.74330(-1)	1.30864	1.61670	a_0/Å
a_2	1.36811	1.37767(-1)	1.43846	1.37683	a_0/Å^2
b_1	2.7644(-1)	2.8225(-2)	2.952(-2)	2.0363(-1)	a_0/Å
b_2	2.1886(-1)	-1.8569(-2)	-3.3506(-1)	-1.1457(-1)	a_0/Å^2
c_1	-1.92(-3)	2.931(-3)	1.73(-3)	-6.4(-4)	a_0/Deg
c_2	-2(-5)	-3.9(-5)	2(-5)	-3(-5)	a_0/Deg^2
d_1	1.01786	-9.4229(-2)	1.314(-2)	3.5857(-1)	a_0/Å^2
d_2	1.87(-3)	6.101(-3)	1.66(-3)	-3.18(-3)	a_0/DegÅ
d_3	9.66(-3)	1.545(-3)	2.68(-3)	3.26(-3)	a_0/DegÅ

To sum up the discussion in appendix B.1.2,; the models constructed using eq. 3.10 were found to be in good agreement with the LoProp results. Furthermore it was found that quite a lot of the terms in eq. 3.10 were without any significant influence for the atomic polarizability of hydrogen. For the oxygen atom it was found that no simplification of eq. 3.10 should be made. It was also found that the atomic polarizability of oxygen was

Table 3.7: Coefficients for the various components of the polarizability of the oxygen atom

	α_{xx}	α_{xy}	α_{yy}	α_{zz}	*Unit*
Constant	2.34104	4.8152(-1)	2.54533	3.46180	a_0
a_1	6.2608(-1)	6.9109(-1)	1.04195	1.92799	a_0/Å
a_2	-6.8014(-1)	4.5372(-1)	-2.85748	-1.70663	a_0/Å^2
b_1	6.2608(-1)	6.9109(-1)	1.04195	1.92799	a_0/Å
b_2	-6.8014(-1)	4.5372(-1)	-2.85748	-1.70663	a_0/Å^2
c_1	-1.231(-2)	1.100(-2)	8.04(-3)	3.21(-3)	a_0/Deg
c_2	-2.3(-4)	-9(-5)	2.1(-4)	4(-4)	a_0/Deg^2
d_1	-3.63068	-1.60016	1.73289	1.32501	a_0/Å^2
d_2	-4.204(-2)	8.78(-3)	-1.123(-2)	-1.31(-3)	a_0/DegÅ
d_3	-4.204(-2)	8.78(-3)	-1.123(-2)	-1.31(-3)	a_0/DegÅ

significantly more difficult to describe than that of oxygen, the reason for this is two folded; 1) the electronic structure of oxygen is more complex than that of hydrogen and introduces more degrees of freedom. 2) LoProp defines the change of charges imposed due to an external electric field to go into the bonds. As the oxygen atom has two bonds compared to the single bond of the hydrogen, this increases the degrees of freedom in the system.

3.5 Repulsion- and dispersion energy parameters

3.5.1 The repulsion energy

A Boltzmann weighting on the error with a temperature of 6000 K was used in the calculations. In the SA-NR routine an initial temperature[1] of 1 where used and the temperature scheme was the following;

$$T = T_0 \left(1 - \frac{n}{n_{tot}}\right) \tag{3.11}$$

where T_0 was the initial temperature, n the step number and n_{tot} the total number of steps. A total of 50000 MC steps were taken at each temperature, and 50 different temperatures were used. In the SA-M routine the initial temperature was set to 10000 K, the number of steps at each temperature was the same as in the SA-NR routine and the following temperature scheme was used;

$$T = T_0 \left(1 - \frac{n}{n_{tot}}\right)^2 \tag{3.12}$$

The maximum displacement was updated for each 500th step, and only one degree of freedom (randomly selected) was perturbed at each MC step. When using the GA algorithm rounds of 1200 seconds were used, each generating around 2000 generations. The number of rounds where set to 114.

The results from the various optimization routines are presented in table 3.8. As might be seen from this table, both the SA algorithms ended in the same minima, whereas the GA did not find this minima. The best fit was found using the SA-M algorithm, which performed slightly better (but slower) than the SA-NR algorithm. The unweighted RMSE value is slightly lower for the SA-NR search, but this value cannot be used to compare the two methods, as it was not the objective of the methods to optimize this value. The values found by the SA-M algorithm were used in the force field.

The differences between the weighted RMSE values and the unweighted values in table 3.8 are not very large. This indicates that the chosen temperature of 6000 K used for the boltzmanns weighting of the error lies at an appropriate level.

There might be several reasons that prevented the GA algorithm to perform as good as the SA algorithms, but it most probably arises from the relative short time-interval used and an increase in this time should improve the performance. One indication of this is that one can get quite low RMSE values from the GA algorithm after only 96 generation, starting from a random set of points. No similar ability to find a minima that good in so short time interval was found for any of the SA algorithms, thus indicating that the GA algorithm seems to be more robust than the SA algorithms. The time used on one GA run (1200 s) was also considerably lower than the time used on the SA algorithms (12 h), and an increase in the run time of the GA algorithm might increase its performance. One last

[1] in the SA-NR algorithm, temperature is unitless

point to note is the fact that the GA algorithm originally was designed for studies of the inflow in oil wells, and therefore not especially designed for the task of finding force field parameters.

Table 3.8: Repulsion energy parameters

Result	*SA-NR*	*SA-M*	*GA (1935 generations)*	*Unit*
RMSE	1.8645(-1)	1.865(-1)	1.2212	kcal/mol
RMSE Weighted	1.865(-1)	1.506(-1)	3.5508(-1)	kcal/mol
α_{OO}	-3.6079	-3.6048	-3.0200	Å^{-1}
α_{HO}	-5.3304	-5.3051	-5.0000	Å^{-1}
α_{HH}	-3.4131	-3.4038	-3.7100	Å^{-1}
κ_{OO}	3.8172(4)	3.7923(4)	1.0486(4)	kcal/mol
κ_{HO}	1.5754(4)	1.5188(4)	9.1750(3)	kcal/mol
κ_{HH}	8.5791(2)	8.4353(2)	1.2865(3)	kcal/mol

3.5.2 Dispersion energy

A combination of the SA-NR algorithm and simplex search was used to model the various parameters in the dampening term of the dispersion energy (eq. 2.40). The SA-NR algorithm was used with 5000 MC steps at each temperature and with 10 different temperatures. The initial temperature was chosen to be 8, and the temperature scheme in eq. 3.11 was used. After the SA-NR algorithm was finished, a simplex search was performed with the results from the SA-NR as the initial value.

Table 3.9: Dampening parameters for the dispersion energy, the RMSE was 1.886(-1) kcal/mol

Atom pair	b_{ij} [Å^{-1}]	B_{ij} [kcalÅ$^3/a_0^2$mol]
OO	6.41164(-6)	-2.5344(5)
OH	4.50922(-4)	-2.0895(3)
HH	1.28677(-6)	-4.81654(5)

The result from the combined SA-NR and simplex search is presented in table 3.9, and judging by the RMSE value, the fit seems to be satisfying[2].

[2] In these results B_{ij} also contains conversion to Protomols internal units

Chapter 4

Computational details

The force field described in the two previous chapters was implemented in the Protomol [76] framework. This is an object-oriented framework written in C++, primarily for MD simulations. The force field originally implemented in Protomol is the CHARMM force field [9]. Because of the difference between CHARMM and the force field constructed here, several algorithms had to be implemented. Pseudo-code of these implementations can be found in appendix C.

Protomol is an open source framework, and can be downloaded from;
http://protomol.sourceforge.net/

4.1 Simulation procedure

The simulation procedure used is presented in figure 4.1. After it has been verified that the total number of MC steps has not been reached, the system is perturbed according to the scheme described in section 4.2. This is followed by a calculation of the various atomic properties (for pseudo-code see appendix C), as described by the models in chapter 3. The various energy contributions are then computed (see description in chapter 2 and 3), and the Metropolis test is performed. If the result from the Metropolis test is true, then the new configuration is accepted, if not it is discarded and the system returned to the old configuration. It is then tested if the point should be sampled or not (as described in section 4.5).

The simulation were also divided into macrosteps, one macrosteps consisting of 10000 MC steps. Macrosteps were used to give an increased flexibility in the simulations.

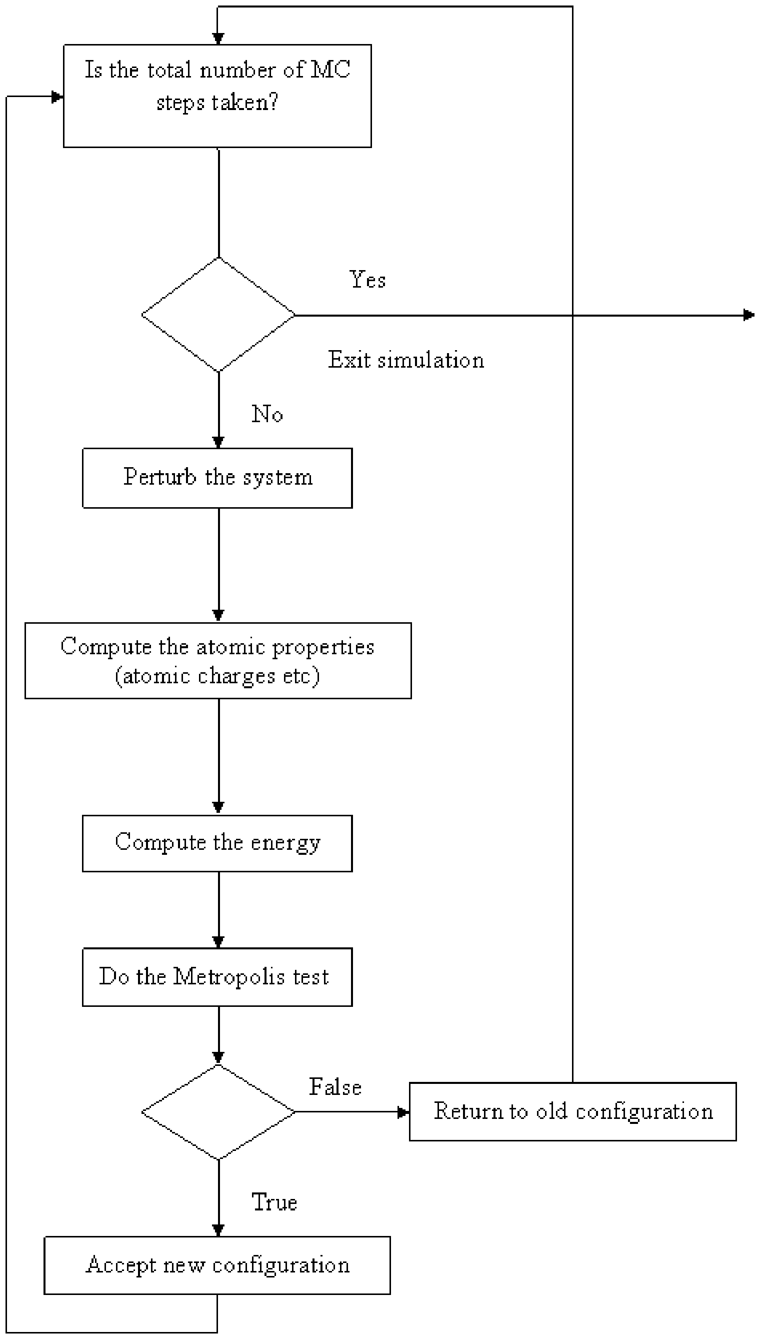

Is the total number of MC steps taken?
Yes
Exit simulation
No
Perturb the system
Compute the atomic properties (atomic charges etc)
Compute the energy
Do the Metropolis test
False
Return to old configuration
True
Accept new configuration

Figure 4.1: Flowsheet for the simulation

4.2 Perturbation scheme

It was chosen not to perturb the cartesian coordinates of the various atoms in the system, as often done in regular Metropolis MC simulation, but rather to perturb the various degrees of freedom in the system. A pseudo code of the used perturbation scheme is presented in figure 4.2. The reason for choosing to perturb the was that we; 1) wants to study the distribution of the bond lengths and angles and 2) it makes it possible to study rigid system easy, so as to make a comparison between rigid and flexible systems. Figure 4.2 also describes how the perturbation is performed, except for the perturbation of the angle, which is described in appendix A.3.

```
1    Chose the degree of freedom to perturb randomly
     (bonds, angles, rotation or translation)
2    *if* bonds:
3        Perturb all bonds according to the following equation;
         $r_i^{new} = r_i^{old} + dr^{max} random(-1, 1)$
4        **Compute** the unity vectors, $\vec{i}_i$ between the H and O atom
5        **Compute** the new position of the hydrogen atom as;
         $R_H^i = R_O + \vec{i}_i r_i^{new}$, where $R_O$ is the oxygen on the molecule
6    *if* angle:
         Perturb all angles according to the following equation;
         $\theta_i^{new} = \theta_i^{old} + d\theta^{max} random(-1, 1)$
7        **Compute** the new positions of the hydrogen atom according to the
         procedure described in appendix A.3
8    *if* rotation:
9        Chose an euler angle randomly
         Perturb that euler angle for all molecules according to the
         following equation;
         $A_i^{new} = A_i^{old} + dA_{max} random(-1, 1)$,
         where $A$ is $\psi$, $\phi$ or $\cos\theta$ (see Leach [1], page 421)
10       Construct the rotation matrix for the molecule
11       Rotate the molecule with the oxygen atom as the origo
12   *if* translation:
13       Perturb the cartesian coordinates of the oxygen according to
         the following equation;
         $q_i^{new} = q_i^{old} + dr_{max}^{trans} random(-1, 1)$,
         where $q \in [x, y, z]$
14       Move the oxygen to the new position, and the positions of
         the hydrogen atoms as;
         $R_H^i = R_O^{new} + \vec{R}_{OH}^{old}$
```

Figure 4.2: Pseudo code for the perturbation scheme used

Which degree of freedom to perturb was chosen from a uniform distribution with 80% probability of choosing either bonds or angles, and 20% probability of choosing one of the external degrees of freedom. This set of probabilities was deduced from an article on multiple time-step methods by Watanabe and Karplus [77], and reflects the short time-step they use for the internal degrees of freedom.

4.3 About the system

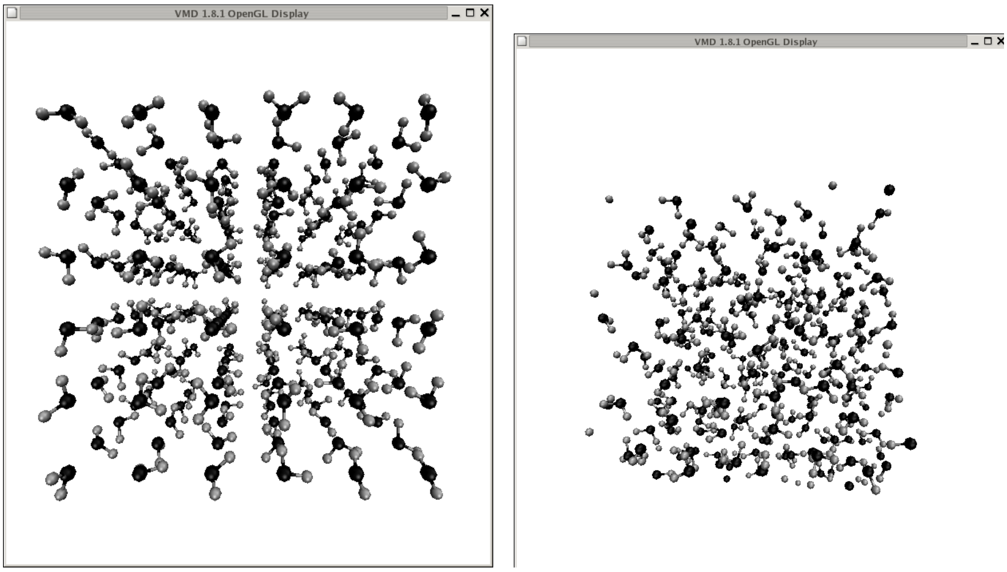
The simulations where done on a system containing 216 water molecules in a cubic box of length 18.6 Å (this is the same setting as in Åstrand et al. [20]). Periodic boundary conditions were used, and the coulombic interactions had a cut-off radius of 8.5 Å. The temperature during the simulations was 298.15 K.

4.4 Relaxation

The initial system contained 216 randomly rotated water molecules with equal spacing between the oxygen atoms. This system was then relaxed through a series MD simulations. These MD simulations where done using the CHARMM22 force field and a leapfrog integrator. The number of MD steps taken where 10000, and a time step of 1 fs where used. The seed used was 1234, otherwise the general settings were those described in section 4.3. After these MD steps a series of MC 8000 steps using the force field constructed in this study were performed. The structure of the system before relaxation is presented in figure 4.3(a), and after relaxation in figure 4.3(b).

In figure 4.4 the kinetic-, potential- and total energy are plotted as a function of the number of MD steps for the first 1000 steps. From this figure we can clearly see that the system has melted well before the first 1000 steps have been taken.

A similar procedure was done for the rigid molecules, were a long MD simulation using the Leapfrog integrator was done with small time-steps (0.05 fs) and very large (1000 times larger than CHARMMs default values) force constants. For the rigid molecules, the equilibrium geometry presented in table 3.1 were used.



(a) Initial structure of the system (figure made using the vmd program [78])

(b) Snap shoot of the system after relaxation (figure made using the vmd program [78])

Figure 4.3: Snap shoots of the system

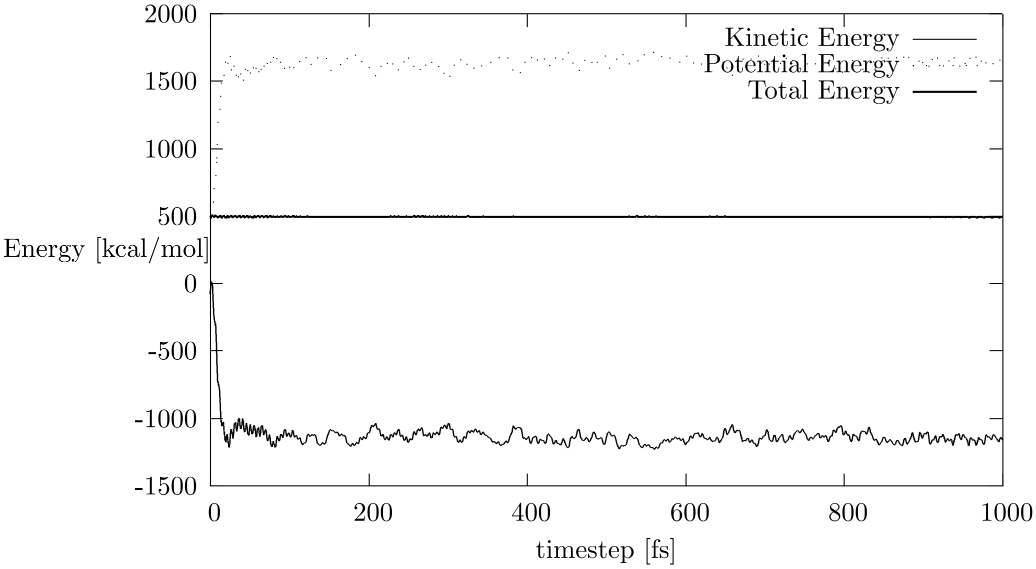


Figure 4.4: Energetics during the MD relaxation

4.5 Sampling procedure

The sampling of the radial distribution, bond lengths and angles where done for every 10th steps, as recommended by Allan and Tildesley [2]. The reason for not sampling all points encountered despite the fact that they are a part of phase space, arises from high correlation between the consecutive steps in a simulation.

4.5.1 Error estimation

According to Allan and Tildesley [2] the errors in structural properties, such as the radial distribution function, can be calculated by the standard deviation of the bins in the simulation. The reason that gaussian statistics can be used comes from the central limit theorem. In this report the standard deviation for each bin for the macrosteps were calculated, and the mean of these standard deviations (over all bins) reported for the properties.

4.5.2 Gas phase

For comparison gas phase results were calculated. These results were computed by running simulations on a single water molecule, given the same conditions as the liquid phase simulations, thus giving an ensemble average for the single molecule.

Chapter 5

Results from simulations

This chapter contains the results from the MC simulations described in chapter 4. The results reported are the vibrational spectra of water, the water geometry and the liquid structure. The results presented here will be discussed in chapter 6.

5.1 Vibrational spectra

The vibrational spectra obtained for water are presented in figure 5.1, for liquid and gas phase. The liquid phase spectrum was obtained using 150000 MC steps, while the gas phase spectrum was obtained taking 200000 MC steps on an isolated molecule, as described in the previous chapter. In figure 5.2 the vibrational spectra calculated, when the ZPV energy is not included in the Metropolis test, is presented. This spectrum was obtained taking 30000 MC steps for the liquid phase. In the corresponding gas phase spectrum 200000 MC steps were taken. The frequencies were sampled in bins of 10 cm^{-1}.

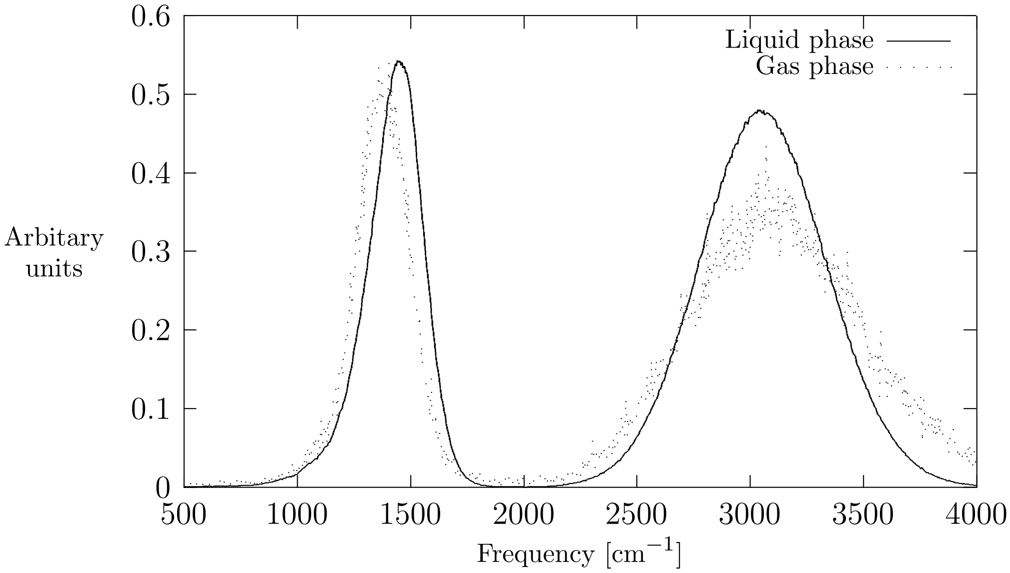


Figure 5.1: Vibrational spectra of water with ZPV energy

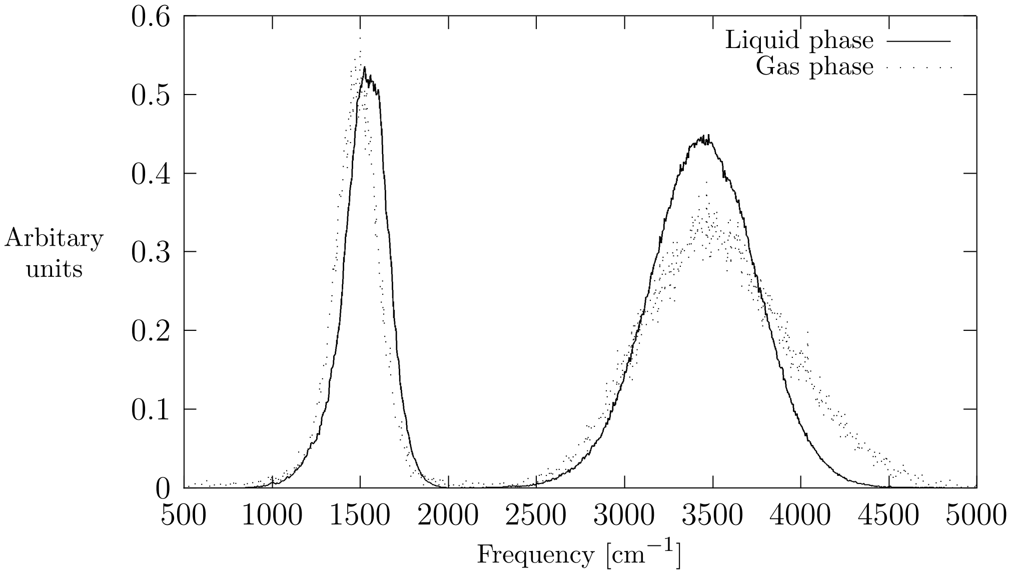


Figure 5.2: Vibrational spectra of water, when ZPV energy is excluded from Metropolis test

5.2 The water geometry

The distribution of the angle and bond lengths are presented in figure 5.3 and 5.4 respectively. The bond and angle distribution were sampled for simulations with and without inclusion of the zero-point vibrational energy. The mean bond length and angle are presented in table 5.1. Figure 5.5(a) and 5.5(b) report the probability of a combination of the two bond lengths. Figure 5.5(c) and 5.5(d) report the probability of a combination of a bond and an angle. The variance of the two properties are also presented in this table. The dipole moment at the expectation values of the geometry for the various simulations are presented in table 5.3. Calculations were done with and without inclusion of the ZPV energy in order to study the effect of this term on the internal coordinates of the molecule.

The distributions were found using 260000 MC steps for the simulations with ZPV energy and 240000 MC steps for the simulations without the ZPV energy.

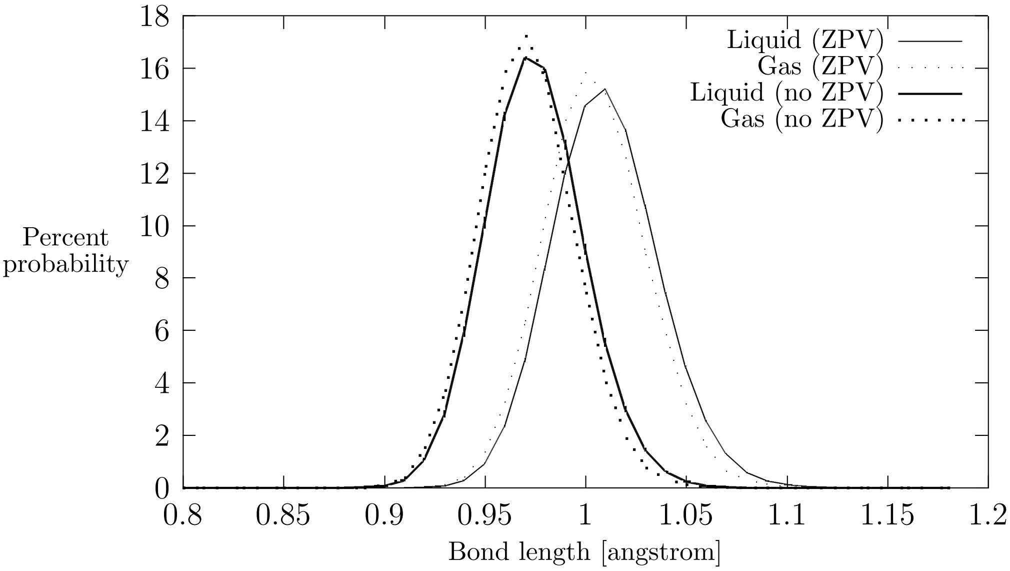


Figure 5.3: Distribution of bond lengths

Table 5.1: Mean and variance of the geometrical properties in liquid phase

	With ZPV		*Without ZPV*	
	Expectation value	Variance	Expectation value	Variance
Angle [Deg]	107.6	44.2	106.9	37.6
Bond [Å]	1.0101	0.0007	0.9751	0.0006

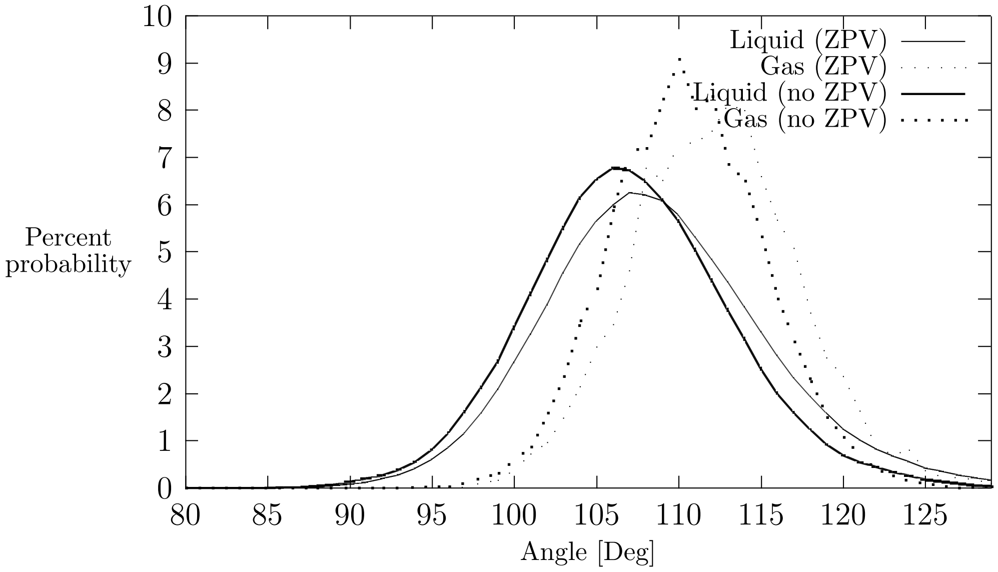


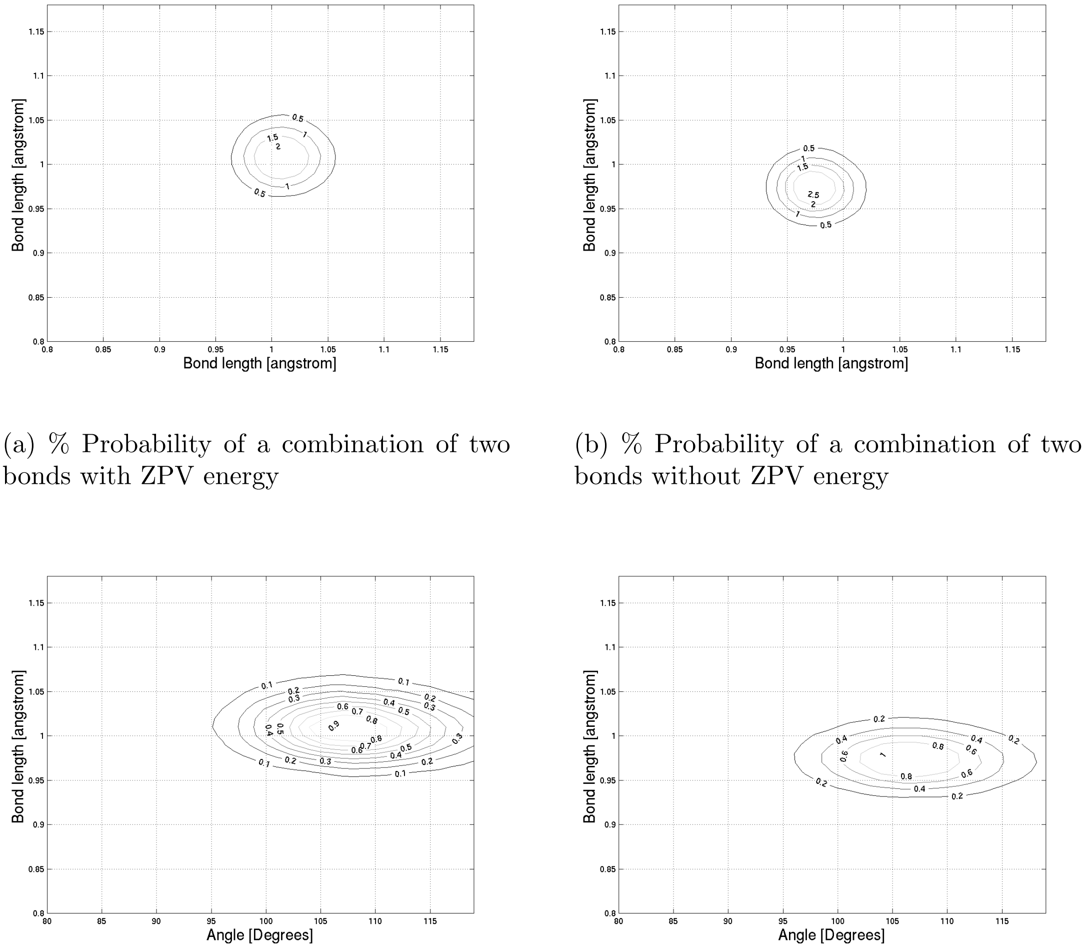
Figure 5.4: Distribution of the angle

Table 5.2: Mean and variance of the geometrical properties in gas phase

	With ZPV		*Without ZPV*	
	Expectation value	Variance	Expectation value	Variance
Angle [Deg]	112.1	25.3	110.4	21.2
Bond [Å]	1.0040	0.0006	0.9700	0.0005

Table 5.3: Norm of the dipole at the mean geometry

	Dipole moment [D]	
	Gas	Liquid
With ZPV energy	2.4057	2.4629
Without ZPV energy	2.3590	2.4032



(a) % Probability of a combination of two bonds with ZPV energy

(b) % Probability of a combination of two bonds without ZPV energy

(c) % Probability of a combination of a bond and an angle with ZPV energy

(d) % Probability of a combination of a bond and an angle without ZPV energy

Figure 5.5: % Probability of combinations of bond length and angle

5.3 Liquid structure

The radial distribution function for simulations of rigid molecules and flexible molecules with and without the ZPV energy term are presented in figures 5.6 to 5.8. The results were sampled using 270000, 260000 and 240000 MC steps respectively for the rigid molecules, the molecules with ZPV energy and the molecules without. The rigid molecules were simulated having the equilibrium geometry. The experimental results are taken from Soper et al. [79].

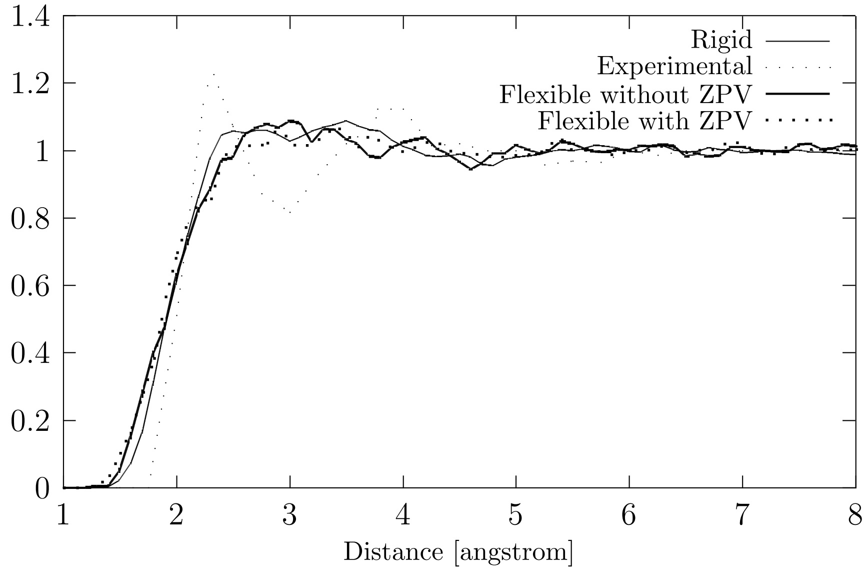


Figure 5.6: HH Radial distribution function for rigid molecules and flexible molecules with and without the ZPV energy

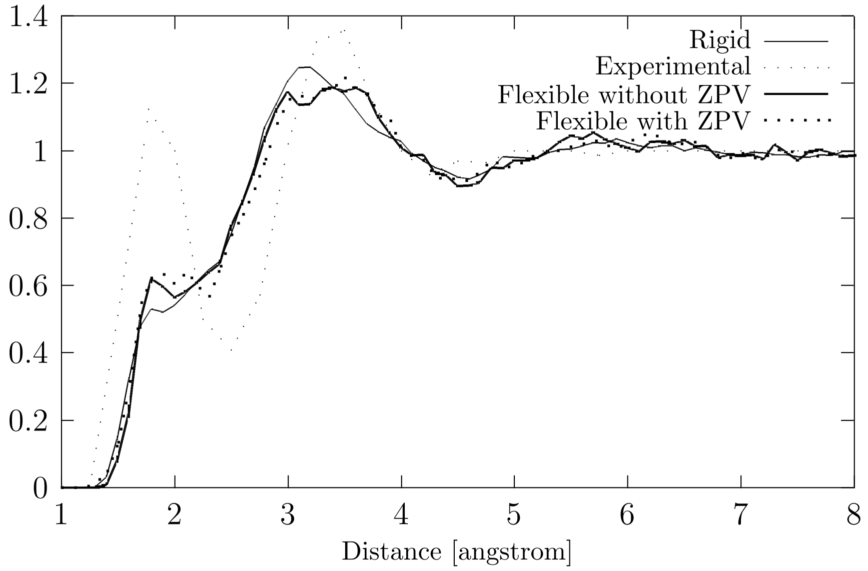


Figure 5.7: OH Radial distribution function for rigid molecules and flexible molecules with and without the ZPV energy

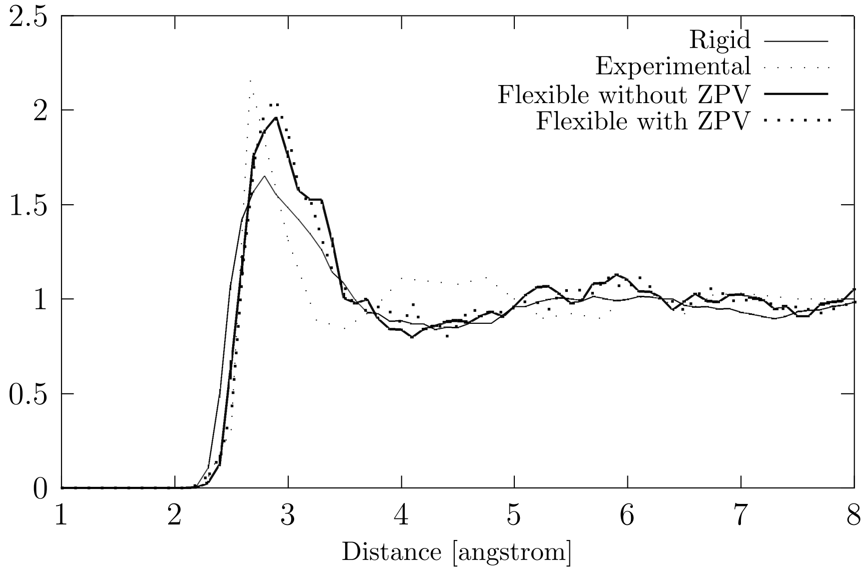


Figure 5.8: OO Radial distribution function for rigid molecules and flexible molecules with and without the ZPV energy

5.4 Error estimates

In table 5.4 and 5.5 the mean standard deviation for the bins, as described in chapter 4, are reported for the simulations. For the gas phase simulations all standard deviations were in the range between 10^{-6}-10^{-8}.

Table 5.4: Mean standard deviation for the vibrational specta, bond and angle distribution

Simulation	*Spectra*	*Bond*	*Angle*
Liquid, ZPV included	0.0083	0.0835	0.1904
Liquid, ZPV not included	0.0134	0.1288	0.1539

Table 5.5: Mean standard deviation for the radial distribution functions

Simulation	*HH*	*OH*	*OO*
Rigid molecules	0.0297	0.0323	0.0637
Flexible with ZPV energy	0.0308	0.0369	0.0580
Flexible without ZPV energy	0.0319	0.0364	0.0557

Chapter 6

Discussion of simulation results

This chapter contains a discussion of the results from the simulation presented in chapter 5. The discussion will focus on the behavior of the force field compared to experimental and other theoretical results.

6.1 Vibrational spectra

6.1.1 Gas phase spectra

The gas phase spectra of water presented in figure 5.1 and 5.2 we are unable to separate the two high frequency peaks in the ensemble average, and instead there is broad peak consisting of the two merged peaks [80]. If we compare the obtained spectra with the literature values presented in table 6.1, the peaks in the spectrum from the simulations incorporating the ZPV energy in the Metropolis test are located at lower values than both the theoretical and the experimental values. In contrast to the quantum mechanical methods this simulation gives systematically lower values of frequencies than the experimental values. If we regard the spectrum obtained in the simulation where the ZPV energy was not included in the Metropolis test, the peaks have been shifted to the right of the spectrum. The shift for the angle bending (the first peak) are about 100 cm^{-1} and 400 cm^{-1} for the bond stretches. This leads to values closer to the experimental values, and results within the same range of accuracy as the quantum mechanical results in table 6.1. This indicates that inclusion of the ZPV energy in the Metropolis test, in the manner done in this study, leads to a molecule which is more easy to perturb. The melting of the two peaks observed in the two gas phase spectra might come from the resolution chosen in the bins, as we can see from table 6.1 these peaks are only 100 cm^{-1} apart. In experimental measurements the problem of two overlapping peaks is solved by deconvolution [80], however this is beyond the scope of this study.

Table 6.1: ZPV frequencies from the literature, all values are in cm^{-1}. The CASSCF values are calculated at the equilibrium geometry

Vibrational mode	*QM ([81])*	*QM [82]*	*CASSCF*	*Experimental ([1])*
Angel bend	1667.80	1657.44	1671.00	1595
Symmetric stretch	3854.35	3848.89	3809.51	3652
Anti-symmetric stretch	3972.02	3965.33	3926.11	3756

6.1.2 Solvent effects

Comparing the difference between the peaks for gas and liquid phase water in figure 5.1, we are able to study the effect of the solvent on the vibrational frequencies. For figure 5.1 the shift of the angle bending is about 70 cm^{-1}, which is somewhat higher than the results from QM and experimental values presented in table 6.2, but still of the same magnitude and with the same sign. The shift in the bond stretch terms are more difficult to determine a choice of reference for the gas phase are almost impossible to get exact. It is therefore more appropriate to express this shift to be within a range of -35 to -180 cm^{-1}. This is of course very crude, however the change has the same sign as the values in table 6.2, even if only the upper part has the same magnitude.

If one compares the spectrum of the liquid in figure 5.1 with the spectra obtained by Wallqvist and Teleman [14], that the peaks in figure 5.1 are much broader than the peaks of Wallqvist and Teleman. For the angle bending peak in figure 5.1 there is a peak in the same area as the peaks of Wallqvist and Teleman, and they have also, for their anharmonic potential, a melting of the two bond stretch peaks. The bond stretch peaks of Wallqvist and Teleman are located at higher frequencies than those in figure 5.1.

Table 6.2: Shift in ZPV frequencies from the literature, all values are in cm^{-1}

Vibrational mode	*QM ([81])*	*Experimental ([83])*
Angel bend	40	50
Symmetric stretch	-401	-265
Anti-symmetric stretch	-387	-275

6.1.3 The effect of including V_{zpv} in the Metropolis test

If the vibrational spectrum is calculated, but the ZPV energy not included into Metropolis test we get the result presented in figure 5.2. If we first of all compare this results with the results in figure 5.1 for the liquid phase, it is clear that the major difference between the results is manifested in the bond stretch vibrations. This peak has been shifted to the right, with about 500 cm^{-1}. This is the same effect as we saw for the gas phase. The results from these simulations therefore seem to be closer in agreement with the experimental and theoretical values. The effect on the angle bending vibration is small, giving a peak at a slightly higher frequency. This suggests that the effect of the ZPV energy on the intramolecular degrees of freedom mostly affects the bond stretch part, giving a more flexible molecular model. The shift from gas to liquid phase for the spectra when the ZPV energy is excluded from the Metropolis test is very similar to those found when this energy has been included. For the angle bending mode the shift is about 70 cm^{-1} and for the bond stretch modes approximately -50 cm^{-1}, which is quite small compared to the experimental and theoretical results in table 6.2.

6.2 The water geometry

The distributions of bond lengths for simulations with and without the ZPV energy is presented in figure 5.3, and it is clear from this figure that introducing the ZPV energy shifts the distribution towards longer bond lengths. This confirms the results from the vibrational spectra. Inclusion of ZPV energy also gives a broader distribution. The distributions in figure 5.3 are almost gaussian, and are comparable in both size and location to the ones obtained by Hess et al. [84].

If we compare the expectation value of the bond length presented in table 5.1 with literature values in table 6.3, the value obtained when including the ZPV energy is around 0.01-0.05 Å larger than most of the literature values, and more specifically 0.04 Å larger than the experimental value. The results obtained without the ZPV energy are in a closer agreement with both theoretical and experimental values, and the discrepancy between the expectation value for this simulation and the experimental results are only 0.005 Å (or a half percent of the total, experimental value).

Compared with the bond length, the angle distributions in figure 5.4 are very wide, something which is also reflected in the large variance of the angle in table 5.1. The distribution of the angle also have the typical bell shape of a gaussian distribution. The effect of introducing the ZPV energy on the angle is relatively small. We have already seen that the frequency of the angle bending vibration only shows a small shift when including the ZPV energy. In figure 5.4 we also observe a quite small change in the angle distribution towards larger angles, when including the ZPV energy. This suggests that the ZPV energy plays little role when regarding the angle. The discrepancy between the angle and the bond lengths, when it comes to the sensibility of the vibrations, comes from two effects. First of all the frequencies involved in the angle bending are much lower than those for the bond stretch vibrations. The energy involved in this vibration is therefore smaller than the bond stretch vibrations. Furthermore there are two terms involved in stretching the bonds. These two effects leads to a larger amount of energy involved in the stretching of the bonds, and hence a larger effect of the ZPV energy. The expectation values of the angle are in good agreement with the experimental results deviating only with around 1^o. The theoretical values presented in the literature are spread, and in light of the broad distributions in figure 5.4, this is not very strange. However, the expectation values calculated for the angle in the simulations done in this study are within the same order of magnitude as the theoretical results in table 6.3.

Table 6.3: Overview of theoretically predicted and experimental measured water geometry

	Angle [deg]	*Bond [Å]*
QM ([82])	106.72	0.9884
QM/MM ([85])	110.20	0.951
MD ([84])	102.7	0.982
MC ([86])	102.7	0.982
Experimental ([87])	106.06	0.970

The expectation values for bond length and angle in gas phase presented in table 5.2 deviates somewhat from the equilibrium geometry values. The reason for this is the anharmonicity in the model, and the gradient in direction of the local maxima in the angle term. The anharmonicity gives a steeper gradient when going in the direction of a value which is lower than the equilibrium value (this can clearly be seen in figure 3.3), and hence it is more likely to have a bond length that is larger than the equilibrium value compared to a value thats smaller. The anharmonicity leads to a change from the equilibrium value to the expectation value of 0.007 Å for the simulation not including the ZPV energy, which interestingly is half the numerical difference between the equilibrium and the “effective” geometry used by Åstrand et al. [81]. For the simulation including the ZPV energy the difference are even larger, due to the additional effect of the inclusion of the ZPV energy. A similar effect as the anharmonicity can be attributed to the cosine term of the angle. It is clear from figure 3.4 that the gradient in direction of the local maxima included to give a more physically sound model is lower than the gradient in the opposite direction. The effect of this on the angle is much more dramatic than for the bond, giving an angle several degrees larger than the equilibrium value. The values are even larger than the values in the liquid phase. This is again an indication (along with the low value of the angle bend vibration) that the angle bending term in the potential are too “sloppy”, and should be parameterized with more data points having a larger value than those in the current parameterization.

In figures 5.5(a) to 5.5(d) the probability of a combination of two bond lengths or a certain bond length and an angle (from here on called the bond-bond distribution and bond-angle distribution) are presented. If we first of all compare the two bond-bond distributions in figure 5.5(a) and 5.5(b), the whole distribution are shifted towards larger values when including the ZPV energy. Furthermore the distribution from the simulation where the ZPV energy has been included (figure 5.5(a)) is broader than the one where its excluded, therefore increasing the total number of attainable configurations. One interesting conclusion we might draw from the bond-bond distribution is that configurations having a large or small bond length on one bond are only possible when the other bond have a length corresponding to the expectation value. The same trend described above is also valid for the bond-angle distribution, where the ZPV energy shifts the distribution towards larger values and gives a wider distribution (see figures 5.5(c) and 5.5(d)). From the bond-angle distributions we also see that extreme values of the angle are only attainable when the bond length is around the expectation value, this is especially the case when the ZPV energy is included. The opposite effect does not seem to be present.

According to Leach [1] the dipole moment of water in liquid state is approximately 2.6 D. If we compare this value with the values in table 5.3, the norm of the dipole moment at the expectation values of the angle and the bond length shows a good agreement with the experimental value. However if one compares the gas phase values to the experimental value which is around 1.85 D the results are not satisfactory. This discrepancy probably arises by virtue of construction as the equilibrium distance attained from the CASSCF geometry optimization deviates from the experimentally measured geometry with regard to the bond length (0.963039 Å vs 0.9572 Å [88]). The trend of increasing the dipole moment going from a gas to a liquid state is still present, although the change is not numerically comparable to that of real life. The change of the geometry which occur as a consequence of the introduction of the ZPV energy gives rise to an increased dipole moment as could be expected. On a final note it must be pointed out that dipole values computed at the expectation value of the bond length and angle are not per definition the correct expectation value of the dipole, which is given by;

$$\vec{\mu}_{exp} = \int\int\int \vec{\mu}(r_1, r_2, \theta)\mathcal{P}(r_1, r_2, \theta)dr_1dr_2d\theta \tag{6.1}$$

however as the distributions are quite symmetric, the values should be correct to a good approximation.

In the discussion above the focus have been on the constructed force fields ability to reproduce the experimentally measured values for the geometry. It is shown that the resulting geometry and dipole moment for the liquid state both are in good agreement with the experimental results. It is also shown that some of this good agreement must be ascribed to the expansion point chosen. A better test of the methods ability to describe the liquid state is the shift from gas to liquid phase. If we compare the experimentally measured bond length in gas and liquid phase, there is a difference of approximately 0.0128 Å. The difference in expectation values obtained from the simulations, is 0.0061 Å when the ZPV energy as been included and 0.0051 Å when it is excluded. Numerically this is half the value of the experimentally measured shift, however taking into account the uncertainty in the experimentally measures for the liquid phase which is 0.005 Å, the results for the bond length must be said to be within a good agreement with the experimentally measured values with a slightly better result when the ZPV energy is included. For the angle the situation is somewhat different. Comparing the experimentally measured values where the shift from gas to liquid phase gives an angle change of 1.54 ± 1.8^o, a value different from the -4.5^o and -3.5^o is obtained when including and excluding the ZPV energy respectively. This again suggest a too “shallow well” for the angle bending term in the intramolecular potential.

6.2.1 Rigid empirical models in light of the obtained distributions

In light of the distributions obtained from the simulations it is interesting to discuss the approximation of rigidity in empirical models. The geometry of the most common empirical water models are presented in table 6.4, and if we first of all look at the choice of bond length in these models they all lie within the distributions in figure 5.3. However the concept of rigidity must be seen in light of the variance and width of the distribution. The variance for the bond length distributions are quite low and the difference between the maximum and minimum bond length is only 0.15 Å. This indicates that having a constant bond length is not a very large approximation, as long as the choice of bond length is within a reasonable size.

The variance of the angle is on the contrary to the bond length quite large, and the molecule therefore seems to be much more flexible in the angle than in the bonds. From table 6.4 most of the common water models have an angle with a value in the vicinity of the expectation values in table 5.1, however given the large variance of the distributions in figure 5.4, the approximation of a constant angle cannot be said to be a good approximation.

Table 6.4: Water geometries for various rigid water models [1]

Model	*Angle*	*Bond [Å]*
TIP3P, TIP3P	104.52	0.9572
SPC, SPC/E, ST2	109.47	1.0
BF	105.7	0.96

6.3 The liquid structure

The liquid structure in the form of radial distribution functions tells us about the potentials ability to describe the intermolecular interactions correctly. As the distribution function is normalized to an ideal gas, it will go towards one as the interaction energy between the molecules goes towards zero [1].

In the HH radial distribution function (from now on called RDF) in figure 5.6 there is a slight peak, and quit broad, around 2.5-4 Å for all the three simulations, followed by values around 1. If we compare the simulation results with the experimental RDF, the curves from the simulations starts to increase before the experimental curve. Furthermore, whereas the simulations results have one broad peak, the experimental curve has two distinct maxima's, both larger in magnitude.

[1] an ideal gas is characterized by having not interaction between the particles and neglectable particle volume [89]

The lack of structure in the HH RDF might be an effect of the volume or size of the atom being too small or large, insufficient modeling of the interactions between the hydrogen atoms or a combination of both these effects. As mentioned the HH RDF from the simulations starts to rise somewhat earlier than the experimental RDF (at around 1.3 Å compared to 1.9 Å). This indicates that the size of the hydrogen atom might be too small, and that this might be one of the reasons for the lack of structure. The size of an atom, given the functional form of the repulsion energy used this study (eq. 2.33), is decided by the size of the prefactor κ_{ij}. Reducing this atom pair property would reduce the size of the atom (with respect to the other atom) and increasing it would increase the size. This factor seems thus to be the culprit when it comes to the lack of structure in the HH RDF. This would also have the effect of reducing the intermolecular interactions between the atoms, which would consist of electrostatic, induction and dispersion forces in a range were the repulsion should be dominating. Two factors influence the size of κ_{ij}; the sample points and the Boltzmanns weight. Insufficient sampling of high hydrogen-hydrogen repulsion parts of phase space would lead to a set of parameters describing the low energy states well, but neglecting the higher energy states. Boltzmanns weighting of the error has the same effect, as described in chapter 2. In this study a Boltzmanns weighting at a temperature of 6000 K was used, reducing the error of the fit for a point having an energy of 20 kcal/mol to 20% of the original value. This is not a very harsh weighting, and as discussed in chapter 3 the difference between the weighted RMSE and the actual RMSE is quite low, indicating that the temperature has been chosen at an appropriate level, however a short study of the effect of this weighting would be preferable for future work. This leaves the points in phase space chosen to parameterized the repulsion energy. For the HH repulsion the high energy points are found for dimer calculations where the hydrogen atoms points towards each other. Of the 50 HFSCF calculations performed to model the repulsion energy, 16 of these had the orientation described above. This number should be sufficient, however the majority of these points are positioned at low energy areas, and an increased number of configurations (with focus on the high energy conformations) should be used in a future parameterization.

The broadness of the peak in the HH RDF suggest that the $\alpha_{i,j}$ parameter in eq. 2.33 is numerically too small. Another sign of this is the slope of the HH RDF which is not as steep as the experimental HH RDF in the region from 1-2.4 Å. A numerically too small $\alpha_{i,j}$ would give a sphere which is too soft compared to the real world, and this lead to a potential allowing for moves in a region where the repulsion energy would not be dominating, but rather act as a dampening of the attractive forces (the dispersion and induction energy). This is an effect similar to the one studied by Nymand et al. [90], where they modified the size of the dipole interactions resulting from the induction energy. When the induced dipole was removed completely they also lost most of the structure in the HH RDF. This is again an indication of the need for a new parameterization as described above. To sum up the structure in the HH RDF have disappeared as a consequence of a combination of a too soft potential and too small hydrogen atoms.

From the OH RDF in figure 5.7 it is clear that both of the flexible simulations have a larger and more obvious first peak than the rigid simulations. This peak is located around 2-2.5 Å. The second peak in this distribution, and indeed the rest of the distribution, are sheared by the three simulations and is located at approximately 3.2 Å. The first peak in the OH RDF occurs because of the hydrogen bond between the oxygen and hydrogen atom. The height of this peak has sometimes been used as a measure of the amount of hydrogen bonding in a force field, but there are some controversy around this [79], and this has therefore not been calculated in this study. The main contribution to the hydrogen bonding comes from the electrostatic energy [91], and the fact this peak is located at the same position, for the flexible molecules, does indeed give credit to the flexible charges and dipoles from LoProp. This might also be the reason for the failure of the rigid molecules to give a clear peak at this point. As the geometry chosen for the rigid molecules was the equilibrium geometry, the molecule might have an incorrect geometry with regards to the electrostatic interactions. The height of the first peak in the OH RDF are far from the height of the experimental value. Again this form of RDF is similar to the one found by Nymand et al. [90] when they removed the induced dipole moment completely. This suggest that there is too much dampening due to the repulsion energy. Another hint for this is that the OH RDF from the simulations starts to go upwards later than the experimental curve. This suggest that the repulsion between the oxygen and hydrogen atoms rises too early, giving too large atoms, killing of attractive interactions.

The OO RDF presented in figure 5.8 has a clear, high first peak located at 2.9 Å followed by a steep decent into a local minima at around 4 Å. This trend is sheared by all the three simulations. There seem to be another local maxima around 6 Å, however this peak is very broad and low. The first peak in the OO RDF is positioned almost at the same place as the experimental value and has approximately the same value for the flexible molecules, and somewhat lower value for the rigid. The second peak in OO RDF has, however, disappeared. In a study of the dependence of the OO RDF, Brdarski et al.[91] found that by modifying the potential slightly by making the size of the oxygen became larger, they where able to make the second peak in the OO RDF vanish completely. This also seem to be the effect we encounters, and this also account for the slower descend of the simulation curves compared to the experimental results. Again this is a question of parameterization.

If we compare the results from the simulation were the ZPV energy is included with the results from the simulation where it is excluded, the difference between the to scenarios are small. It can therefore be concluded that inclusion of the ZPV energy has little effect on the liquid structure.

When discussing the force fields ability to predict the water geometry it was found that the shift from gas to liquid phase gave a shift in the bond length which was about half of that measured from the experiments. It is clear that a modification of the repulsion energy as described above would increase the shift, as there attractive forces in the system would have a larger influence. This would of course also influence the vibrational spectra.

Simulations were done to confirm the trends described above. This was done by modifying the model of the repulsion energy. Such simulations would be inconsistent as they would not have correct dispersive interactions (as the parameterization of the dampening term depends on the repulsion energy), but they should be able to capture trends. Unfortunately these simulations failed due to polarization catastrophy.

6.4 Error analysis

The mean standard deviation of the bins for the various properties is presented in table 5.4 and 5.5. The mean standard deviation for the vibrational spectra, bond lengths and angle distributions in table 5.4 are low, indicating that a sufficient number of MC steps have been taken to calculate these properties accurately. The exception from this is the vibrational spectrum of the liquid when the ZPV energy is excluded from the Metropolis test. Here the error is still a good deal lower than the values in figure 5.2, but it would be preferable to lower this value even more by increasing the total number of MC steps calculating this property.

For the RDFs the errors are low for all the simulations, again indicating that an appropriate number of MC steps been taken. Despite this there is some parts of RDF which does not show a smoothness (see figures 5.6 to 5.8). This could probably be changed using curve smoothening as suggested by Allen and Tildesley [2].

Using mean standard deviation over all bins only gives a crude measurement of the general uncertainty of a simulation. In future work it might therefore be preferable to have a more detailed and thorough study of the uncertainty. This could include a study of the uncertainty as a function of the free variable and the behavior of the standard deviation as a function of the total number of MC steps.

6.5 Further work

In the discussion above we have seen that the well of the angle was not parameterized strong enough, and that a new parameterization of the intramolecular energy with inclusion of more data points with a larger angle would be preferable. To remove the hetroscedastisity described in chapter 3, writing the angle term as;

$$k\frac{(\cos\theta - \cos\theta_e)^2}{r_1 + r_2} \tag{6.2}$$

or

$$k\frac{(\cos\theta - \cos\theta_e)^2}{r_{1,3}} \tag{6.3}$$

which would provide a potential for the angle going towards zero for large separations between the hydrogen atoms.

As discussed above the repulsive part of the intermolecular potential are not satisfactory, and hence should be reparameterized. When doing this it would, as mentioned, be best to both include more points at the high repulsion parts of phase space and to study the effect of the Boltzmanns weight in more details.

In future work implementation of a dampening factor for the induction energy would be preferable. Such a dampening would prevent polarization catastrophy, and would thus enable the simulation and the parameterization (as the induction energy enters as a term when deciding the repulsion energy and dampening term of the dispersion energy) to avoid this. Such a dampening term might for instance be the Thole modified interaction tensor used by Brdarski et al. [92].

The basis set used in the quantum mechanical computations were chosen on the background of the basis set saturation of the atomic charges, and the diffuse function were used to give an improved description of the polarizability. This way of choosing a basis set might not be the most appropriate for a study of this nature, and it might be preferable to study the choice of basis set and as well as the choice of active space with regard to other properties.

The speed of a simulation is alway an interesting thing, and work should be done to improve the speed of the simulation. Here the most important point would probably be to improve the convergence of the iterative procedure used in the calculation of the induction energy.

Chapter 7

Conclusion

In this study a polarizable model for flexible water molecules was constructed based on quantum chemical computations. Atomic charges, dipole moments and polarizability tensors as described by the LoProp method were used to describe the electrostatic, induction and dispersion energy of the system. The charges, dipole moments and polarizability were all modeled as function of the internal coordinates of the water molecule. An intramolecular potential for the water molecule was also constructed using Simon-Parr-Finland, cosine, Urey-Bradley and cross terms. Both the atomic properties and the intramolecular energy were fitted to data from CASSCF level of theory using PLS regression.

Using data from dimer computation at the HFSCF and MP2 level of theory, the repulsion energy and a dampening term for the dispersion energy were fitted to functional forms. This was done using two simulated annealing algorithms and one genetic algorithm. In a short comparison of these, it was found that the best fit were found using a simulated annealing algorithm based on the Metropolis algorithm, however the genetic algorithm was found to be more robust than the other algorithms.

The force field constructed where implemented in the Protomol framework, and a Metropolis Monte Carlo simulation were run using rigid molecules and flexible molecules both with and without the ZPV energy. In these simulations it was found that the effect of the ZPV energy on the molecular geometry was to give larger bond lengths and and a larger angle. Inclusion of the ZPV energy also shifted the spectra towards lower frequencies. It was found that the angle bending term used in the intramolecular potential was too softly parameterized, and should be reparameterized. The shift of the bond length was found to be in agreement with experimental values, with the best agreement for the simulation where the ZPV energy was included. This agreement is expected to be even better with improvement of the intermolecular energy. The results found for the liquid structure were in reasonable agreement with the experimental results, however it was found that repulsion energy term should be improved. Improvement of this energy term can be done by including a dampening term in the induction energy and through an increased number of dimer calculations.

Bibliography

[1] Andrew R. Leach. *Molecular Modelling, principles and applications.* Prentice Hall, second edition, 2001.

[2] M.P. Allen and D.J. Tildesley. *Computer Simulation of Liquids.* Oxford University Press, 1987.

[3] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21(6):1087–1092, June 1953.

[4] C.H. Cho, S. Singh, and G.W. Robinson. Understanding all of water's anomalies with a nonlocal potential. *J. Chem. Phys.*, 107(18):7979–7987, 1997.

[5] T.A. Halgren and W. Damm. Polarizable force fields. *Current Opinion in Structural Biology*, 11:236–242, 2001.

[6] Michael D. Fayer. *Elements of Quantum Mechanics.* Oxford university press, first edition, 2001.

[7] Robert B. Leighton. *Principles of Modern Physics.* McGraw-Hill Book company inc, first edition, 1959.

[8] Alan Isaacs, editor. *A Dictionary of Physics.* Oxford University Press, 4th edition, 2003.

[9] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comp. Chem.*, 4(2):187–217, 1983.

[10] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Com.*, 92:1–41, 1995.

[11] W.L. Jorgensen, J.D. Chandrasekhar, R.W. Impey, and M.L. Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79:926–935, 1983.

[12] H.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, and J. Hermans. *Intermolecular Forces*, chapter Interaction Model for Water in Relation to Protein Hydration, pages 331–342. Dordrecht, Reidel, 1981. Pullman, B. (editor).

[13] F.H. Stillinger and T.A. Weber. Improved Simulation of Liquid Water by Molecular Dynamics. *J. Chem. Phys.*, 60:1545–1557, 1974.

[14] A. Wallqvist and O. Teleman. Properties of flexible water models. *Mol. Phys.*, 74:515–533, 1991.

[15] A. Wallquist. Incorporating intramolecular degrees of freedom in simulations of polarizable liquid water. *Chem. Phys.*, 148:439–449, 1990.

[16] H. Saint-Martin, J. Hernández-Cobos, I. Ortega-Blake, and H.J.C. Berendsen. A mobile charge densities in harmonic oscillators (MCDHO) molecular model for numerical simulations: The water-water interaction. *J. Chem. Phys.*, 113(24):10899–10911, 2000.

[17] J.M. Hermida-Ramón, S. Brdarski, G. Karlström, and U. Berg. Inter- and Intramolecular Potential for the N-Formylglycinamide-Water System. A Comparison between Theoretical and Empirical Force Fields. *J. Comp. Chem.*, 24:161–176, 2002.

[18] O. Engkvist, N. Forsberg, M. Schütz, and G. Karlström. A comparison between the NEMO intermolecular water potensial and *ab initio* quantum chemical calculations for the water trimer and tetramer. *Mol. Phys.*, 90:277–287, 1997.

[19] P.-O. Åstrand, A. Wallqvist, and G. Karlström. Nonempirical intermolecular potential for urea-water systems. *J. Chem. Phys.*, 100(2):1262–1273, 1994.

[20] P.-O. Åstrand, P. Linse, and G. Karlström. Molecular dynamics study of water adopting a potential function with explicit atomic dipole moments and anisotropic polarizabilities. *Chem. Phys.*, 191:195–202, 1995.

[21] S. Brdarski and G. Karlström. Modeling of the Exchange Repultion Energy. *J. Phys. Chem. A*, 102:8182–8192, 1998.

[22] J-C. Soetens and C. Millot. Effect of distributing multipoles and polarizabilities on molecular dynamics simulations of water. *Chem. Phys. Lett.*, 235:22–30, 1995.

[23] A.J. Stone. Distributed Multipole Analysis or how to describe a molecular charge distribution. *Chem. Phys. Lett.*, 83:233–239, 1981.

[24] Richard F.W. Bader. *Atoms in molecules : a quantum theory.* The International series of monographs on chemistry; 22. Clarendon Press, 1990.

[25] S.W. Rick, S.J. Stuart, J.S. Bader, and B.J. Berne. Fluctuating Charge Force Fields for Aqueous Solutions. *J. Mol. Liq.*, 65/66:31–40, 1995.

[26] H.A. Stern, F. Rittner, B.J. Berne, and R.A. Friesner. Combined fluctuating charge and polarizable dipole models: Application to a five-site water potential function. *J. Chem. Phys.*, 115(5):2237–2251, August 2001.

[27] H.A. Stern, G.A. Kaminski, J.L. Banks, R. Zhou, B.J. Berne, and R.A. Friesner. Fluctuating Charge, Polarizable Dipole, and Combined Models: Parameterization from ab Initio Quantum Chemistry. *J. Phys. Chem. B*, 103:4730–4737, 1999.

[28] P. Jedlovszky and J. Richardi. Comparison of different water models from ambient to supercritical condition: A Monte Carlo simulation and molecular Ornstein-Zernike study. *J. Chem. Phys.*, 110(16):8019–8031, 1999.

[29] P.W. Atkins and R.S. Friedman. *Molecular Quantum Mechanics.* Oxford University Press, 3rd edition, 1997.

[30] Guy H. Grant and W. Graham Richards. *Computational Chemistry.* Number 29 in Oxford chemistry primers. Oxford Science Publications, first edition, 1998.

[31] P.-O. Löwdin. A classic review on electron correlation. *Adv. Chem. Phys.*, 2:207, 1959.

[32] B.O. Roos, P.R. Taylor, and E.M. Siegbahm. A Complete Active Space SCF Method (CASSCF) Using Density Matrix Formulated Super-CI Approach. *Chem. Phys.*, 48:157–173, 1980.

[33] C. Møller and M.S. Plesset. Note on an Approximate Treatment for Many-Electron Systems. *Phys. Rev.*, 46:618–622, 1934.

[34] K. Ruud, P.-O. Åstrand, and P.R. Taylor. Vibrational Effects on Molecular Properties in Large Molecules. *J.Comp.Meth.Sci.Eng.*, 3(1):7–39, 2003.

[35] Roland Lindh. Localized molecular properties: the LoProp Approach. Seminar at Dep. of Theoretical Chemistry, Lund University, March 2004.

[36] K.B. Wiberg and P.R. Rablen. A Comparison of Atomic Charges Derived via Different Procedures. *J. Comp. Chem.*, 14(12):1504–1518, 1993.

[37] J. Cioslowski. General and Unique Partition of Molecular Electronic Properties into Atomic Contributions. *Phys. Rev. Lett.*, 62(13):1469–1471, 1989.

[38] U. Dinur and A.T. Hagler. Determination of atomic point charges and point dipoles from the Cartesian derivatives of molecular dipole moments and second order moments, and from energy second derivatives of planar dimers. I. Theory. *J. Chem. Phys.*, 91:2949–2958, 1989.

[39] J. Cioslowski. A New Population Analysis Based on Atomic Polar Tensors. *J. Am. Chem. Soc.*, 111:8333–8341, 1989.

[40] H. Solheim, K. Ruud, and P.-O. Åstrand. Atomic dipole moments calculated using analytical molecular secondr-moment gradients. *J. Chem. Phys.*, 120:10368–10378, 2004.

[41] L. Gagliardi, R. Lindh, and G. Karlström. Local properties of quantum chemical systems: the LoProp approach. unpublished.

[42] P. Comba and R. Remenyi. Inorganic and bioinorganic molecular mechanics modeling - the problem of force field parameterization. *Coord. Chem. Rec*, 238-239:9–20, 2003.

[43] J.C.A. Boeyens and P. Comba. Molecular mechanics: theoretical basis, rules, scope and limits. *Coord. Chem. Rec*, 212:3–10, 2001.

[44] Kim H. Esbensen. *Multivariate Data Analysis -in practice.* CAMO ASA/CAMO Process AS, 5th edition, 2001.

[45] Ronald E. Walpole, Raymond H. Myers, and Sharon L. Myers. *Probability and statistics for Engineers and Scientists.* Prentice Hall, 6th edition, 1998.

[46] William H. Press, Saul A. Teukolsky, William T. Vetterlig, and Brian P. Flannery. *Numerical Recipes in C.* Cambrigde University Press, 2nd edition, 1992. May be found on http://www.nr.com.

[47] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[48] J.M. Hayes and J.C. Greer. Extraction of analytical potential function parameters from ab inito potential energy surfaces and anaytical forces. *Comp. Phys. Com.*, 147:803–825, 2002.

[49] T.R. Cundari and W. Fu. Genetic algorithm optimization of a molecular mechanics force field for technetium. *Inorganica Chimica Acta*, 300-302:113–124, 1999.

[50] J. Hunger, S. Beyreuther, G. Huttner, K. Allinger, U. Radelof, and Z. Laszlo. How to Derive Force Field Parameters by Genetic Algorithms: Modelling *tripod*-$Mo(CO)_3$ Compunds as an Example. *Eur. J. Inorg. Chem.*, pages 693–702, 1998.

[51] P.-O. Norrby and T. Liljefors. Automated Molecular Mechanics Parameterization with Simultaneous Utilization of Experimental and Quantum Mechanical Data. *J. Comp. Chem.*, 19(10):1146–1166, 1998.

[52] K.-H. Cho, K.T. No, and H.A. Scheraga. A polarizable force field for water using an artificial neural network. *J. Mol. Struct.*, 641:77–91, 2002.

[53] S. Lifson and A. Warshel. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *J. Chem. Phys.*, 49(11):5116–5129, 1968.

[54] P.M. Morse. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys. Rev.*, 34:57–64, 1929.

[55] G. Simons, R.G. Parr, and J.M. Finlan. New alternative to the Dunham potential for diatomic molecules. *J. Chem. Phys.*, 59(6):3229–3224, 1973.

[56] P.F. Fougere and R.K. Nesbet. Eletronic Structure of C_2. *J. Chem. Phys.*, 44:285, 1966.

[57] A.J. Thakkard. A new generalized expansion for the potential energy curves of diatomic molecules. *J. Chem. Phys.*, 62(5):1693–1701, 1975.

[58] Stevica Brdarski. *Modeling of intra- and intermolecular potentials.* PhD thesis, Department of Theoretical Chemistry, Lund University, 1999.

[59] Per-Olof Åstrand. Point polarizabilities in an external field. Lecture notes, March 2004.

[60] S.F. Boys and F. Bernardi. The Calculation of Small Molecular Interaction by the Difference of Separate Total Energies. Some Procedures with Reduced Error. *Mol. Phys.*, 19:553–566, 1970.

[61] E.B. jr Wilson, J.C. Decius, and P.C. Cross. *Molecular Vibrations, The Theory of Infrared and Raman Vibrational Spectra.* McGraw-Hill Book Company, 1955.

[62] F.A. Cotton. *Chemical Applications of Group Theory.* Wiley-interscience, 2nd edition, 1971.

[63] Roland Kjellander. *The basis of statistical thermodynamics or My favorite path to termodynamics and beyond.* Dept. of Physical Chemistry, University of Göteborg, Sweden, 1991.

[64] John M. Seddon and Julian D. Gale. *Thermodynamics and Statistical Mechanics.* The Royal Society of Chemistry, 2001.

[65] P.W. Atkins. *Physical Chemistry.* Oxford University Press, sixth edition, 1999.

[66] Raymond Chang. *Physical Chemistry for the Chemical and Biological Sciences.* University science books, third edition, 2000.

[67] Terrell L. Hill. *An Introduction to Statistical Thermodynamics.* Dover Publications, Inc., 1986.

[68] Asbjørn Holt. A new adaptive umbrella potential sampling for free energy calculations. Technical report, Dep. of Chemistry, NTNU, Januar 2003. Report from Physical Chemistry, Specialization (TKJ4700).

[69] J.P. Valleau. Monte Carlo: Choosing which game to play. In M. Meyer and P. Vassilis, editors, *Computer Simulation in Material Science*, volume 205 of *NATO ASI Series, Series E: Applied Science*, pages 67–84. NATO, Kluwer Academic Publishers, 1991.

[70] M.H. Kalos and P.A. Whitlock. *Monte Carlo Methods, Volum I: Basics.* John Wiley & Sons, Inc., 1986.

[71] G. Karlström, R. Lindh, P.A. Malmqvist, B.O. Roos, U. Ryde, V. Veryazov, P-O. Widmark, M. Cossi, B. Schimmelpfennig, P. Neogrady, and L. Seijo. Molcas: a program package for computational chemistry. *Computational Materials Science*, 28(2):222–239, 2003.

[72] T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90:1007–1023, 1988.

[73] Inc. Minitab. Minitab statistical software, release 14 for windows, 2003. State College, Pennsylvania. Minitab(r) is a registered trademark of Minitab Inc.

[74] Per Kristian Holt. Implementasjon og verifisering av en genetisk algoritme for å identifisere innstrømning i brønnmodell. Master's thesis, Høgskolen i Telemark, 2004.

[75] C.H. Edwards Jr. and D.E. Penny. *Elementary Linear Algebra.* Prentice Hall, 1988.

[76] T. Matthey and J.A. Izaguirre. Protomol: A molecular dynamics framework with incremental parallelization. SIAM Conference on Parallel Processing for Scientific Computing, 2001.

[77] M. Watanabe and M. Karplus. Dynamics of molecules with internal degrees of freedom by multiple time-step methods. *J. Chem. Phys.*, 99(10):8063–8074, 1993.

[78] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics . *J. Molec. Graphics*, 14.1:33–38, 1996.

[79] A.K. Soper, F. Bruni, and M.A. Ricci. Site-site correlation functions of water from 25-400^{o}C: Revised analysis of new and old diffraction data. *J. Chem. Phys.*, 106(1):247–254, 1997.

[80] B. Buttingsrud. Deconvolution in analytical chemistry. Master's thesis, NTNU, 2003.

[81] P-O. Åstrand, K. Ruud, and P.R. Taylor. Calculation of the vibrational wavefunction of polyatomic molecules. *J. Chem. Phys.*, 112(6):2655–2667, 2000.

[82] T. Nymand and P.-O. Åstrand. Calculation of the Geometry of the Water Molecule in Liquid Water. *J. Phys. Chem. A*, 101:10039–10044, 1997.

[83] T.A. Ford and M. Falk. *Can. J. Chem.*, 46:3579–,286, 1968.

[84] B. Hess, S.-M. Humberto, and H.J.C. Berendsen. Flexible constraints: An adiabatic treatment of quantum degrees of freedom, with applications to the flexible and polarizable mobile charge densities in harmonic oscillators model for water. *J. Chem. Phys.*, 116(22):9602–9610, 2002.

[85] N.W. Moriarty and G. Karlström. Geometry optimization of a water molecule in water. A combined quantum chemical and statistical mechanics treatment. *J. Chem. Phys.*, 106(15):6470–6474, 1997.

[86] S.-M. Humberto, B. Hess, and H.J.C. Berendsen. An application of flexible constraints in Monte Carlo simulations of the isobaric-isothermal ensemble of liquid water and ice lh with the polarizable and flexible mobile charge densities in harmonic oscillators model. *J. Chem. Phys.*, 120(23):11133–11143, 2004.

[87] K. Ichikawa, Y. Kameda, T. Yamaguchi, and H. Wakita. Neutron-diffraction investigation of the intramolecular structure of a water molecule in liquid phase at high temperature. *Mol. Phys.*, 73:79, 1991.

[88] W.S. Benedict, N. Gailar, and E.K. Plyler. *J. Chem. Phys.*, 24:1139, 1965.

[89] Alan Isaacs, John Daintith, and Elizabeth Martin, editors. *A Dictionary of Science.* Oxford University Press, 4th edition, 2003.

[90] T.M. Nymand, P. Linse, and P.-O. Åstrand. A comparison of effective and polarizable intermolecular potentials in simulations: liquid water as a test case. *Mol. Phys.*, 99(4):335–348, 2001.

[91] S. Brdarski, P.-O. Åstrand, and G.W Karlström. Dependence of the repulsive part of the intermolecular potential on the structure and diffusion of liquid water. Paper IV in PhD thesis of Brdarski [58], 1999.

[92] S. Brdarski, P.-O. Åstrand, and G. Karlström. The inclusion of electron correlation in intermolecular potentials: application to the formamide dimer and liquid formamide. *Theor. Chem. Acc.*, 105:7–14, 2000.

[93] T. Helgaker, H. J. Aa. Jensen, P. Joergensen, J. Olsen, H. Aagren K. Ruud, A.A. Auer, K.L. Bak, V. Bakken, O. Christiansen, S. Coriani, P. Dahle, E. K. Dalskov, T. Enevoldsen, B. Fernandez, C. Haettig, K. Hald, A. Halkier, H. Heiberg, H. Hettema, D. Jonsson, S. Kirpekar, R. Kobayashi, H. Koch, K. V. Mikkelsen, P. Norman, M. J. Packer, T. B. Pedersen, T. A. Ruden, A. Sanchez, S. P. A. Sauer T. Saue, B. Schimmelpfennig, K. O. Sylvester-Hvid, P. R. Taylor, and O. Vahtras. Dalton, a molecular electronic structure program, Release 1.2 . 2001.

Appendix A

Further details to theory

A.1 Second derivatives of the intramolecular potential

The second derivatives of the intramolecular energy (eq. 2.22) is;

$$
\begin{aligned}
\frac{\partial V_{intra}}{\partial r_1 \partial r_1} &= B_0 \left[-2(2+3b_1)\frac{r_e}{r_1^3} + 6(1+3b_1)\frac{r_e^2}{r_1^4} - 12b_1 \frac{r_e^3}{r_1^5} \right] \\
&+ C_1 \left[-4\frac{r_e}{r_1^3} + 6\frac{r_e^2}{r_1^4} + 8\frac{r_e^2}{r_2 r_1^3} - 4\frac{r_e^3}{r_2^2 r_1^3} - 12\frac{r_e^3}{r_2 r_1^4} + 6\frac{r_e^4}{r_1^4 r_2^2} \right] \\
&+ 2k_\theta \left(\frac{r_3^2}{r_1^3 r_2} - \frac{r_1}{r_1^3} \right) \\
&+ k_{UB}\frac{(2r_1 - 2r_2\cos\theta)^2}{2r_3^4} \\
&- 5k_{UB}\frac{(r_3 - r_{UB})(2r_1 - 2r_2\cos\theta)^2}{2r_3^5} \\
&+ 2k_{UB}\frac{r_3 - r_{UB}}{r_3^3} \\
&+ 2k_{UB}\frac{(r_3 - r_{UB})^2(r_1 - 2r_2\cos\theta)^2}{r_3^6} \\
&+ 2k_{UB}\frac{(r_3 - r_{UB})}{r_3^4}
\end{aligned}
\tag{A.1}
$$

$$\begin{aligned}
\frac{\partial V_{intra}}{\partial r_1 \partial r_2} &= 4C_1 \left[\frac{r_e^2}{r_1^2 r_2^2} - \frac{r_e^3}{r_1^2 r_2^3} - \frac{r_e^3}{r_1^3 r_2^2} + \frac{r_e^4}{r_1^3 r_2^3} \right] \\
&+ k_\theta \left(\frac{r_3^2}{r_1^2 r_2^2} + \frac{1}{r_1^2} + \frac{1}{r_2^2} \right) \\
&+ k_{UB} \frac{(2r_1 - 2_2 \cos\theta)(2r_1 - 2r_2 \cos\theta)}{2r_3^4} \\
&- 5k_{UB} \frac{(r_3 - r_{UB})(2r_1 - 2r_2 \cos\theta)(2r_2 - 2_1 \cos\theta)}{2r_3^5} \\
&- 2k_{UB} \frac{(r_3 - r_{UB}) \cos\theta}{r_3^3} \\
&+ 2k_{UB} \frac{(r_3 - r_{UB})^2 (2r_1 - 2r_2 \cos\theta)(2r_2 - 2r_1 \cos\theta)}{r_3^6} \\
&+ 2k_{UB} \frac{(r_3 - r_{UB})^2 \cos\theta}{r_3^4}
\end{aligned} \tag{A.2}$$

$$\begin{aligned}
\frac{\partial V_{intra}}{\partial r_1 \partial \theta} &= k_{UB} \frac{r_1 r_2 \sin\theta (2r_1 - 2r_2 \cos\theta)}{r_3^4} \\
&- 5k_{UB} \frac{(r_3 - r_{UB})(2r_1 - 2r_2 \cos\theta) r_1 r_2 \sin\theta}{r_3^5} \\
&- 2k_{UB} \frac{(r_3 - r_{UB}) r_2 \sin\theta}{r_3^3} \\
&+ 4k_{UB} \frac{(r_3 - r_{UB})^2 (2r_1 - 2r_2 \cos\theta) r_1 r_2 \sin\theta}{r_3^6} \\
&- 2k_{UB} \frac{(r_3 - r_{UB}) r_1 \sin\theta}{r_3^4}
\end{aligned} \tag{A.3}$$

$$\begin{aligned}
\frac{\partial V_{intra}}{\partial \theta \partial \theta} &= 2k_{UB} \frac{r_1^2 r_2^2 \sin^2\theta}{2r_3^4} \\
&- 10k_{UB} \frac{(r_3 - r_{UB}) r_1^2 r_2^2 \sin^2\theta}{r_3^5} \\
&+ 2k_{UB} \frac{(r_3 - r_{UB}) r_1 r_2 \cos\theta}{r_3^3} \\
&+ 8k_{UB} \frac{(r_3 - r_{UB}) r_1^2 r_2^2 \sin^2\theta}{r_3^6} \\
&- 2k_{UB} \frac{(r_3 - r_{UB})^2}{r_3^4} \\
&+ 2k_\theta \sin^2\theta - 2k_\theta (\cos\theta - \cos\theta_e) \cos\theta
\end{aligned} \tag{A.4}$$

A.2 Change of basis

A.2.1 Vector

The derivation below follows the one given in Edwards and Penny [75].

Any vector $\vec{v}$ may be expressed as a sum of basis vectors such that;

$$\vec{v} = x_1\vec{e}_1 + x_2\vec{e}_2 + \cdots + x_n\vec{e}_n \tag{A.5}$$

in the vector space $\mathbf{V}^n$. Here x_1, x_2 etc is called the coordinates of the vector $\vec{v}$, and

$$\vec{v}_e = (x_1, x_2, \ldots, x_n) \tag{A.6}$$

is called the coordinates vector. Given $\vec{v}_e$ in a Euclidean vector space $\mathbf{e}^e$, the transformation of $\vec{v}_e$ into a new Euclidean vector space $\mathbf{e}^m$ is given by;

$$\vec{v}_m = \mathbf{M}_e^{-1}\mathbf{M}_m\vec{v}_e \tag{A.7}$$

where $\vec{v}_m$ is the coordinate vector in the new basis. Here $\mathbf{M}_e$ and $\mathbf{M}_m$ are matrices with the basis vectors as columns. Given a orthonormal basis the expression above can be written as;

$$\vec{v}_m = \mathbf{M}_e^T\mathbf{M}_m\vec{v}_e = \mathbf{Q}\vec{v}_e \tag{A.8}$$

where;

$$\mathbf{Q} = \begin{bmatrix} \vec{m}_1\cdot\vec{e}_1 & \vec{m}_1\cdot\vec{e}_2 & \vec{m}_1\cdot\vec{e}_3 \\ \vec{m}_2\cdot\vec{e}_1 & \vec{m}_2\cdot\vec{e}_2 & \vec{m}_2\cdot\vec{e}_3 \\ \vec{m}_3\cdot\vec{e}_1 & \vec{m}_3\cdot\vec{e}_2 & \vec{m}_3\cdot\vec{e}_3 \end{bmatrix} \tag{A.9}$$

for the three dimensional case.

A.2.2 Tensor

The change from a cartesian basis $\{\vec{e}_1, \vec{e}_2, \vec{e}_3\}$ to a new cartesian basis $\{\vec{m}_1, \vec{m}_2, \vec{m}_3\}$, for a tensor $\mathbf{S}$ is given by the following equation;

$$\mathbf{S}^{(m)} = \mathbf{Q}\mathbf{S}^{(e)}\mathbf{Q}^T \tag{A.10}$$

where $\mathbf{Q}$ is;

$$\mathbf{Q} = \begin{bmatrix} \vec{m}_1\cdot\vec{e}_1 & \vec{m}_1\cdot\vec{e}_2 & \vec{m}_1\cdot\vec{e}_3 \\ \vec{m}_2\cdot\vec{e}_1 & \vec{m}_2\cdot\vec{e}_2 & \vec{m}_2\cdot\vec{e}_3 \\ \vec{m}_3\cdot\vec{e}_1 & \vec{m}_3\cdot\vec{e}_2 & \vec{m}_3\cdot\vec{e}_3 \end{bmatrix} \tag{A.11}$$

Proof; let $\vec{u}$ and $\vec{v}$ be vectors satisfying;

$$\vec{v} = \mathbf{S}\vec{u} \tag{A.12}$$

denote the components of $\vec{u}$ and $\vec{v}$ in the two basis as $\vec{u}^{(e)}$, $\vec{u}^{(m)}$, $\vec{v}^{(e)}$ and $\vec{v}^{(m)}$. We know that these are related through the following equations;

$$\begin{array}{ll} \vec{u}^{(m)} = \mathbf{Q}\vec{u}^{(e)} & \vec{u}^{(e)} = \mathbf{Q}^T\vec{u}^{(m)} \\ \vec{v}^{(m)} = \mathbf{Q}\vec{v}^{(e)} & \vec{v}^{(e)} = \mathbf{Q}^T\vec{v}^{(m)} \end{array} \tag{A.13}$$

the tensor-vector product now becomes;

$$\vec{v}^{(m)} = \mathbf{S}^{(m)}\vec{u}^{(m)} \tag{A.14}$$

$$\vec{v}^{(e)} = \mathbf{S}^{(e)}\vec{u}^{(e)} \tag{A.15}$$

substituting for $\vec{u}^{(e)}$ and $\vec{v}^{(e)}$ in A.15 from A.13, we get;

$$\mathbf{Q}^T\vec{v}^{(m)} = \mathbf{S}^{(e)}\mathbf{Q}^T\vec{u}^{(m)} \tag{A.16}$$

right multiplying the equation above with $\mathbf{Q}$ and comparing with eq. A.14 we get;

$$\mathbf{S}^{(m)} = \mathbf{Q}\mathbf{S}^{(e)}\mathbf{Q}^T \tag{A.17}$$

A.3 A geometrical approach to the perturbation of an angle

To perturb the angle in a water molecule, some geometrical considerations must be made. In general one has, given a vector of length $|l|$, an infinite number of possible solutions lying on the surface of a sphere as illustrated in figure A.1(a). For water one might impose such change of the angle to the molecular plane. This is done by demanding the following equation to be fulfilled;

$$\vec{N} \cdot \vec{l} = 0 \tag{A.18}$$

where $\vec{N}$ is the normal vector to the plane. This narrows our choice to all possible choices on a circle with radius $|l|$. When perturbing an angle we know that the difference between the new angle and the old one is given by;

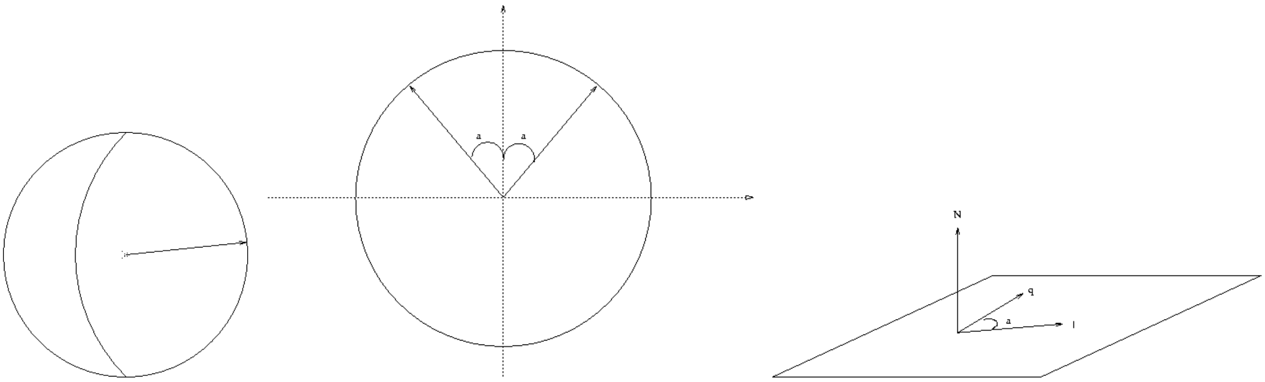
$$a = \theta_{new} - \theta_{old} \tag{A.19}$$

giving us the following equation;

$$\vec{l}^{new} \cdot \vec{l}^{old} = |\vec{l}^{new}||\vec{l}^{old}| \cos a \tag{A.20}$$

and the two possible solutions illustrated in figure A.1(b).

It is obvious that only one of these two solutions is correct (the other having the opposite effect on the total angle), and hence to solve this problem we need to know another vector in the plane such that one solution can be found (as illustrated in figure A.1(c)). This



(a) All possible solutions known only the length

(b) The two possible solutions knowing the length, plane and angle a

(c) The one possible solutions knowing the length, plane, angle a and another vector q

Figure A.1: Illustrations of the possible choices of a vector of given length

might be done if one introduce a vector $\vec{q}$ such as illustrated in figure A.2. This gives us the following equation ($\vec{q}$ might be found from the previous values);

$$\vec{l}^{new} \cdot \vec{q} = |\vec{l}^{new}||\vec{q}| \cos \frac{\theta_{new}}{2} \tag{A.21}$$

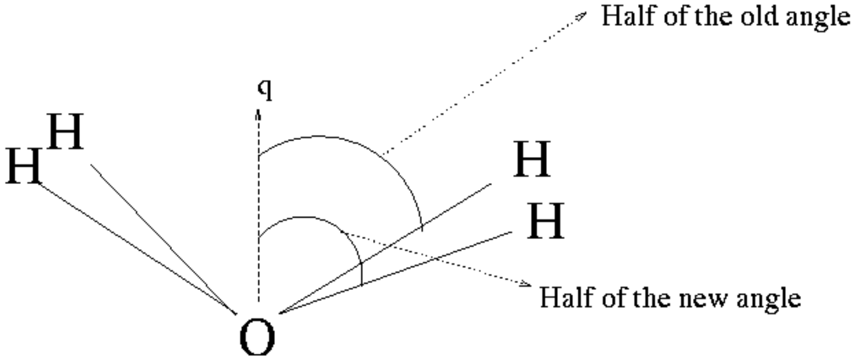


Figure A.2: A possible choice of a vector in the plane

As the scalar product is in general;

$$\vec{n} \cdot \vec{m} = n_x m_x + n_y m_y + n_z m_z \tag{A.22}$$

equation A.18, A.20 and A.21 gives us a linear set of equation for the new vector;

$$\begin{array}{ccccccc} n_x l_x & + & n_y l_y & + & n_z l_z & = & 0 \\ l_x^{old} l_x & + & l_y^{old} l_y & + & l_z^{old} l_z & = & |\vec{l}^{new}||\vec{l}^{old}| \cos a \\ q_x l_x & + & q_y l y & + & q_z l_z & = & |\vec{l}^{new}||\vec{q}| \cos \frac{\theta_{new}}{2} \end{array}$$

and we can get the new vector $\vec{l}$ through the following matrix equation;

$$\vec{l} = \mathbf{A}^{-1} \vec{y} \tag{A.23}$$

Appendix B

Further details to results

B.1 Results from modeling the atomic dipole and polarizabilities

In this section the regressional results from the modeling of the atomic dipole moment and the atomic polarizability will be presented. As this corresponds to 12 different curve fittings, only a some of the results are presented here, as presenting them all would be too space consuming. The following results have therefore been chosen for this discussion;

- Score and Loadings plot
- Standardized coefficients plot
- Residual normal probability plot
- Response plot

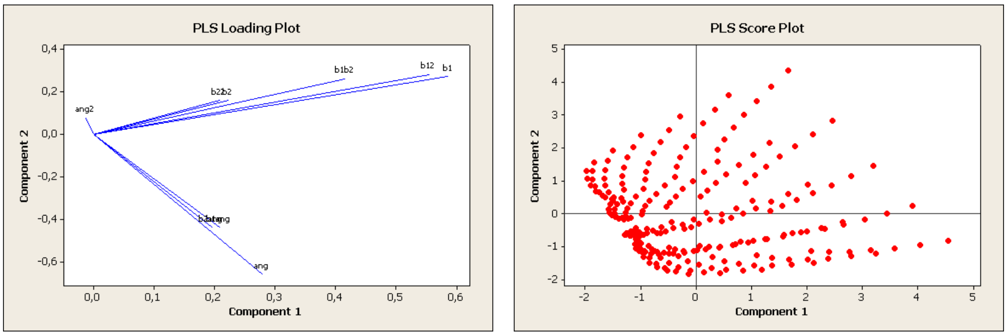
The reason for this choice is as follows; we want to see which variables and samples that influence the given property most (hence the score and loadings plot), we want to find out which variable that is the most influential on the model (therefore the standardized coefficients plot), the distribution of the residuals must be tested to verify the validity of the model (therefore the residual normal probability plot) and we want to see how good the fit to the model is (which is shown in the response plot).

B.1.1 Atomic dipole moments

The atomic dipole moments were fitted to eq. 3.9 for both the oxygen and hydrogen atom.

Hydrogen atom

The loadings and score plot from the PLS regression on the x- and y-components of the atomic dipole moment of hydrogen are presented in figures B.1, the regressional results are presented in figures B.2 and B.3 for the x and y component respectively. These results are discussed below.



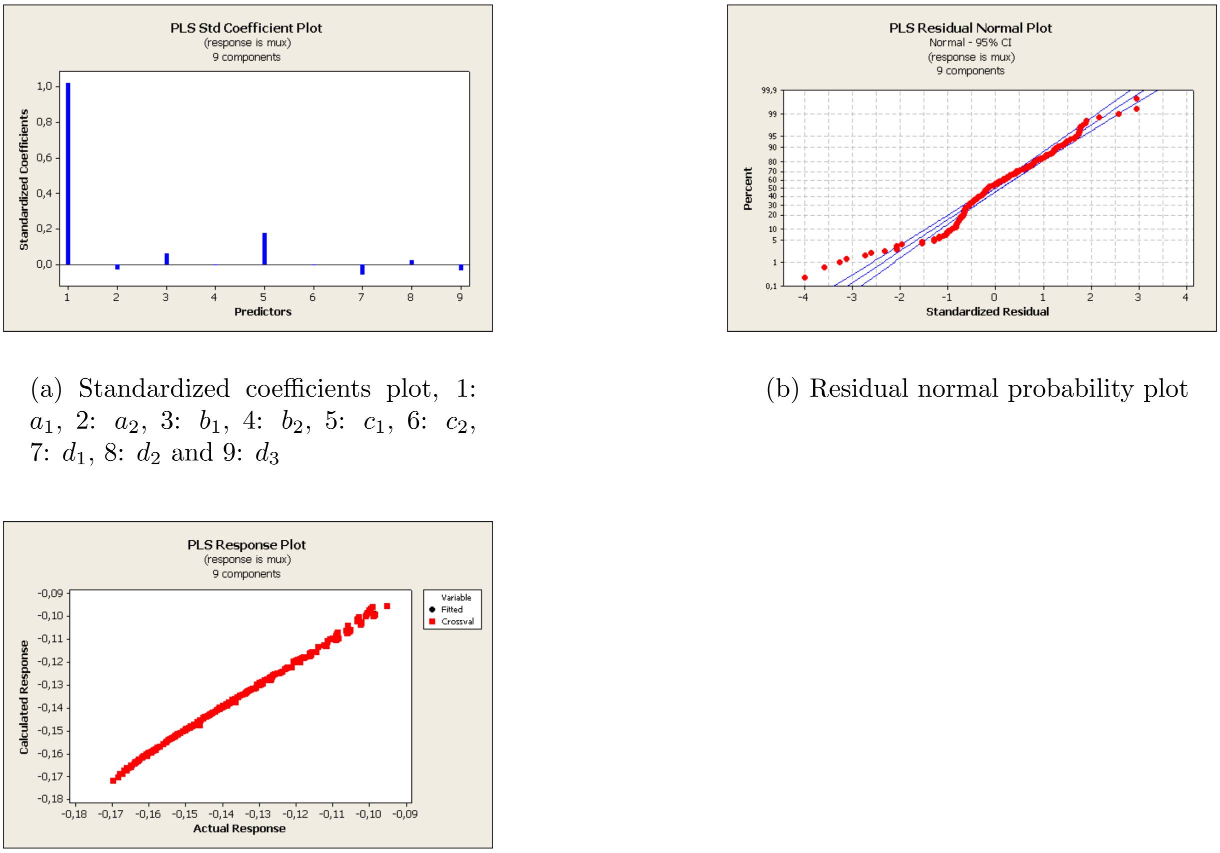
(a) Loadings plot

(b) Score plot, high number indicates long bond length

Figure B.1: Score and Loadings plot for the PLS regression on the atomic dipole moment of hydrogen

The score and loadings plot are presented in figure B.1, and from the loadings plot (figure B.1(a)) it is interesting to see that both the bond lengths and the angle are spanning out both the first and the second PLS component. With this said, it is clear that the bond lengths are more influential on the first component, whereas the angle (and cross terms between the angle and the bond lengths) is most influential on the second component. Let us also note that there is a pattern in the score plot (figure B.1(b)) which to a large degree divides the data into the various data series, but that the major amount of points are gathered to the left of the plot. This is also the area where even the square of the angle has some influence.

In figure B.2 and B.3 the results of the model for the x- and y-component of the dipole moment (please remember that this model of the dipole of a hydrogen in the xy-plane with lies along the x-axis) are presented. The results for these to models will be discussed separately, however they are both connected to the loadings and score plot above.



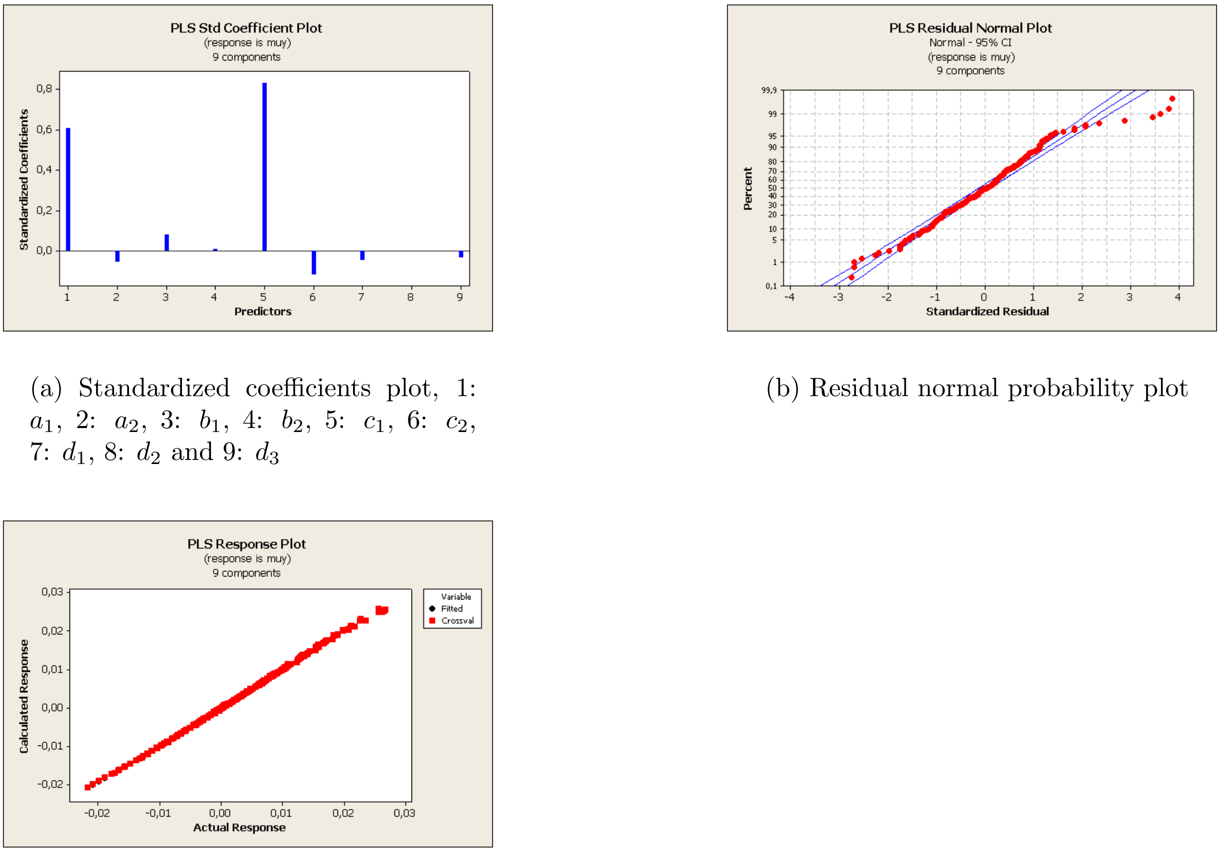
(a) Standardized coefficients plot, 1: a_1, 2: a_2, 3: b_1, 4: b_2, 5: c_1, 6: c_2, 7: d_1, 8: d_2 and 9: d_3

(b) Residual normal probability plot

(c) Response plot, R^2=0.9986

Figure B.2: Results from PLS regression on the x-component of the dipole of hydrogen

The standardized coefficience plot for the model of the x-component of the dipole moment of hydrogen in figure B.2(a) shows that the most important variable for determining the size of this is the bond length, r_1. This is also directly the position of the atom along the x-axis, so it is not surprising that this should be the most influential contribution, but note that it shows a linear dependence along this axis, as the quadric contribution is almost neglectable. This is followed by the angle of the molecule, which also shows a linear dependence. The other contributions seems to be neglectable or to give very small corrections to the model. The residual normal probability plot in figure B.2(b) shows a small tail at the lower end of the plot, suggesting that most, but not all, residuals follow a normal distribution. According to the manual of Minitab [73], such deviation indicated some skewness (deviation from a symmetrical distribution around the mean), and moderate departures from normality can usually be ignored given a sufficiently large data set. The response plot shows that the model gives a very good fit to the data set (a very large R^2 value), although there seems to be some deviation for the largest values of the x-component. The properties described above indicate that a simpler model of the dipole moment (at least the x-component) might be more suitable. However a model selection study showed that the optimum number of components corresponds to 9, which also is the number of variables introduced in this study, and it was therefore decided to use the model described above.



(a) Standardized coefficients plot, 1: a_1, 2: a_2, 3: b_1, 4: b_2, 5: c_1, 6: c_2, 7: d_1, 8: d_2 and 9: d_3

(b) Residual normal probability plot

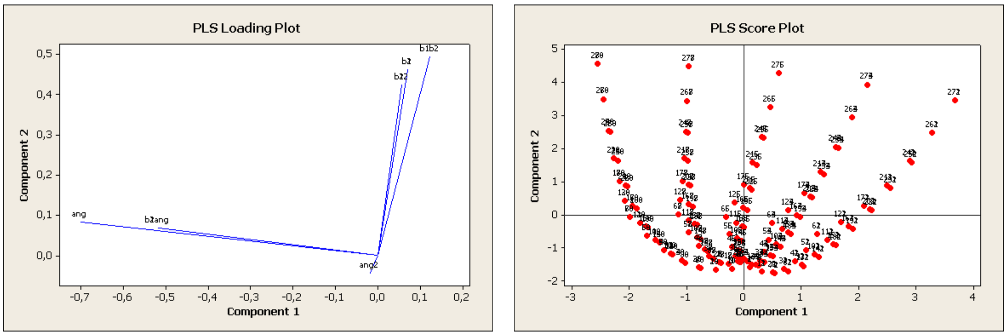
(c) Response plot, R^2=0.9992

Figure B.3: Results from PLS regression on the y-component of the dipole of hydrogen

For the y-component we see in figure B.3(a) that the most important coefficient is the linear angle term. This is followed by the linear term of the bond length. As opposed to the results for the x-component, the quadric term of the angle is of some importance, suggesting some curvature in the y-component with respect to this variable. The residual normal probability plot in figure B.3(b) suggest that the residuals are following a normal distribution and the response plot in figure B.3(c) shows that the model does give a very good fit to the data set.

Oxygen atom

The results from the regressional model of the dipole of oxygen are presented and discussed below. In figure B.4 the score and loadings plot are presented, and in figure B.5 and B.6 the results x and y component are presented.



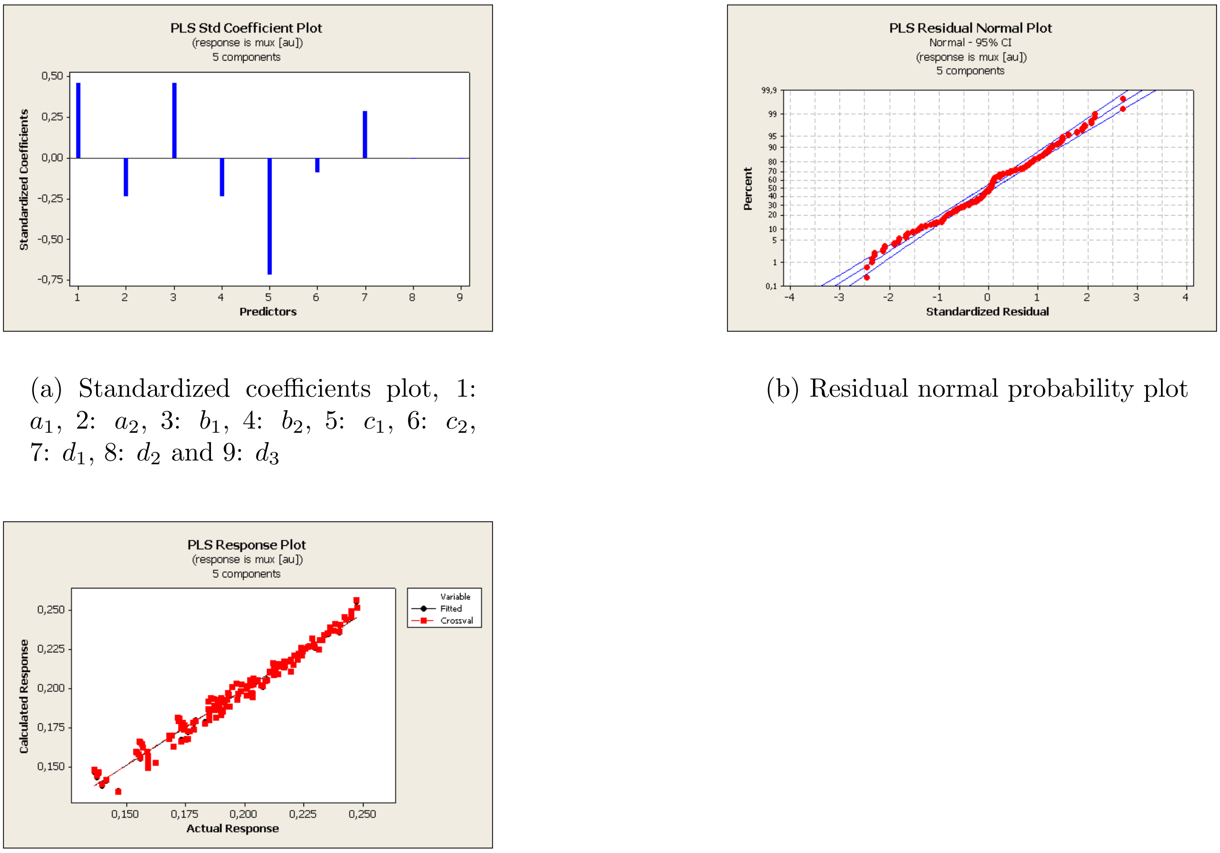
(a) Loadings plot

(b) Score plot, high number indicates long bond length

Figure B.4: Score and Loadings plot for the PLS regression on the atomic dipole moment of oxygen

The importance of the angle in describing the variance in the data set when modeling the dipole for the oxygen atoms seems to be very large when looking at the loadings plot in figure B.4(a). The angle is (together with the cross terms involving the angle) spanning out the first PLS component, and in the score plot in figure B.4(b) shows the five different angles used in the ab initio calculations appears in a very obvious pattern. The various bond terms (please remember that there is no difference in importance between the two bonds in the water molecule when regarding the whole dipole moment of oxygen) spans out the second PLS component giving the large number of points in the bottom of the score plot.

The results for the models of the x- and y-component of the dipole moment of oxygen are presented in figures B.5 and B.6. As for the hydrogen atom, these will be discussed separately.



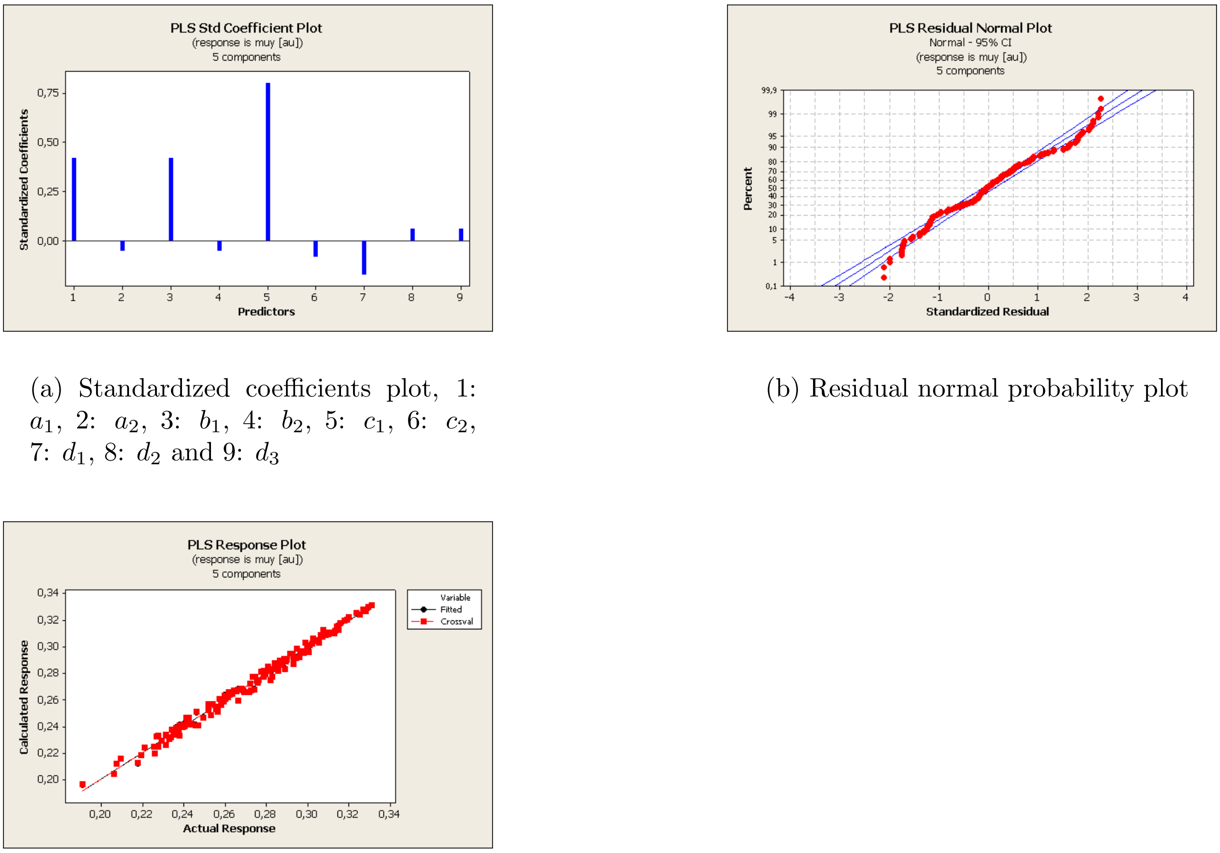
(a) Standardized coefficients plot, 1: a_1, 2: a_2, 3: b_1, 4: b_2, 5: c_1, 6: c_2, 7: d_1, 8: d_2 and 9: d_3

(b) Residual normal probability plot

(c) Response plot, R^2=0.9731

Figure B.5: Results from PLS regression on the x-component of the dipole of oxygen

The standardized coefficients plot in figure B.5(a) tells us that the x-component of the dipole moment for oxygen is more difficult to describe by the use of geometry than that of hydrogen. This is not surprising as the electron distribution around the oxygen atom is significantly more complex than that of the hydrogen, thus introducing several more degrees of freedom for the local property. The most important contribution seems to come from the linear terms of the bonds and the angle, however there also seems to be significant contributions from the quadric terms in the model, and from the cross term between the bonds. The only terms which seems to be insignificant in the model is the cross terms between the angle and the bond lengths. Turning the attention back to the loadings plot, these two lies along the position of the angle, and this might explain their redundancy, as the angle is the variable describing the variance in this direction. Turning the attention to the residual normal probability plot (figure B.5(b)) all the residuals follows a straight line, and is therefore normal distributed. In the response plot presented in figure B.5(c) the plot lies along the straight line, however there seems to be some hetroscedastic noise in the system, manifesting itself by giving systematically larger deviation for small values of the component. This noise starts for values somewhat distant from the equilibrium geometry, and might arise from the models inability to describe such situations. The trend is not very large, and as the R^2 value is relatively large, the model was used.



(a) Standardized coefficients plot, 1: a_1, 2: a_2, 3: b_1, 4: b_2, 5: c_1, 6: c_2, 7: d_1, 8: d_2 and 9: d_3

(b) Residual normal probability plot

(c) Response plot, R^2=0.9909

Figure B.6: Results from PLS regression on the y-component of the dipole of oxygen

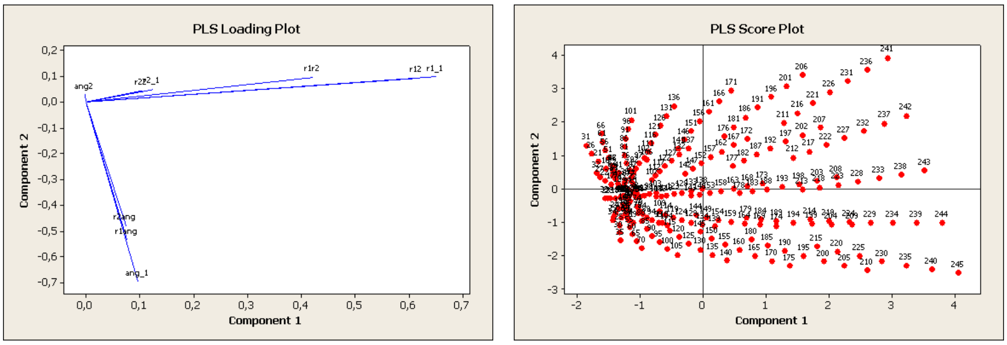
The same variables which were important for the x-component of the dipole moment, namely the linear terms of the two bond lengths and the angle, are also the most important for the y-component. The quadric terms however, seems to be much less important for the y-component, and together with the cross terms they're only giving minor corrections to the model (see the standardized coefficient plot in figure B.6(a)). A slightly hetroscedastic tendency is observed in the response plot in figure B.6(c) as it was for the x-component, however the judging by the R^2 value the fit is still very good. In the residual normal probability plot (figure B.6(b)), there is only three points which lies outside the 95% confidence interval line, and the residual must therefore be said to follow a normal distribution.

B.1.2 Atomic polarizabilities

The xx-, yy-, xy- and zz-components of atomic polarizabilities were all fitted to eq. 3.10 for both the oxygen and hydrogen atom.

Hydrogen atom

The score and loadings plot from the PLS regression on the atomic polarizability of hydrogen are presented in figure B.7, and the results for the various components are presented in figure B.8 to B.11.

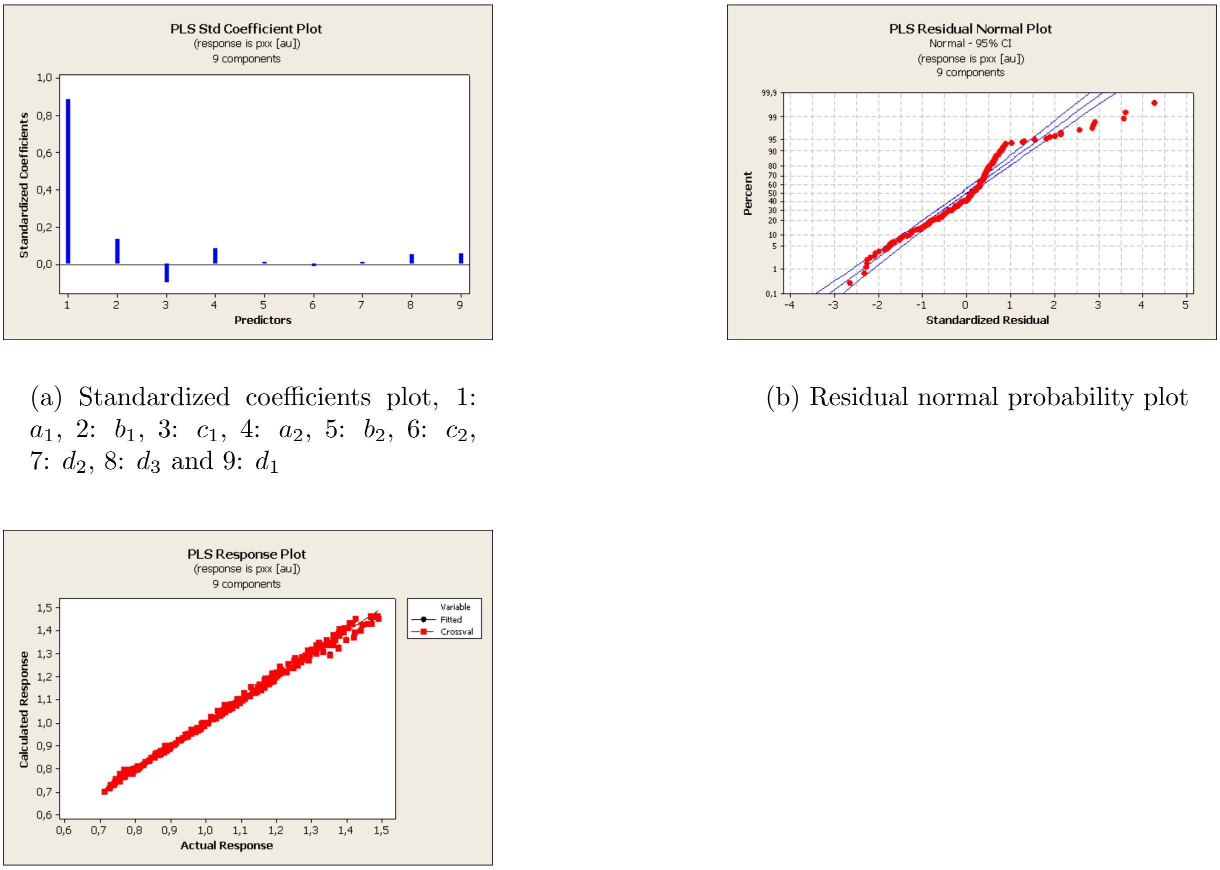


(a) Loadings plot

(b) Score plot, high number indicates long bond length

Figure B.7: Loadings and score plot for the atomic polarizability of hydrogen

In the loadings plot (figure B.7(a)) the first PLS component is span out by the bond length between the H-atom and the O-atom. This is not surprising as the charges in the LoProp scheme are suppose to go to the bonds in the system, when perturbed by an external field [35, 41]. The second component is span out by the angle, and as we shall see later, this is due to the off-diagonal components of the polarizability. If we look at the score plot in figure B.7(b) the major part of the variance come from the parts of the system where the bond lengths and angles are relatively large, and that the major part of the data points are gathered to the left part of the plot.



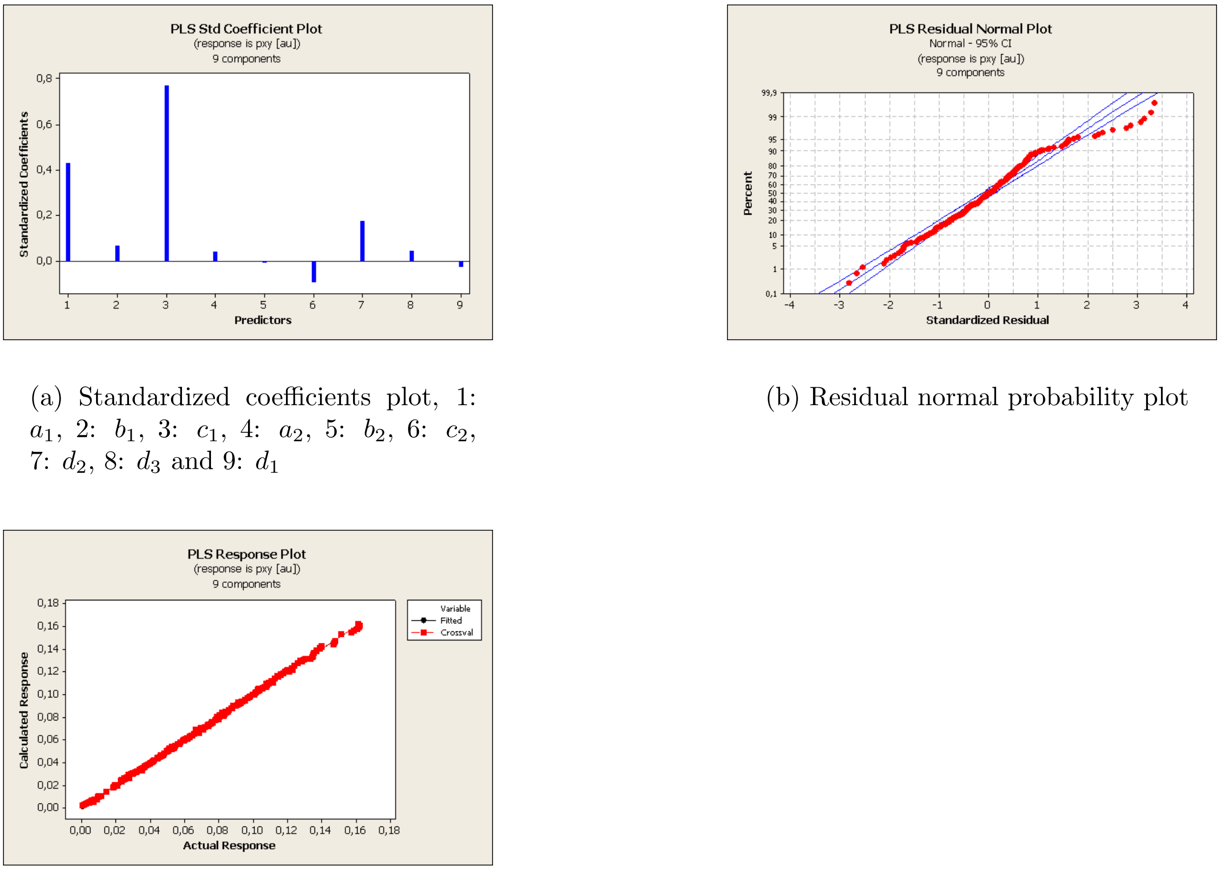
(a) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: a_2, 5: b_2, 6: c_2, 7: d_2, 8: d_3 and 9: d_1

(b) Residual normal probability plot

(c) Response plot, R^2=0.9962

Figure B.8: Results from PLS regression on the xx-component of the polarizability of hydrogen

The by far largest contribution to the xx-component can in figure B.8(a) be identified to be the bond length between the H-atom and the O-atom. This is of course a combination of the fact that LoProp (as already mentioned) directs the charges into the bond when subject to an external field and the geometry used in the calculations. The other hydrogens bond and the angle together with the quadric term of r_1 have some, but very little influence on the component, and the other terms are neglectable. The residual normal probability plot in figure B.8(b) tells us that there is a slight skewness in the data, but nothing serious, and the response plot in figure B.8(c) reports a good fit to the data, except for some points with a larger value (or at least they are not as good fitted as the others).



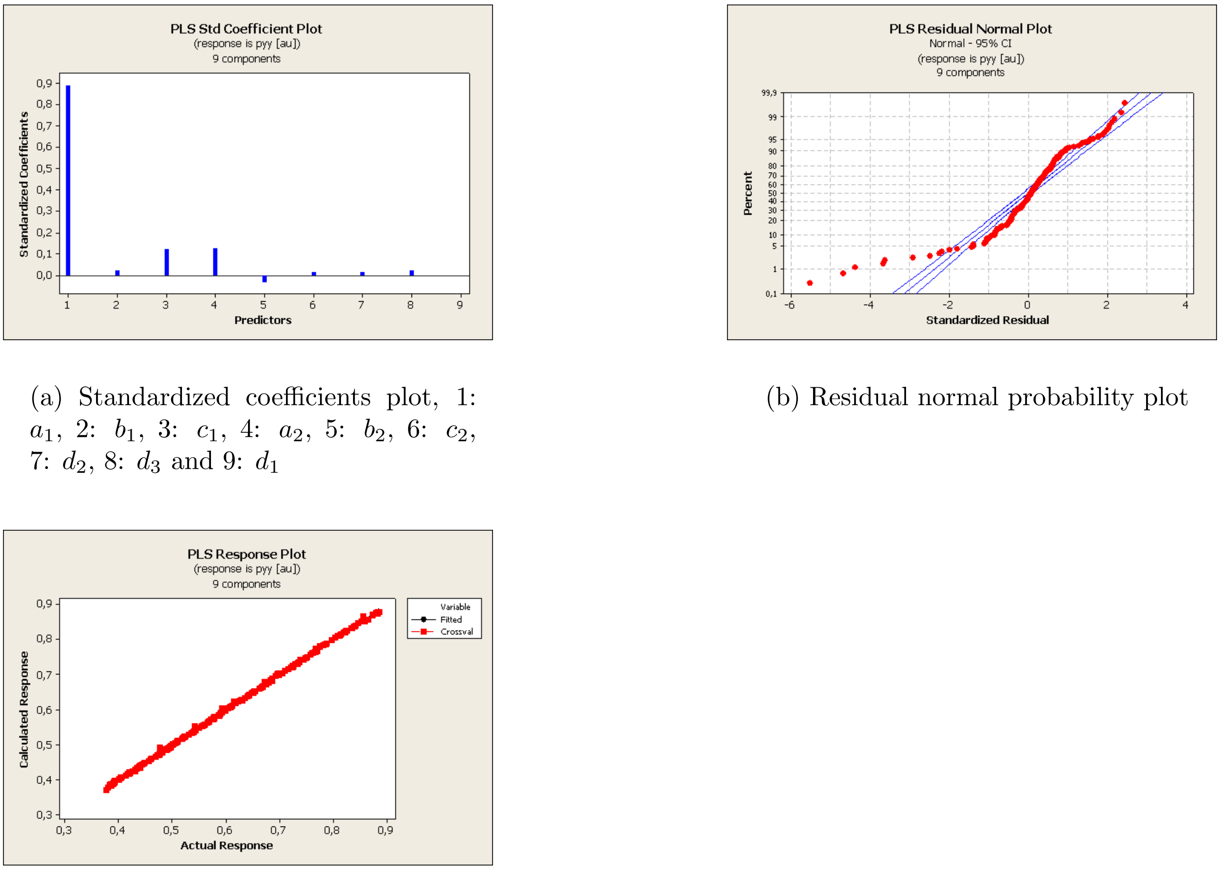
(a) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: a_2, 5: b_2, 6: c_2, 7: d_2, 8: d_3 and 9: d_1

(b) Residual normal probability plot

(c) Response plot, R^2=0.9996

Figure B.9: Results from PLS regression on the xy-component of the polarizability of hydrogen

Figure B.9(a) shows that the angle in the water molecule is of importance when determining the off-diagonal component of the hydrogens polarizability, and hence the reason for it to span out the second PLS component in figure B.7(a). As can be seen in figure B.9(a) the coefficient of the angle has the largest influence on the model for this component, followed by the bond length r_1. There is also some curve linearity in the model, furthermore the cross term between the angle and r_1 actually has some influence. From this we might conclude that the distance between the two hydrogens (as described by the angle) is the most influential factor on the xy-component of the polarizability. From figure B.9(b) and B.9(c) there is not much to report other than a slight tail in the residual normal probability plot and a very good fit in the response plot.



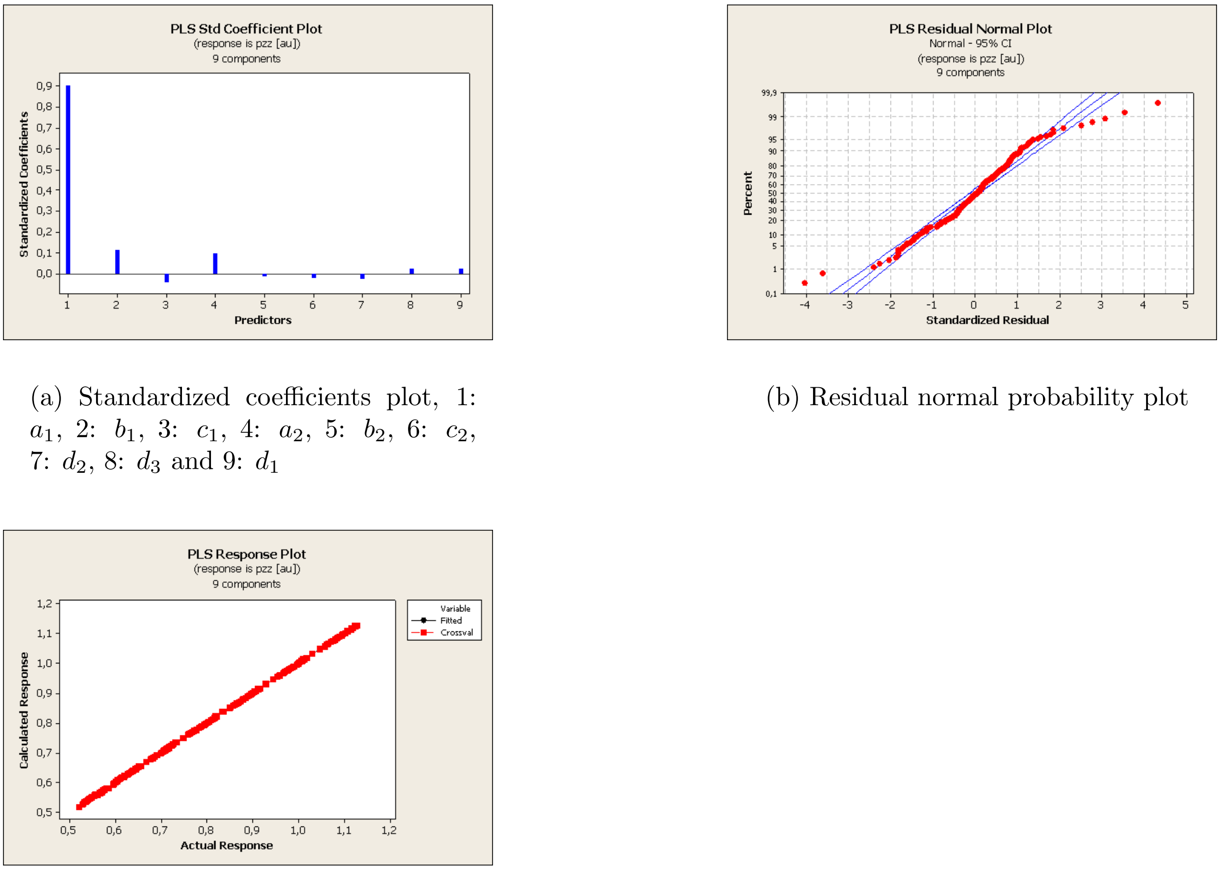
(a) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: a_2, 5: b_2, 6: c_2, 7: d_2, 8: d_3 and 9: d_1

(b) Residual normal probability plot

(c) Response plot, R^2=0.9997

Figure B.10: Results from PLS regression on the yy-component of the polarizability of hydrogen

The linear term of r_1 is by far the most influential term for the yy-component also (see figure B.10(a)), although some curve linearity in this term is also present. These two terms together with the linear angle term are the only ones of importance when modeling the yy-component. In the residual normal probability plot in figure B.10(b) there is a quite long tail, and a lot of the residuals seems to lie slightly outside the 95% confidence interval line. This is somewhat worrying, as it indicates that there is some deviation from normality in the residuals. The response plot does, however, tell us that we have a very good fit (see figure B.10(c)) and the model is therefore used.



(a) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: a_2, 5: b_2, 6: c_2, 7: d_2, 8: d_3 and 9: d_1

(b) Residual normal probability plot

(c) Response plot, R^2=0.9999

Figure B.11: Results from PLS regression on the zz-component of the polarizability of hydrogen

Figure B.11(a) tells us that also for the zz-component of the polarizability the r_1 term is the most important. The two other contributions with some influence are the quadric r_1 term and the bond length of the other H-atom. The other terms seems to be neglectable. The fit described in the response plot (figure B.11(c)) is very good, eventhough there seems to be an indication of skewness in figure B.11(b).

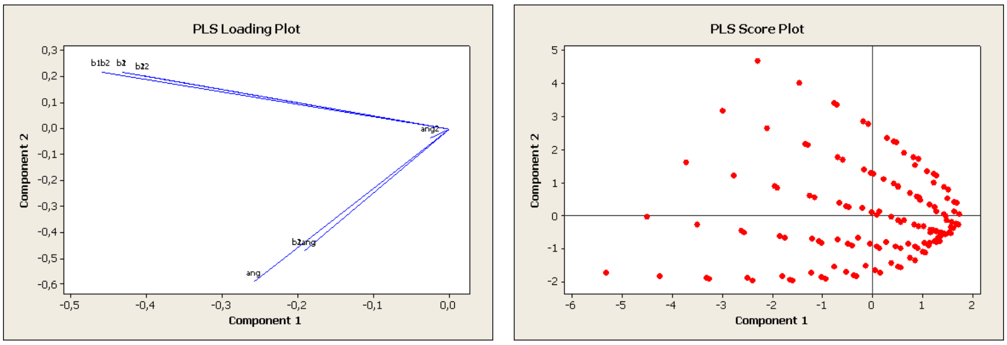
The results discussed above show that quite a lot of the terms included in the model are without any significant influence. This tells us that a simpler model might be more appropriate to use. Taking the important contributions into account one should use the following equation;

$$\begin{aligned} p_{A_{ij}} &= a_1(r_1 - r_e) + a_2(r_1 - r_e)^2 \\ &+ b_1(r_2 - r_e) \\ &+ c_1(\theta - \theta_e) + c_2(\theta - \theta_e)^2 \\ &+ d_1(r_1 - r_e)(r_2 - r_e) + const \end{aligned} \tag{B.1}$$

instead of eq. 3.10 to model the atomic polarizability of hydrogen.

Oxygen atom

Loadings and score plot from the PLS regression on the atomic polarizability of oxygen are presented in figures B.12, and the results for the four components are presented in figure B.8 to B.11. These results are discussed below.

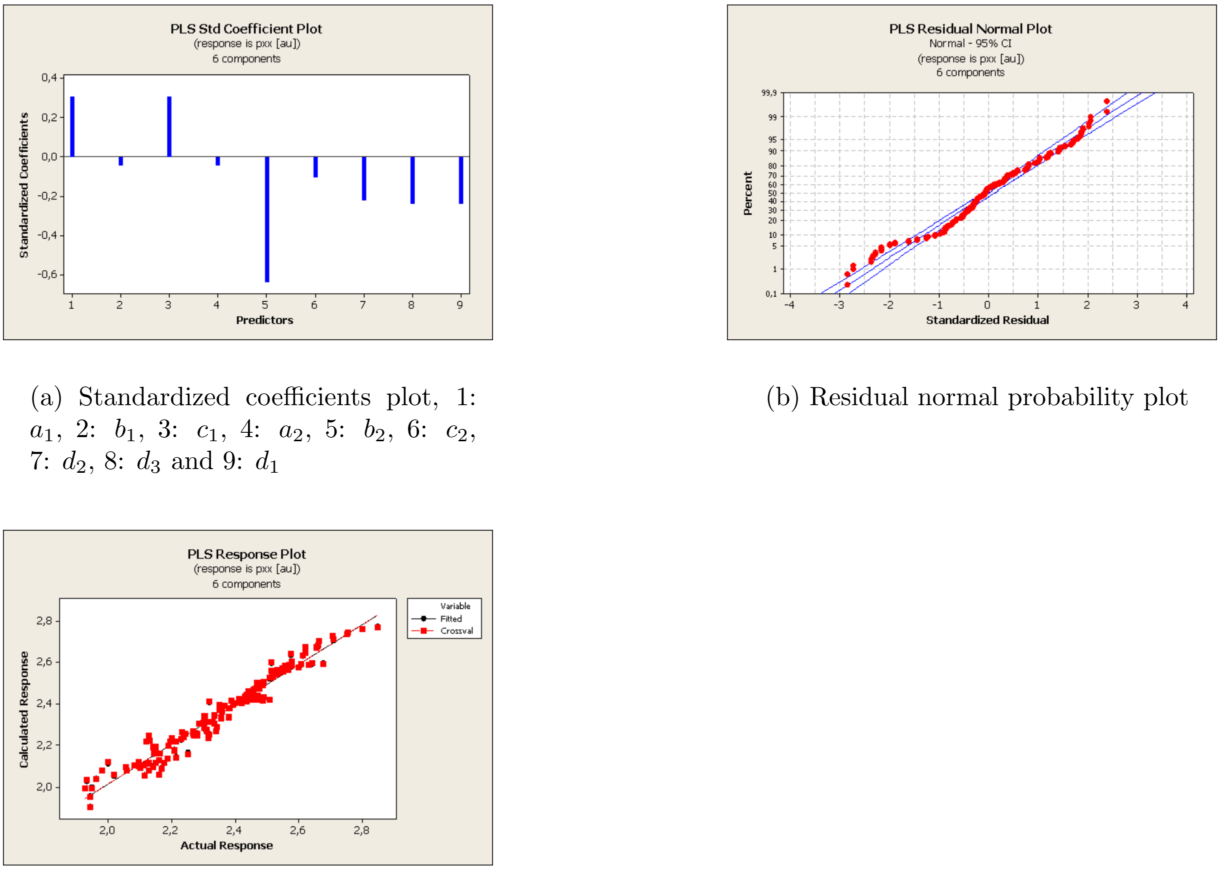


(a) Loadings plot

(b) Score plot, high number indicates long bond length

Figure B.12: Loadings and score plot for the polarizability of oxygen

The loadings and score plot for the modeling of the polarizability of oxygen are presented in figure B.12, and we can see that both the bond terms and the linear angle term expands the first two PLS components (loadings plot in figure B.12(a)), however the first PLS component are mainly spanned out by the bond terms which seem to be quite correlated. The linear angle term is the most influential for the second PLS component. It is also clear that the quadric angle term does not describe much of the variance in the system. The score plot in figure B.12(b) demonstrate that it is the more extreme points that gives the largest contributions to the variance.



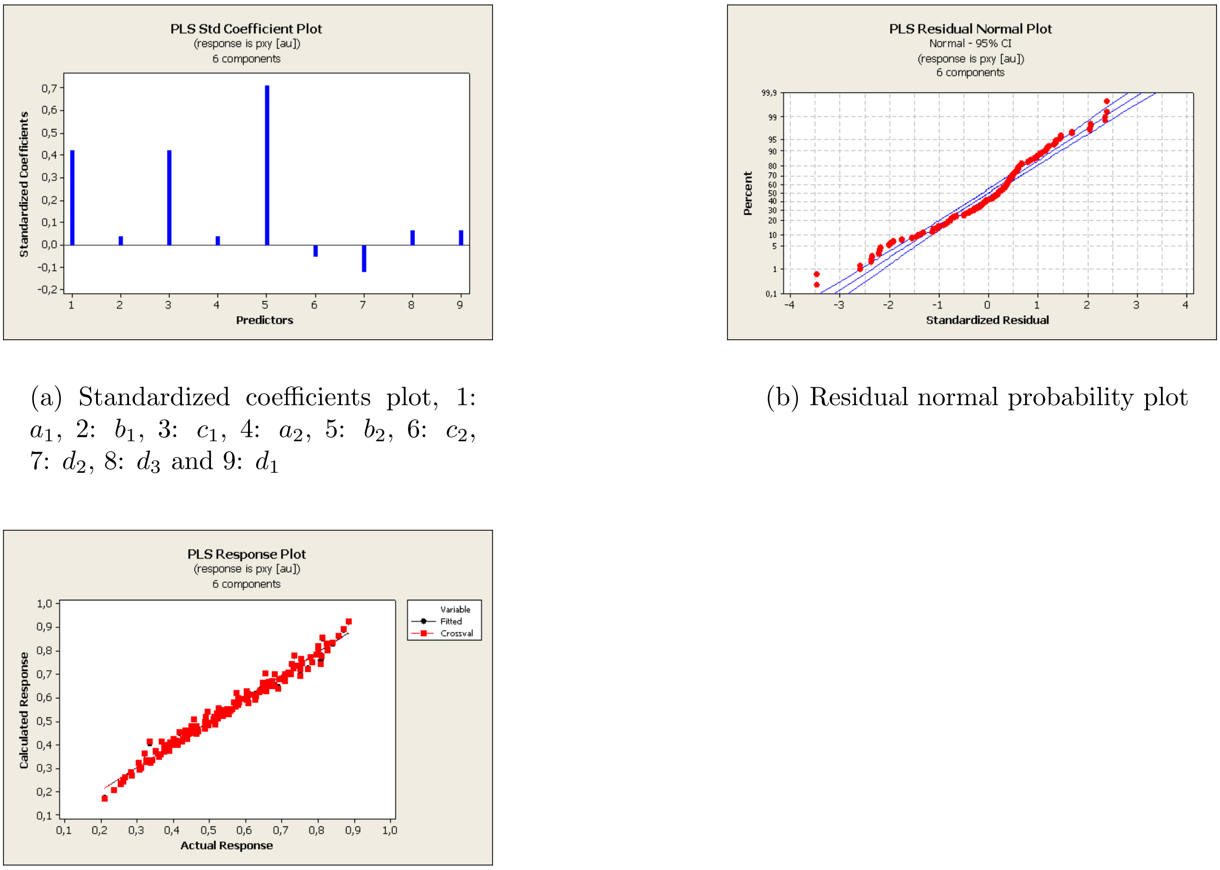
(a) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: a_2, 5: b_2, 6: c_2, 7: d_2, 8: d_3 and 9: d_1

(b) Residual normal probability plot

(c) Response plot, R^2=0.9576

Figure B.13: Results from PLS regression on the xx-component of the polarizability of oxygen

The results for the model of the xx-component of oxygens polarizability are presented in figure B.13, and from the standardized coefficients plot in figure B.13(a) the most influential variable for this component is the linear angle term. Furthermore the linear bond terms and the three cross terms are important variables. There seems to be a slight curve linear dependence on the angle, and no curve linear dependence on the bond terms. The residuals in figure B.13(b) appears to be normal distributed. In figure B.13(c) the data points are evenly spread, and lies on a straight line, indicating a good fit and homoscedasticity in the residuals.



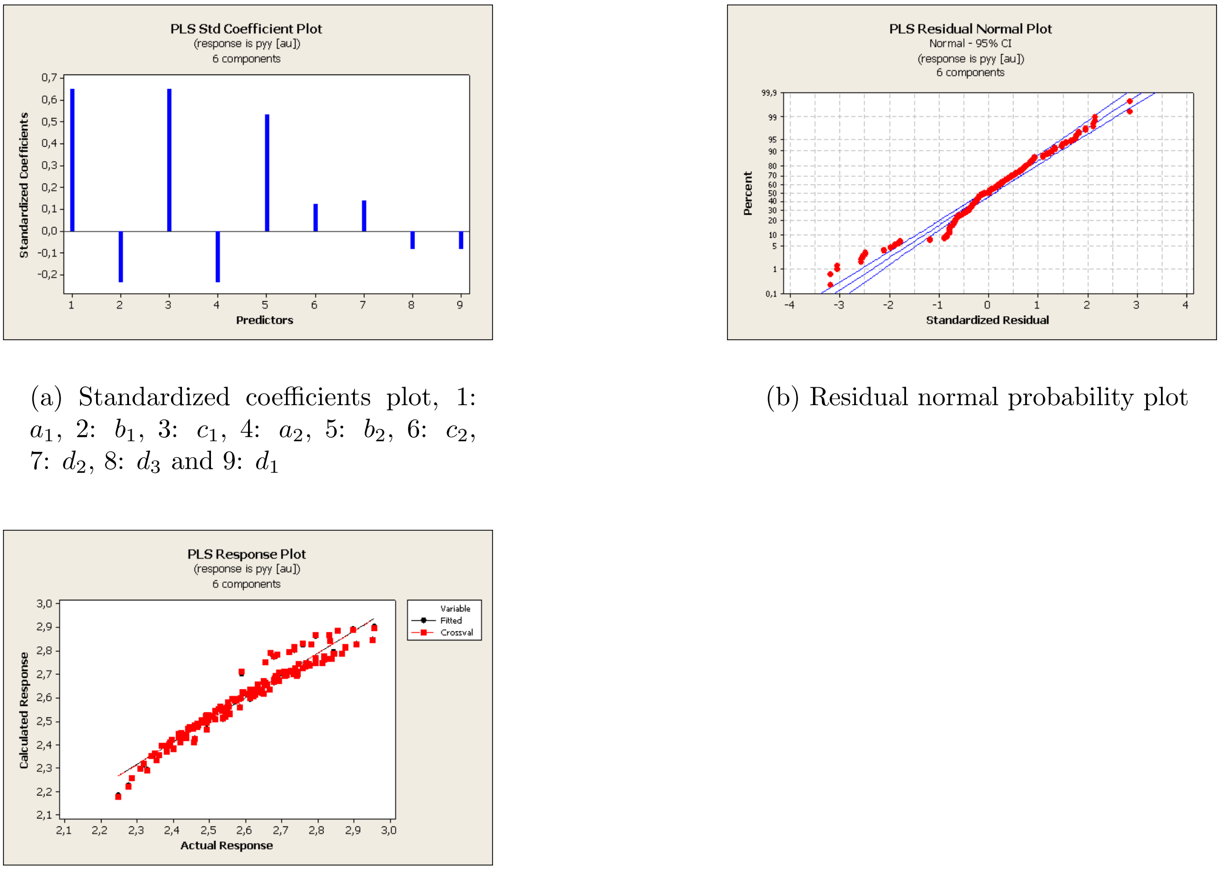
(a) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: a_2, 5: b_2, 6: c_2, 7: d_2, 8: d_3 and 9: d_1

(b) Residual normal probability plot

(c) Response plot, R^2=0.9837

Figure B.14: Results from PLS regression on the xy-component of the polarizability of oxygen

For the xy-component we see from the results presented in figure B.14(a) shows that the linear terms are the most influential on the components, with small and almost neglectable contribution from the other terms. The residuals presented in figure B.14(b) seem to have a slight but neglectable skewness, and are thus satisfying. The response plot in figure B.14(c) tells us that we have a model which describes the property good, with only some deviations for the smallest and largest values.



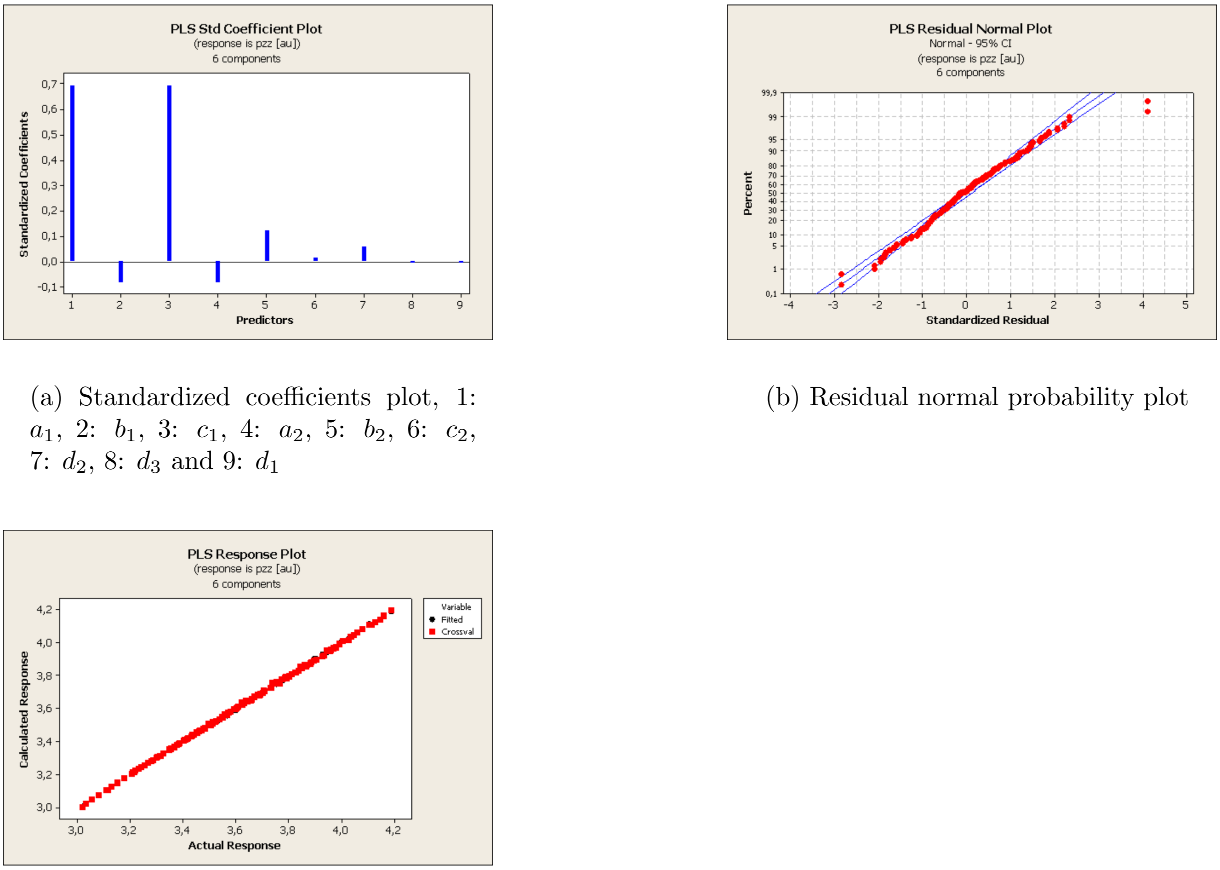
(a) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: a_2, 5: b_2, 6: c_2, 7: d_2, 8: d_3 and 9: d_1

(b) Residual normal probability plot

(c) Response plot, R^2=0.9451

Figure B.15: Results from PLS regression on the yy-component of the polarizability of oxygen

For the yy-compontent in figure B.15(a) there is a curve linear dependence in all terms, however also for this component the linear terms are the most important. For the yy-component the linear bond terms are more influential than the angle term, in contrast to the two previously described components. Disregarding a tail at the lower end of the residual normal probability plot (figure B.15(b)) indicating some skewness, the residuals follows a normal distribution. The response plot in figure B.15(c) tells us that some of the data points with lowest value are overestimated and some with higher value are underestimated. The fit however is satisfying, as it describes with an R^2 over 0.94.



(a) Standardized coefficients plot, 1: a_1, 2: b_1, 3: c_1, 4: a_2, 5: b_2, 6: c_2, 7: d_2, 8: d_3 and 9: d_1

(b) Residual normal probability plot

(c) Response plot, R^2=0.9980

Figure B.16: Results from PLS regression on the zz-component of the polarizability of oxygen

The results for the zz-component of the atomic polarizability of oxygen are presented in figure B.16, and we can observe from the standardized coefficients plot in figure B.16(a) that the by far most influential terms here are the linear bond terms. There are also contributions from the linear angel term and quadric bond terms, however these are quite small compared to the linear bond terms. There are two points in the normal residual probability plot in figure B.16(b) deviating from the others, indicating that they might be outliers. A Principal Component Analysis (PCA) where done to investigate this, and no outliers were found. These points were therefore used in the model. The fit described in figure B.16(c) is very good, and it can be concluded that the model of the zz-component is very satisfying.

The model of the atomic polarizability of oxygen seem to be satisfying and should not be simplified. The polarizability of oxygen is clearly more difficult to describe than that of hydrogen. The reasons for this is the more complex electronic structure of oxygen compared to hydrogen, and because the oxygen atom is connected to two atoms, whereas the hydrogens are only connected to one.

B.2 Correlation between Dinur-Hagler charges and LoProp charges

To investigate the performance of the LoProp charges, the Dinur-Halger (DH) charges for the water monomer was calculated for the same configurations and level of theory as the LoProp charges. This was done using the Dalton [93] software program. The DH charges are closely related to the Atomic Polar Tensor (APT) charges of Cioslowski [39], but instead of the mean value of the derivate of the dipole in cartesian coordinates, the DH charges uses the derivative of the dipole moment along the axis perpendicular to the plane of the molecule. Given for instance a water molecule in the xy-plane the DH charges becomes;

$$q_i = \frac{\partial \mu_i}{\partial z} \tag{B.2}$$

To see how the two methods behave compared to each other, the difference between the atomic charges was taken for both elements, and it was tested if the difference between the two methods could be related to the internal degrees of freedom in the molecule. To do this a PLS regression was done, approximating the difference to the following expression;

$$q_{\textit{Dinur-Hagler}} - q_{\textit{LoProp}} = a_1 r_1 + a_2 r_2 + a_3 \theta \tag{B.3}$$

for both the hydrogen- and oxygen charges. A good regressional fit would thus suggest that the difference between the two methods arise from their ability to describe the perturbation from the equilibrium geometry, and not other sources of error.

B.2.1 Results and Conclusion

In table B.1 the correlation between the DH- and LoProp charges is presented. As can be observed from this table there is a strong correlation between the two methods. This suggests that there is relatively little difference between the charges produced by the two methods.

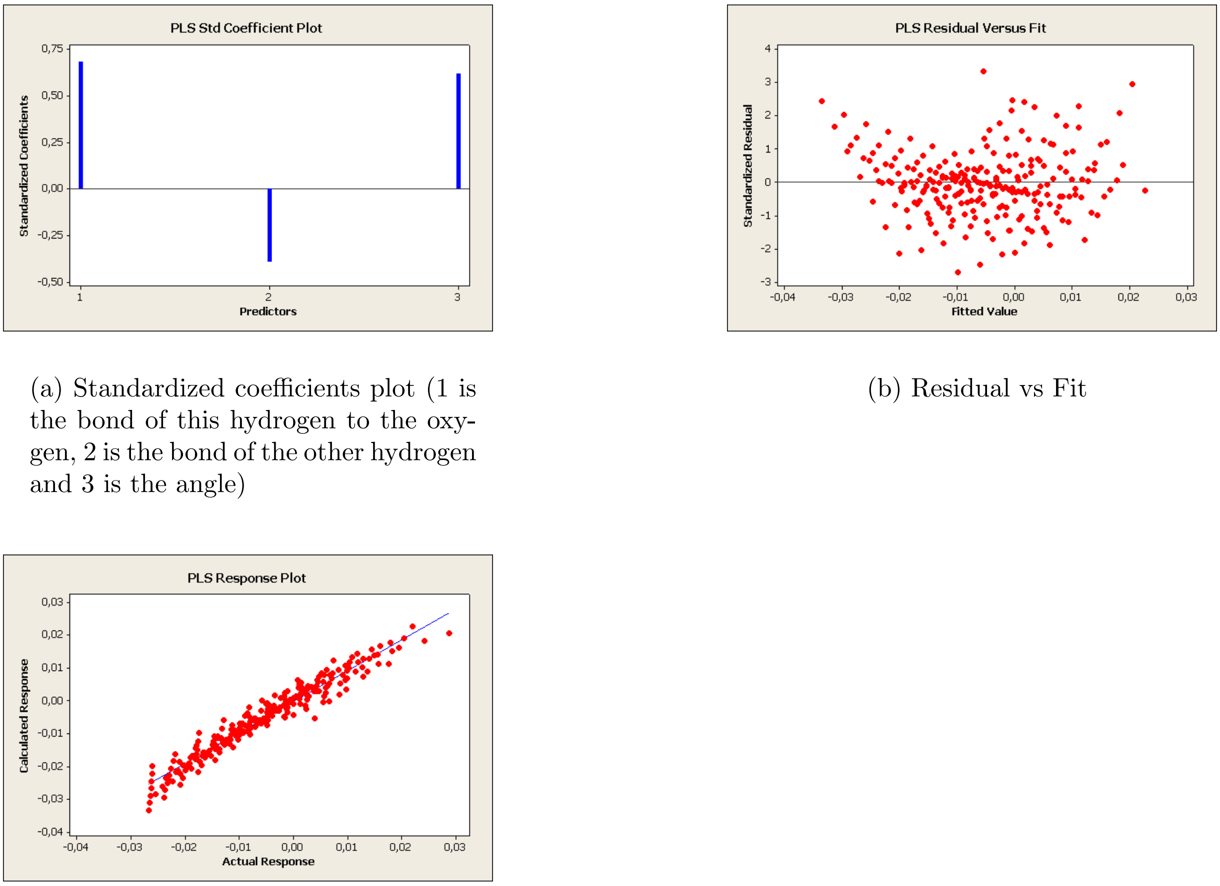
Table B.1: Correlation between Dinur-Hagler and LoProp charges

Atom	*Correlation*
Hydrogen	0.94132
Oxygen	0.91630

In table B.2 some statistics on the difference between DH and LoProp are presented. As we can see the mean deviation between the two methods are relatively low (around 14-15% of the mean numerical value of the charges). The standard deviation (STD) of the difference is relatively large compared to the mean, this is furthermore reflected in the max-min difference, which has a spread around 0.05547 for hydrogen and 0.13338 for oxygen.

Table B.2: Basic statistics on the difference between Dinur-Hagler and LoProp charges

	Atom	
	Hydrogen	Oxygen
Mean	-0.00535	0.01069
STD	0.01124	0.01543
Max	0.02877	0.05198
Min	-0.02659	-0.01314

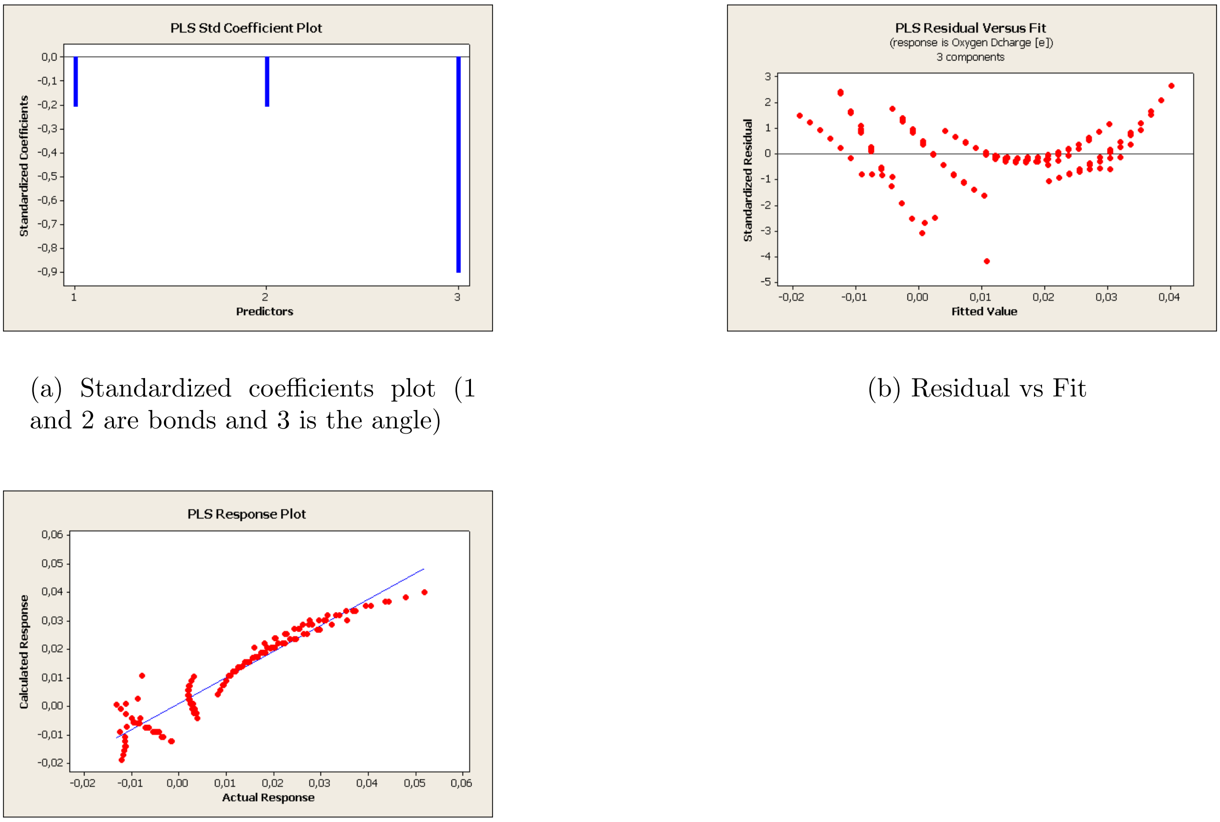


(a) Standardized coefficients plot (1 is the bond of this hydrogen to the oxygen, 2 is the bond of the other hydrogen and 3 is the angle)

(b) Residual vs Fit

(c) Response plot, R^2=0.93838

Figure B.17: Results from PLS regression on the difference in atomic charge of hydrogen between Dinur-Hagler and LoProp



(a) Standardized coefficients plot (1 and 2 are bonds and 3 is the angle)

(b) Residual vs Fit

(c) Response plot, R^2=0.91462

Figure B.18: Results from PLS regression on the difference in atomic charge of oxygen between Dinur-Hagler and LoProp

The results from the PLS regression are presented in figure B.17 and B.18. In figure B.17(c) and B.18(c) it can be observed that the regression is able to explain a significant amount of the difference between the two methods. For hydrogen the R^2 is 0.93838 and oxygen 0.91462 and hence it is possible to describe much of the variation between the two methods with a simple linear model. The residual vs. fit plots in figure B.17(b) seems to give an even spread of the standardized residuals, while the residual vs fit plot in figure B.18(b) suggests that a linear model might not be the best solution for oxygen. However as the aim of the regression was not to give a best fit to the data, but to see if some of the discrepancies between the methods could be described through the internal degrees of freedom, this is not relevant. The standardized coefficients plotted in figure B.17(a) tells us that for the hydrogen atoms the major part of the difference can be described by the hydrogens bond to the oxygen and the angle, while the other hydrogens bond is of less importance. For oxygen the angle is the most important variable (see figure B.18(a)).

The conclusion of this short comparison of DH charges and LoProp charges is that the two methods give very similar charges under the same conditions (basis-set etc). The difference between them for hydrogen mainly arises from how they describe changes in bonds to the hydrogen and the angle of the molecule, while it for oxygen mainly arises because of differences in how they handle changes of the angle.

Appendix C

Further details to implementation

In this appendix the various algorithms implemented are presented in the form of pseudo-codes.

C.1 Atomic properties

```
for i in all molecules:
    Compute $r_1$, $r_2$ and $\theta$
    Compute the charges, dipoles and polarizabilities with respect to the
    orientation of the reference state.
    Select one hydrogen as the reference ($H_1$) and rotate the other hydrogen
    with $\theta$ degrees around the z-axis.
    Compute the rotational matrix for the whole molecule, $\mathbf{Q}$
    Rotate the whole molecule using matrix $\mathbf{Q}$
```

Figure C.1: Pseudo code for calculation of atomic properties

C.2 Induction energy

```
for i in all molecules:
    for j in all atoms in molecule $i$:
        for k in all molecules, except $i$:
            for l in all atoms in molecule $k$:
                Compute distance between atom $l$ and $j$, $\vec{R}_{jl}$
                Compute the electric field on atom $l$ from $j$;
                $\vec{E}_{jl} = \frac{q_l^{perm}\vec{R}_{jl}}{|\vec{R}_{jl}|^3} + \frac{\vec{R}_{jl}(\vec{R}_{jl}\cdot\vec{\mu}_l^{perm})}{|\vec{R}_{jl}|^5}$
                Add $\vec{E}_{jl}$ to the total, external electric field on atom $j$;
                $\vec{E}_j = \vec{E}_j + \vec{E}_{jl}$
        Compute $\vec{\mu}_j^{ext} = \alpha_j\vec{E}_j$
if first simulation step:
    Use $\vec{\mu}_j^{ext}$ as initial guess of $\vec{\mu}_j^{ind}$
else:
    Use $\vec{\mu}_j^{ind}$ from the previous simulation step as initial guess
while 1:
    for i in all molecules:
        for j in all atoms in molecule $i$:
            for k in all molecules, except $i$:
                for l in all atoms in molecule $k$:
                    Compute distance between atom $l$ and $j$, $\vec{R}_{jl}$
                    Compute $\mathbf{T_{jl}}$;
                    $\mathbf{T_{jl}} = \frac{3}{|\vec{R}_{jl}|^5}\vec{R}_{jl}^T\vec{R}_{jl} - \mathbf{I}\frac{1}{|\vec{R}_{jl}|^3}$
                    Compute the contribution to the induced electric field
                    from the induced dipole on atom $l$ on atom $j$;
                    $\vec{E}_{jl}^{ind} = \mathbf{T}_{jl}\mu_l^{ind,\ n}$
                    Add $\vec{E}_{jl}^{ind}$ to the total, induced electric field on atom $j$;
                    $\vec{E}_j^{ind} = \vec{E}_j^{ind} + \vec{E}_{jl}^{ind}$
            Compute the new induced dipole of atom $j$;
            $\vec{\mu}_j^{ind} = \vec{\mu}_j^{ext} + \alpha_j\vec{E}_j$
    for i in all atoms:
        Compute the largest relative difference in length of the induced
        dipole between this and the previous iteration
    if max deviation =< convergence criteria:
        break
    else:
        Set all $\vec{\mu}_l^{(ind,\ n)}$ equal to $\vec{\mu}_l^{(ind,\ n+1)}$
for i in all atoms:
    Compute the induced energy;
        $V_{ind} = V_{ind} + \frac{1}{2}\vec{\mu}_j^{ind}\vec{E}_j$
```

Figure C.2: Pseudo code for calculation of the induction energy

C.3 Repulsion energy

```
for i in all molecules:
    for j in all atoms in molecule $i$:
        for k in all other molecules:
            for l in all atoms in molecule $k$:
                Identify type of atom for atom $j$ and atom $l$
                Compute distance between atom $l$ and $j$, $\vec{R}_{jl}$
                Compute contribution to repulsion energy from
                current atom pair;
                $V_{rep} = V_{rep} + \kappa_{jl} e^{\alpha_{jl} R_{jl}}$
```

Figure C.3: Pseudo code for calculation of the repulsion energy

C.4 Dispersion energy

```
for i in all molecules:
    for j in all atoms in molecule $i$:
        for k in all other molecules:
            for l in all atoms in molecule $k$:
                Identify type of atom for atom $j$ and atom $l$
                Compute distance between atom $l$ and $j$, $\vec{R}_{jl}$
                Compute dampening term of the current atom pair;
                $\epsilon = B_{jl}(1 - e^{-b_{jl} R_{jl}})$
                Compute $\mathbf{T_{jl}}$;
                $\mathbf{T_{jl}} = \frac{3}{|\vec{R}_{jl}|^5} \vec{R}_{jl}^T \vec{R}_{jl} - \mathbf{I} \frac{1}{|\vec{R}_{jl}|^3}$
                $\beta = 0$
                for m in range(0,2):
                    for n in range(0,2):
                        for o in range(0,2):
                            for p in range(0,2):
                                Compute;
                                $\beta = \beta + \alpha^j[m][n]\alpha^l[o][p]\mathbf{T_{jl}}[m][o]\mathbf{T_{jl}}[n][p]$
                Compute the energy contribution of the current pair;
                $V_{disp} = V_{disp} + \epsilon\beta$
```

Figure C.4: Pseudo code for calculation of the dispersion energy

C.5 Zero point vibrational corrections

```
for i in all molecules:
    Compute $r_1$, $r_2$ and $\theta$
    Compute the second derivatives
    Construct Hessian matrix
    Compute the F-matrix and the G-matrix
    Compute eigenvalues ($\nu_i$) of $\mathbf{FG}$ using Jacobi algorithm of Press et al. [46].
    Compute ZPV energy;
    $V_{ZPV} = \sum_1^3 \frac{\nu_i}{2}$
```

Figure C.5: Pseudo code for calculation of the ZPV energy