UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Faculty of Science and Technology

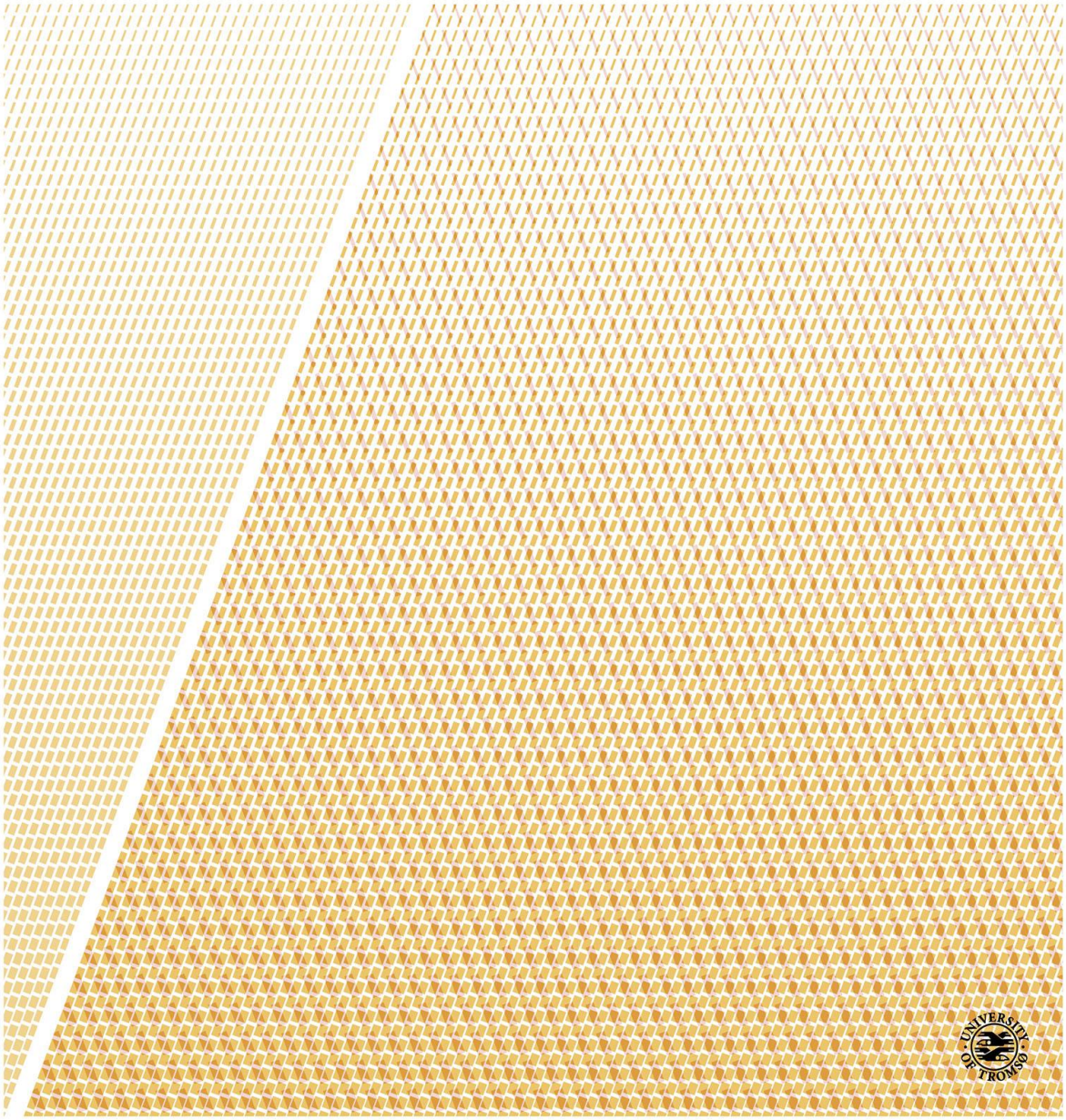
Department of Chemistry

Calculating molecular properties in realistic environments

—

Maarten T. P. Beerepoot

A dissertation for the degree of Philosophiae Doctor – May 2016



Abstract

This thesis focuses on how absorption properties of molecules are influenced by their environment and how this can be calculated accurately. Calculations have been performed with a polarizable embedding (PE) multiscale model. The environment is described classically by charges and electric multipoles for the permanent electrostatics and polarizabilities for polarization interactions. Density-functional theory (DFT) and approximate singles and doubles coupled-cluster theory (CC2) are used to describe the electronic structure of the molecules. The results indicate that the effects of environmental polarization on electronic and vibrational properties are significant and that the employed PE model is accurate in cases where electrostatic interactions dominate. A large part of the environment needs to be described explicitly for converged molecular properties, especially since polarization interactions range over a long distance. However, accurate embedding parameters for the electrostatic and polarization interactions are important mainly for the closest environment of a chromophore. This enables a reduction of the computational cost of obtaining embedding potentials without sacrificing accuracy. For localized properties, PE is to be preferred over a cluster approach because the latter is severely limited by the possible size of the molecular system. For calculation of two-photon absorption (TPA), DFT and CC2 give qualitatively but not quantitatively similar results. Finally, it is shown that the comparison between calculated TPA cross sections and other experimental or theoretical work is challenging. The presented works contribute to the *realistic* description of a molecular environment with the accurate prediction of molecular properties in chemical environments as ultimate goal.

Contents

List of Papers

This thesis is based on the following seven peer-reviewed publications, which are referred to as **Papers I-VII** in the text.

I **M. T. P. Beerepoot**, A. H. Steindal, J. Kongsted, B. O. Brandsdal, L. Frediani, K. Ruud and J. M. H. Olsen, "A polarizable embedding DFT study of one-photon absorption in fluorescent proteins", *Phys. Chem. Chem. Phys.* **15** (2013), 4735–4743.

II T. Schwabe, **M. T. P. Beerepoot**, J. M. H. Olsen and J. Kongsted, "Analysis of computational models for an accurate study of electronic excitations in GFP", *Phys. Chem. Chem. Phys.* **17** (2015), 2582–2588.

III **M. T. P. Beerepoot**, A. H. Steindal, K. Ruud, J. M. H. Olsen and J. Kongsted, "Convergence of environment polarization effects in multiscale modeling of excitation energies", *Comp. Theor. Chem.* **1040** (2014), 304–311.

IV **M. T. P. Beerepoot**, A. H. Steindal, N. H. List, J. Kongsted and J. M. H. Olsen, "Averaged solvent embedding potential parameters for multiscale modeling of molecular properties", *J. Chem. Theory Comput.* **12** (2016), 1684–1695.

V **M. T. P. Beerepoot**, D. H. Friese and K. Ruud, "Intermolecular charge transfer enhances two-photon absorption in yellow fluorescent protein", *Phys. Chem. Chem. Phys.* **16** (2014), 5958–5964.

VI **M. T. P. Beerepoot**, D. H. Friese, N. H. List, J. Kongsted and K. Ruud, "Benchmarking two-photon absorption cross sections: Performance of CC2 and CAM-B3LYP", *Phys. Chem. Chem. Phys.* **17** (2015), 19306–19314.

VII N. H. List, **M. T. P. Beerepoot**, J. M. H. Olsen, B. Gao, K. Ruud, H. J. Aa. Jensen and J. Kongsted, "Molecular quantum mechanical gradients within the polarizable embedding approach—Application to the internal vibrational Stark shift of acetophenone", *J. Chem. Phys.* **142** (2015), 034119.

The following related and unrelated peer-reviewed publications have not been included in the thesis.

A. Pikulska, A. H. Steindal, **M. T. P. Beerepoot** and M. Pecul, "Electronic circular dichroism of fluorescent proteins: A computational study", *J. Phys. Chem. B* **119** (2015), 3377–3386.

D. H. Friese, **M. T. P. Beerepoot** and K. Ruud, "Rotational averaging of multiphoton absorption cross sections", *J. Chem. Phys.* **141** (2014), 214103.

D. H. Friese, **M. T. P. Beerepoot**, M. Ringholm and K. Ruud, "Open-ended recursive approach for the calculation of multiphoton absorption matrix elements", *J. Chem. Theory Comput.* **11** (2015), 1129–1144.

P. K. Wawrziniak, **M. T. P. Beerepoot**, H. J. M. de Groot and F. Buda, "Acetyl group orientation modulates the electronic ground-state asymmetry of the special pair in purple bacterial reaction centers", *Phys. Chem. Chem. Phys.* **13** (2012), 10270–10272.

Acknowledgements

First of all, I acknowledge the contributions of Arnfinn, Bin Gao, Bjørn Olav, Daniel, Hans Jørgen, Jacob, Jógvan Magnus, Luca, Kenneth, Nanna and Tobias to the papers that form the basis of this thesis. I thank my supervisor Kenneth for sharing his expertise and his network, for creating an excellent atmosphere for collaboration, for motivating independent thought and ideas and for showing that an academic career is much more than science alone. I thank my co-supervisors Bjørn Olav and Luca for discussion on widely varying topics. Your availability and open doors are much appreciated. Special thanks to Arnfinn for help and collaboration from the very beginning to the very end and for showing that a healthy work–life balance is the key to happiness. I thank Jacob, Jógvan Magnus and Nanna for excellent collaboration and inspiring discussion. I feel privileged having the chance to work together with you. Thanks also Roberto and Anna for proof-reading (a part of) this thesis and coming with helpful suggestions. In addition to the people mentioned above, I also thank Magnus, Marius and Yann for discussion on various topics.

I thank all present and past members of the Centre for Theoretical and Computational Chemistry (CTCC) for the inspiring working environment and Stig for all his help with organizational issues. A big thanks goes to Espen, Steinar, Dan, Roy and the other members of the HPC group in Tromsø for their (often instant) technical support. I am grateful to the members of the research groups of Jacob Kongsted at the University of Southern Denmark in Odense and of Geert-Jan Kroes at the University of Leiden for a pleasant atmosphere during my visits abroad.

I have had great benefit from the schools I visited during my PhD period and would like to thank the excellent teachers and organizers of the school on multiscale modeling (Stockholm, 2012, organized by Hans Ågren), the European summer school on quantum chemistry (Sicily, 2013, organized by Trond Saue), the winter school on theoretical spectroscopy (Helsinki, 2013, organized by Dage Sundholm) and the workshop on biomolecular modeling (Odense, 2015, organized by Ilia Solov'yov and Jacob Kongsted). I have equally benefited from the excellent courses on writing, presenting and teaching organized by the BioStruct graduate school, the High North Academy and the Centre for Teaching, Learning and Technology (Result)

at the University of Tromsø.

My position was financed by a grant from the European Research Council (Grant 279619) at the CTCC, a centre of excellence financed by the Research Council of Norway (Grant 179568/V30). Furthermore, I acknowledge travel grants from BioStruct, the national graduate school in structural biology, and computer time through the Norwegian Supercomputer Program (Grant NN4654K).

Last but not least, thank you Anna for all your love, patience and support in the past years.

Abbreviations

The following abbreviations are all defined in the text. This list is to allow for easy reference.

CC	coupled-cluster
CC2	approximate singles and doubles coupled-cluster
CCSD	coupled-cluster singles and doubles
CI	configuration interaction
DFT	density-functional theory
DMA	distributed multipole analysis
ESP	electrostatic potential
GFP	green fluorescent protein
GGA	generalized gradient approximation
HF	Hartree–Fock
LDA	local density approximation
LJ	Lennard–Jones
KS	Kohn–Sham
MC-SCF	multi-configuration self-consistent field
MD	molecular dynamics
MFCC	molecular fractionation with conjugate caps
MO	molecular orbital
MPA	multiphoton absorption
OPA	one-photon absorption
PCM	polarizable continuum model
PDE	polarizable density embedding
PE	polarizable embedding
PNA	*para*-nitroaniline
QM	quantum-mechanical
QM/MM	quantum mechanics / molecular mechanics
RMSD	root-mean-square deviation
RESP	restrained electrostatic potential
SCF	self-consistent field
TPA	two-photon absorption
vdW	van der Waals

Introduction

The goal of theoretical chemistry is to develop methods for the calculation of molecular properties and to apply these to chemically relevant problems. In general, these properties can be energies, structures or the response of the molecule to some sort of perturbation. In this work, the emphasis lies on the absorption of light by a molecule. A molecule can absorb photons with a specific energy, leading to a change in the electronic structure of the molecule: an *electronic excitation*. The specific energies at which this happens are called excitation energies and are intrinsic properties of the molecule. Since the most intense absorption of many organic molecules is in the ultraviolet or visible region of the electromagnetic spectrum, this process is often referred to as UV–VIS absorption. In the most simple case, one photon carries the energy needed to excite a molecule. In the context of this thesis, this process is referred to as *one-photon absorption* (OPA). Electronic excitations can also be accomplished by the simultaneous absorption of two photons that together carry exactly enough energy to excite the molecule. This process is called *two-photon absorption* (TPA). The calculation of excitation energies and the intensity of OPA and TPA processes is central in this thesis.

Theoretical chemistry is closely related to other parts of chemistry. A theoretical method that is accurate enough to reproduce experimentally measured quantities can be used to predict the properties of other molecules that have not been synthesized (yet). For instance, when calculations on fluorescent proteins can reproduce certain experimental trends in excitation energies that result from structural differences in the proteins (**Paper I**), the same models can be used to predict the effect of a mutation on the excitation energy of a similar protein. In addition to predictions, calculations can give microscopic and mechanistic insights that are challenging or impossible to

obtain from experiments. This can be illustrated by theoretical calculations on the yellow fluorescent protein, where the intensity of TPA is found to be enhanced as a result of intermolecular charge-transfer transitions (**Paper V**). This highlights the role of theoretical chemistry as a complement to experimental work.

Choosing the right quantum-mechanical (QM) method for the calculation of a property is crucial to obtain a result with reasonable accuracy. As an example, density-functional theory (DFT) can give a good estimate of the intensity of an OPA process, while the absolute value of the TPA intensity can significantly deviate from more accurate methods (**Paper VI**).

However, the properties of a molecule are not only determined by the molecule's structure, but also by its environment, which should thus also be included in the calculations. One of the many examples where the environment tunes molecular properties is the vibrational frequency of the carbonyl group in acetophenone. This frequency depends critically on the electric field generated by its environment and thus varies in different solvents (**Paper VII**).

It is often necessary to use a molecular system of thousands of atoms to describe the influence of the environment on the properties of a molecule. A central problem in theoretical chemistry is to describe this molecular environment in an accurate way while keeping the computational cost manageable. One of the ways to do this is to describe the molecule by QM methods and its environment by more approximate and cheaper classical methods. Such a *multiscale* model is used in this thesis. It is often a rather small part of the total molecular system for which the properties need to be calculated. This is the case for an active site in a protein or a solute in an organic solvent and can be illustrated by the electronic excitation in the green fluorescent protein (GFP). Description of only the chromophore by QM methods is accurate enough as long as electrostatic and polarization effects in the rest of the protein are accounted for (**Paper II**). **Paper IV** describes how the molecular environment can be represented in an accurate way using classical parameters while keeping the computational cost low.

The required accuracy in the description of the molecular environment depends not only on the specific question and specific property under study, but also on the type of environment. As an example, the molecular system that is modeled in a calculation needs to be larger in a protein

than in a homogeneous solvent (**Paper III**). An additional challenge of large molecular systems is that also dynamic effects need to be taken into account. In this thesis, molecular properties are calculated in a sequential approach: first, molecular structures are obtained from e.g. a molecular dynamics (MD) simulation at a given temperature. Second, the properties of one or more of these structures are calculated.

The aim of this thesis is the accurate calculation of molecular properties in realistic environments. The emphasis is on the representation of the molecular environment rather than on the methods to calculate the molecular properties. With the exception of **Paper VI**, all papers are concerned with molecules in an explicit environment. The types of molecular environments considered are homogeneous solvents and proteins and the molecular properties are mainly restricted to absorption properties. The introductory chapters do not only serve as a background to the papers but also as an introduction to the computational procedure to calculate molecular properties of chemical systems. The text is written with a master student in Chemistry in mind and should as such be useful for a PhD student entering the field.

Different aspects of the calculation of molecular properties in realistic environments are discussed in the introductory chapters: methods to calculate the molecular properties (Chapter 1), methods to include a molecular environment in the calculations (Chapter 2), the modeling of the molecular structure (Chapter 3) and a classical parametrization of the environment, which we call an embedding potential (Chapter 4). Chapter 5 contains a short summary of the work, the main limitations in the used methods and a perspective on possible directions for future research.

Chapter 1

Electronic structure theory

The goal of this chapter is to introduce the QM methods used in this thesis to calculate molecular properties of closed-shell molecules and discuss some of the different factors determining the accuracy of such calculations. Section 1.1 will introduce important concepts from molecular quantum mechanics: the Schrödinger equation, Hartree–Fock (HF) theory, correlation methods and the calculation of molecular properties. Section 1.2 will introduce DFT and in particular different types of exchange–correlation functionals.

1.1 Molecular quantum mechanics

The central idea in molecular quantum mechanics is that the energy and properties of a system are determined by a wave function $\mathbf{\Psi}$. Section 1.1.1 will introduce the most fundamental equation in wave-function theory, namely the Schrödinger equation, as well as the important concepts of the Born–Oppenheimer approximation and the variational principle. Section 1.1.2 will introduce HF theory, basis sets and self-consistent field (SCF) theory. Section 1.1.3 will introduce ways to include electron correlation with particular emphasis on coupled-cluster (CC) theory.

1.1.1 The Schrödinger equation

The energy of a molecular system E can be obtained by acting on the wave function with the Hamiltonian operator $\mathcal{H}$ and solving the resulting eigenvalue problem, as stated by the time-independent Schödinger equation,

$$\mathcal{H}\mathbf{\Psi} = E\mathbf{\Psi}. \tag{1.1}$$

The total Hamiltonian of a molecular system may be written as[1]

$$\mathcal{H} = T_\mathrm{n} + T_\mathrm{e} + V_\mathrm{ne} + V_\mathrm{ee} + V_\mathrm{nn} \tag{1.2}$$

and contains the kinetic energy of the nuclei (T_n) and the electrons (T_e), the attraction between nuclei and electrons (V_ne) and the repulsion between electrons (V_ee) and between nuclei (V_nn). Two approximations are made at this point: relativistic effects are neglected and the nuclei are treated as point charges. These approximations are unproblematic for the molecular systems treated in this work. The operators for the kinetic energy are given as[1]

$$T_n = -\sum_{I=1}^{N} \frac{\nabla_I^2}{2m_I} \tag{1.3}$$

$$T_e = -\sum_{i=1}^{n} \frac{\nabla_i^2}{2} \tag{1.4}$$

with m_I the mass of nucleus I, N the number of nuclei and n the number of electrons and

$$\nabla_i^2 = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}. \tag{1.5}$$

Eqs. 1.3 and 1.4 are given in Hartree atomic units,[2,3] in which the magnitude of the electronic mass, elementary charge, reduced Planck constant and Coulomb constant are set to one. These units will be used throughout this thesis unless otherwise specified. The kinetic energy operator T with ∇^2 (Eq. 1.5) makes the Schrödinger equation in Eq. 1.1 a second-order differential equation.

The nucleus–electron attraction and electron–electron repulsion are given as

$$V_{\mathrm{ne}} = -\sum_{i=1}^{n}\sum_{I=1}^{N}\frac{Z_I}{|\boldsymbol{r}_i - \mathbf{R}_I|} \tag{1.6}$$

$$V_{\mathrm{ee}} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} \tag{1.7}$$

with Z_I the charge of nucleus I and $|\boldsymbol{r}_i - \mathbf{R}_I|$ the distance between electron i and nucleus I.

The mass of a particle appears in the denominator of the kinetic energy operators in Eqs. 1.3 and 1.4 (m_{e}=1). This means that the kinetic energy of electrons (m_e= $9.1{\cdot}10^{-31}$ kg) is around $1.8{\cdot}10^3$ times higher than the kinetic energy of protons and neutrons (m= $1.7{\cdot}10^{-27}$ kg). Thus, the electrons move much faster than the nuclei. This naturally leads to the Born–Oppenheimer approximation, which states that the motion of electrons and nuclei can be separated due to their different masses.[4] The implication of the Born–Oppenheimer approximation is that every nuclear geometry has its associated electronic energy, giving rise to the concept of the potential energy surface. The electronic wave function is thus a function of the electronic coordinates while depending parametrically on the nuclear coordinates: $\boldsymbol{\Psi}_e(\boldsymbol{r};\mathbf{R})$. The electronic Hamiltonian can be written in the Born–Oppenheimer approximation as[4]

$$\mathcal{H}_{\mathrm{e}} = T_{\mathrm{e}} + V_{\mathrm{ne}} + V_{\mathrm{ee}} + V_{\mathrm{nn}}. \tag{1.8}$$

The nucleus–nucleus attraction V_{nn} is a constant given a particular nuclear geometry.

The electronic Schrödinger equation can only be solved exactly for a system with one electron. In all other cases, approximations and numerical methods are needed to find the wave function $\boldsymbol{\Psi}$. One useful tool to find

the best approximation to the exact ground-state wave function $\mathbf{\Psi}$ is the *variational theorem*,[5] stating that any trial wave function has an energy equal to or higher than the ground-state energy. Thus, when parametrizing the wave function and minimizing the energy with respect to the set of parameters, one can converge towards the exact wave function.

1.1.2 Hartree–Fock theory

The variation theorem does not specify what the wave function should look like, thus any mathematically convenient form can in principle be chosen. One way to write the many-electron wave function $\mathbf{\Psi}_e$ is as a product of one-electron wave functions ψ_i,

$$\mathbf{\Psi}_e = \psi_1 \psi_2 ... \psi_n, \tag{1.9}$$

which is known as a Hartree product.[6] The Hartree product turns out to be a bad choice for the wave function because it violates the *Pauli exclusion principle*: the wave function must change sign when the coordinates of two electrons are interchanged.[6] A solution to this—originally introduced by Slater[7]—is to write the wave function as a determinant,

$$\mathbf{\Phi}_{\mathrm{HF}} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\boldsymbol{r}_1) & \psi_2(\boldsymbol{r}_1) & ... & \psi_n(\boldsymbol{r}_1) \\ \psi_1(\boldsymbol{r}_2) & \psi_2(\boldsymbol{r}_2) & ... & \psi_n(\boldsymbol{r}_2) \\ ... & ... & ... & ... \\ \psi_1(\boldsymbol{r}_n) & \psi_2(\boldsymbol{r}_n) & ... & \psi_n(\boldsymbol{r}_n) \end{vmatrix}, \tag{1.10}$$

which is the wave function used in HF theory.

Each of the one-electron wave functions ψ_i in Eq. 1.10 can be written as a linear combination of basis functions χ_α,[1]

$$\psi_i = \sum_{\alpha=1}^{M} c_{i\alpha} \chi_\alpha, \tag{1.11}$$

with M the total number of basis functions. The coefficients $c_{i\alpha}$ in Eq. 1.11 are called molecular orbital (MO) coefficients and are the parameters that can be minimized in a variational procedure. The basis functions are mathematical objects and the expansion in Eq. 1.11 is exact for an infinite number of basis functions. In practice, however, the choice of basis functions is guided by computational efficiency of integral calculations and convergence

of the properties with the number of basis functions.[1] These requirements are fulfilled by *Gaussian functions* placed on the nuclei, which are used throughout this thesis.

Fock was the first to combine Hartree's SCF procedure[2] with Slater's determinantal wave function.[7] In this way, electron i can be described by a one-electron operator F_i (called the *Fock operator*), which depends on the mean field of all other electrons $j \neq i$. The Fock operator contains a kinetic operator, nuclear attraction as well as Coulomb and exchange operators J_{ij} and K_{ij}, respectively,[1]

$$F_i = -\frac{\nabla_i^2}{2} - \sum_{I=1}^{N} \frac{Z_I}{|\boldsymbol{r}_i - \mathbf{R}_I|} + \sum_{j>i}^{n} (J_{ij} - K_{ij}) \tag{1.12}$$

with J_{ij} and K_{ij} leading to two-electron integrals over the basis functions χ_α and χ_β,

$$J_{ij} = \int\int \chi_\alpha(1)\chi_\beta(2) \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} \chi_\alpha(1)\chi_\beta(2) d\boldsymbol{r}_i d\boldsymbol{r}_j, \tag{1.13}$$

$$K_{ij} = \int\int \chi_\alpha(1)\chi_\beta(2) \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} \chi_\alpha(2)\chi_\beta(1) d\boldsymbol{r}_i d\boldsymbol{r}_j. \tag{1.14}$$

The Coulomb operator represents the classical repulsion between two electrons. The non-classical exchange operator arises from the antisymmetry requirement of the wave function. Comparing Eqs. 1.8 and 1.12, one notes that the Coulomb and exchange operators J_{ij} and K_{ij} together make up the electron–electron repulsion V_{ee}. This procedure leads to the HF equations,

$$\left[-\frac{\nabla_i^2}{2} - \sum_{I=1}^{N} \frac{Z_I}{|\boldsymbol{r}_i - \mathbf{R}_I|} + \sum_{j>i}^{n} (J_{ij} - K_{ij}) \right] \psi_i = \varepsilon_i \psi_i, \tag{1.15}$$

from which the one-electron functions ψ_i can be obtained. A problem, however, is that the Fock operator depends on all one-electron functions through the Coulomb and exchange terms. In other words, all one-electron functions ψ_j with $j \neq i$ are needed to calculate ψ_i, meaning that a *self-consistent field* procedure[2] is needed. Starting from an initial guess, new one-electron functions ψ_i are calculated every step using the HF equations in Eq. 1.15. This procedure continues until the change (measured by some suitable criterion) from one iteration to another is below a certain threshold. At that point the wave function is said to be converged.

1.1.3 Correlation methods

In the HF approach, electrons move in the mean field generated by all other electrons. In reality, however, the motion of the electrons is correlated. This section describes methods that go beyond the HF approach by introducing explicit electron correlation.

The HF equations allow for the calculation of the one-electron functions ψ_i, from which the total energy may be calculated. This procedure gives the lowest energy possible for one determinant $\mathbf{\Phi}_{\mathrm{HF}}$ given the basis set. To obtain a further lowering of the energy, additional Slater determinants $\mathbf{\Phi}$ can be added with associated coefficients a_i,

$$\mathbf{\Psi} = a_0\mathbf{\Phi}_{\mathrm{HF}} + a_1\mathbf{\Phi}_1 + a_2\mathbf{\Phi}_2 + ... \tag{1.16}$$

While the basis set determines the quality of the one-electron wave functions, the number and type of determinants in Eq. 1.16 (also called the *many-electron basis*) determines the quality of the description of electron correlation.[8] The expansion in Eq. 1.16 allows for systematic improvement of the wave function, which is an important advantage of wave-function methods over DFT (Section 1.2).

In the HF Slater determinant $\mathbf{\Phi}_{\mathrm{HF}}$ in Eq. 1.10, the electrons are placed in the orbitals with the lowest energies. For closed-shell systems around the equilibrium geometry, there is usually only one way of doing this. At a geometry further from equilibrium, other configurations may start to play a role, giving rise to *non-dynamical* (also called *static*) correlation.[8] This can be dealt with by including more than one ground-state configuration in Eq. 1.16 and optimizing simultaneously the coefficients of the determinants a_i in Eq. 1.16 and the MO coefficients $c_{i\alpha}$ in Eq. 1.11 for a given determinant. This approach is called *multi-configuration* SCF (MC-SCF).[9]

The remaining part of the correlation is called *dynamical* correlation[8] and is related to the $\frac{1}{\boldsymbol{r}_i - \boldsymbol{r}_j}$ term in the electron–electron repulsion V_{ee} (Eq. 1.7), which becomes singular for $\boldsymbol{r}_i = \boldsymbol{r}_j$. There is, however, no rigid separation possible between static and dynamical correlation.[8]

A systematic way to construct new determinants in Eq. 1.16 is to excite electrons from occupied to virtual orbitals starting from the HF determinant $\mathbf{\Phi}_{\mathrm{HF}}$. Exciting one electron gives rise to a singly-excited Slater

determinant (*single* excitations), exciting two electrons simultaneously gives rise to doubly-excited Slater determinants (*double* excitations), *etc.* If all possible excitations (up to n-fold excitations with n the number of electrons) are taken into account, the method is called *full configuration interaction* (full CI). Given the quick increase in the number of Slater determinants with the number of electrons n, full CI is not feasible but for the smallest systems. Truncation of the excitations at some order—such as CISD with all single and double excitations—leads to a variational procedure that however suffers from the lack of size-consistency, *i.e.*, the sum of the energy of two separate calculations on fragments A and B is not equal to the energy of system A+B.[1]

In CC theory,[10] the additional Slater determinants in Eq. 1.16 are obtained by operating with an exponential operator $e^{\mathbf{T}}$ on the HF wave function $\mathbf{\Phi}_{\mathrm{HF}}$ as[10]

$$\mathbf{\Psi}_{\mathrm{CC}} = e^{\mathbf{T}}\mathbf{\Phi}_{\mathrm{HF}} \tag{1.17}$$

$$= \left(\mathbf{1} + \mathbf{T} + \frac{1}{2}\mathbf{T}^2 + \frac{1}{6}\mathbf{T}^3 + ...\right)\mathbf{\Phi}_{\mathrm{HF}} \tag{1.18}$$

with the cluster operator $\mathbf{T}$ defined as

$$\mathbf{T} = \mathbf{T}_1 + \mathbf{T}_2 + ... + \mathbf{T}_n. \tag{1.19}$$

The k-electron excitation operator $\mathbf{T}_k$ excites k electrons from occupied (occ) orbitals i, j to virtual (virt) orbitals a, b according to[10]

$$\mathbf{T}_1\mathbf{\Phi}_{\mathrm{HF}} = \sum_{i}^{\mathrm{occ}}\sum_{a}^{\mathrm{virt}} t_i^a \mathbf{\Phi}_i^a, \tag{1.20}$$

$$\mathbf{T}_2\mathbf{\Phi}_{\mathrm{HF}} = \sum_{i<j}^{\mathrm{occ}}\sum_{a<b}^{\mathrm{virt}} t_{ij}^{ab} \mathbf{\Phi}_{ij}^{ab}, \tag{1.21}$$

with the coefficients t usually referred to as *amplitudes*. Eqs. 1.20 and 1.21 generate single excitations $\mathbf{\Phi}_i^a$ and double excitations $\mathbf{\Phi}_{ij}^{ab}$. The amplitudes t are determined from the CC equations, which arise by inserting Eq. 1.17 in the Schrödinger equation and projecting on excited-state determinants.[1] The CC energy can be obtained by inserting Eq. 1.17 in the Schrödinger

equation (Eq. 1.1) and projecting onto the HF determinant,

$$\begin{aligned}\langle \mathbf{\Phi}_{\mathrm{HF}}|\mathcal{H}e^{\mathbf{T}}|\mathbf{\Phi}_{\mathrm{HF}}\rangle &= E_{\mathrm{CC}}\langle \mathbf{\Phi}_{\mathrm{HF}}|e^{\mathbf{T}}\mathbf{\Phi}_{\mathrm{HF}}\rangle \\ &= E_{\mathrm{CC}},\end{aligned} \tag{1.22}$$

assuming intermediate normalization.

The excitation operators define a large number of new Slater determinants. To reduce the number of excited Slater determinants, a common approximation (also used in **Paper II**, **Paper V** and **Paper VI**) is the *frozen core approximation*, in which only valence electrons are excited. This reduces the sum over the occupied orbitals. A further approximation (investigated also in **Paper II**) is the *reduced virtual space* approach, in which the highest virtual orbitals (given by some threshold on the energy) are also neglected. This reduces the sum over the virtual orbitals.

In the exact treatment, the cluster operator contains all possible excitations, meaning that the sum in Eq. 1.19 is up to $\mathbf{T}_n$ with n the number of electrons. The cluster operator $\mathbf{T}$ is truncated at a given order in practical CC calculations, allowing for systematic improvement of the wave function. In coupled-cluster singles and doubles (CCSD), the cluster operator is $\mathbf{T}$=$\mathbf{T}_1$+$\mathbf{T}_2$, giving an exponential operator

$$e^{\mathbf{T}} = \mathbf{1} + \mathbf{T}_1 + \underbrace{\mathbf{T}_2 + \frac{1}{2}\mathbf{T}_1^2}_{\text{doubles}} + \underbrace{\mathbf{T}_2\mathbf{T}_1 + \frac{1}{6}\mathbf{T}_1^3}_{\text{triples}} + ... \tag{1.23}$$

It is clear from Eq. 1.23 that CCSD contains not only single ($\mathbf{T}_1$) and double ($\mathbf{T}_2 + \frac{1}{2}\mathbf{T}_1^2$) excitations, but also triple excitations (and quadruple excitations, *etc.*) that arise from a combination of $\mathbf{T}_1$ and $\mathbf{T}_2$ (so-called *disconnected* triples[6]). The definition of the exponential operator makes truncated CC methods size-consistent, which is an advantage over CI methods. A disadvantage of CC theory is that it is not a variational method.

The approximate coupled-cluster singles and doubles model (CC2)[11] is derived from CCSD by including only some of the contributions from double excitations and expressing the coefficients t_{ij}^{ab} in Eq. 1.21 in terms of the coefficients t_i^a in Eq. 1.20. The reduced number of amplitudes leads to a more favourable scaling with the number of basis functions in comparison to CCSD, allowing much larger molecular systems to be treated such as those in **Paper II** (OPA; N≤161) and **Paper V** (TPA; N=62) with N

the number of atoms. The advantage of CC2 over DFT (Section 1.2)—which is computationally more efficient and thus allows for larger molecular systems—is that it systematically improvable.

1.1.4 Molecular properties

The optimized wave function $\mathbf{\Psi}$ from e.g. HF theory (Section 1.1.2) can be used to calculate the energy and properties of a molecular system. We will in this section focus on the calculation of molecular properties from a variational wave function for exact-state theory. Similar strategies can be used for non-variational methods.[12]

For a normalized wave function, the energy can be obtained as the expectation value of the Hamiltonian operator by multiplying the Schrödinger equation (Eq. 1.1) on the left with $\mathbf{\Psi}^*$ and inserting the normalization condition $\langle\mathbf{\Psi}|\mathbf{\Psi}\rangle$=1,

$$E = \langle\mathbf{\Psi}|\mathcal{H}|\mathbf{\Psi}\rangle. \tag{1.24}$$

Many static molecular properties can be calculated from the response of the energy or wave function to some perturbation.[13] This perturbation can be e.g. an external static (electric or magnetic) field or a geometrical distortion of the molecule. Molecular properties can be formulated as derivatives of the energy with respect to a perturbation evaluated at zero perturbation strength.[13] The assumption is that the change in energy as a result of the perturbation is small in comparison to the total energy, so that the energy can be written as a Taylor expansion. For a general perturbation parameter λ we can write

$$E(\lambda) = E(\lambda{=}0) + \left(\frac{\partial E}{\partial \lambda}\right)_{\lambda=0} \lambda + \frac{1}{2}\left(\frac{\partial^2 E}{\partial \lambda^2}\right)_{\lambda=0} \lambda^2 + \ldots, \tag{1.25}$$

where $E(\lambda{=}0)$ is the energy of the unperturbed system and $\left(\frac{\partial E}{\partial \lambda}\right)_{\lambda=0}$ and $\left(\frac{\partial^2 E}{\partial \lambda^2}\right)_{\lambda=0}$ correspond to molecular properties that are first and second order in the energy, respectively.

We first consider the example of a nuclear displacement $dx{=}x{-}x_0$ as perturbation, which gives[13]

$$E(x) = E(x_0) + \left(\frac{\partial E}{\partial x}\right)_{x_0} dx + \frac{1}{2}\left(\frac{\partial^2 E}{\partial x^2}\right)_{x_0} dx^2 + \ldots, \tag{1.26}$$

where x is one of the Cartesian coordinates x, y or z for a given nucleus. The first-order change in the energy with a nuclear displacement is called the *gradient*

$$g_x = \left(\frac{\partial E}{\partial x}\right)_{x_0} \tag{1.27}$$

along Cartesian coordinate x and contains the negative force acting on a nucleus in that direction. Thus, the force F_x on that nucleus is given as

$$F_x = -g_x = -\left(\frac{\partial E}{\partial x}\right)_{x_0} \tag{1.28}$$

when evaluated at a reference geometry $x{=}x_0$. All $3N$ components (with N the number of atoms) of the molecular gradient are zero at the equilibrium geometry, which is a criterion used in geometry optimization (Section 3.2).

The second-order change in the energy with a nuclear displacement is called the *Hessian*

$$H_{xx} = \left(\frac{\partial^2 E}{\partial x^2}\right)_{x_0}. \tag{1.29}$$

The molecular Hessian has $(3N)^2$ components with N the number of atoms and is symmetric ($H_{xy}{=}H_{yx}$). The gradient and Hessian play a role in e.g. geometry optimizations (Section 3.2).

When the perturbation λ is a static electric field along Cartesian coordinate x, F_x, Eq. 1.25 becomes[5]

$$E(F_x) = E(F_x{=}0) + \left(\frac{\partial E}{\partial F_x}\right)_{F_x=0} F_x + \frac{1}{2}\left(\frac{\partial^2 E}{\partial F_x^2}\right)_{F_x=0} F_x^2 + \ldots. \tag{1.30}$$

The energy of molecules with a permanent dipole moment $\boldsymbol{\mu}$ and polarizability $\boldsymbol{\alpha}$ will be lowered as a result of the applied electric field and can be described phenomenologically as

$$E(F_x) = E(F_x{=}0) - \mu_x F_x - \frac{1}{2}\alpha_{xx} F_x^2 - \ldots. \tag{1.31}$$

Comparison of Eqs. 1.30 and 1.31 gives expressions for the components of the permanent molecular dipole moment $\boldsymbol{\mu}$,

$$\mu_x = -\left(\frac{\partial E}{\partial F_x}\right)_{F_x=0}, \tag{1.32}$$

and the components of the electric dipole–dipole polarizability $\boldsymbol{\alpha}$,

$$\alpha_{xx} = -\left(\frac{\partial^2 E}{\partial F_x^2}\right)_{F_x=0}. \tag{1.33}$$

The permanent dipole moment is thus the first-order change in energy as a result of applying an electric field. The dipole moment will also change as a result of the applied electric field. The difference is referred to as the *induced dipole moment*. Induced dipole moments and polarizabilities play in important role in the polarizable embedding (PE) method used in this thesis (Section 2.2).

There are different ways to calculate the energy and the derivatives. The derivative may be calculated either *analytically* or *numerically*.[13] The analytical approach requires a differentiable form of the energy expression and is often associated with considerable programming effort. The advantages—especially for the molecular gradient, where there are $3N$ perturbations—include greater speed and precision over the numerical approach.[13] In numerical differentiation, the energy is calculated explicitly for the perturbed molecular system. This requires calculations at many displaced molecular coordinates in the case of the gradient. The derivative can then be calculated by e.g. finite difference. The analytical molecular gradient is available within the PE framework as described in **Paper VII**.

The dependence of the Hamiltonian on the perturbation can be written as

$$\mathcal{H} = \mathcal{H}^{(0)} + \lambda\mathcal{H}^{(1)} + \lambda^2\mathcal{H}^{(2)} \tag{1.34}$$

where $\mathcal{H}^{(0)}$ is the unperturbed Hamiltonian (Eq. 1.2) and $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ are first and second order perturbation Hamiltonians. Knowledge of $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ is crucial because they allow the calculation of the energy of the perturbed system from knowledge of the unperturbed system.[5]

The first-order derivative of the energy can be obtained by differentiation of Eq. 1.24,

$$\frac{\partial E}{\partial \lambda} = \left\langle \frac{\partial \Psi}{\partial \lambda} | \mathcal{H} | \boldsymbol{\Psi} \right\rangle + \left\langle \boldsymbol{\Psi} | \frac{\partial \mathcal{H}}{\partial \lambda} | \boldsymbol{\Psi} \right\rangle + \left\langle \boldsymbol{\Psi} | \mathcal{H} | \frac{\partial \Psi}{\partial \lambda} \right\rangle. \tag{1.35}$$

The first and third terms on the right-hand side are the same for real wave functions.[1] The wave-function can depend on the perturbation through the basis functions χ and through wave function coefficients c[1]—e.g. MO coefficients (Eq. 1.11)—and can be written as

$$\frac{\partial \Psi}{\partial \lambda} = \frac{\partial \Psi}{\partial \chi}\frac{\partial \chi}{\partial \lambda} + \frac{\partial \Psi}{\partial c}\frac{\partial c}{\partial \lambda}. \tag{1.36}$$

The basis functions χ do not depend on the perturbation parameter for electric perturbations ($\frac{\partial \chi}{\partial \lambda}$=0), so the first term on the right-hand side in Eq. 1.36 is zero in that case. Moreover the energy is minimized with respect to all coefficients c ($\frac{\partial \Psi}{\partial c}$=0) for variational wave functions, so that Eq. 1.36 reduces to zero. Eq. 1.35 then reduces to the *Hellmann–Feynman* theorem[14,15]

$$\frac{\partial E}{\partial \lambda} = \langle \mathbf{\Psi} | \frac{\partial \mathcal{H}}{\partial \lambda} | \mathbf{\Psi} \rangle, \tag{1.37}$$

which thus only holds for variational wave functions. Letting the perturbation strength go to zero ($\lambda \gg \lambda^2$), we see that the first-order energy derivative can be obtained as the expectation value of $\mathcal{H}^{(1)}$ as

$$\left(\frac{\partial E}{\partial \lambda}\right)_{\lambda=0} = \langle \mathbf{\Psi} | \mathcal{H}^{(1)} | \mathbf{\Psi} \rangle. \tag{1.38}$$

Thus, knowledge of the unperturbed wave function $\mathbf{\Psi}$ can be used to compute first-order properties, *i.e.*, first-order derivatives of the energy. More generally, there is a $2n$+1 rule stating that a property (energy derivative) to order $2n$+1 requires the n'th derivative of the wave function.[5,12,16] This derivation does not hold for non-variational wave functions, but a similar rule can still be used with a Lagrangian approach.[12,16]

A component of the permanent electric dipole moment can be calculated using the Hellmann–Feynman theorem as an expectation value over the dipole moment operator in the unperturbed system,[5]

$$\mu_x = -\langle \mathbf{\Psi} | \mu_x | \mathbf{\Psi} \rangle. \tag{1.39}$$

using[5]

$$\mathcal{H}^{(1)} = -\boldsymbol{\mu} \tag{1.40}$$

with $\mathcal{H}^{(1)}$ defined in Eq. 1.34. Components of the electric dipole–dipole polarization tensor (Eq. 1.33) can be obtained in exact-state theory with $\mathcal{H}^{(1)}$ from Eq. 1.40 and $\mathcal{H}^{(2)}$=0 as[5]

$$\alpha_{ab} = -2 \sum_{f \neq 0} \frac{\langle 0|\mu_a|f\rangle\langle f|\mu_b|0\rangle}{\omega_{0f}}, \tag{1.41}$$

which is a sum over excited states f with excitation energies ω_{0f}=$E_0 - E_f$.

Until this moment, we have focused on time-independent perturbations such as nuclear displacements and static electric fields. Absorption properties result from the interaction between a molecule and an oscillating electromagnetic field. This requires the use of the time-dependent Schrödinger equation,[5]

$$\mathcal{H}\mathbf{\Psi}(t) = i\frac{\partial}{\partial t}\mathbf{\Psi}(t). \tag{1.42}$$

The calculation of molecular properties as derivatives of the energy cannot straightforwardly be applied here because of energy exchange between the molecule and the field.[17] An alternative is offered by *response theory*,[18] which is based on time-dependent perturbation theory. Since the magnitude of the applied electromagnetic radiation is much smaller than the local molecular electric field even for strong lasers, perturbation theory can be applied.[17]

The starting point for time-dependent perturbation theory is the time-dependent Schrödinger equation where the Hamiltonian $\mathcal{H}$ consists of a time-independent part $\mathcal{H}^{(0)}$ and a time-dependent perturbation $\mathcal{H}^{(1)}(t)$,

$$\mathcal{H} = \mathcal{H}^{(0)} + \mathcal{H}^{(1)}(t). \tag{1.43}$$

The time-dependent perturbation can be expressed using frequency components V^{ω} as[18]

$$\mathcal{H}^{(1)}(t) = \int_{-\infty}^{+\infty} V^{\omega} e^{-i\omega t} d\omega. \tag{1.44}$$

In order to calculate molecular properties, the expectation value of a time-independent operator $\mathcal{A}$ over a time-dependent wave function is expanded in orders of the perturbation as

$$\langle t|\mathcal{A}|t\rangle = \langle t|\mathcal{A}|t\rangle^{(0)} + \langle t|\mathcal{A}|t\rangle^{(1)} + \langle t|\mathcal{A}|t\rangle^{(2)} + ..., \tag{1.45}$$

where we write $|t\rangle$ for the time-dependent wave function $\mathbf{\Psi}(t)$ and where $\langle\mathbf{\Psi}(t)|\mathcal{A}|\mathbf{\Psi}(t)\rangle^{(1)}$=$\langle\mathbf{\Psi}^{(1)}(t)|\mathcal{A}|\mathbf{\Psi}(t)\rangle + \langle\mathbf{\Psi}(t)|\mathcal{A}|\mathbf{\Psi}^{(1)}(t)\rangle$, *etc.* The first term in the expansion is equal to the expectation value of the unperturbed system over the time-independent wave function[17]

$$\langle t|\mathcal{A}|t\rangle^{(0)} = \langle\mathbf{\Psi}|\mathcal{A}|\mathbf{\Psi}\rangle. \tag{1.46}$$

The terms that are first and second order in the perturbation can be written as Fourier transforms in terms of so-called *response functions*,[18]

$$\langle t|\mathcal{A}|t\rangle^{(1)} = \int \langle\langle\mathcal{A}; V^{\omega}\rangle\rangle_{\omega} e^{-i\omega t} d\omega \tag{1.47}$$

$$\langle t|\mathcal{A}|t\rangle^{(2)} = \frac{1}{2}\iint \langle\langle\mathcal{A}; V^{\omega_1}, V^{\omega_2}\rangle\rangle_{\omega_1,\omega_2} e^{-i(\omega_1+\omega_2)t} d\omega_1 d\omega_2, \tag{1.48}$$

where $\langle\langle\mathcal{A}; V^{\omega}\rangle\rangle_{\omega}$ and $\langle\langle\mathcal{A}; V^{\omega_1}, V^{\omega_2}\rangle\rangle_{\omega_1,\omega_2}$ are the *linear* and *quadratic* response functions, respectively.

The linear response function for electric dipole perturbations determines absorption properties and can in exact-state theory be written as[18]

$$\langle\langle\mu_a; \mu_b\rangle\rangle_{\omega} = \sum_{f\neq 0}\left(\frac{\langle 0|\mu_a|f\rangle\langle f|\mu_b|0\rangle}{\omega - \omega_{0f}} - \frac{\langle 0|\mu_a|f\rangle\langle f|\mu_b|0\rangle}{\omega + \omega_{0f}}\right), \tag{1.49}$$

where $|0\rangle$ is used for the ground state $\mathbf{\Psi}_0$, $|f\rangle$ for an excited state $\mathbf{\Psi}_f$ and ω_{0f}=$E_0 - E_f$ as previously. The linear response function in Eq. 1.49 is thus expressed in terms of eigenstates of the unperturbed system and is also called the frequency-dependent polarizability.[18] Eq. 1.49 reduces to the frequency-independent polarizability in Eq. 1.41 for ω_{0f}=0. Thus, response theory can also be used to calculate time-independent molecular properties.[18]

The linear response function in Eq. 1.49 has a pole for ω=ω_{0f}, *i.e.*, the response function diverges at the excitation energies ω_{0f} of the molecule. Thus, excitation energies can be calculated from the poles of the linear response function. Transition moments for absorption processes can be calculated from *residues* of the response functions. The residue of the linear response function at ω=ω_{0f} is defined as[18]

$$\lim_{\omega\to\omega_{0\mathrm{f}}} (\omega - \omega_{0\mathrm{f}})\langle\langle\mu_a; \mu_b\rangle\rangle_{\omega} \tag{1.50}$$

and gives the intensity of the OPA process as $\langle 0|\mu_a|f\rangle\langle f|\mu_a|0\rangle$. From this, the OPA transition moment S_a can be formulated as

$$S_a = \langle 0|\mu_a|f\rangle. \tag{1.51}$$

The single residue of the quadratic response function can be obtained in a similar way as[18]

$$\lim_{\omega_1 \to \omega_{0f}/2} (\omega_1 - \omega_{0f}/2) \langle\langle \mu_a; \mu_a, \mu_b \rangle\rangle_{\omega_1, \omega_2} \tag{1.52}$$

and can be used to derive the TPA transition moment S_{ab} as

$$S_{ab} = \sum_{n \neq 0} \left(\frac{\langle 0|\mu_a|n\rangle \langle n|\bar{\mu}_b|f\rangle}{\omega_{0n} - \omega_{0f}/2} + \frac{\langle 0|\mu_b|n\rangle \langle n|\bar{\mu}_a|f\rangle}{\omega_{0n} - \omega_{0f}/2} \right) \tag{1.53}$$

for the degenerate case $\omega_1{=}\omega_2{=}\omega_{0f}/2$. Here, $\langle n|\bar{\mu}_a|f\rangle$ is a difference dipole moment for $n{=}f$ ($\langle f|\mu_a|f\rangle - \langle 0|\mu_a|0\rangle$) and a transition moment between two excited states if $n{\neq}f$.

Response theory is used in this thesis to calculate excitation energies and one- and two-photon absorption strengths. Response theory gives calculated transition moments S_a and S_{ab} for a particular orientation of the molecule. Experiments, however, are carried out in isotropic samples, *i.e.*, comprising molecules in many different orientations. Rotational averaging over all possible orientations is thus necessary for a meaningful comparison between calculated and measured values. The rotationally averaged one- and two-photon transition strengths $\langle\delta\rangle$ can be obtained from the transition moments S_a and S_{ab} (with complex conjugates $\bar{S}_a$ and $\bar{S}_{ab}$) as[19]

$$\langle \delta^{\mathrm{OPA}} \rangle = \frac{1}{3} \sum_a S_a \bar{S}_a, \tag{1.54}$$

$$\langle \delta^{\mathrm{TPA}} \rangle = \frac{1}{15} \sum_{ab} \left(2 S_{ab} \bar{S}_{ab} + S_{aa} \bar{S}_{bb} \right), \tag{1.55}$$

with a and b Cartesian coordinates x, y or z.

The dimensionless oscillator strength f is often used for the OPA probability and can be calculated from the rotationally averaged OPA transition strength $\langle\delta^{\mathrm{OPA}}\rangle$ and the excitation energy ω as

$$f = 2\omega \cdot \langle \delta^{\mathrm{OPA}} \rangle. \tag{1.56}$$

The TPA cross section σ^{TPA} in *centimetre–gram–second* units can be obtained from the TPA strength $\langle\delta^{\mathrm{TPA}}\rangle$ in Eq. 1.55 as

$$\sigma^{\mathrm{TPA}} = \frac{N \pi^3 \alpha a_0^5 \omega^2}{c} g(\omega) \langle \delta^{\mathrm{TPA}} \rangle, \tag{1.57}$$

where N is an integer value (see **Paper VI** for the choice of N), α is the fine structure constant, a_0 the bohr radius, ω the photon energy, c the speed of light and $g(\omega)$ the lineshape function describing spectral broadening effects.

1.2 Density-functional theory

In DFT, it is the electron density ρ rather than the wave function $\mathbf{\Psi}$ that is used to compute the energy and properties of a system. Important milestones in the development of DFT are the Hohenberg–Kohn theorems[20] (Section 1.2.1) and Kohn–Sham (KS) theory[21] (Section 1.2.2). A critical factor in DFT calculations is the exchange–correlation functional, which is discussed in Section 1.2.3.

1.2.1 The Hohenberg–Kohn theorems

The idea behind DFT is to express the electronic energy $E[\rho(\boldsymbol{r})]$ as a functional of the electron density $\rho(\boldsymbol{r})$. The energy functional can be written as a sum of the kinetic energy $T[\rho]$, the electron–electron Coulomb repulsion $J[\rho]$, the electron–electron exchange repulsion $K[\rho]$ and the external energy $E_{\text{ext}}[\rho]$ as[1]

$$E[\rho] = T[\rho] + J[\rho] + K[\rho] + E_{\text{ext}}[\rho]. \tag{1.58}$$

The external energy $E_{\text{ext}}[\rho]$ contains the nuclear–electron attraction through the interaction of the electron density with the external potential $v_{\text{ext}}(\boldsymbol{r})$,

$$E_{\text{ext}}[\rho] = \int \rho v_{\text{ext}}(\boldsymbol{r}) d\boldsymbol{r} \tag{1.59}$$

with the external potential due to the nuclear–electron attraction given as

$$v_{\text{ext}}(\boldsymbol{r}) = -\sum_{I=1}^{N} \frac{Z_I}{|\mathbf{R}_I - \boldsymbol{r}|}, \tag{1.60}$$

summing over all nuclei I.

Apart from this nuclear–electron attraction in $E_{\text{ext}}[\rho]$ (Eq. 1.59), only the electron–electron repulsion $J[\rho]$ can be expressed exactly in the electron density ρ as[20]

$$J[\rho] = \frac{1}{2} \int \int \frac{\rho(\boldsymbol{r}_i)\rho(\boldsymbol{r}_j)}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} d\boldsymbol{r}_i d\boldsymbol{r}_j. \tag{1.61}$$

Attempts to express the kinetic energy and exchange energy as functionals of the electron density have been made since the 1920s, but with little success in chemistry because early models were not able to predict molecular binding.[22] The proof that the electronic energy can be uniquely determined by the electron density was given in 1964 by Hohenberg and Kohn.[20] They proved that it is impossible for two different external potentials to describe the same electron density, concluding that the electron density uniquely determines the Hamiltonian, the electronic energy and indeed all properties of a molecular system. Furthermore, Hohenberg and Kohn established a variational principle for the electron density, which gives a procedure to choose the best electron density by evaluating its energy. This is done by introducing a Lagrange multiplier μ, which ensures that the integral over the electron density sums up to the total number of electrons n in the system. The electron density with the lowest energy can then be found using the functional derivative

$$\frac{\delta}{\delta\rho}\left[E[\rho] - \mu\left[\int \rho d\boldsymbol{r} - n\right]\right] = 0, \tag{1.62}$$

which leads to the Euler–Lagrange equation[22]

$$\mu = v_{\text{ext}} + \frac{\delta F[\rho]}{\delta\rho}, \tag{1.63}$$

with v_{ext} defined in Eq. 1.60 and $F[\rho]$ the universal functional containing contributions from the kinetic energy and electron–electron interaction,[20]

$$F[\rho] = T[\rho] + J[\rho] + K[\rho]. \tag{1.64}$$

1.2.2 Kohn–Sham DFT

The Hohenberg–Kohn theorems state that the electron density can be used to determine the energy and properties of a molecular system, but does not give a form for the energy functional $E[\rho]$. An approach to calculate the electron density was formulated in 1965 by Kohn and Sham.[21] They proposed to substitute the exact but unknown kinetic energy functional $T[\rho]$ with the known kinetic energy functional for a system of non-interacting particles, $T_s[\rho]$. The relatively small correction to the kinetic energy ($T[\rho] - T_s[\rho]$) is then taken together with the exchange and the correlation parts

in the exchange–correlation functional $E_{\text{xc}}[\rho]$.[21] The energy functional $E[\rho]$ may then be written as

$$E[\rho] = T_s[\rho] + J[\rho] + E_{\text{ext}}[\rho] + E_{\text{xc}}[\rho]. \tag{1.65}$$

Kohn and Sham introduced one-electron functions ψ_i called KS orbitals from which the electron density can be calculated as

$$\rho = \sum_{i=1}^{n} |\psi_i|^2, \tag{1.66}$$

where the sum is over all occupied KS orbitals i. The kinetic energy of the non-interacting system can be calculated exactly using KS orbitals as

$$T_s[\rho] = \sum_{i=1}^{n} \langle \psi_i | -\frac{1}{2}\nabla_i^2 | \psi_i \rangle. \tag{1.67}$$

The KS orbitals can be obtained by solving the KS equations,[21]

$$\left[-\frac{\nabla_i^2}{2} + \int \frac{\rho(\boldsymbol{r}_j)}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} d\boldsymbol{r}_j + v_{\text{ext}}(\boldsymbol{r}_i) + v_{\text{xc}}(\boldsymbol{r}_i) \right] \psi_i(\boldsymbol{r}_i) = \varepsilon_i \psi_i(\boldsymbol{r}_i), \tag{1.68}$$

with the exchange–correlation potential $v_{\text{xc}}[\rho]$ defined as a functional derivative

$$v_{\text{xc}}[\rho] = \frac{\delta E_{\text{xc}}[\rho]}{\delta \rho}. \tag{1.69}$$

Note the similarity between the HF equations in Eq. 1.15 and the KS equations in Eq. 1.68. The three last terms on the left-hand side of Eq. 1.68 are collectively referred to as the *effective potential* or the KS potential,

$$v_{\text{eff}}(\boldsymbol{r}_i) = \int \frac{\rho(\boldsymbol{r}_j)}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} d\boldsymbol{r}_j + v_{\text{ext}}(\boldsymbol{r}_i) + v_{\text{xc}}(\boldsymbol{r}_i). \tag{1.70}$$

The KS approach thus gives a practical way of calculating the electron density from the KS orbitals similar to the calculation of HF orbitals (Section 1.1.2). First, a set of KS orbitals is chosen as an initial guess. Next, a new set of KS orbitals is obtained by solving Eq. 1.68 using an appropriate exchange–correlation functional $E_{\text{xc}}[\rho]$. This SCF procedure is repeated until convergence. The optimized electron density can then be calculated from

the occupied KS orbitals using Eq. 1.66. If the exact form of the exchange–correlation functional $E_{\mathrm{xc}}[\rho]$ were known, the KS approach would give the exact electronic energy including electron correlation. Various approximations to $E_{\mathrm{xc}}[\rho]$ have been proposed and the quest towards an exact density functional is an important area of research in theoretical chemistry. The density functionals that are most important to the work in this thesis will be presented in the next section.

1.2.3 Exchange–correlation functionals

Whereas a systematic way of improving DFT calculations to an exact limit does not exist, there is a range of different exchange–correlation functionals available. The choice of the density functional depends on e.g. the type of calculation and the desired accuracy. This section discusses the local density approximation (Section 1.2.3.1), the generalized gradient approximation (Section 1.2.3.2), hybrid functionals (Section 1.2.3.3) and long-range corrected functionals (Section 1.2.3.4).

1.2.3.1 Local density approximation

In the local density approximation (LDA), the exchange–correlation functional is chosen to depend only on the electronic density without taking its derivatives into account. The reason for this choice is that one can obtain an exact solution when choosing a uniform electron gas as a model system. The most common LDA approach is SVWN, which means using Slater's X_α approach (S) for the exchange energy[23] in combination with the correlation energy of Vosko, Wilk and Nusair (VWN).[24] Dirac obtained the exchange energy for a uniform electron gas,[25]

$$E_{\mathrm{x}}^{\mathrm{LDA}}[\rho] = C_{\mathrm{x}} \int \rho^{4/3} d\boldsymbol{r} \tag{1.71}$$

with $C_{\mathrm{x}} = -\frac{3}{4}(\frac{3}{\pi})^{1/3}$. In Slater's original formulation of the X_α approach,[23] the correlation part was neglected and Dirac's exchange energy was scaled to approximate the correlation energy. The VWN correlation functional was obtained using quantum Monte Carlo by subtraction of the kinetic and exchange energies from the total energies and interpolation from different densities to an analytical form.[24] LDA overestimates correlation energies

and therefore also bonding energies,[1,22] which makes it necessary to go beyond LDA for most chemical purposes.

1.2.3.2 Generalized gradient approximation

The generalized gradient approximation (GGA) adds information about the density gradient at a particular point to the functional by introducing the dimensionless reduced density gradient,[22,26]

$$x = \frac{\nabla\rho}{\rho^{4/3}}. \tag{1.72}$$

Perdew and Wang (PW86)[26] modified the LDA exchange functional to include x with different exponents and three constants a, b and c,[22]

$$E_{\mathrm{x}}^{\mathrm{PW86}}[\rho] = E_{\mathrm{x}}^{\mathrm{LDA}}[\rho]\left(1 + ax^2 + bx^4 + cx^6\right)^{1/15}, \tag{1.73}$$

with x defined in Eq. 1.72. A popular GGA exchange functional is Becke's one-parameter functional known as B88,[27]

$$E_{\mathrm{x}}^{\mathrm{B88}}[\rho] = \int \rho^{4/3}\left(C_x + \frac{\beta x^2}{1 + 6\beta x \cdot \operatorname{arcsinh}(x)}\right) d\boldsymbol{r}, \tag{1.74}$$

which yields an electron density with the proper asymptotic limit. The single parameter β was obtained by fitting to exchange energies of noble gas atoms.[27] When comparing the B88 exchange functional with Eq. 1.71, one can see that it is made up of $E_{\mathrm{x}}^{\mathrm{LDA}}[\rho]$ plus a correction that can be called $\Delta E_{\mathrm{x}}^{\mathrm{B88}}$.

Lee, Yang and Parr (LYP) proposed a GGA correlation functional with four parameters that were determined by fitting to the helium atom.[28] The B88 exchange and LYP correlation functionals are semi-empirical in the sense that they have parameters that are fit to experimental data. Alternatively, a functional can be *ab initio* by satisfying theoretically exact conditions.[22]

Also higher-order derivatives of the density can be introduced in the density functional, giving rise to what is known as a *meta*-GGA functional. This is usually accompanied by a large increase in the number of semi-empirical parameters.[22]

1.2.3.3 Hybrid exchange–correlation functionals

The performance of GGA functionals can be greatly improved by introducing a fraction of exact (*i.e.*, HF) exchange (Eq. 1.14), giving rise to *hybrid* functionals. The most common way to do this is using Becke's three-parameter (B3) form,[29]

$$E_{\mathrm{xc}}^{\mathrm{B3}} = aE_{\mathrm{x}}^{\mathrm{exact}} + (1-a)E_{\mathrm{x}}^{\mathrm{LDA}} + b\Delta E_{\mathrm{x}}^{\mathrm{B88}} + E_{\mathrm{c}}^{\mathrm{LDA}} + c\Delta E_{\mathrm{c}}^{\mathrm{GGA}}. \qquad (1.75)$$

In Becke's original work,[29] the PW91 correlation energy was used and the parameters were optimized to a=0.2, b=0.72 and c=0.81. The combination of Becke's three-parameter expression with the PW91 correlation functional is called B3PW91. Better known is the modification B3LYP by Stephens *et al.*,[30] which uses the LYP correlation energy as $E_{\mathrm{c}}^{\mathrm{GGA}}$ together with the same parameters a, b and c. Hybrid functionals typically give good geometries and perform well for most molecular properties.[6] Thus, B3LYP is in this thesis used for geometry optimizations in **Paper I**, **Paper II**, **Paper III**, **Paper V** and **Paper VII**, for frequency calculations in **Paper VII**, for electrostatic potentials (ESPs) in **Paper IV** and to calculate localized embedding parameters (see Chapter 4).

1.2.3.4 Long-range corrected functionals

Hybrid functionals are successful in describing short-range interactions between electrons, which is good enough for many molecular properties. For some properties, however, it is also important to describe long-range electron–electron interactions correctly. This can be done by including more exact exchange for long-range interactions using so-called *range separation*, in which the fraction of exact exchange gradually increases with distance at the cost of LDA exchange. A popular way to do this is proposed by Yanai *et al.*,[31] in which the Coulomb repulsion operator is split up in a short-range and a long-range part,

$$\frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} = \underbrace{\frac{1 - [\alpha + \beta \cdot \mathrm{erf}(\mu|\boldsymbol{r}_i - \boldsymbol{r}_j|)]}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}}_{\text{short range}} + \underbrace{\frac{\alpha + \beta \cdot \mathrm{erf}(\mu|\boldsymbol{r}_i - \boldsymbol{r}_j|)}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}}_{\text{long range}}. \qquad (1.76)$$

The parameter α controls the amount of short-range exact exchange and $\alpha+\beta$ is the amount of exact exchange at infinite separation. The original

CAM-B3LYP functional by Yanai *et al.* has α=0.19, β=0.46 and μ=0.33.[31] Introducing range separation gives a much better performance for excitation energy calculations involving charge-transfer transitions.[32] This is the reason the CAM-B3LYP functional is used to describe charge-transfer transitions in **Paper I**, **Paper III**, **Paper IV**, **Paper V** and **Paper VI**.

Chapter 2

Embedding methods

Chapter 1 has described how QM methods can be used to calculate molecular properties. The molecular systems that are treated in this thesis—proteins and organic molecules in homogeneous solvents—are however far too large to be described by QM methods at the level of accuracy needed for molecular property calculations. A solution to this problem is given by so-called focused embedding models: the central molecule or molecular fragment is described with an electronic structure method while its environment is described at a lower level of theory. The goal of this chapter is to introduce different embedding methods (Section 2.1) and in particular the PE model that is used in this thesis (Section 2.2). The PE model allows for mutual polarization between the central molecule and its environment and is therefore well-suited to describe the response of the environment to an electronic excitation in a chromophore.

2.1 Overview of embedding methods

The environment of a molecule or molecular fragment can be incorporated in a calculation in different ways. The environment can either be described *explicitly* by its atomic coordinates or *implicitly* as a dielectric medium. A prominent example of an implicit embedding method is the *polarizable continuum model* (PCM),[33] which is also used in **Paper VII**. Only the atomic coordinates of the molecule of interest are needed for molecular property calculations with implicit embedding methods. This usually requires a geometry optimization of the molecule, a procedure that is described in Section 3.2. The molecular modeling for explicit embedding methods is much more involved since the atomic coordinates of the whole molecular environment are needed. Chapter 3 describes some strategies to do this. In this chapter, we focus on explicit methods to incorporate the effect a molecular environment in a QM calculation.

In a so-called *cluster approach*, an enlarged molecular system is treated with one QM method. Alternatively, the environment can be incorporated classically[34] (quantum mechanics/molecular mechanics; QM/MM) or using another QM description (QM/QM).[35] The cluster approach is in principle the most exact. However, it is limited to rather small molecular systems due to the unfavourable scaling of QM methods with system size. It is demonstrated in **Paper II** for CC2 calculations on the GFP chromophore that the size of a molecular cluster is *by far* not large enough to sufficiently incorporate the complete effect of a protein environment. Thus, a lower-level approach is needed to describe the rest of the environment. In this thesis, molecular properties are calculated for a relatively small molecule in a large environment that affects the properties of the central molecule, making QM/MM methods a good choice. QM/MM methods are used in **Paper I**, **Paper II**, **Paper III**, **Paper IV**, **Paper V** and **Paper VII**.

The total energy of the system, $E_{\mathrm{QM/MM}}$, can be partitioned in contributions from the QM and classical regions as well as an interaction term,[34]

$$E_{\mathrm{QM/MM}} = E_{\mathrm{QM}} + E_{\mathrm{MM}} + E_{\mathrm{QM\text{–}MM}}. \tag{2.1}$$

Expressions for the energy of the QM region, E_{QM}, have been described in Chapter 1. The energy of the classical environment, E_{MM}, may be described

by force fields methods (Chapter 3). However, when the geometry of the classical region is frozen, E_{MM} is a constant and can be disregarded. The interaction energy between the QM and classical regions, $E_{\mathrm{QM–MM}}$, is the main focus of this chapter. In general, it can be written as[36]

$$E_{\mathrm{QM–MM}} = E^{\mathrm{bond}}_{\mathrm{QM–MM}} + E^{\mathrm{el}}_{\mathrm{QM–MM}} + E^{\mathrm{vdW}}_{\mathrm{QM–MM}}, \tag{2.2}$$

where $E^{\mathrm{bond}}_{\mathrm{QM–MM}}$ contains interactions between bonded atoms, $E^{\mathrm{el}}_{\mathrm{QM–MM}}$ contains classical electrostatic and polarization interactions and the van der Waals (vdW) term $E^{\mathrm{vdW}}_{\mathrm{QM–MM}}$ contains dispersion and exchange–repulsion interactions. Different ways to describe the intermolecular interactions give rise to different embedding methods. Intermolecular interactions can be divided into classical (electrostatic and polarization) and non-classical (dispersion and exchange–repulsion) interactions.[37] The main contribution to the exchange–repulsion interaction comes from the Pauli antisymmetry requirement.[38] This contribution is not included in all QM/MM schemes due to its non-classical and short-range character. Instead, the non-classical term in Eq. 2.2 is often neglected or treated classically, in which case it does not affect the electronic structure of the QM region and enters only at the energy level.[36] Different choices for $E^{\mathrm{bond}}_{\mathrm{QM–MM}}$ and $E^{\mathrm{el}}_{\mathrm{QM–MM}}$ are discussed below.

The easiest way to treat the bonding between the QM and classical regions ($E^{\mathrm{bond}}_{\mathrm{QM–MM}}$ in Eq. 2.2) is to avoid having a covalent bond through the boundary, in which case the interaction energy in Eq. 2.2 only contains electrostatic and vdW terms. This is why QM/MM calculations on solute–solvent systems are less problematic than calculations on proteins with a covalently bound region of interest. The location of the boundary should be chosen such that the perturbation of the chemical system is minimal. This means that the cut should be far enough away from the region of interest and preferably in a single bond. In the backbone of a protein, for instance, it is better to cut around a C_α than across an amide bond. Popular approaches to treat a covalent QM–MM boundary are the link atom approach and frozen orbitals.[36,39] In the link atom approach, hydrogen atoms that are not part of the molecular system are introduced as caps to the QM region to avoid dangling bonds. This introduces an extra charge very close to some of the atoms in the classical region, which

could easily cause overpolarization. One of the solutions to this is deleting or redistributing the electrostatic parameters in the affected part of the classical region.[36] The link atom approach is used in the PE model by Olsen *et al.*[40,41] and is therefore used in the PE calculations in **Paper I**, **Paper II** and **Paper III**. In the frozen orbital approach, the QM region is capped with an orbital that is kept frozen and localized to keep the link with the classical region stable. Even though the frozen orbital approach is theoretically more sound, it requires different orbitals for different chemical species and QM methods, requiring heavy parametrization and making it more challenging to implement and use.[36] Parameters for all 20 amino acids are available[42] in the frozen orbital approach by Friesner and co-workers,[43] enabling its use in the QM/MM geometry optimizations in **Paper I**, **Paper II**, **Paper III** and **Paper V**.

The electrostatic part of the QM–MM interaction ($E^{\text{el}}_{\text{QM–MM}}$ in Eq. 2.2) is the most important part because electrostatic interactions extend over a much longer range than bonded or vdW terms. In the simplest treatment, known as *mechanical embedding*,[36] the energy of the interaction is calculated classically at the same level as the energy of the classical region, E_{MM}, in Eq. 2.1. While this does affect the total energy of a system, it does not change the electronic structure of the QM region and is therefore of no consequence for the calculation of molecular properties.

The electronic charge distribution of the environment can also be directly included in the Hamiltonian in what is called *electrostatic embedding*,[36]

$$E^{\text{el}}_{\text{QM-MM}} = -\sum_{i=1}^{n}\sum_{s=1}^{S}\frac{q_s}{|\boldsymbol{r}_i - \mathbf{R}_s|} + \sum_{I=1}^{N}\sum_{s=1}^{S}\frac{q_s Z_I}{|\mathbf{R}_I - \mathbf{R}_s|}. \tag{2.3}$$

The first term contains the interaction between all electrons with position $\boldsymbol{r}_i$ and with charge -1 in the QM region and all sites with position $\mathbf{R}_s$ and charge q_s in the environment. This term is a one-electron operator in the Hamiltonian. The second term contains the interaction between all nuclei with position R_I and charge Z_I in the QM region and all sites in the environment. This term is a constant given a particular geometry.

Electrostatic embedding (Eq. 2.3) allows for the charge distribution of the environment to polarize (*i.e.*, affect the charge distribution of) the QM region but not the other way around. In *polarizable embedding*, the environ-

ment is also polarized by the QM region by adding additional polarization interactions.[36] A PE approach may have permanent electrostatic interactions that are described in a similar way to those from electrostatic embedding in Eq. 2.3. However, the charges used may well be different since polarization effects are often included implicitly in electronic embedding by scaling the charges (see Section 4.2). In others words, there should be a consistent treatment of permanent electrostatic and polarizable interactions. The importance of polarization was already emphasized in the first QM/MM study by Warshel and Levitt.[34]

An example of a mechanical embedding approach is the original formulation of the ONIOM scheme.[35] ONIOM is a so-called *subtractive* scheme,[36] meaning that the total QM/MM energy is calculated not by adding a term $E_{\text{QM–MM}}$ (Eq. 2.1), but instead adding the *difference* between the classical energy of the total system and the classical energy of subsystem I to the QM energy of the central subsystem,

$$E_{\text{QM/MM}} = E^{\text{I}}_{\text{QM}} + E^{\text{tot}}_{\text{MM}} - E^{\text{I}}_{\text{MM}}. \tag{2.4}$$

A clear advantage of a subtractive scheme is that there is no need for an explicit description of the QM–MM boundary. A disadvantage, however, is that even though the energy of the central region is modified to take into account its environment (and can thus be used in a geometry optimization), the electron density and hence most molecular properties are not affected. A modification of the ONIOM scheme has been presented that includes also the charge distribution of the environment in the Hamiltonian,[44] which allows for the calculation of molecular properties through electrostatic embedding.

An accurate and in principle exact way to incorporate the electrostatic environment is to describe it in terms of its frozen electron density, which is the central idea in *frozen density embedding*.[45] With a density-based description of the environment also short-range exchange–repulsion effects can be described at a QM level. Polarization between different subsystems can be introduced by optimizing not only the density of the central subsystem but also the density of other fragments in the environment in so-called *freeze-thaw-cycles*.[46,47] This treatment, however, quickly becomes very expensive when increasing the number of subsystems. Therefore, ground-state gas-phase densities are often used for at least part of the molecular system.[48]

Frozen density embedding potentials can also be constructed separately for ground and excited states with so-called *state-specific* embedding potentials, which allow for the description of differential polarization effects between two different electronic states.[49]

Other PE approaches include the polarization effects classically. One of the ways to do this is by using *induced dipoles*, which respond to the electric field from the surrounding and create an additional electric field themselves. Examples of this approach are the discrete reaction field by Jensen, van Duijnen and Snijders,[50] the PE model by Olsen *et al.*[40,41] (Section 2.2) and the MMpol model by Curutchet *et al.*[51]

In the *effective fragment potential* approach, classical electrostatic and polarization as well as non-classical exchange–repulsion effects from a fragment in the environment are modeled by effective potentials that are included in the Hamiltonian as one-electron operators.[52] These potentials also allow for the treatment of covalent bonds by frozen orbitals in a buffer region around the central subsystem.[52]

Frozen density embedding can be considered a QM/QM (or rather DFT/DFT) method since the environment is described by its electron density. The subtractive scheme used in the ONIOM method (Eq. 2.4) can also be used as a QM/QM method with the higher-level QM method used to calculate the energy of subsystem I ($E^I_{\text{QM-high}}$) and the lower-level QM method to calculate the energy of both subsystem I ($E^I_{\text{QM-low}}$) and the total molecular system ($E^{\text{tot}}_{\text{QM-low}}$). The effective fragment potential approach and the PE method discussed in the next section are strictly speaking QM/MM methods since they use classical parameters, which can however be derived from QM methods.

2.2 Polarizable embedding

In the PE method by Olsen *et al.*,[40,41] the environment is represented by a collection of classical sites that carry electric multipoles to represent the permanent electrostatics in addition to polarizabilities, which allow for polarization through the use of induced dipoles. Many-body polarization effects are included trough an SCF optimization of the induced dipoles, in which the classical region polarizes the QM region as well as the other way around.

The total energy of the embedded molecule is written as a sum of the energy of the isolated QM region, E_{QM}, and the energy contribution from the interaction between the QM region and its environment, $E_{\mathrm{QM–PE}}$,

$$E = E_{\mathrm{QM}} + E_{\mathrm{QM–PE}}. \tag{2.5}$$

Upon comparison with the more general Eq. 2.1, we note that $E_{\mathrm{QM–PE}}$ is the part of $E_{\mathrm{QM–MM}}$ that affects the embedded molecule and that the energy of the classical environment E_{MM} is not evaluated because it does not affect the energy and properties of the QM region. Any choice of QM method can in principle be used and the PE method has been implemented for HF,[40] SOPPA,[53] DFT,[40] CC2, CCSD and CCSDR(3),[54] RI-CC2[55] and MC-SCF.[56] The interaction energy between the QM region and its surroundings can further be divided into different contributions,

$$E_{\mathrm{QM–PE}} = E_{\mathrm{es}} + E_{\mathrm{ind}} + E_{\mathrm{LJ}}, \tag{2.6}$$

where E_{es} and E_{ind} are the electrostatic interaction energies of the quantum region with the permanent and induced charge distributions of the environment, respectively. E_{LJ} is the Lennard–Jones (LJ) interaction energy, which is an approximation to the dispersion and exchange–repulsion contributions.

For practical purposes, the electrostatic interaction energy E_{es} can be divided into the interaction energy of the static multipoles in the environment with the nuclear ($E_{\mathrm{mul,n}}$) and electronic ($E_{\mathrm{mul,e}}$) coordinates of the QM region,

$$E_{\mathrm{es}} = E_{\mathrm{mul,n}} + E_{\mathrm{mul,e}}, \tag{2.7}$$

where $E_{\mathrm{mul,n}}$ depends on the nuclear point charges and $E_{\mathrm{mul,e}}$ on the electron density of the QM region. In this thesis, the permanent multipoles include charges q, dipoles $\boldsymbol{\mu}$ and quadrupoles $\mathbf{Q}$. The interaction energy between the permanent multipoles in the environment and the nuclear charges in the QM region is a sum over all nuclei in the QM region and all classical sites,

$$E_{\mathrm{mul,n}} = \sum_{I=1}^{N} Z_I \sum_{s=1}^{S} \left[T_{Is}^{(0)} q_s - T_{Is,a}^{(1)} \mu_{s,a} + \frac{1}{2} T_{Is,ab}^{(2)} Q_{s,ab} \right] \tag{2.8}$$

with $\mu_{s,a}$ and $Q_{s,ab}$ Cartesian components of the dipole and quadrupole on site s and using Einstein's summation convention over repeated Cartesian

indices a and b. The interaction tensor $\mathbf{T}_{Is}^{(k)}$ of order k has one component for charges (k=0), three for dipoles (k=1) and nine for quadrupoles (k=2). Its components are given as[38]

$$T_{Is}^{(0)} = \frac{1}{|\mathbf{R}_I - \mathbf{R}_s|}, \tag{2.9}$$

$$T_{Is,a}^{(1)} = \nabla_a \frac{1}{|\mathbf{R}_I - \mathbf{R}_s|}, \tag{2.10}$$

$$T_{Is,ab}^{(2)} = \nabla_a \nabla_b \frac{1}{|\mathbf{R}_I - \mathbf{R}_s|}. \tag{2.11}$$

For the origin of the interaction tensor, see Section 4.1.1.1 on the electric multipole expansion.

The expression for the interaction between the permanent multipoles and the electrons i is similar,

$$E_{\text{mul,e}} = -\sum_{i=1}^{n}\sum_{s=1}^{S}\left[t_{is}^{(0)} q_s - t_{is,a}^{(1)}\mu_{s,a} + \frac{1}{2}t_{is,ab}^{(2)} Q_{s,ab}\right] \tag{2.12}$$

with a sum over all electrons instead of over all nuclei and

$$\mathbf{t}_{is}^{(k)} = \int \mathbf{T}_{is}^{(k)} \rho(\boldsymbol{r}_i) d\boldsymbol{r} \tag{2.13}$$

with ρ the electron density and the interaction tensor $\mathbf{T}_{is}^{(k)}$ given in Eq. 2.9. The total electrostatic energy E_{es} in the PE model is equivalent to Eq. 2.3 for simple point-charge electrostatic embedding when Eqs. 2.8 and 2.12 are truncated at k=0.

The induction energy (E_{ind} in Eq. 2.6) includes the polarization of the environment by itself and by the QM region. This polarization is truncated at first order in the current version of the PE model, leading to[40]

$$E_{\text{ind}} = -\frac{1}{2}\sum_{s=1}^{S} \boldsymbol{\mu}_s^{\text{ind}} \cdot \left(\mathbf{F}_{\text{mul}}^s + \mathbf{F}_{\text{e}}^s + \mathbf{F}_{\text{n}}^s\right), \tag{2.14}$$

where $\boldsymbol{\mu}_s^{\text{ind}}$ contains the x-, y and z-components of the induced dipole on site s in the environment. $\mathbf{F}_{\text{mul}}^s$, $\mathbf{F}_{\text{e}}^s$ and $\mathbf{F}_{s,\text{n}}$ contain the x-, y and z-components of the electric field at site s due to the multipole moments in the environment and the electrons and nuclei in the QM region, respectively. This linear response to the field $\mathbf{F}$ does not necessarily hold for high

electric fields. In principle, one can use hyperpolarizabilities to describe the quadratic response to the field and higher-order polarizabilities for even higher responses, but this is not made use of in the current version of the PE model. The induced dipole $\boldsymbol{\mu}_s^{\text{ind}}$ on site s can be calculated as

$$\boldsymbol{\mu}_s^{\text{ind}} = \boldsymbol{\alpha}_s \mathbf{F}_{\text{tot}}^s \tag{2.15}$$

with $\boldsymbol{\alpha}_s$ being the anisotropic dipole–dipole polarizability of site s, which is a symmetric 3×3 matrix with six unique components. One can in principle extend the description to include also dipole–quadrupole polarizabilities[57] to improve the description of the linear response to the electric field. $\mathbf{F}_{\text{tot}}^s$ in Eq. 2.15 is the total electric field at site s, containing contributions from the electrons and nuclei in the QM region and the static multipoles and other induced dipoles in the environment,

$$\mathbf{F}_{\text{tot}}^s = \mathbf{F}_{\text{e}}^s + \mathbf{F}_{\text{n}}^s + \mathbf{F}_{\text{mul}}^s + \mathbf{F}_{\text{ind}}^s. \tag{2.16}$$

Alternatively, the induced dipole at site s can be calculated from a scalar isotropic polarizability α_s^{iso} as

$$\boldsymbol{\mu}_s^{\text{ind}} = \alpha_s^{\text{iso}} \mathbf{F}_{\text{tot}}^s. \tag{2.17}$$

Since the induced dipole at every site s depends on the total electric field at site s, $\mathbf{F}_{\text{tot}}^s$ (see Eqs. 2.15 and 2.17), which in turn depends on the induced dipoles on other sites (Eq. 2.16), the induced dipoles need to be obtained from an SCF procedure.[40] The problem can then be rewritten as a matrix–vector equation to calculate all induced dipoles in the environment, $\boldsymbol{\mu}_{\text{ind}}$, as[40]

$$\boldsymbol{\mu}_{\text{ind}} = \mathbf{B} \left(\mathbf{F}_{\text{mul}} + \mathbf{F}_{\text{e}} + \mathbf{F}_{\text{n}} \right), \tag{2.18}$$

where $\boldsymbol{\mu}_{\text{ind}}$ is a vector containing all $3S$ induced dipoles and $\mathbf{B}$ the classical linear response matrix with dimension $3S \times 3S$, which is defined as[40]

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\alpha}_1^{-1} & -\mathbf{T}_{12}^{(2)} & \ldots & -\mathbf{T}_{1S}^{(2)} \\ -\mathbf{T}_{21}^{(2)} & \boldsymbol{\alpha}_2^{-1} & \ldots & -\mathbf{T}_{2S}^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ -\mathbf{T}_{S1}^{(2)} & -\mathbf{T}_{S2}^{(2)} & \ldots & \boldsymbol{\alpha}_S^{-1} \end{pmatrix}^{-\mathbf{1}} \tag{2.19}$$

and contains the polarizabilities $\boldsymbol{\alpha}_s$ at all sites and the interaction tensors between all pairs of induced dipoles $\mathbf{T}_{ij}^{(2)}$ with $\mathbf{T}_{ij}^{(k)}$ defined in Eq. 2.9. The optimized set of induced dipoles depends on the electron density or wave function of the QM region through $\mathbf{F}_\mathrm{e}$ in Eq. 2.18.

The LJ interaction energy (E_LJ in Eq. 2.6) can be calculated as a 12–6 potential as[58]

$$E_\mathrm{LJ} = \sum_{s=1}^{S} \sum_{I=1}^{N} \epsilon_{Is} \left[\left(\frac{r_{Is}}{|\mathbf{R}_I - \mathbf{R}_s|} \right)^{12} - 2 \left(\frac{r_{Is}}{|\mathbf{R}_I - \mathbf{R}_s|} \right)^{6} \right], \tag{2.20}$$

where the summation runs over all interactions between QM nuclei and classical sites and where ϵ_{Is} is the well depth, r_{Is} is the equilibrium bond length and $\mathbf{R}_I$ and $\mathbf{R}_s$ are the coordinates of the atoms in the QM and classical region, respectively. The parameters ϵ and r need to be specified for every atom in the QM region and classical regions. The interaction parameters ϵ_{Is} and r_{Is} in Eq. 2.20 are computed by Berthelot's combination rule[59] for ϵ_{Is} and Lorentz' combination rule[60] for r_{Is},

$$\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}, \tag{2.21}$$

$$r_{ij} = \frac{r_i + r_j}{2}. \tag{2.22}$$

The LJ energy in Eq. 2.20 depends only on nuclear coordinates and is therefore of no importance to the calculation of the electron density and hence to molecular properties calculated from it.[40] It does, however, play a role in geometry optimization using the PE model (**Paper VII**). It is important to bear in mind that the LJ energy is a rather crude approximation to the real exchange–repulsion and dispersion effects of the environment, as also discussed in **Paper VII**. A much better description is given in polarizable density embedding (PDE), in which the electrostatic part of the closest environment is described by an exact ground-state electron density.[61] This allows not only for density-based non-electrostatic repulsion, but also for an improvement of the electrostatic interaction energy E_es.

The QM–PE contribution can be included in the KS equations (Eq. 1.68) by adding an operator $v_\mathrm{QM–PE}$ to the effective potential (Eq. 1.70) with all terms that depend on the electron density. Alternatively, a similar term can

be introduced in the HF equations in Eq. 1.15. Thus, $v_{\text{QM–PE}}$ includes the electronic part of the electrostatic interaction (Eq. 2.12) and the interaction through induced dipoles, both of which enter as a one-electron operator in the Hamiltonian. The nuclear part of the electrostatic interaction (Eq. 2.8) and the LJ interaction (Eq. 2.20) are independent of the electron density and therefore not included in $v_{\text{QM–PE}}$. The electron density (or, equivalently, wave function) is then optimized in an SCF procedure. The induced dipoles are optimized at every step, ensuring mutual polarization between the QM region and its environment.

In order to evaluate the different energy contributions to $E_{\text{QM–PE}}$ (Eq. 2.6), the classical region needs to be defined using a set of parameters, collectively referred to as the *embedding potential*. The coordinates of all sites are required for all energy contributions. Electric multipoles of order $k \leq K$ are needed in the calculation of the electrostatic energy E_{es} in Eqs. 2.8 and 2.12. Polarizabilities $\boldsymbol{\alpha}$ are needed for the calculation of the induced dipoles and can be either anisotropic (Eq. 2.15) or isotropic (Eq. 2.17). LJ parameters r and ϵ (for E_{LJ} in Eq. 2.20) for all atoms are only required if nuclear displacements are involved, e.g. in a geometry optimization (**Paper VII**). In this thesis, the parameters are put exclusively on atoms in the classical region, whereas it is also possible to place parameters on e.g. bond midpoints.[40] The collection of sites s does not need to be the same for the different contributions. It is therefore possible to place polarizabilities only on a subset $S_1 \leq S$ of the atoms in the classical region (**Paper III**) or to truncate the electric multipoles at a different order K for different parts of the classical region (**Paper IV**). Chapter 4 describes how the parameters for the classical region can be obtained.

Chapter 3

Molecular modeling

Molecular modeling in the context of this thesis is understood as all steps required to obtain the molecular structures used in the calculation of molecular properties. For QM/MM calculations, this includes both the central subsystem and the classical region around it. This chapter aims to describe the methods used for molecular modeling in this thesis and will therefore focus on proteins (**Paper I**, **Paper II**, **Paper III** and **Paper V**) and small molecules in solvents (**Paper III**, **Paper IV** and **Paper VII**) using examples from those works. Since molecular properties are as a rule very sensitive to the particular geometry of the molecule under study, comparison between different methods is best done with *exactly* the same molecular structure. Thus, the molecular structures in (part of) **Paper II** and **Paper VI**—where the performance of different theoretical methods is compared—are taken from other works. Section 3.1 describes the evaluation of the forces acting on a molecular system using classical force fields. These forces are fundamental in both energy minimizations and MD simulations, which will be treated in Sections 3.2 and 3.3, respectively. Section 3.4 describes the extra steps required to prepare the molecular structure of a protein starting from a crystal structure.

3.1 Classical force fields

The set of forces acting on each atom in a molecular system plays an important role in molecular modeling, in particular to minimize the energy of a structure (Section 3.2) and to propagate a molecular system in time (Section 3.3). The forces can be obtained by evaluating the molecular gradient, which is the first-order change of the energy with respect to a nuclear displacement (Section 1.1.4). The energy expression can be evaluated in different ways. The QM calculation of the energy has been described in Chapter 1 and is the most accurate. For large molecular systems such as the proteins or solute–solvent systems, it quickly becomes too expensive to take into account the electronic structure. An alternative and much faster way of evaluating the energy of and forces on a molecular system is given by molecular force fields. The two methods can also be combined using the QM/MM calculations described in Chapter 2.

A *force field* is a mathematical expression for the energy of a molecular system as a function of the nuclear coordinates, $E(\mathbf{R})$, together with a set of parameters to describe the interactions between the atoms. Thus, all electronic effects are neglected, making force fields purely classical. The dependence of the energy on $\mathbf{R}$ is not shown explicitly in the following. A fundamental basis of force fields is the empirical observation that interaction between similar atoms can be described with similar parameters, *i.e.*, the parameters are *transferable*. This means that a limited set of parameters can describe a large set of molecules. The functional expression of most common force fields is pairwise additive, meaning that the total energy is obtained by summing over interaction between pairs of atoms.

Every force field for molecules contains at least an expression for bond stretching, angle bending, dihedral torsion, electrostatic and vdW interactions,

$$E = \underbrace{E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}}}_{\text{bonded terms}} + \underbrace{E_{\text{el}} + E_{\text{vdW}}}_{\text{non-bonded terms}} . \tag{3.1}$$

Specific force fields may have extra terms such as a hydrogen bonding energy.[62] Hydrogen bonding can however also be modeled by appropriate parametrization of the other non-bonded terms. The forces can be obtained by differentiation of Eq. 3.1 with respect to the nuclear coordinates $\mathbf{R}$ as shown in Eq. 1.28. The most common functional form of each of the

terms in Eq. 3.1 is discussed below with reference to three common force fields: OPLS,[63–65] AMBER[66] and CHARMM.[62] The OPLS force field is partially based on the AMBER force field and was originally introduced for liquid simulations,[63,64] which is why it is used for the solute–solvent MD simulations in **Paper III**, **Paper IV** and **Paper VII**. The CHARMM force field is used for GFP in **Paper I** because a parametrization of GFP was available for that force field.[67]

The bond stretching and angle bending energies E_{bond} and E_{angle} are usually modeled as harmonic potentials,

$$E_{\text{bond}} = \sum_{\text{bonds}} \frac{1}{2} k_b \left(R - R_0\right)^2 \tag{3.2}$$

$$E_{\text{angle}} = \sum_{\text{angles}} \frac{1}{2} k_\theta \left(\theta - \theta_0\right)^2 \tag{3.3}$$

with k_b and k_θ being force constants, R and θ the bond length and angle as calculated from the structure, and R_0 and θ_0 the equilibrium bond length and angle. The 'equilibrium' values for R_0 and θ_0 do not necessarily represent the bond length and angle at the equilibrium structure, but instead constitute a set of parameters that leads to the equilibrium structure when used in combination with the rest of the force field.[1] The harmonic approximation is simple, but not valid at larger displacements. Indeed, bond breaking cannot be described when the energy increases quadratically with the distance between two atoms. Improvement is possible to give the right dissociation energy (the energy at infinite separation) in a *Morse potential*, which adds one more parameter. Most common force fields such as AMBER, OPLS and CHARMM, however, use harmonic potentials because they are computationally more efficient and because displacements from the equilibrium geometry are small at room temperature.[62]

The dihedral energy often requires periodicity to be included in the energy expression. For instance, rotating one methyl group in ethane around the dihedral gives three minima (and three maxima) in the energy at 120° separation. Periodicity may be accomplished by a sum of cosine func-

tions,[62,66]

$$E_{\text{dihedral}} = \sum_{\text{dihedrals}} \frac{1}{2} \sum_{n=1}^{N} [1 \pm V_n \cos(n\omega)] \tag{3.4}$$

with ω the dihedral angle and V_n the barrier for rotating around the dihedral with periodicity $\frac{360^{\text{o}}}{n}$. The Fourier series in Eq. 3.4 is in principle able to describe any torsional profile exactly using an unlimited number of terms N. In practice, however, the number of terms is limited and the fit is far from exact. The dihedral potentials in AMBER, OPLS and CHARMM are all variations of Eq. 3.4. The number and type of terms n to include depends on the type of dihedral to be modeled. For instance, the dihedral term in ethane can be modeled with N=3 and V_1=0 and V_2=0 to give a periodicity of 120°. *Proper* dihedrals consist of four atoms that are linked through bonds. Alternatively, *improper* dihedrals may be defined in a similar manner to impose a penalty on non-planar geometries around a sp^2-hybridized atom.[39] Improper dihedrals are used with the OPLS force field in **Paper III**, **Paper IV** and **Paper VII** to keep the geometries of several aromatic molecules (uracil, acetophenone, benzene, phenol, toluene) planar. In CHARMM (and hence in **Paper I**), improper dihedrals are defined as harmonic potentials.[62]

The electrostatic energy E_{el} between nuclei I and J is usually modeled as a Coulomb interaction between their partial charges q_I and q_J,

$$E_{\text{el}} = \frac{q_I q_J}{\varepsilon |\mathbf{R}_I - \mathbf{R}_J|}, \tag{3.5}$$

with ε an effective dielectric constant. The total electrostatic energy is obtained by summing over all pairs of atoms, thus neglecting many-body polarization effects.[39] The partial charges are often determined by fitting to the electrostatic potential (ESP) of equilibrium geometries as explained in Chapter 4. Average polarization effects are usually included implicitly through increased values for the partial charges. The partial charges often sum up to zero for a molecular fragment (a so-called *charge group*), which allows for transferability of the parameters from one molecule to another. For instance, the partial charges in *para*-nitroaniline (PNA) in **Paper IV** sum up to zero separately for the nitro group, the amine group and each CH group as shown in Figure 3.1.

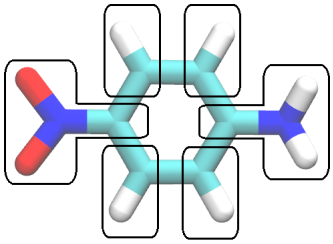


Figure 3.1. The molecular structure of PNA with the charge groups used in **Paper IV** shown in black shapes. Carbon atoms are shown in cyan, nitrogen in blue, oxygen in red and hydrogen in white. The molecular structure is made with VMD.[68]

The charges that are obtained in this way can be improved upon for a specific molecule at the cost of the transferability of the parameters.

The vdW energy E_{vdW} between atoms I and J is usually modeled as an LJ potential (*cf.* Eq. 2.20),[58]

$$E_{\mathrm{LJ}} = \epsilon_{IJ}\left[\left(\frac{r_{IJ}}{|\mathbf{R}_I - \mathbf{R}_J|}\right)^{12} - 2\left(\frac{r_{IJ}}{|\mathbf{R}_I - \mathbf{R}_J|}\right)^{6}\right], \tag{3.6}$$

with ϵ_{IJ} defined in Eq. 2.21 and r_{IJ} defined either as in Eq. 2.22 or as $r_{IJ}{=}\sqrt{r_I r_J}$. The LJ potential has the correct R^{-6} scaling for the attractive part, while the repulsive part scales as $(R^{-6})^2$ for computational efficiency. Improvements on the repulsive part lead to more parameters and higher computational cost, but can also lead to improvements for systems where the contribution of the vdW energy is high.[1] The LJ parameters in the OPLS force field were developed by evaluating *macroscopic* characteristics, namely the density and enthalpy of vaporization of organic molecules from Monte Carlo simulations.[63] This successful approach was later taken over in the AMBER force field.[66]

The total non-bonded energy can be calculated by summing Eqs. 3.5 and 3.6 over all atom pairs in the molecular system. Interactions between bonded atoms (1,2-interactions) and between atoms two bonds away from each other (1,3-interactions) are excluded from this sum, since their interactions are already described by E_{bond} (Eq. 3.2) and E_{angle} (Eq. 3.3). 1,4-interactions are usually scaled down by a factor between 0.5 and 1.[1]

For all but the smallest systems, the non-bonded interactions use most of the computational time in the energy calculation.[39] Indeed, the number of bonded interactions in a molecular system scales more or less linearly with the number of atoms, while the number of non-bonded interactions increases roughly quadratically.[39] To overcome this scaling problem and to allow for large molecular systems to be used, the non-bonded interactions are only calculated explicitly up to a certain threshold. Figure 3.2 shows the decay of a Coulomb interaction between two charges (R^{-1}) and of the London dispersion interaction between two induced dipoles (R^{-6}) at a distance R.

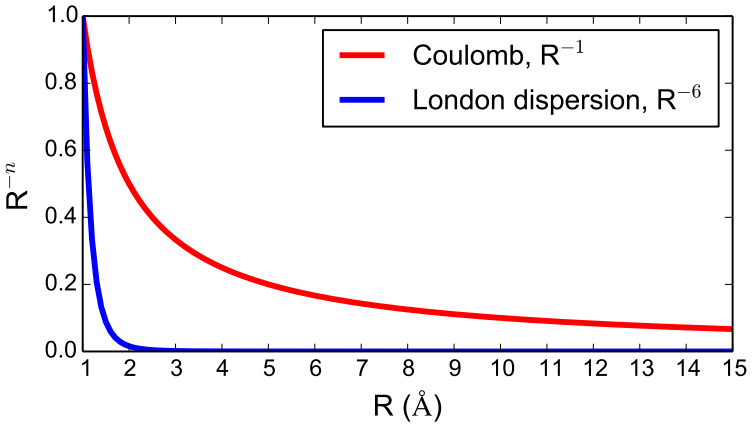


Figure 3.2. Decay of Coulomb and London dispersion interactions as a function of the distance between two particles R.

It is clear that the truncation of electrostatic interactions at a distance of 10 or 15 Å leads to a severe error, while the London dispersion interaction (and thus also the repulsive part of the vdW interaction, which decays even more quickly) is effectively zero at those points. This is the reason why only electrostatic interactions need to be treated beyond the non-bonded cut-off. This can for instance be done with the particle mesh Ewald approach,[69] which scales more favourably than the explicit calculation of all atom pairs. It is important to avoid calculating the distance between all atom pairs every time the energy is evaluated. Indeed, if the distances between all atom pairs needed to be calculated at every step to check whether they are inside or outside the non-bonded threshold, having this threshold would hardly speed up to calculation. A solution to this is the *Verlet neighbour list*, which

stores all atom pairs that are within or just outside the non-bonded cut-off and which is updated after a number of steps.[70] Only distances between pairs of atoms in the neighbour list are evaluated at every step, leading to a significant decrease of the number of interactions to calculate. The choice of the correct update frequency is the key to accurate and efficient calculation of non-bonded interactions.[39] Efficient calculation of the energy is important because energy minimization (Section 3.2) or MD (Section 3.3) require a large number of energy evaluations.

3.2 Energy minimization

The potential energy of a molecular system can be reduced by displacing the atomic coordinates along the force acting on each nucleus. The energy can be expression in different ways, such as with QM methods (Chapter 1) or using molecular force fields (Section 3.1). This section discusses different algorithms to displace the molecular structure to minimize the potential energy. In the context of this thesis, the minimization algorithms can be divided into *classical minimization* to prepare a molecular system for an MD simulation (Section 3.2.1) and *QM geometry optimization* to obtain an equilibrium structure for a QM calculation (Section 3.2.2).

3.2.1 Classical energy minimization

The goal of classical energy minimization is often to reduce the potential energy of a starting structure before performing an MD simulation using the same force field that will be used in the MD simulation. The starting structure might have specific bond lengths, angles or dihedrals that lead to a high energy for the given force field parametrization. Since kinetic energy will be introduced in the subsequent molecular energy simulation at a finite temperature—leading to a redistribution of the energy over kinetic and potential energy—it is not important to reach an absolute minimum in the potential energy.

The simplest way to achieve a lowering of the energy is to take a step along the negative molecular gradient, *i.e.*, following the force acting on each nucleus. This approach of minimizing the energy is known as the *steepest descent* method. The force vector $\mathbf{F}$ of length $3N$ contains one value for

each of the Cartesian components for each atom in the molecular system (Eq. 1.28). One way to do the minimization is to take one step with a certain step size and then re-evaluate the force again. The step size can either be fixed or changed depending on the force vector. Another way to do the minimization is to evaluate the energy for different step sizes and take the step that leads to the lowest energy, in which case the approach is referred to as a *line search*. Figure 3.3 shows the energy per molecule in a box of 1000 ethanol molecules during a steepest descent minimization (blue line) as a function of the step. The step size in the example is 5 pm for the direction with the largest force and scaled according to the magnitude of the force for all other coordinates.[71]

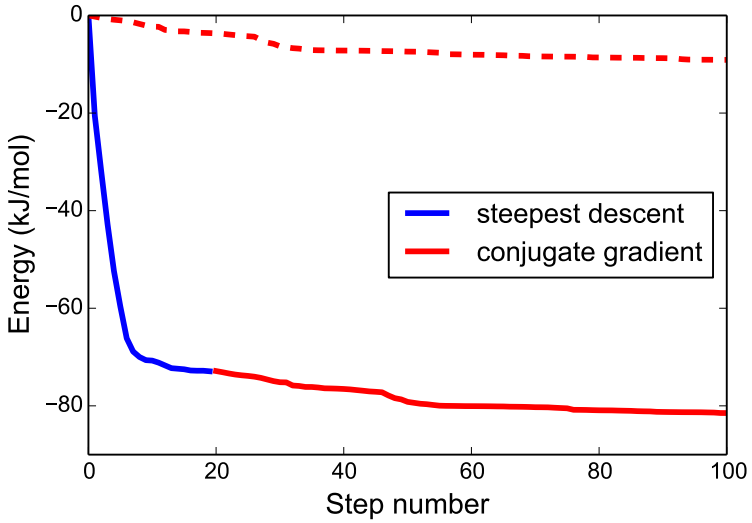


Figure 3.3. Potential energy per molecule in a box of 1000 ethanol molecules during classical minimization with 20 steps steepest descent (in blue) followed by 80 steps conjugate gradient (in red) or 100 steps of conjugate gradient (red, dashed). The energy of the starting structure is set to 0 kJ/mol.

Figure 3.3 illustrates how the steepest descent algorithm is very efficient in minimizing the energy in the first steps (far from the minimum, *i.e.*, in the *global region*), but slows down after a while (closer to the minimum, *i.e.*, in the *local region*). Indeed, the energy flattens out already after 7 steps in the example in Figure 3.3. Every step in the steepest descent is partially

reverting the previous since the direction is chosen again for every step.[1] Different algorithms can solve this problem. One of the simplest approaches is the *conjugate gradient* approach, in which the direction of the step is taken by considering not only the gradient of the current step, but also the gradient of the previous one. Figure 3.3 illustrates how the conjugate gradient algorithm (red line) is able to minimize the energy further from the point where the steepest descent algorithm almost has a stationary energy. However, using the conjugate gradient method from the starting point far from the minimum (red dashed line) leads to a much slower convergence compared to starting with the steepest descent method (blue line), which is why a combination of the two has been chosen for the classical minimization of solvent systems in **Paper III**, **Paper IV** and **Paper VII**: 20 steps of steepest descent followed by 1000 steps of conjugate gradient.

3.2.2 QM geometry optimization

The goal of QM energy minimizations is often to find the structure with the lowest energy (the equilibrium structure), *i.e.*, performing a *geometry optimization*. In contrast to the classical energy minimization described in the previous section, it is for this purpose also required to test whether the potential energy is in a minimum. These requirements are that the first derivative of the energy is zero and that the second derivative is positive with respect to all nuclear displacements,[72]

$$g_x = \left(\frac{\partial E}{\partial x}\right) = 0, \tag{3.7}$$

$$H_{xx} = \left(\frac{\partial^2 E}{\partial x^2}\right) > 0 \tag{3.8}$$

for all Cartesian components x of the molecular gradient $\mathbf{g}$ and molecular Hessian $\mathbf{H}$ (Section 1.1.4). The molecular gradient $\mathbf{g}$ is a vector of length $3N$ with N the number of atoms in the region to be optimized. The molecular Hessian in Eq. 3.8 is a second-order tensor of dimensions $3N \times 3N$. Thus, to find the equilibrium geometry it is necessary to calculate both the molecular gradient and (an approximation to) the Hessian. The number of negative eigenvalues of the Hessian is referred to as the *Hessian index* and is zero for a minimum of the potential energy.[72]

In analogy to the molecular gradient, the Hessian can be calculated analytically or numerically. In **Paper VII**, the Hessian (and from it the vibrational frequencies) is calculated numerically from an analytical gradient using atomic displacements of 0.01 a.u. This is rather expensive since it requires $3N$ calculations on displaced nuclear geometries in both directions.

The Newton–Raphson method uses the gradient $\mathbf{g}$ and the Hessian $\mathbf{H}$ to determine a step towards the mimimum energy structure. First, a local second-order model of the energy is constructed,[1]

$$E(\mathbf{x}) = E(\mathbf{x}_0) + \mathbf{g}d\mathbf{x} + \frac{1}{2}\mathbf{H}(d\mathbf{x})^2, \tag{3.9}$$

which expands the energy around $\mathbf{x}_0$ using a displacement vector $d\mathbf{x}{=}\mathbf{x}-\mathbf{x}_0$, the molecular gradient $\mathbf{g}$ and diagonalized Hessian $\mathbf{H}$ at $\mathbf{x}_0$. Since the change in the local second-order model is zero at the minimum,

$$\mathbf{g}d\mathbf{x} + \frac{1}{2}\mathbf{H}(d\mathbf{x})^2 = 0, \tag{3.10}$$

and a step towards the minimum of the local model is given by

$$d\mathbf{x} = -\mathbf{g}\mathbf{H}^{-1}, \tag{3.11}$$

where $d\mathbf{x}$ contains the displacement of all nuclear coordinates. A step using Eq. 3.11 is referred to as a *Newton step*. Since the real function is not a second-order function, several steps are needed to approach the minimum. The Newton–Raphson method is unbound, meaning there is no upper limit to the step size. This is a particular problem when one of the Hessian eigenvalues is close to zero, causing a large step in one direction.[1] A maximum displacement known as the *trust radius* is usually set so that the second-order approximation in Eq. 3.9 is reasonable.[72,73]

Alternatively, an approximate Hessian can be used instead of the exact Hessian, in which case the method is referred to as *quasi-Newton* optimization. This is much faster in practise and is therefore used in most geometry optimizations including those in **Paper VII** and in all other QM/MM geometry optimizations in this thesis. The Hessian at step $i{+}1$ is approximated using the approximate Hessian at step i, the gradient difference between i and $i{+}1$ and the size of the step between i and $i{+}1$.[72] A common algorithm to do this is the Broyden–Fletcher–Goldfarb–Shanno scheme,[74] which ensures that the quadratic model has a minimum.[72] Quasi-Newton methods require an initial Hessian, which can be the exact Hessian or an initial guess.

The geometry optimization usually continues until some convergence criteria are met. This is usually done by setting a lower limit to the change in energy, step size or norm of the gradient. In the geometry optimization in **Paper VII**, two out of three of these criteria should be met for the optimization to end. Figure 3.4 shows the potential energy and Hessian index of a molecule of PNA in water in the course of a PE geometry optimization. In the first part of the optimization (the *global region*), the energy quickly

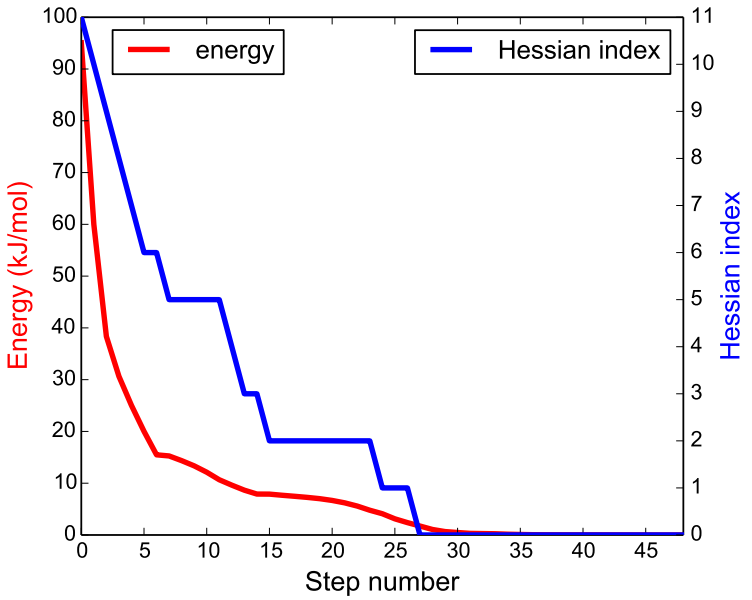


Figure 3.4. Potential energy and Hessian index of a molecule of PNA during a PE geometry optimization in a frozen water box. The energy of the optimized structure is set to 0 kJ/mol.

decreases and the Hessian index reduces to to zero. Another 15 to 20 steps are needed to locate the exact minimum to satisfy the chosen convergence criteria (the *local region*). An optimization algorithm should be efficient in both parts of the optimization.[72]

For a molecule without environment, the equilibrium structure is usually used to calculate a molecular property. This is also the strategy used in **Paper VI**. For a molecule in an environment, however, the choice of whether to do geometry optimization is less straightforward, especially when temperature effects are introduced through an MD simulation. Indeed, geometry

optimization of (a part of) the molecular system leads to a reduction of the energy and thus the temperature effects are (partially) lost in the optimized region. This is illustrated in **Paper I**, where the broadening of the absorption peak in the calculated spectrum of fluorescent proteins is reduced by a factor of three as a result of geometry optimization of the chromophore in the presence of the frozen protein. In the calculation of vibrational properties such as in **Paper VII**, however, geometry optimization is necessary for the Hessian index to be zero, which allows for the calculation of vibrational frequencies. For optical properties, the requirements are less harsh and QM/MM geometry optimization is only necessary when the starting structure of the central subsystem is not good enough for the calculation of the molecular property. This is investigated in **Paper I**, where trends between different fluorescent proteins are not well reproduced on unoptimized crystal structures, clearly indicating that these structures are not good enough for QM/MM calculations. Also, structural changes as a result of QM/MM geometry optimization of MD snapshots are small for e.g. GFP—for which the force field was tailor-made—and larger for e.g. BFP—which is not as well described by the force field used. We note that QM/MM MD can partially resolve these issues because it allows for temperature effects in a simulation while simultaneously allowing the structure to be in a minimum of the QM potential energy landscape.

3.3 Molecular dynamics

In an energy minimization (Section 3.2), the potential energy is reduced by displacing the nuclear coordinates in an appropriate direction. This is useful to find *one* structure with a low energy, but not to find *multiple* structures to sample a larger energy landscape or indeed to overcome barriers to locate another low-energy region. This is however possible by *molecular dynamics*, in which also kinetic energy and thus temperature is added to the simulation.

It is not possible to simulate a macroscopic system to sample all possible configurations at once due to the size of such as system. Indeed, the molecular systems used in this thesis have a dimension on the order of tens of nanometers, which is a factor of 10^7 smaller than the size of a typical sample in a laboratory (the volume is then smaller by a factor of 10^{21}). In other words, it is not possible to obtain an *ensemble average* from one molecular

structure. MD provides an alternative way to sample the most important conformations, *i.e.*, the most important part of *phase space*. The *ergodic hypothesis* states that the time average of a property of a representative system is equal to its ensemble average.[6,39] This is the basic idea of an MD simulation: a representative microscopic system is propagated in time using Newton's equation of motion to sample phase space. The ergodic hypothesis implies that all possible conformations of a molecular system (*i.e.*, the complete phase space) can be sampled regardless of the starting point.[6] In practice, however, the total time of a molecular simulation is limited by the computational time and resources available for the calculation, leading to a statistical error that can be estimated from the calculated properties in addition to an unknown systematic error in the calculation related to how representative the sampling is.[1] The ergodic hypothesis can be applied to any property, which is the basis of the sequential approach (MD followed by property calculations) used in this thesis. It is important to realize that the statistical error in the calculated property—estimated for instance by the standard error of a number of values obtained from different structures—only relates to the number of structures used and not to the systematic error of the calculations. The systematic error is related to the method used to calculate the property and the quality of the underlying structures and may well be much larger than the statistical error.

The same principles of MD may be applied regardless of how the forces on the atoms are calculated. The forces may be calculated by QM methods to give rise to *ab-initio* MD. Alternatively, the forces on a part of the system may be calculated using a combination of QM and classical methods (*cf.* Chapter 2) while using classical forces on the rest of the system, giving rise to QM/MM MD. The MD simulations in this thesis have all been performed using classical force fields (Section 3.1).

Apart from the forces on the nuclei, an MD simulation also needs a way to propagate the coordinates of the nuclei in time using Newton's equations (Section 3.3.1) and a molecular system in equilibrium as the starting point (Section 3.3.2).

3.3.1 Integration of Newton's equations

Newton's second equation of motion can be used to calculate the acceleration $\mathbf{a}_I$ on atom I from the force $\mathbf{F}_I$ acting on the atom,

$$\mathbf{a}_I = \frac{d^2\mathbf{x}_I}{dt^2} = \frac{\mathbf{F}_I}{m}, \tag{3.12}$$

with $\mathbf{a}_I$, $\mathbf{x}_I$ and $\mathbf{F}_I$ vectors containing x-, y- and z-components. Thus, the change in position of the atoms can be obtained by integration of the force $\mathbf{F}_I$. Using an analytical expression for $\mathbf{F}_I$, the positions of the atoms can in principle be propagated in time. However, this procedure gives a many-body problem that cannot be solved analytically for real systems.[39]

The practical alternative is given by numerical integration with a finite time step Δt, within which the forces are assumed to be constant. Positions and velocities at time $t + \Delta t$ can be calculated using a Taylor expansion that is usually truncated at second order. Different algorithms exist to do this, differing among other things in the memory requirements and the conservation of physical properties such as energy and momentum.[39] One common way to do this is the *Velocity Verlet* algorithm,[75]

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)(\Delta t)^2, \tag{3.13}$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\mathbf{a}(t) + \mathbf{a}(t + \Delta t)}{2}\Delta t, \tag{3.14}$$

in which the next positions are calculated from the positions $\mathbf{x}(t)$, velocities $\mathbf{v}(t)$ and accelerations $\mathbf{a}(t)$ at time t in the first step (Eq. 3.13). The next acceleration $\mathbf{a}(t+\Delta t)$ can then be calculated from the old positions $\mathbf{x}(t)$ and the new positions $\mathbf{x}(t+\Delta t)$ by numerical differentiation using Newton's second law (Eq. 3.12) in the second step. The next velocities can be calculated from the velocities $\mathbf{v}(t)$ at time t and the acceleration at time t and $t+\Delta t$ in the third step (Eq. 3.14).

An alternative scheme is the leap-frog algorithm,[76] in which the positions are calculated at time t, $t+\Delta t$, $t+2\Delta t$, *etc.*, while the velocities are calculated at time $t+\frac{1}{2}\Delta t$, $t+\frac{3}{2}\Delta t$, $t+\frac{5}{2}\Delta t$, *etc.* The leap-frog algorithm is the standard algorithm used in the GROMACS MD software[77] and is used for the MD simulations in **Paper III**, **Paper IV** and **Paper VII**.

The choice of the time step is crucial for the MD simulation. A time step that is too short leads to a slow propagation in time and hence a poor

sampling. A time step that is too long leads to atoms coming too close to each other and to instabilities in the algorithms.[39] The time step should be able to describe all motions in the molecular system. The fastest motion in most systems is usually the vibration of hydrogen atoms at a frequency of around 10^{14} Hz. A time step of 1 fs (10^{-15} s) enables the description of such a vibration in ten steps, which is enough to ensure numerical stability. Restraining the vibration of hydrogen (such as with the RATTLE algorithm[78]) allows for a slightly longer time step to be chosen.

The part of the MD procedure requiring most computational effort, however, is not the integration of Newton's equation, but the calculation of the forces. The calculation can thus be accelerated by calculating the forces between some atoms only every n'th time step. This is the idea behind the RESPA algorithm,[79] which is used in **Paper I** to increase the speed of the MD simulations. Forces between bonded and nearby atoms are evaluated at every time step, while forces between atom pairs further away are evaluated every n'th step and are assumed to be constant in between.

3.3.2 Equilibration of the molecular system

The first step in the preparation of an MD simulation is usually a classical minimization of the potential energy as described in Section 3.2. In the next step, usually referred to as the *equilibration phase*, the kinetic energy is added and the density and other physical properties of the system are optimized. The goal of the equilibration phase is to evolve from the starting structure to an equilibrium state, from which a simulation can be started. Physical properties such as energy, temperature, pressure or structural properties may be monitored to see whether the system has reached equilibrium.[39]

To start an MD simulation, the velocities at time t=0 are required to obtain the new coordinates and new velocities at time t=Δt (Eqs. 3.13 and 3.14). Initial velocities are usually assigned randomly by choosing values for v_x, v_y and v_z from a Gaussian distribution that is usually referred to as the Maxwell–Boltzmann distribution,

$$f(v_x) = \sqrt{\frac{m}{2\pi kT}} e^{-\frac{mv_x^2}{2kT}}, \tag{3.15}$$

with $f(v_x)$ the fraction of molecules with a velocity v_x in the x-direction.

This ensures that the molecular speed $v=\sqrt{v_x^2+v_y^2+v_z^2}$ is distributed according to the Maxwell distribution of speed,

$$f(v)=\sqrt{\left(\frac{m}{2\pi kT}\right)^3}4\pi v^2 e^{-\frac{mv^2}{2kT}}. \tag{3.16}$$

In this way, the kinetic energy is introduced as a function of the chosen temperature T. Figure 3.5 shows the distribution of the molecular speed for molecular ensembles at different temperatures T.

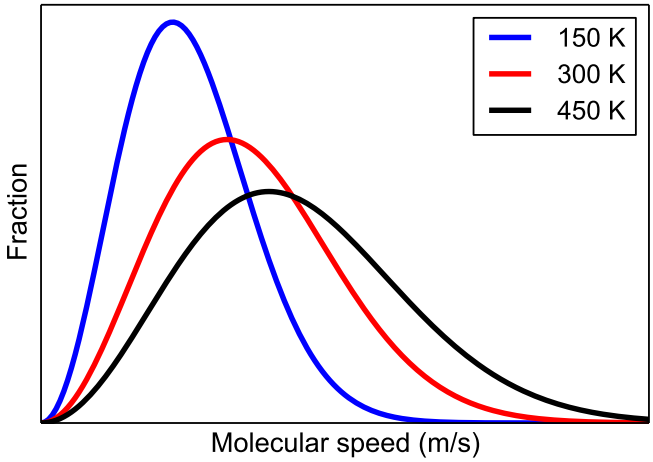


Figure 3.5. Maxwell molecular speed distribution of molecular ensembles at different temperatures. The plot shows the fraction of molecules with a certain molecular speed.

Often the initial configuration of the molecular structure does not have the correct density. The density of the molecular system can be optimized by allowing the volume V to change during the equilibration, e.g. by using the NPT ensemble. The pressure $p(t)$ can be evaluated at every time step and compared to a reference pressure. The pressure can be adjusted by rescaling all atomic coordinates and thus also the total volume of the system. Rescaling all atomic coordinates by adjusting the volume to the target value leads to the same pressure during the whole simulation and does not allow the pressure of the system to fluctuate. This fluctuation can be ensured by coupling the pressure of the system to an external bath at pressure p_{bath} using Berendsen's pressure coupling scheme.[80] In this approach, the

pressure is not set to its target value at every time step, but changed at a speed given by a relaxation constant τ_p as

$$\frac{dp(t)}{dt} = \frac{1}{\tau_p}\left(p_{\text{bath}} - p(t)\right). \tag{3.17}$$

The relaxation constant has units of time and can be changed to determine the strength of the coupling to the bath. Choosing τ_p equal to the time step of the MD simulation is equivalent to instant rescaling of the atomic coordinates. A higher value for τ_p leads to a weaker coupling to the external bath.[39]

Figure 3.6 illustrates how the density of a box of ethanol molecules adjusts to a value close to the experimental density with two different relaxation times τ_p.

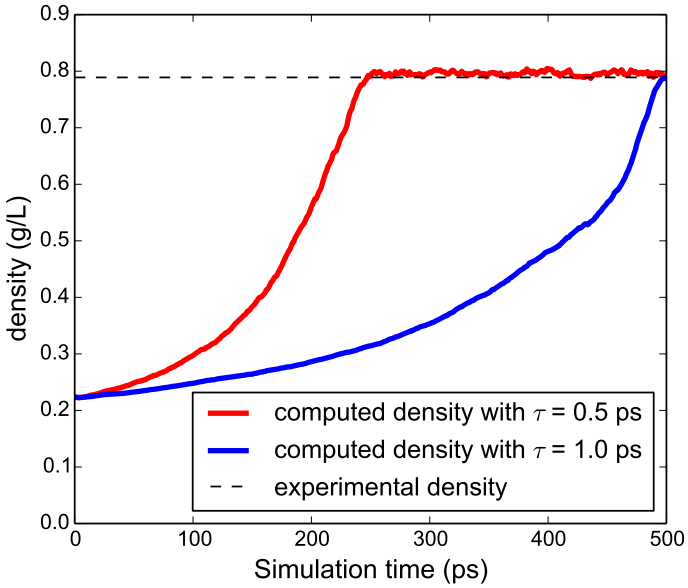


Figure 3.6. Density of an ethanol solvent box (described by the OPLS force field[64]) during NPT equilibration starting from a low density. The density is compared to the experimental density of 0.789 g/L at 293 K (black, dashed).[81] The density of the simulation is controlled by coupling to an external bath[80] with a pressure of 1 bar and relaxation constant 0.5 ps (red) and 1.0 ps (blue). The data for τ_p=0.5 ps are taken from a simulation in **Paper IV**.

The initial density of the system is much lower than the equilibrium value, which is reached after around 250 ps with τ_p=0.5 ps (red line). Equilibration with a weaker coupling (τ_p=1.0 ps, blue line) leads to a slower convergence towards the equilibrated density and requires a longer equilibration time. Equilibration of the density of the solvent systems in **Paper III**, **Paper IV** and **Paper VII** is done using Berendsen's approach with a relaxation constant of 0.5 ps. The density is monitored as illustrated in Figure 3.6 to ensure that the equilibration time is long enough. Berendsen's approach is also one of the ways in which the temperature of the system can be controlled in an MD simulation.

Before analyzing the trajectory of the MD simulation (e.g. by taking snapshots or calculating physical properties of the system from the trajectory), it is important to equilibrate the system with exactly the same settings as in the MD simulation. In other words, the *production phase* is preceded by an *equilibration phase*. In **Paper I**, a 15 ns NPT run on fluorescent proteins was preceded by 10 ns of equilibration. In **Paper III**, **Paper IV** and **Paper VII**, the NVT production phase on the solvated molecule was preceded by a 2 ns equilibration in the NVT ensemble that came *after* the 500 ps equilibration of the density in the NPT ensemble.

3.4 Protein crystal structures

The previous sections have described how molecular structures can be obtained from energy minimizations (Section 3.2) and MD simulations (Section 3.3) from some starting structure. The step of obtaining a high-quality starting structure is fundamental for accurate molecular property calculation and this step is more involved for proteins than for solute–solvent structures. One way to obtain a protein starting structure is to prepare the protein from a crystal structure from the protein data bank,[82] which is the strategy used for all protein structures in this thesis. This section describes the main challenges of this approach, many of which are in some way related to the lack of hydrogen atoms in crystal structures. The most important steps are illustrated by the preparation of the 1U19 crystal structure[83] of bovine rhodopsin that is used in **Paper III**. Rhodopsin is a transmembrane protein in rod photoreceptor cells in the eye, which are extremely sensitive to light. Its chromophore retinal undergoes a cis–trans isomerization when

light is absorbed, triggering a process responsible for vision.

The protein structures deposited in the protein data bank [82] are obtained by fitting the structure to an electron density map, which in turn is obtained from X-ray crystallography experiments. The wavelength of X-rays is approximately of the same order as the dimensions of the atoms in the protein (approx. 1 Å). The most challenging part of X-ray experiments is to obtain a regular crystal of the protein. The resulting electron density map may be constructed from the diffraction pattern of the X-ray experiment, *i.e.*, the way in which the crystal scatters the electromagnetic radiation. The *resolution* of the crystal structure gives the minimal distance at which two objects can clearly be separated and is ultimately determined by the quality of the crystal. [84,85] A *low* value thus means a *high* resolution. The 1U19 crystal structure has a resolution of 2.2 Å. This means that the position of hydrogen atoms cannot be located in the electron density map and that protonation states of side chains cannot be determined, which holds for all but the most accurate X-ray crystal structures. [84,85] Not all parts of the protein have the same resolution in the electron density map, however, and the reported resolution refers to the part of the protein that has the highest order in the crystal. The *temperature factor* of an atom or atom group gives the disorder of that region of the electron density map with a high value meaning a large uncertainty in the position. [84,85] Structure files from the protein data bank usually contain the temperature factor (also called B-factor) for all atoms. The 1U19 crystal structure has high temperature factors (above 100 $Å^2$) on the side of the protein that also contains the C-terminus as shown in Figure 3.7.

The temperature factors of the atoms on the other side of the protein are much lower (40 to 50 $Å^2$). One should be very careful interpreting structural characteristics of a part of the protein that is not well-resolved. In the case of rhodopsin in **Paper III**, this is not particularly problematic because the region that is not resolved well is far away from the retinal chromophore, where the absorption process takes place. The N- and C-termini of the proteins or certain side chains can be too flexible for the structure to be determined at all and crystal structures therefore often do not contain all residues and side groups.

A first step in the preparation of a crystal structure is often selecting the right molecules in the crystal structure. Sometimes not all crystallized

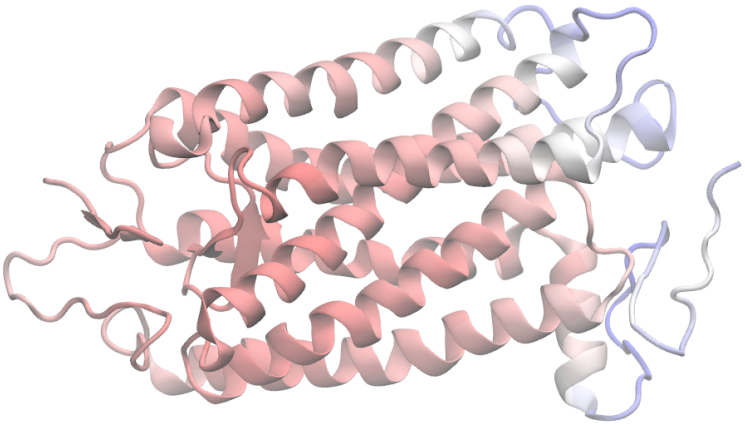


Figure 3.7. Structure of chain A of bovine rhodopsin from crystal structure 1U19.[83] The colouring is done according to the temperature factor (B-factor) of every amino acid with red for low and blue for high values. The N-terminus is on the left and the C-terminus on the right of the figure. The figure is made in VMD.[68]

molecules are relevant to include. In **Paper III**, only chain A was selected from the dimer in the 1U19 crystal structure because the focus was on the chromophore and its close environment and not on the interaction between the two monomers.

A next step is the addition of hydrogen atoms, which is not always trivial since several amino acid residues can have different protonation states and several things need to be taken into account. In reality, amino acid residues can be protonated and deprotonated over time. In a classical simulation, however, the input structure determines the protonation state in the whole simulation, underlining the importance of a sensible starting structure. The pKa value of a side chain in solution is a good first approximation of its protonation state in the protein. Glutamic acid (pKa 4.07) and aspartic acid (pKa 3.9) are usually charged, but can be protonated in particular chemical surroundings. Special software can be used to find the most likely protonation state based on the local surrounding.[86] Lysine (pKa 10.54) and tyrosine (pKa 10.46) are rarely deprotonated at a neutral pH, but can in principle also have different protonation states. By far the most problematic residue in this respect is histidine (pKa 6.04). Histidine (Figure 3.8) can not only be positively charged or neutral, but its neutral form can be protonated

in two different places, either at N_δ or at N_ϵ.

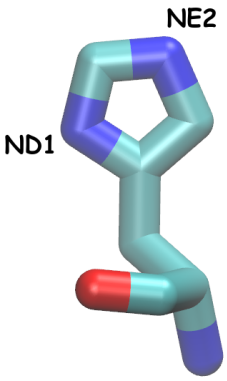


Figure 3.8. Histidine residue from a crystal structure with N_δ (ND1) and N_ϵ (NE2) indicated. Carbon atoms are shown in cyan, nitrogen in blue and oxygen in red. The figure is made in VMD.[68]

Since histidine residues are often involved in hydrogen-bonding, the crystal structure may contain evidence as to which of the protonation states is most likely. Examples of this are shown in Figure 3.9.

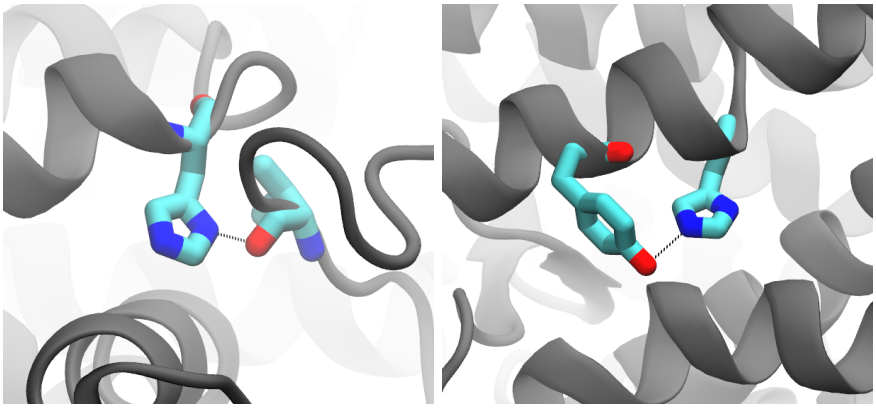


Figure 3.9. Evidence for a N_δ-protonated histidine (left) and a N_ϵ-protonated histidine (right) from the crystal structure of rhodopsin 1U19.[83] The distance of the black dashed lines are 2.18 Å between His65 ND1 and Val337 O (left) and 2.92 Å between His211 NE2 and Tyr206 OH (right). Carbon atoms are shown in cyan, nitrogen in blue, oxygen in red and the protein backbone is shown in grey. The figures are made in VMD.[68]

Cysteine residues (pKa 8.37) are usually protonated, but two cysteine residues can form a sulphur bridge, in which case hydrogens should not be added. Sulphur bridges are usually easy to detect in crystal structures as illustrated for the link between Cys110 and Cys187 in rhodopsin in Figure 3.10.

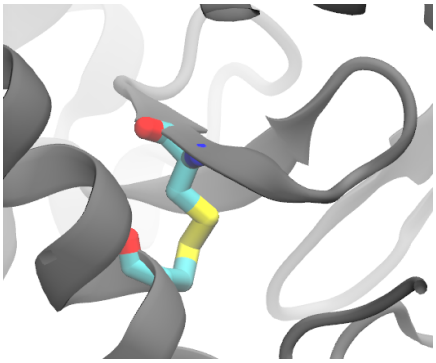


Figure 3.10. Evidence for a sulphur bond between Cys110 and Cys187 in chain A of the crystal structure of rhodopsin 1U19[83], linking two different parts of the protein together. Carbon atoms are shown in cyan, nitrogen in blue, oxygen in red, sulphur in yellow and the protein backbone is shown in grey. The figure is made in VMD.[68]

It is important to mention the protonation states used in the protein model, since there is not always consensus on the protonation state and because it may influence the resulting calculations. Indeed, the protein environment in rhodopsin or fluorescent proteins (**Paper I**) tunes the absorption maximum, thus correct modeling of the electrostatics is fundamental to reproduce (relative) excitation energies. After adding the hydrogen atoms, their positions can be optimized using the strategies described in Section 3.2.1, often with constraints on the positions of the heavy atoms.

Chapter 4

Embedding potentials

For an accurate calculation of the property of a molecule in a molecular environment, an accurate description of its surroundings is fundamental. Chapter 2 has described how a molecular environment can be included classically in the QM calculation of a molecular property. In PE calculations, the atoms in the environment need to be described at least by their coordinates, a parametrization of their ESP (e.g. in an electric multipole expansion) and their response to an electric field (with polarizabilities). The collection of all these parameters is called the *embedding potential*. Strategies to obtain the coordinates of all these atoms (the molecular structure) have been described in Chapter 3. This chapter describes how the other parameters can be obtained. The distinction will be made between two very different strategies to obtain the parameters: either from QM calculations on molecules or molecular fragments (Section 4.1), or from a database containing parameters for different atom types (Section 4.2). This is followed by a discussion of the accuracy (Section 4.3) and the computational cost (Section 4.4) of the different procedures. The emphasis is on embedding potentials for solvents (used in **Paper III**, **Paper VII** and **Paper IV**) and proteins (used in **Paper I**, **Paper II** and **Paper III**).

4.1 QM-based parameters

A simple example of an embedding potential of two water molecules is shown in Figure 4.1. This potential contains the same parameters for all six atoms: coordinates, atom-based multipoles of order 0 (charges), 1 (dipoles) and 2 (quadrupoles), anisotropic dipole–dipole polarizabilities and an exclusion list to exclude polarization interactions within each molecule.

```
@COORDINATES
6
AA
O  29.816  29.141  25.136     1
H  29.526  29.061  24.226     2
H  29.006  29.051  25.646     3
O  27.286  28.771  26.656     4
H  26.456  28.821  27.126     5
H  27.866  28.271  27.226     6
@MULTIPOLES
ORDER 0
6
1  -0.736
2   0.368
3   0.368
4  -0.739
5   0.370
6   0.370
ORDER 1
6
1  -0.181 -0.028 -0.066
2   0.073  0.019  0.208
3   0.190  0.021 -0.112
4  -0.042 -0.076  0.175
5   0.188 -0.007 -0.114
6  -0.128  0.116 -0.137
ORDER 2
6
1  -4.157  0.092 -0.221 -4.874  0.024 -3.632
2  -0.419  0.011  0.143 -0.454  0.038 -0.052
3  -0.125  0.037 -0.206 -0.454 -0.024 -0.347
4  -3.694 -0.399 -0.042 -4.611 -0.257 -4.359
5  -0.125 -0.013 -0.201 -0.458  0.013 -0.344
6  -0.305 -0.140  0.168 -0.331 -0.145 -0.290
@POLARIZABILITIES
ORDER 1 1
6
1   5.309 -0.036  0.068  5.582 -0.012  5.122
2   1.178 -0.030  0.770  1.646  0.177  3.261
3   2.895  0.126 -1.098  1.653 -0.156  1.566
4   5.127  0.147  0.022  5.466  0.110  5.349
5   2.886  0.153 -1.151  1.420  0.351  1.740
6   1.832 -0.654  1.127  2.159 -0.614  2.050
EXCLISTS
6 3
1   2   3
2   1   3
3   1   2
4   5   6
5   4   6
6   4   5
```

Figure 4.1. Example of an embedding potential of two water molecules. The embedding potential is based on an atomic electric multipole expansion up to quadrupoles and anisotropic dipole–dipole polarizabilities.

The goal of this section is to describe how the parameters for the embedding potential can be obtained from QM calculations. The parameters to reproduce the ESP and the polarizabilities will be treated in Sections 4.1.1 and 4.1.2, respectively, assuming that the classical region consists of separate molecules that allow for one QM calculation on each molecule. Section 4.1.3 describes strategies to obtain these parameters for larger molecules, for which one QM calculation is not possible or not desirable.

4.1.1 The electrostatic potential

The electrostatic energy in the PE method is calculated as a sum of the interaction of the classical sites with the nuclei (Eq. 2.8) and the electrons (Eq. 2.12) in the QM region. The electron density of the QM region thus interacts with classical multipoles rather than with the electron density of molecules or fragments in the classical region, for reasons of efficiency. It is, however, important that the parameters placed on the classical sites reproduce the ESP of the molecule or fragment in the best possible way. The potential $V(\mathbf{r}')$ is the force on a unit of positive charge located at $\mathbf{r}'$ and can be calculated from the wave function Ψ as a sum over contributions from the molecule's nuclei and electrons,[6]

$$V(\mathbf{r}') = \sum_{I=1}^{N} \frac{Z_I}{|\mathbf{r}' - \mathbf{R_I}|} - \int \Psi^*(\mathbf{r}) \frac{1}{|\mathbf{r}' - \mathbf{r}|} \Psi(\mathbf{r}) d\mathbf{r} \tag{4.1}$$

or equivalently from the electron density ρ as

$$V(\mathbf{r}') = \sum_{I=1}^{N} \frac{Z_I}{|\mathbf{r}' - \mathbf{R_I}|} - \int \frac{\rho(\mathbf{r})}{|\mathbf{r}' - \mathbf{r}|} d\mathbf{r}. \tag{4.2}$$

The electric multipoles are classical parameters that are meant to reproduce the ESP. They can be obtained in different ways, the most important ones being a multipole expansion (Section 4.1.1.1) and fitting to the ESP (Section 4.1.1.2). In the context of this thesis, the former is used to generate multipoles up to quadrupoles and the latter to generate charges only.

4.1.1.1 The electric multipole expansion

Electric multipoles can easily be calculated for a set of discrete charges as[39]

$$q^{\text{tot}} = \sum_{I=1}^{N} q_I, \tag{4.3}$$

$$\mu_a^{\text{tot}} = \sum_{I=1}^{N} q_I r_a, \tag{4.4}$$

$$Q_{ab}^{\text{tot}} = \sum_{I=1}^{N} q_I r_a r_b, \tag{4.5}$$

where q^{tot} is the total monopole moment (*i.e.*, charge), μ_a^{tot} is a component of the dipole moment, Q_{ab}^{tot} is a component of the quadrupole moment, a and b are the Cartesian coordinates and r_a and r_b the distance of I from the origin along coordinate a. The multipoles are thus defined with respect to an origin (the expansion centre) where only the first non-zero moment is origin-independent.[38,39] In principle, electric multipoles can be generated up to any order and an infinite expansion gives an exact representation of the ESP of the set of charges. A truncated multipole expansion is accurate at long distances because the lowest-order multipoles are the ones with the slowest decay. Indeed, charge–charge interactions decay as r^{-1}, charge–dipole interactions as r^{-2} and charge–quadrupole interactions as r^{-3} (*cf.* Figure 3.2), so at large distances the electric monopole dominates.[38] At short range, truncation of the multipole expansion is less successful. The expansion breaks down completely when the wave function of two molecules overlap due to the neglect of exchange–repulsion effects.[38]

Representing the ESP of a molecule rather than a set of discrete charges, however, is more difficult since the electron density is more complicated than a set of point charges (the ESP can, however, be *represented* by partial charges, see Section 4.1.1.2). Cartesian components of the molecular multipole moments can be expressed as a function of the nuclear charges Z_I and electron density $\rho(\boldsymbol{r})$ of an n-electron system, again with r_a and r_b relative to an origin, as[6]

$$\begin{aligned} q^{\text{tot}} &= \sum_{I=1}^{N} Z_I - \int \rho(\boldsymbol{r}) d\boldsymbol{r} \\ &= \sum_{I}^{N} Z_I - n \end{aligned} \tag{4.6}$$

$$\mu_a^{tot} = \sum_{I=1}^{N} Z_I r_a - \int \rho(\boldsymbol{r}) r_a d\boldsymbol{r} \tag{4.7}$$

$$Q_{ab}^{\text{tot}} = \sum_{I=1}^{N} Z_I r_a r_b - \int \rho(\boldsymbol{r}) r_a r_b d\boldsymbol{r} \tag{4.8}$$

with N and n the number of nuclei and electrons, respectively, a and b Cartesian coordinates and r_a and r_b distances along those coordinates.

To simplify the evaluation of Eq. 4.1 and to represent the ESP as a function of the electric multipoles, it is necessary to expand $\frac{1}{|\boldsymbol{r}'-\boldsymbol{r}|}$ in a

Taylor series around the centre $\boldsymbol{r}=\boldsymbol{r}_0$ (here truncated at second order) as

$$\begin{aligned}\frac{1}{|\boldsymbol{r}'-\boldsymbol{r}|} = &\frac{1}{|\boldsymbol{r}'-\boldsymbol{r}_0|} \\ &+\sum_a \left(\frac{\partial}{\partial a}\frac{1}{|\boldsymbol{r}'-\boldsymbol{r}_0|}\right)_{r_0} r_a \\ &+\frac{1}{2}\sum_a\sum_b \left(\frac{\partial^2}{\partial a \partial b}\frac{1}{|\boldsymbol{r}'-\boldsymbol{r}_0|}\right)_{r_0} r_a r_b,\end{aligned} \tag{4.9}$$

where a and b can be the Cartesian coordinates x, y and z. The ESP in Eq. 4.1 can now be written as a function of the electric multipoles (Eqs. 4.6 to 4.8) as[38]

$$V(\mathbf{r}') = T^{(0)}q^{\text{tot}} + T_a^{(1)}\mu_a^{\text{tot}} + \frac{1}{2}T_{ab}^{(2)}Q_{ab}^{\text{tot}}, \tag{4.10}$$

using Einstein summation over repeated indices and with the components of the interaction operator (*cf.* Eqs. 2.9 to 2.11) written as[38]

$$T^{(0)} = \frac{1}{|\mathbf{r}'-\mathbf{r}_0|}, \tag{4.11}$$

$$T_a^{(1)} = \nabla_a \frac{1}{|\mathbf{r}'-\mathbf{r}_0|}, \tag{4.12}$$

$$T_{ab}^{(2)} = \nabla_a\nabla_b \frac{1}{|\mathbf{r}'-\mathbf{r}_0|}. \tag{4.13}$$

Eq. 4.1 is now written classically in Eq. 4.10, allowing for a very efficient evaluation of the ESP of a molecule or fragment.

The use of electric multipoles in a PE calculation requires localization of the multipoles on e.g. the nuclei of the atoms in the classical region. Partial charges and multipoles positioned on nuclei are not well-defined physically, in contrast to molecular multipole moments and the ESP.[6] Thus, the localized properties should represent the total ESP of the molecule or molecular fragment in the best possible way. A classic approach to obtain localized partial charges is the approach by Mulliken.[87] These charges are based on an orbital population analysis, where each pair of basis functions is used to represent a charge distribution. In the Mulliken approach, this charge distribution is divided equally over the atoms on which the basis functions are placed (which can be the same atom as well) without taking into account the

electronegativity of those atoms or the diffuseness of the basis function.[87] This leads to a problem with Mulliken charges, namely their high sensitivity to the basis set and the lack of basis set convergence.[38] Mulliken's approach for localized charges has later been generalized to higher-order multipoles by Karlström.[88] Another popular way to localize multipoles is the distributed multipole analysis (DMA) by Stone,[89] which is also based on products of basis functions to define a charge distribution. The DMA scheme is not limited to bonds or bond midpoints and can be used with any set of expansion centres.[39]

A disadvantage of all these approaches is the sensitivity to the choice of basis set. The LoProp approach by Gagliardi *et al.*[90] is meant to overcome this problem by using atomic orbitals as basis functions. A localized basis set is used to calculate localized values as expectation values in a series of four consecutive steps. The LoProp approach can be used with bonds or bond midpoints as expansion centres. An important advantage of the LoProp approach over other approaches—apart from the reduced sensitivity to the basis set—is the transferability of the parameters, *i.e.*, the magnitude of the multipoles is similar for chemically equivalent atoms.[90,91] This is an important prerequisite for the use of electric multipoles in classical force fields.

In this thesis, the terminology of Olsen, Aidas and Kongsted[40] is used to refer to embedding potentials based on electric multipoles with M0 meaning only electric monopoles (charges), M1 electric monopoles and dipoles, *etc.*

4.1.1.2 ESP-fitted charges

Another approach to reproduce the molecular ESP in Eq. 4.1 classically is to define a set of charges that is fit to represent the ESP in the best possible way.[92,93] In this way, the resulting charges are much less sensitive to the basis set than e.g. Mulliken charges. *Atomic* charges are usually chosen for this procedure, *i.e.*, charges are placed on nuclei only. Thus, one has to find a set of atomic charges q_I on nuclei I that minimizes the error between the QM ESP in Equation 4.1 and the potential generated by this set of atomic charges[6]

$$V(\mathbf{r}') = \sum_{I=1}^{N} \frac{q_I}{|\mathbf{r}' - \mathbf{R_I}|}. \tag{4.14}$$

The error of $V(\mathbf{r}')$ is calculated for points on a grid and minimized with respect to Eq. 4.1 by some fitting procedure to give a set of optimized charges. Charges obtained in this way are here referred to as M* following Schwabe *et al.*[94] to distinguish them from electric monopoles derived from a multipole expansion (M0).

The different schemes to obtain ESP-fitted charges differ mainly in the choice of the grid on which the ESP is evaluated, the algorithm used for the fitting and the use of additional constraints,[39] resulting in very different partial charges in some cases.[95] Momany used a simple grid and a least-squares fitting procedure to derive ESP-fitted charges for formamide, methanol and formic acid, constraining the total charge to be zero and the molecular dipole to its experimental value.[92] Cox and Williams derived ESP-fitted charges for 14 small molecules and ions using points on a regular grid between 1.2 Å and 2.2 Å from the vdW surface of the molecules and using a similar iterative least-squares procedure for the fitting.[93] Singh and Kollman[96] used the least-squares fitting of Cox and Williams but introduced another grid to evaluate the ESP on. This grid consists of shells at surfaces defined by the vdW radii times a factor, originally four shells at 1.4, 1.6, 1.8 and 2.0 times the vdW radii.[96] Chirlian and Francl used a spherical grid and introduced a computationally efficient scheme to do the fitting using a Lagrange multiplier, constraining the total charge to zero.[97] Charges obtained with their method are referred to as CHELP charges. While the fitting scheme of Chirlian and Francl proved to be efficient, the charges vary when molecules are re-oriented due to the irregular grid used.[39] An improved procedure—known as CHelpG—was developed by Breneman and Wiberg using a regular grid with points between the vdW surface and 2.8 Å from any atom and the Lagrange multiplier approach of Chirlian and Francl for the fitting.[98] A more recent scheme by Hu, Lu and Yang (HLY) allows for ESP fitting of partial charges, higher electric multipoles and atomic polarizabilities. This method gives increased numerical stability by using the entire molecular space for the fitting rather than discrete grid points, with a weighting function to prioritize the chemically important region between 1.4 and 2.0 times the vdW distance.[99]

Some of the problems associated with ESP-fitted charges are the conformational dependence of the charges, the low degree of transferability and the poorly determined charges especially innermost in the molecule, all of

which are in some way related to the statistical nature of the fitting process.[100] Kollman and co-workers have developed the restrained electrostatic potential (RESP) fitting procedure for the AMBER force field to overcome these problems.[100] The RESP approach introduces penalty functions that reduce the magnitude of the fitted charges. Moreover, it uses a scheme to ensure chemically equivalent atoms get the same charge. The charges are determined using a least-squares fitting procedure on the grid defined by Besler, Merz and Kollman,[101] whose fitting procedure is known as MK.

All fitting schemes can in principle be used with constraints on e.g. the total molecular dipole moment. Moreover, it is rather straightforward to increase the density of fitting points when doing the fitting procedure. A technique that is related to ESP-fitting is to fit partial atomic charges to (calculated or experimental) electric multipoles, which can be more accurate than standard ESP-fitting schemes especially far away from the molecule.[95] The quality of the ESP-fitted charges depends crucially on the method used to calculate the QM ESP, as does the quality of the parameters derived from an electric multipole expansion.

4.1.2 Polarizabilities

The representation of the ESP as described in the previous section is based on calculations *in vacuo*. In reality, the electric charge distribution of molecular fragments in the classical region is influenced by the electric field generated by other molecular fragments. The linear response of a molecule (or atom) to an applied electric field is determined by the polarizability, from which an induced dipole can be calculated using Eq. 2.15. The molecular polarizability has been defined in Eq. 1.33 as the second-order derivative of the energy with respect to a perturbing electric field, evaluated at zero field strength. In this thesis (and in the current version of the PE model) only dipole–dipole polarizabilities are used, while also dipole–quadrupole *etc.* polarizabilities can be formulated.[38] The ab-component of the dipole–dipole polarizability $\boldsymbol{\alpha}$ can be defined as the change in the a-component of the dipole moment $\boldsymbol{\mu}$ by the b-component of an applied electric field $\mathbf{F}$ with a and b Cartesian coordinates,[6]

$$\alpha_{ab} = \left(\frac{\partial \mu_a}{\partial F_b}\right)_{\mathbf{F}=0}. \tag{4.15}$$

Since the dipole moment $\boldsymbol{\mu}$ and the electric field $\mathbf{F}$ are first-order tensors, the dipole–dipole polarizability $\boldsymbol{\alpha}$ is a second-order tensor with nine components. Due to the symmetry of the polarizability tensor (α_{xy}=α_{yx}, α_{xz}=α_{zx}, α_{yz}=α_{zy}), only six components need to be specified (see Figure 4.1): α_{xx}, α_{xy}, α_{xz}, α_{yy}, α_{yz} and α_{zz}. One can also define an isotropic polarizability α^{iso} (also called mean polarizability[5]), which is calculated from the diagonal elements of the second-order tensor $\boldsymbol{\alpha}$ as

$$\alpha^{\text{iso}} = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}\right). \tag{4.16}$$

The use of isotropic polarizabilities is computationally faster because the calculation of the induced dipole only requires a multiplication of a zeroth- and first-order tensor (Eq. 2.15: $\boldsymbol{\mu}_{\text{ind}}$=$\alpha^{\text{iso}} \cdot \mathbf{F}$), whereas the use of dipole–dipole polarizabilities requires the multiplication of a second- and first-order tensor (Eq. 2.17: $\boldsymbol{\mu}_{\text{ind}}$=$\boldsymbol{\alpha}\cdot\mathbf{F}$). Furthermore, it is more straightforward to use atomic isotropic polarizabilities when considering the structural variation of a molecule (**Paper IV**), as they are rotationally invariant. Embedding parameters based on isotropic and anisotropic dipole–dipole polarizabilities are here referred to as P1 and P2, respectively, following Olsen, Aidas and Kongsted.[40]

The localization of polarizabilities to nuclei or bond midpoints is technically more involved than the localization of electric multipoles, but can be done using some of the same methods such as LoProp[90] and DMA.[38] The LoProp approach is used to generate localized polarizabilities for the embedding potentials in this thesis.

The polarizabilities and induced dipoles introduce many-body effects in the PE calculations. In this way, they correct for the lack of molecular environment in the calculation of the electrostatic embedding parameters of a molecule. It is therefore relevant that polarization interactions are only calculated for interactions between atoms that were not included in the same QM calculation of the electrostatic parameters.[102] Every atom thus has a list of other atoms for which polarization interactions have to be excluded. This list is called the *exclusion list* and is part of the section with polarizabilities in the embedding potential (Figure 4.1).

4.1.3 Strategies for large molecules

We have so far assumed that a QM calculation can be performed on each molecule to obtain the embedding parameters, which is what is done for small (solvent) molecules. Indeed, the largest solvent molecule in **Paper III** and **Paper IV** is hexane with 20 atoms, which can easily be treated by QM methods. The size of biomolecules, however, often does not permit treatment of the whole molecule in one QM calculation. If this were possible, one would not need to use multiscale methods to describe its properties. The model of *wild-type* GFP in **Paper I**, **Paper II** and **Paper III** contains 3566 atoms excluding the water molecules present in the embedding potential. Dividing the large molecules into smaller fragments not only makes the QM calculation of parameters computationally cheaper, but also allows for intramolecular polarization effects in the PE calculation. Thus, fragmentation strategies are needed for large molecules.

One strategy to split the large QM calculation on a protein into parts is the molecular fractionation with conjugate caps (MFCC) approach by Zhang and Zhang.[103] The MFCC method was originally formulated to calculate the interaction energy between a protein and another smaller molecule (such as a drug molecule) by *ab initio* rather than QM/MM methods. Instead of calculating the interaction energy between protein and molecule directly, the interaction energy is calculated as the sum of the interaction energy of one amino acid residue and the molecule. Each amino acid is cut out and an appropriate cap is added on both sides. This introduces artificial interactions between the small molecule and the caps. To correct for this, interaction energies between the molecule and the caps need to be subtracted from the sum. These calculations are done using a pair of *conjugate caps* consisting of the cap on the C-terminus of residue number i and the cap on the N-terminus of residue i+1. These conjugate caps together form a small molecule for which the interaction energy with the small molecule M is calculated. The total *ab initio* interaction energy thus consists of a sum of all interactions energies between capped amino acids with the molecule M minus the interaction energy of the conjugate caps with the molecule M. The computational cost of this approach scales linearly with the size of the protein. An additional advantage of this procedure is the easy parallelization of the approach. A drawback of the original formulation of MFCC is

that it does not include many-body effects, *i.e.*, only pairwise interactions are included.

Söderhjelm and Ryde have applied the MFCC procedure to localized properties such as multipoles and polarizabilities from a LoProp calculation.[102] The property P^k on site k can be calculated from its properties P_i^k in the fragment and conjugate cap calculations as

$$P^k = \sum_{i=1}^{n} c_i P_i^k, \tag{4.17}$$

in which c_i is 1 for a normal fragment and -1 for a conjugate cap. The sum is over all fragments i that include site k. This approach allows for the calculation of localized parameters of a large (fragmented) molecule, which is why it is used for proteins in **Paper I**, **Paper II** and **Paper III**.

4.2 Parameters from a database

QM-derived embedding parameters (Section 4.1) take into account the specific orientation of a molecule or molecular fragment and can be very accurate. An alternative way of obtaining atom-based parameters is to extract them from a database. This approach relies on defining an *atom type* for each atom and taking parameters for that atom type from a database. These databases can be e.g. molecular force fields (Section 3.1) or tailor-made parameter sets for embedding potentials such as the one in **Paper IV**.

The division of atoms into atom types varies between different databases. This is illustrated for the carbon atoms in toluene in Figure 4.2. The OPLS

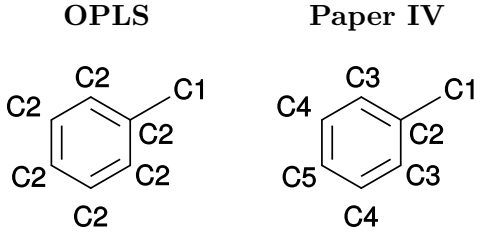


Figure 4.2. Atom types of the carbon atoms of toluene in the OPLS force field (left) and in **Paper IV** (right).

force field[64] distinguishes here between aromatic (opls_145; C2 in Figure

4.2) and aliphatic (opls_148; C1 in 4.2) carbon atoms. The same atom types are also used for other hydrocarbons such as the aromatic carbons in benzene or the aliphatic carbons in hexane. The parameter set from **Paper IV**, on the contrary, distinguishes all chemically non-equivalent atoms in the molecule. Thus, the methyl carbon (C1), its neighbour (C2) and the atoms in *ortho*- (C3), *meta*- (C4) and *para*-positions (C5) all have their own atom types and own associated parameters. All these atom types are different from those used for e.g. benzene and hexane.

Choosing fewer general atom types rather than many specific parameters has both advantages and disadvantages. An advantage of fewer general parameters such as in classical force fields is that no new parameters are needed when a new molecule is used. Indeed, a force field contains atom types for most functional groups and usually there is a straightforward match. A disadvantage is that general parameters cannot take into account the different chemical environments experienced by similar but non-equivalent atoms. This is illustrated in Figure 4.3 and Table 4.1 for toluene. Figure 4.3 shows

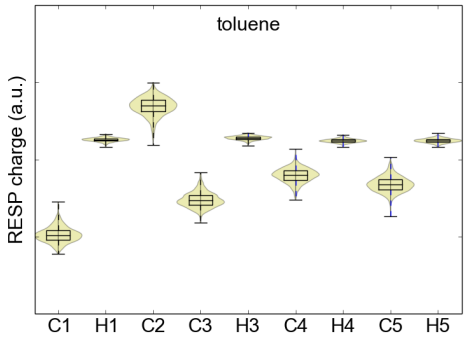


Figure 4.3. Variation in calculated B3LYP/aug-cc-pVTZ RESP charges for toluene. The diagram shows the variation in geometry-specific charges for 1000 different geometries. The numbering of the carbon atoms is shown in Figure 4.2 (right). From **Paper IV**.

the variation in RESP charges of the atoms in toluene for 1000 different geometries of the molecule. It is clear that the aromatic carbon atoms C2–C5 have different charges with especially the carbon attached to the methyl group (C2) differing from the others. The averaged charges are tabulated in Table 4.1 and compared to the charges for the different atoms in the OPLS force field, which assigns a charge of -0.115 to all aromatic carbon atoms.

Table 4.1. Different atom types and charges for the atoms of toluene in **Paper IV** and the OPLS force field.[64] See Figure 4.2 (right) for the atom numbering used.

atom	**Paper IV**	atom type	OPLS
C1	-0.48679	opls_148	-0.065
C2	0.34481	opls_145	-0.115
C3	-0.26155	opls_145	-0.115
C4	-0.10270	opls_145	-0.115
C5	-0.16049	opls_145	-0.115
H1	0.12857	opls_140	0.060
H3	0.13992	opls_146	0.115
H4	0.12153	opls_146	0.115
H5	0.12236	opls_146	0.115

The molecule-specific charges used in **Paper IV** ensure a more accurate ESP compared to OPLS with the more general parameters. The root-mean-square deviation (RMSD) of the ESP at twice the vdW distance of toluene has been calculated for 10 toluene molecules in **Paper IV**. The average RMSD for the averaged B3LYP/aug-cc-pVTZ RESP charges compared to QM is 2.4 kJ/mol. When the same calculation is done using the OPLS charges, the result is 4.3 kJ/mol. Thus, molecule-specific parameters are more accurate at the cost of being less generally applicable.

Isotropic parameters allow for easy transfer between different geometries. It is however also possible to use the same anisotropic parameters for different molecules as long as the geometry is kept fixed. In this way, molecular properties for different snapshots can be used in combination with an MD simulation with fixed molecular geometries.[40,94,104] B3LYP/aug-cc-pVTZ solvent embedding parameters have been published for methanol, water, acetonitrile and carbon tetrachloride in Ref. 94 (M2P3) and for methanol, water, dichloromethane and ethanol in Ref. 104 (M2P2). These parameters can be used in embedding potentials as long as the same geometry is used, but require re-orientation to fit the coordinate system of every molecule.

Both **Paper IV** and the mentioned anisotropic embedding parameter sets contain electrostatic parameters as well as polarizabilities, whereas the OPLS, AMBER and CHARMM force fields only contain partial charges. For water, many more different models exist, both with or without polarizabilities.[39] These water models can also be used in embedding potentials. O and H charges are given in Table 4.2 for water models that can be used in embedding potentials.

Table 4.2. Charges for oxygen and hydrogen atoms in water from TIP3P[105], SPC[106], Ahlström[107], **Paper IV** and CHelpG from Ref. 104.

atom	TIP3P	SPC	Ahlström	**Paper IV**	CHelpG
O	-0.834	-0.82	-0.6690	-0.674	-0.656
H	0.417	0.41	0.3345	0.337	0.328

TIP3P[105] and SPC[106] differ mainly in the structure and have very similar charges. Since these two water models do not include polarizabilities, their charges are higher to include some polarization effects implicitly. The water model of Ahlström *et al.* is based on the SPC structure but has a polarizability on oxygen in addition, hence its name *polarizable* SPC or PSPC.[107] Its charges (-0.669 for oxygen) are very similar to the ESP-fitted charges from **Paper IV** (-0.674 using B3LYP/aug-cc-pVTZ RESP fitting) and from Schwabe[104] (-0.656 using B3LYP/aug-cc-pVTZ CHelpG fitting). All of these water models can in principle be used in embedding potentials as long as polarization effects are either treated implicitly (TIP3P or SPC, electrostatic embedding) or explicitly (Ahlström or ESP-fitted charges, PE) but not both.

Paper IV contains ESP-fitted charges and isotropic polarizabilities for common solvent molecules. Similar data sets exist for proteins.[108–110] Genheden, Söderhjelm and Ryde have presented transferable ESP-fitted charges for all common protein residues,[109] building on previous work.[108] The charges are derived using the MK scheme and averaged over different conformations to remove the conformational dependence. They have shown that the resulting charges are independent of the specific protein used to de-

rive them and are thus transferable from one protein to another.[109] Söderhjelm, Kongsted and Ryde have presented a set of transferable isotropic polarizabilities for proteins.[110]

Some care should be exercised with the exclusion list when using polarizabilities obtained in different ways.[102] For QM-derived averaged parameters such as the ones in **Paper IV**, the same rule applies as discussed in Section 4.1.2: atoms that are present in the same QM calculation should be excluded from polarizing each other. Thus, atoms from the same solvent molecules are excluded since the QM calculations from which the average parameters were obtained contained one complete solvent molecule.

Finally, also the LJ parameters r and ϵ in an embedding potential can be obtained from force fields. Since the LJ interaction in Eq. 2.20 is a classical interaction between atom pairs in the QM and classical region, also the atoms in the QM region need to be parametrized. One should be careful that the parameters for the QM and classical region can be used together.[43,111] In **Paper VII**, compatibility has been ensured by choosing LJ parameters from the AMBER force field, which have been published for all AMBER atom types by Cornell *et al.*[66] and modified to be used in QM/MM calculations by Freindorf *et al.*[111] Other sets of LJ parameters for QM/MM calculations exist as well,[43] and there is some evidence that the geometry around the QM region is sensitive to the choice of LJ parameters.[36]

4.3 Accuracy of the parameters

Several possible choices to obtain parameters for an embedding potential have been discussed in Sections 4.1 and 4.2. The choice of parameters has profound implications on both the accuracy of the resulting potential (reviewed in this section) and on the computational time that is spent on obtaining the potential (discussed in Section 4.4). Results from the literature on solute–solvent systems[40,57,94] and proteins[91,110,112,113] will be discussed separately in Section 4.3.1 and Section 4.3.2, respectively. The focus is here on the accuracy of potentials that can be used in PE calculations. Where possible, emphasis is on tests of the ESP rather than other properties because the ESP enters the PE calculation (Section 2.2) and is thus the most direct way to assess the accuracy of an embedding potential.

4.3.1 Solvent systems

Holt and Karlström have investigated the effect of dipole–quadrupole polarizabilities in addition to dipole–dipole polarizabilities to describe the response of small molecules to an applied electric field.[57] They found that the error compared to a QM reference decreases by as much as a factor of two as a result of including also dipole–quadrupole polarizabilities when a homogeneous electric field is applied. The difference is smaller but still present for an inhomogeneous electric field.

Olsen, Aidas and Kongsted have investigated the accuracy of the molecular ESP of a water molecule generated by different sets of embedding parameters compared to a QM reference.[40] They found that the LoProp electric multipole expansion is more or less converged when including quadrupoles (M2). Adding parameters not only to atoms but also to bond midpoints was shown to give only a very small increase in accuracy. The ESP generated by charges only (Ahlström's model[107] or LoProp monopoles, M0) had a larger error especially when the ESP is evaluated at a surface close to the molecule. A LoProp electric multipole expansion up to dipoles (M1), however, was shown to give even larger errors. This unexplained observation—found in other works as well[94,112]—points to a weakness in the LoProp approach, namely that there is no consistent improvement when increasing the level of truncation in the multipole expansion. The accuracy of the polarizabilities was investigated in the same work by calculating the induced ESP of water caused by an applied electric field and comparing to a QM reference.[40] In this way, it was found that the use of polarizabilities on all atoms is better than the one molecular polarizability located on the oxygen atom as used in Ahlström's force field. Going from isotropic (P1) to anisotropic dipole–dipole polarizabilities (P2) had have a small effect only, explained by the low anisotropy of water. The accuracy of excitation energies for organic molecules in water followed the same trends as the accuracy of the ESP of one water molecule, showing that the M2P2 potential is an accurate choice for the water molecule and that the minimum requirement is charges and isotropic polarizabilities.[40]

Schwabe *et al.* extended the analysis of solvent ESPs to include also acetonitrile, methanol and carbon tetrachloride.[94] As for the electrostatic part, the same trend was observed as in Ref. 40: convergence of the LoProp

multipole expansion at quadrupoles with dipoles performing worse than monopoles, especially at short distances. Interestingly, ESP-fitted charges (CHelpG in this case) showed a performance almost as good as the LoProp M2 parameters with the exception of water. The results in Ref. 94 also indicate that dipole–dipole polarizabilities lead to a more accurate response to an applied electric field for acetonitrile and carbon tetrachloride—where the polarizabilities have large magnitudes—whereas differences are small for water and methanol.

We have calculated ESP errors for M* and LoProp M2 parameters over a much broader range of solvents in **Paper IV**, averaging over different conformations rather than focusing on a single one. We found that on average the M2 parameters lead to a more accurate description of the QM ESP than the ESP-fitted charges when evaluated at a surface at twice the vdW radii of the solvent atoms. Notable exceptions are formamide and the chloromethanes—rather isotropic molecules—where the difference in error between M* and M2 is very small. The effect of using averaged ESP-fitted charges was also investigated since the transferability of these parameters can be used to construct embedding potentials without the need for explicit QM calculations. The use of averaged parameters of course lowers the accuracy with respect to a geometry-specific QM ESP, but the error as a result of the averaging was much lower than the error resulting from the use of M* parameters (rather than M2) for most molecules. Interestingly, a test on ethanol showed that the error of averaging ESP-fitted charges was lower for RESP charges (2.43 kJ/mol, averaged over 10 geometries) than for the other ESP-fitting schemes MK (4.32 kJ/mol), HLY (4.47 kJ/mol) and CHelpG (6.27 kJ/mol) by comparing the ESP of averaged ESP-fitted charges to the ESP of geometry-specific ESP-fitted charges. Even though the additional constraints in RESP introduce a larger error than for other fitting schemes, the combined result was that averaged RESP charges are more accurate than averaged ESP-fitted charges with other fitting schemes. Possibly, even smaller errors can be obtained by carefully testing which constraints should be used in the ESP-fitting, and which are automatically taken care of by the averaging procedure. One can also consider to include more than one geometry in the fitting procedure, thus to obtain *one* set of fitted charges by minimizing the error of the ESP of *several* molecules simultaneously. **Paper IV** also contains a test of the *averaged* isotropic

polarizabilities by comparing the induced ESP at an applied electric field to a QM reference. The error for toluene, phenol and benzene was be notably higher than for the other solvents. This was found to be a result not of the use of averaged parameters, but of the use of isotropic values. Indeed, also the (isotropic) ESP-fitted charges were found to give a much larger error than the (anisotropic) M2 parameters for these three molecules.

Taken together, these results indicate that in general an M2P2 potential is an accurate choice with isotropic M*P1 potentials as a good alternative for those molecules that do not have an unusually high anisotropy.

4.3.2 Proteins

The size of proteins makes it necessary to fragment the molecule into smaller parts (Section 4.1.3), which introduces additional errors and requires additional tests.

Olsen *et al.* have investigated the accuracy of amino acid ESPs for different choices of embedding potential.[112] Analysis of four single amino acids reveals some of the same trends in electrostatic parameters as found for solvents: the LoProp multipole expansion is approximately converged at M2 (with the error decreasing as M1 > M0 > M2 > M3) and embedding potentials based on ESP-fitted charges have a moderately (but consistently) higher error than M2 potentials. For dipeptides, it was found that polarization largely compensates for the error introduced by the fragmentation scheme with anisotropic polarizabilities giving only slightly lower errors than isotropic polarizabilities. Calculations on a complete insulin protein showed that a fragmentation scheme in combination with an M2P2 embedding potential gives a better protein ESP than potentials based on ESP-fitted charges with MP2 calculations on the whole protein as reference. Also, a full-structure B3LYP calculation gave a much larger error (43.9 kJ/mol) than the M2P2 (4.7 kJ/mol), M2P1 (6.0 kJ/mol) or CAM-B3LYP (7.7 kJ/mol) potentials,[112] which was attributed to the self-interaction error in KS DFT as described by Jakobsen *et al.*[114] The numbers show that DFT-based potentials (M2P2 and M2P1) on a fragmented protein give lower errors than full-structure DFT calculations (B3LYP and CAM-B3LYP), which is due to the addition of intramolecular polarization effects.

Söderhjelm *et al.* have looked at the basis set dependence of the error of

LoProp multipoles for different models of capped amino acids.[91] Interestingly, they found that the error is larger for diffuse basis sets than for basis sets without diffuse functions in comparison to a QM reference calculated with the same basis set. Thus, while the ESP might be more accurate with a diffuse basis set, the price to pay is a larger error in the ESP from the multipole expansion. The observation was explained by the diffuse character of the resulting wave function (and electron density), for which the multipole expansion converges at a longer distance.[91] In the same study, a multipole expansion up to quadrupoles was found to be more or less converged for most molecules, but few non-polar residues needed higher multipoles to be properly described. At an applied field of 0.01 a.u. in one of the Cartesian directions, the error of induced ESPs from LoProp polarizabilities was comparable or lower than that of a multipole expansion up to quadrupoles, which seems to indicate that the error in the induced ESP is lower than the error of the electrostatic ESP for realistic field strengths.[91]

A systematic study of the ESP gives the most direct way to evaluate the accuracy of an embedding potential,[112] but other studies on molecular properties can give additional insights.[110,113] Söderhjelm has investigated different polarization models and their influence on the accuracy of protein–ligand interaction energies.[113] One interesting conclusion is that the error in using averaged rather than geometry-specific charges is lower when (isotropic) polarizabilities are used in addition to the charges, *i.e.*, the inclusion of polarization leads to better transferability of parameters. The error of the interaction energies was 50 to 65 % larger when using isotropic rather than anisotropic polarizabilities.[113] Söderhjelm, Kongsted and Ryde have studied the conformational dependence of isotropic polarizabilities in proteins.[110] They found that the magnitude of the isotropic polarizabilities is converged for the aug-cc-pVTZ basis set, with the aug-cc-pVDZ basis set having rather small deviations. An important conclusion is that good transferability can be obtained only when the parameters are assigned based on specific atoms in specific amino acids instead of averaging over elements or atom types.[110] In this way, the variation between parameters obtained from different proteins is small, making this a successful strategy to develop transferable embedding parameters for proteins.

The reviewed works on embedding parameters on proteins reveal that the same principles govern the accuracy of embedding parameters for solvent

molecules and proteins, but that polarizabilities serve an additional role in proteins, namely compensating for the error introduced by the fragmentation of the protein. The removal of the anisotropy of the polarizabilities gives an error that depends on e.g. the molecule and the property of interest, but can be a price worth paying when one is interested in transferable embedding parameters.

4.4 Computational cost of embedding potentials

Several ways to obtain parameters for an embedding potential have been introduced in Sections 4.1 and 4.2. The approaches differ both in the accuracy (Section 4.3) and in the computational cost of their calculation. The computational cost of calculating different types of embedding potentials will be made explicit in this section. In addition, some strategies to reduce computational cost are discussed.

The cost of taking embedding parameters from databases (Section 4.2) and building an embedding potential with these is negligible. Indeed, this is a matter of seconds for embedding potentials of most molecular systems as long as an efficient script is available. Even if anisotropic parameters for fixed molecular geometries have to be rotated to match the orientation of the molecule,[40,94,104] this is still orders of magnitude faster than an explicit QM calculations on the molecular fragments. Of course, one should not forget the computational cost needed to obtain the database parameters as done in **Paper IV**, but the cost of *using* these parameters is minimal.

The cost of calculating embedding parameters from QM calculations depends mainly on the number and the size of the molecules or molecular fragments and the method used for the QM calculation. For embedding potentials of homogeneous solvent systems, this is a straightforward multiplication of the number of solvent molecules with the average time needed to calculate the parameters for one molecule given a particular method to calculate the parameters. Table 4.3 shows the averaged computational time needed to calculate M2P2 parameters for a solvent molecule relative to the time needed to calculate this for the smallest solvent molecule in the set, water.

Table 4.3. Relative computational time (relative to water) for a B3LYP/aug-cc-pVDZ calculation to calculate electric multipoles up to quadrupoles and anisotropic dipole–dipole polarizabilities (M2P2) using the LoProp approach[90] in Molcas.[115,116] The number of atoms (N) in the molecule and the number of contracted basis functions (# BF) are also shown. The numbers are averages over 10 different solvent geometries.

solvent	N	# BF	time
water	3	41	1.00
methanol	6	82	6.94
formamide	6	96	10.8
chloroform	5	113	13.0
ethanol	9	123	24.0
dimethyl sulfoxide	10	144	44.9
benzene	12	186	87.3
diethyl ether	15	205	110
toluene	15	233	165
hexane	20	264	226

It is clear that the computational time increases rapidly with the size of the molecule because of the higher number of basis functions involved in the calculation. However, one also needs to take into account that the number of solvent molecules within a threshold R from a QM region decreases with the size of the solvent molecules. The average number of solvent molecules within a threshold R for different solvent molecules is plotted in Figure 4.4 for the geometries of acetone in different solvents from **Paper III**.

A snapshot of acetone (including all solvent molecules that have at least one atom within a threshold of 15 Å) in water has 5.21 times as many molecules as a snapshot of acetone in hexane (Figure 4.4). One calculation on hexane takes 226 times longer than one calculation on water (Table 4.3). Thus, creating an embedding potential for acetone in hexane is in this case approximately 43 times more expensive than creating an embedding potential for acetone in water.

The dependence of the computational time on basis set and type of embedding potential are illustrated in Tables 4.4 and 4.5, respectively. The variation of computational time with basis set (Table 4.4) is significant, in-

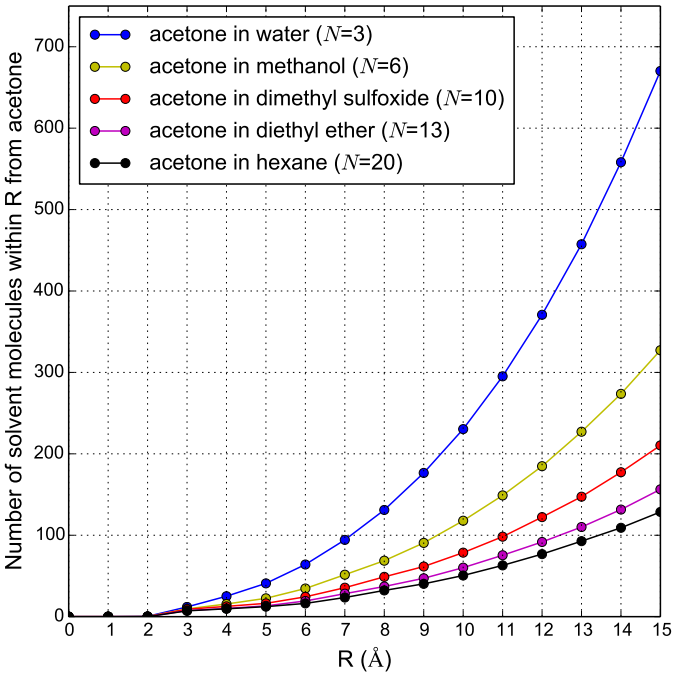


Figure 4.4. Number of solvent molecules with at least one atom within a distance R (in Å) from acetone. N is the number of atoms in the solvent. The numbers are averages over the 50 snapshots of acetone in different solvents used in **Paper III**.

dicating that it is worthwhile to investigate the accuracy of different basis sets before choosing which one to use to calculate embedding potential parameters. Indeed, removing the set of augmented functions of aug-cc-pVDZ saves a factor of 2 to 2.5 in computational time, while going from double-ζ to triple-ζ gives an increase of a factor of 4 to 5. The dependence of the computational time on the basis set depends mainly on the number of basis functions, given by the number and type of all elements in the molecule.

The computational time is more or less independent of the order of truncating the electric multipoles k (Table 4.5). Calculating only electric multipoles and no polarizabilities, however, makes the LoProp calculation around 4 times faster. Indeed, the calculation of LoProp polarizabilities relies on numerical differentiation of multipoles in the presence and absence

Table 4.4. Relative computational time (relative to aug-cc-pVDZ) for a B3LYP calculation to calculate electric multipoles up to quadrupoles and anisotropic dipole–dipole polarizabilities (M2P2) using the LoProp approach[90] in Molcas[115,116] for water and methanol. The number of atoms (N) in the molecule and the number of contracted basis functions (# BF) are also shown. The numbers are averages over 10 different solvent geometries.

water (N=3)			**methanol (N=6)**		
basis set	# BF	time	basis set	# BF	time
cc-pVDZ	24	0.50	cc-pVDZ	48	0.39
aug-cc-pVDZ	41	1.00	aug-cc-pVDZ	82	1.00
cc-pVTZ	58	1.76	cc-pVTZ	116	1.86
aug-cc-pVTZ	92	4.14	aug-cc-pVTZ	184	5.21

of an applied electric field in three Cartesian directions and thus requires six additional multipole calculations. The generation of embedding potentials can thus be made much faster when polarizabilities are calculated only for a subset of sites in the classical region (**Paper III**). Since the neglect of polarization in part of the classical region leads to poor results (as shown in **Paper III**), it would be interesting to investigate the quality of embedding potentials with geometry-specific electrostatic parameters in combination with polarizabilities from a database (such as those from **Paper IV**).

The computational cost of calculating an embedding potential for a protein with the MFCC procedure (Section 4.1.3) is less straightforward to evaluate since the amino acids have different sizes and additional conjugate cap calculations are required. The largest amino acid—tryptophan—requires a QM calculation of the capped residue with 35 atoms (assuming the capping is done as described in Section 4.1.3) and is therefore more expensive than any of the solvent molecules in Table 4.3. Even the smallest amino acid—glycine—requires a QM calculation with 18 atoms and 288 basis functions (based on the aug-cc-pVDZ basis set), which is also more expensive than a calculation on hexane (20 atoms, 264 basis functions with aug-cc-pVDZ). Calculating the embedding potential for a protein with N residues with the MFCC procedure also requires the calculation of $N-1$

Table 4.5. Relative computational time (relative to M2P2) for a B3LYP/aug-cc-pVDZ calculation to calculate electric multipoles up to order k with (MkP2) and without (Mk) anisotropic dipole–dipole polarizabilities using the LoProp approach[90] in Molcas[115,116] for methanol and hexane. The numbers are averages over 10 different solvent geometries.

Mk	relative time	MkP2	time
	methanol		
M0	0.22	M0P2	1.00
M1	0.22	M1P2	1.00
M2	0.22	M2P2	1.00
M3	0.22	M3P2	1.01
M4	0.22	M4P2	1.01
	hexane		
M0	0.27	M0P2	1.00
M1	0.27	M1P2	1.00
M2	0.27	M2P2	1.00
M3	0.27	M3P2	1.00
M4	0.27	M4P2	1.00

conjugate caps. The GFP model used in **Paper I**, **Paper II** and **Paper III** has 230 residues, thus it requires 230 calculations of capped amino acids (18 to 35 atoms) and 229 conjugate caps (12 atoms). In addition, the model contains 142 water molecules (3 atoms), which are however fast to calculate. This makes the calculation of the embedding potentials of 50 snapshots of seven different proteins in **Paper I** very costly and motivates the work to derive general protein embedding parameters.

From the analysis above it is clear that different strategies can be followed to reduce the cost of generating the embedding potentials: reducing the size of the classical region, taking embedding parameters from a database or using a cheaper QM method for the calculation of (a part of) the parameters. Even though a reduction of the total size of the molecular system significantly speeds up the calculation of an embedding potential (*cf.* Figure 4.4), one should carefully test up to which distance the molecular environment influences the molecular properties of the

QM region (**Paper II**, **Paper III** and **Paper IV**) or use a dielectric continuum outside the explicit molecular system.[117] **Paper III** has shown that long-range polarization effects cannot be omitted in the calculation. However, isotropic solvent-specific polarizabilities suffice for the description of molecular fragments further away from the central molecule (**Paper IV**). The strategies and parameters described in **Paper IV** are useful to keep the cost of the calculation of the embedding potentials low without compromising the accuracy. Indeed, **Paper IV** shows that accurate embedding potentials can be obtained by calculating accurate embedding parameters for the solvent molecules closest to the solvated molecule while using averaged parameters for the solvent molecules further away. The use of more approximate parameters for molecules further away—rather than excluding the molecules from the calculation—has previously been shown to be successful.[118,119] This procedure leads to a dramatic reduction of the computational cost compared to embedding potentials where all molecules have geometry-specific parameters, which has been a common procedure for PE calculations. The averaged parameters also allow for PE calculations with *only* averaged parameters—albeit at a reduced accuracy—removing the need for *any* QM-based parameters and making the barrier for using the PE model much lower for new users.

It is important to put the computational time required to calculate an embedding potential in context and consider the *total* computational time required by a procedure to calculate a molecular property. In some cases, the calculation of the embedding potential makes up only a small percentage of the total computational cost of calculating a molecular property. An example of this is the calculation of a property that in itself is expensive to calculate (e.g. three-photon absorption) for a relatively large molecule in a relatively small solvent (e.g. chloroform). In other cases, the calculation of the embedding potentials constitutes the major part of total computational resources spent on a given procedure. Examples of this include the calculation of excitation energies of fluorescent proteins in **Paper I** and a relatively inexpensive QM property calculation such as the dipole moment of a small molecule (PNA, acetone) in a solvent such as dimethyl sulfoxide or propylene carbonate (**Paper IV**). In these cases, it is relevant to consider strategies to obtain the embedding potentials at reduced computational cost.

Chapter 5

Summary and perspective

This chapter presents an overview of the findings of this thesis, as well as its main methodological shortcomings. Moreover, it is shown how the results can contribute to the calculation of accurate multiphoton absorption (MPA) properties in a molecular environment as well as to the calculation of accurate embedding potentials at reduced computational cost. Related to those topics, a perspective is given on important directions for future research on the accurate calculation of molecular properties in realistic environments.

The aim of this thesis is the accurate calculation of molecular properties in realistic environments. It has been shown that the polarization of the environment has a considerable impact on the calculation of molecular properties. This is especially the case for properties related to the absorption of light by a molecule, which causes a reorganization of the charge distribution in the molecule and induces a change in polarization state of the environment. For an accurate description of the influence of large environments on a molecular property, one needs to take into account electrostatic and polarization interactions. The PE method described in Section 2.2 is an accurate way of describing these interactions when electrostatic (including polarization) interactions dominate. For localized properties such as excitation energies, the PE model is preferred over a cluster approach, which can only describe a relatively small part of the environment. For the part of the molecular environment that is further away from the molecule, less accurate embedding parameters suffice, leading to a reduction of the cost of obtaining an accurate embedding potential. Moreover, it is shown that DFT calculations can give qualitatively good results for TPA, while absolute values for the TPA cross section should be evaluated with great care.

The main shortcomings of the works presented in this thesis are the quality of the molecular structures and the neglect of non-electrostatic interactions between the molecule and its environment in the molecular property calculations. The former can be improved by performing QM/MM or *ab initio* rather than classical MD simulations. This also removes the need for QM/MM geometry optimization (Section 3.2.2), thus keeping all dynamical effects from the conformational sampling at a finite temperature. The latter is most problematic for apolar solvents (where non-electrostatic effects are relatively more important) and for molecular systems with explicit boundaries or charge transfer between the QM and classical regions. Including exchange–repulsion interactions can be done by a method like PDE.[61] Since non-electrostatic interactions are short-range, this is especially important for the closest environment of the central subsystem.

Ideally, however, the same (accurate) description of the interactions between the central subsystem and its environment should be used in both the sampling of the conformations and the calculation of the molecular

properties. This would be a considerable improvement over the two-step approach used in this thesis, namely using one method for the conformational sampling and another one for the calculation of molecular properties with a different description of the intermolecular interaction for each of the two steps. Indeed, the MD simulations in this thesis are done with flexible molecules described by classical mechanics without polarization, while the molecular properties are calculated using a mixed quantum and classical method in a polarizable—but frozen—environment.

The results in this thesis motivate further research to investigate the accuracy of calculated TPA and MPA strengths to enable a comparison with experiments. **Paper VI** has shown that the magnitude of TPA strengths calculated with DFT can significantly deviate from CC calculations, which is partially attributed to an underestimation of difference dipole moments by DFT. This is one of the factors that currently limits the comparison of calculated TPA cross sections of medium- and large-sized molecules with experiments. There is however some evidence that *relative* trends between different chromophores and between different solvents are correctly reproduced also by multiscale DFT calculations using the PE method.[120] Building on the work of Hršak *et al.*[120,121] and the work presented in this thesis (**Paper V** and **Paper VI**), one could investigate the quality of TPA strengths calculated with DFT in a polarizable environment. This is made possible by the recently published implementation of PE with CC2 for TPA.[121] In particular, it would be interesting to find out whether difference dipole moments calculated with DFT in a molecular environment are closer to the CC results than in the gas phase and whether the differences found in **Paper VI** are systematic overestimations or average out over different conformations.

In general, the DFT calculations presented in this thesis suffer from an overestimation of the excitation energies and thus wrong prediction of the location of absorption peaks in a spectrum. This makes comparison between calculated and experimental data difficult. The problem is related to the density functional (CAM-B3LYP), which has been chosen on the basis of its *qualitatively* good performance in describing charge-transfer transitions compared to other density functionals.[32] There seems to be no density functional that can reliably describe properties of a wide range of molecules and protonation states both qualitatively and quantitatively (here: excitation

energies *and* intensities). This motivates work in the development of new improved density functionals, for which a clear demonstration of the current problems (such as in **Paper VI**) is helpful. In parallel, benchmark methods such as CC methods are needed to test the performance of DFT methods. One can entirely refrain from using DFT only for small molecular systems. For larger biomolecular systems, however, DFT will always be able to treat larger systems at the same expense, or the same systems at a much lower expense.

Building on the work presented in this thesis, the influence of the molecular environment on MPA properties of (bio)molecules could be investigated in more detail.[122] **Paper I**, **Paper II** and **Paper III** have shown the importance of polarization on the excitation energies of fluorescent proteins. It has also been shown that the polarization significantly modifies the oscillator strength[123–125] and the two-photon absorption cross section.[123,124] It is thus likely that also higher-order MPA strengths are critically dependent on the environment in general and on the polarization of the environment in particular. A DFT implementation for the calculation of MPA strengths to arbitrary order has recently been published.[126] The extensions to PE[122] and PCM[127] are work in progress and will enable the investigation of the effect of the environment on MPA properties.

Several challenges arise in the calculation of higher-order MPA strengths. While the calculation of MPA strengths with DFT is possible for any number of photons,[126] not much is known about the accuracy of these numbers. In fact, transition moments between different excited states become increasingly important for higher-order transition properties and the performance of DFT for those is not well-described but likely rather poor. This makes the comparison of calculated and measured MPA cross sections even more challenging. Currently this is a difficult question to address because of the lack of reference methods beyond TPA[128] for all but the smallest molecular systems.[129]

Another challenge is related to the size of the QM region and the location of the QM–MM boundary in MPA calculations.[122] For a localized property (such as OPA) in a large molecular environment, **Paper II** has shown that a polarizable QM/MM method is to be preferred over a QM cluster model. The electronic transition studied in **Paper II** is a transition between two orbitals that are localized on the conjugated system

of the chromophore. MPA processes, however, are not necessarily localized properties and have many more contributing orbitals. It is thus difficult to choose a sensible location for the QM–MM boundary and larger QM regions are needed for accurate calculations.[122] This motivates further research into flexible and accurate descriptions of covalent links between the QM and classical parts in QM/MM calculations, which is one of the main challenges in the multiscale modeling of biomolecular systems.

Accurate incorporation of the effect of a molecular environment requires a large number of molecular structures to be taken into account. This requires not only a large number of PE calculations, but also a large number of embedding potentials. One strategy to reduce the computational cost of calculating solvent embedding potentials has been presented in **Paper IV**: use geometry-specific parameters for the most important part of the environment and averaged solvent-specific parameters for the rest. One can also think of other combinations of embedding parameters from QM calculations and databases. It is shown in Chapter 4 that the computational time of calculating only electric multipoles is four times less than calculating both electric multipoles and polarizabilities. This makes it worthwhile to investigate the quality of embedding potentials that consist of QM-based electric multipoles (Mk) for all classical sites and averaged (isotropic) polarizabilities on all or a part of all classical sites. Given the results in **Paper IV**, this is likely a cost-effective strategy for most solvent molecules, with the exception of anisotropic molecules such as phenol, toluene and benzene. Research in this direction can make accurate QM/MM calculations attractive for a wider range of users by removing part of the effort and computational cost to generate an accurate parametrization of the classical part.

The advantage of reducing the cost of calculating embedding potentials is not necessarily in the reduction of total cost. Rather, one can increase the quality of the calculations, e.g. by using PDE in the QM/MM calculation or by increasing the number of MD snapshots to reduce the statistical error of averaging properties over different structures. Another option is to investigate possible improvements of the polarization interactions by going beyond the linear response (using hyperpolarizabilities) or by including also dipole–quadrupole polarizabilities to improve the description of the linear response. Thus, one can increase the overall quality of the calculation by

using a given amount of computational resources in the most efficient way. This requires more research into the relative importance of factors such as density-based embedding, non-linear response to an applied field and non-electrostatic effects. Computational studies such as those presented in this thesis are fundamental to show *how* accurate calculations of molecular properties in realistic molecular environments should be performed to obtain accurate results.

Bibliography

[1] F. Jensen. *Introduction to Computational Chemistry.* John Wiley & Sons, Chichester, 1999.

[2] D. R. Hartree. The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods. *Math. Proc. Camb. Phil. Soc.*, 24:89–110, 1928.

[3] D. H. Whiffen. Expression of results in quantum chemistry. *Pure Appl. Chem.*, 50:75–79, 1978.

[4] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Ann. Phys.*, 389:457–484, 1927.

[5] P. Atkins and R. Friedman. *Molecular Quantum Mechanics.* Oxford University Press, Oxford, fourth edition, 2005.

[6] C. J. Cramer. *Essentials of Computational Chemistry. Theories and Models.* John Wiley & Sons, Chichester, second edition, 2004.

[7] J. C. Slater. Note on Hartree's method. *Phys. Rev.*, 35:210, 1930.

[8] P. R. Taylor. Accurate calculations and calibration. In R. Bast and P.-O. Widmark, editors, *European Summerschool in Quantum Chemistry. Book III*, pages 663–739. ESQC committee, eighth edition, 2013.

[9] R. Shepard. The multiconfiguration self-consistent field method. *Adv. Chem. Phys.*, 69:63–200, 1987.

[10] J. Čížek. On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods. *J. Chem. Phys.*, 45:4256–4266, 1966.

[11] O. Christiansen, H. Koch, and P. Jørgensen. The second-order approximate coupled cluster singles and doubles model CC2. *Chem. Phys. Lett.*, 243:409–418, 1995.

[12] T. Helgaker, S. Coriani, P. Jørgensen, K. Kristensen, J. Olsen, and K. Ruud. Recent advances in wave function-based methods of molecular-property calculations. *Chem. Rev.*, 112:543–631, 2012.

[13] T. Helgaker. Analytic calculation of time-independent molecular properties. In R. Bast and P.-O. Widmark, editors, *European Summerschool in Quantum Chemistry. Book II*, pages 411–449. ESQC committee, eighth edition, 2013.

[14] H. Hellmann. *Einführung in die Quantenchemie.* Franz Deuticke, Leipzig, 1937.

[15] R. P. Feynman. Forces in molecules. *Phys. Rev.*, 56:340, 1939.

[16] T. Helgaker and P. Jørgensen. Configuration-interaction energy derivatives in a fully variational formulation. *Theor. Chim. Acta*, 75:111–127, 1989.

[17] P. Norman, K. Ruud, and T. Saue. *Response Properties of Molecular Materials.* Wiley. *In preparation.*

[18] J. Olsen and P. Jørgensen. Linear and nonlinear response functions for an exact state and for an MCSCF state. *J. Chem. Phys.*, 82:3235–3264, 1985.

[19] P. R. Monson and W. M. McClain. Polarization dependence of the two-photon absorption of tumbling molecules with application to liquid 1-chloronaphthalene and benzene. *J. Chem. Phys.*, 53:29–37, 1970.

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

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

[22] D. J. Tozer. Density functional theory. In R. Bast and P.-O. Widmark, editors, *European Summerschool in Quantum Chemistry. Book II*, pages 525–568. ESQC committee, eighth edition, 2013.

[23] J. C. Slater. A simplification of the Hartree–Fock method. *Phys. Rev.*, 81:385–390, 1951.

[24] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis. *Can. J. Phys.*, 58:1200–1211, 1980.

[25] P. A. M. Dirac. Note on exchange phenomena in the Thomas atom. *Math. Proc. Camb. Phil. Soc.*, 26:376–385, 1930.

[26] J. P. Perdew and Y. Wang. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Phys. Rev. B*, 33:8800–8802, 1986.

[27] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098, 1988.

[28] C. Lee, W. Yang, and R. G. Parr. Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, 1988.

[29] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98:5648–5652, 1993.

[30] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. *Ab Initio* calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.*, 98:11623–11627, 1994.

[31] T. Yanai, D. P. Tew, and N. C. Handy. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.*, 393:51–57, 2004.

[32] M. J. G. Peach, P. Benfield, T. Helgaker, and D. J. Tozer. Excitation energies in density functional theory: An evaluation and a diagnostic test. *J. Chem. Phys.*, 128:044118, 2008.

[33] J. Tomasi, B. Mennucci, and R. Cammi. Quantum mechanical continuum solvation models. *Chem. Rev.*, 105:2999–3094, 2005.

[34] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103:227–249, 1976.

[35] M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber, and K. Morokuma. ONIOM: A multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels–Alder reactions and $Pt(P(t\text{-}Bu)_3)_2 + H_2$ oxidative addition. *J. Phys. Chem.*, 100:19357–19363, 1996.

[36] H. M. Senn and W. Thiel. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.*, 48:1198–1229, 2009.

[37] A. D. Buckingham. Permanent and induced molecular moments and long-range intermolecular forces. *Adv. Chem. Phys.*, 12:107–142, 1967.

[38] A. J. Stone. *The Theory of Intermolecular forces.* Oxford University Press, Oxford, second edition, 2013.

[39] A. R. Leach. *Molecular Modelling. Principles and Applications.* Pearson Education, Harlow, second edition, 2001.

[40] J. M. Olsen, K. Aidas, and J. Kongsted. Excited states in solution through polarizable embedding. *J. Chem. Theory Comput.*, 6:3721–3734, 2010.

[41] J. M. H. Olsen and J. Kongsted. Molecular properties through polarizable embedding. *Adv. Quantum Chem.*, 61:107–143, 2011.

[42] D. M. Philipp and R. A. Friesner. Mixed *ab initio* QM/MM modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide. *J. Comput. Chem.*, 20:1468–1494, 1999.

[43] R. B. Murphy, D. M. Philipp, and R. A. Friesner. A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. *J. Comput. Chem.*, 21:1442–1457, 2000.

[44] T. Vreven, K. S. Byun, I. Komáromi, S. Dapprich, J. A. Montgomery Jr., K. Morokuma, and M. J. Frisch. Combining quantum

mechanics methods with molecular mechanics methods in ONIOM. *J. Chem. Theory Comput.*, 2:815–826, 2006.

[45] T. A. Wesolowski and A. Warshel. Frozen density functional approach for *ab initio* calculations of solvated molecules. *J. Phys. Chem.*, 97:8050–8053, 1993.

[46] P. Cortona. Self-consistently determined properties of solids without band-structure calculations. *Phys. Rev. B*, 44:8454–8458, 1991.

[47] T. A. Wesolowski and J. Weber. Kohn–Sham equations with constrained electron density: An iterative evaluation of the ground-state electron density of interacting molecules. *Chem. Phys. Lett.*, 248:71–76, 1996.

[48] Ch. R. Jacob, J. Neugebauer, and L. Visscher. A flexible implementation of frozen-density embedding for use in multilevel simulations. *J. Comput. Chem.*, 29:1011–1018, 2008.

[49] C. Daday, C. König, O. Valsson, J. Neugebauer, and C. Filippi. State-specific embedding potentials for excitation-energy calculations. *J. Chem. Theory Comput.*, 9:2355–2367, 2013.

[50] L. Jensen, P. Th. van Duijnen, and J. G. Snijders. A discrete solvent reaction field model within density functional theory. *J. Chem. Phys.*, 118:514–521, 2003.

[51] C. Curutchet, A. Muñoz-Losa, S. Monti, J. Kongsted, G. D. Scholes, and B. Mennucci. Electronic energy transfer in condensed phase studied by a polarizable QM/MM model. *J. Chem. Theory Comput.*, 5:1838–1848, 2009.

[52] M. S. Gordon, M. A. Freitag, P. Bandyopadhyay, J. H. Jensen, V. Kairys, and W. J. Stevens. The effective fragment potential method: A QM-based MM approach to modeling environmental effects in chemistry. *J. Phys. Chem. A*, 105:293–307, 2001.

[53] J. J. Eriksen, S. P. A. Sauer, K. V. Mikkelsen, H. J. Aa. Jensen, and J. Kongsted. On the importance of excited state dynamic response electron correlation in polarizable embedding methods. *J. Comput. Chem.*, 33:2012–2022, 2012.

[54] K. Sneskov, T. Schwabe, J. Kongsted, and O. Christiansen. The polarizable embedding coupled cluster method. *J. Chem. Phys.*, 134:104108, 2011.

[55] T. Schwabe, K. Sneskov, J. M. Haugaard Olsen, J. Kongsted, O. Christiansen, and C. Hättig. PERI–CC2: A polarizable embedded RI-CC2 method. *J. Chem. Theory Comput.*, 8:3274–3283, 2012.

[56] E. D. Hedegård, N. H. List, H. J. Aa. Jensen, and J. Kongsted. The multi-configuration self-consistent field method within a polarizable embedded framework. *J. Chem. Phys.*, 139:044101, 2013.

[57] A. Holt and G. Karlström. Inclusion of the quadrupole moment when describing polarization. The effect of the dipole–quadrupole polarizability. *J. Comput. Chem.*, 29:2033–2038, 2008.

[58] J. E. Jones. On the determination of molecular fields. II. From the equation of state of a gas. *Proc. R. Soc. Lond. A*, 106:463–477, 1924.

[59] D. Berthelot. Sur le mélange des gaz. *C. R. Acad. Sci.*, 126:1703–1706, 1898.

[60] H. A. Lorentz. Über die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Ann. Phys.*, 248:127–136, 1881.

[61] J. M. H. Olsen, C. Steinmann, K. Ruud, and J. Kongsted. Polarizable density embedding: A new QM/QM/MM-based computational strategy. *J. Phys. Chem. A*, 119:5344–5355, 2015.

[62] 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. Comput. Chem.*, 4:187–217, 1983.

[63] W. L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[64] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational

energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[65] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105:6474–6487, 2001.

[66] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[67] N. Reuter, H. Lin, and W. Thiel. Green fluorescent proteins: Empirical force field for the neutral and deprotonated forms of the chromophore. Molecular dynamics simulations of the wild type and S65T mutant. *J. Phys. Chem. B*, 106:6310–6321, 2002.

[68] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *J. Mol. Graph.*, 14:33–38, 1996.

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

[70] L. Verlet. Computer "experiments" on classical fluids. II. Equilibrium correlation functions. *Phys. Rev.*, 165:201–204, 1968.

[71] D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team. *GROMACS User Manual version 4.6.5.* 2013. www.gromacs.org.

[72] T. Helgaker. Geometry optimizations for minima and saddle points. In R. Bast and P.-O. Widmark, editors, *European Summerschool in Quantum Chemistry. Book II*, pages 493–524. ESQC committee, eighth edition, 2013.

[73] D. C. Sorensen. Newton's method with a model trust region modification. *SIAM J. Numer. Anal.*, 19:409–426, 1982.

[74] R. Fletcher. *Practical Methods of Optimization.* Wiley, New York, 1987.

[75] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76:637–649, 1982.

[76] R. W. Hockney. The potential calculation and some applications. *Meth. Comput. Phys.*, 9:136–211, 1970.

[77] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[78] H. C. Andersen. RATTLE: A "Velocity" version of the SHAKE algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52:24–34, 1983.

[79] M. E. Tuckerman, B. J. Berne, and A. Rossi. Molecular dynamics algorithm for multiple time scales: Systems with disparate masses. *J. Chem. Phys.*, 94:1465–1469, 1991.

[80] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[81] W. M. Haynes, editor. *CRC Handbook of Chemistry and Physics.* CRC Press, 96th edition, 2015.

[82] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucl. Acids Res.*, 28:235–242, 2000.

[83] T. Okada, M. Sugihara, A.-N. Bondar, M. Elstner, P. Entel, and V. Buss. The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *J. Mol. Biol.*, 342:571–583, 2004.

[84] E. Hodis, J. Prilusky, E. Martz, I. Silman, J. Moult, and J. L. Sussman. Proteopedia—a scientific 'wiki' bridging the rift between three-dimensional structure and function of biomacromolecules. *Genome Biol.*, 9:R121, 2008.

[85] J. Prilusky, E. Hodis, D. Canner, W. A. Decatur, K. Oberholser, E. Martz, A. Berchanski, M. Harel, and J. L. Sussman. Proteopedia: A status report on the collaborative, 3D web-encyclopedia of proteins and other biomolecules. *J. Struct. Biol.*, 175:244–252, 2011.

[86] G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.*, 27:221–234, 2013.

[87] R. S. Mulliken. Criteria for the construction of good self-consistent-field molecular orbital wave functions, and the significance of LCAO-MO population analysis. *J. Chem. Phys.*, 36:3428–3439, 1962.

[88] G. Karlström. In P. Th. van Duijnen and W. C. Nieuwpoort, editors, *Proceedings of the fifth seminar on computational methods in quantum chemistry*. Laboratory of Chemical Physics, University of Groningen, The Netherlands, 1982.

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

[90] L. Gagliardi, R. Lindh, and G. Karlström. Local properties of quantum chemical systems: The LoProp approach. *J. Chem. Phys.*, 121:4494–4500, 2004.

[91] P. Söderhjelm, J. W. Krogh, G. Karlström, U. Ryde, and R. Lindh. Accuracy of distributed multipoles and polarizabilities: Comparison between the LoProp and MpProp models. *J. Comput. Chem.*, 28:1083–1090, 2007.

[92] F. A. Momany. Determination of partial atomic charges from *ab initio* molecular electrostatic potentials. Application to formamide, methanol, and formic acid. *J. Phys. Chem.*, 82:592–601, 1978.

[93] S. R. Cox and D. E. Williams. Representation of the molecular electrostatic potential by a net atomic charge model. *J. Comput. Chem.*, 2:304–323, 1981.

[94] T. Schwabe, J. M. H. Olsen, K. Sneskov, J. Kongsted, and O. Christiansen. Solvation effects on electronic transitions: Exploring the performance of advanced solvent potentials in polarizable embedding calculations. *J. Chem. Theory Comput.*, 7:2209–2217, 2011.

[95] E. Sigfridsson and U. Ryde. Comparison of methods for deriving atomic charges from the electrostatic potential and moments. *J. Comput. Chem.*, 19:377–395, 1998.

[96] U. C. Singh and P. A. Kollman. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.*, 5:129–145, 1984.

[97] L. E. Chirlian and M. M. Francl. Atomic charges derived from electrostatic potentials: A detailed study. *J. Comput. Chem.*, 8:894–905, 1987.

[98] C. M. Breneman and K. B. Wiberg. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.*, 11:361–373, 1990.

[99] H. Hu, Z. Lu, and W. Yang. Fitting molecular electrostatic potentials from quantum mechanical calculations. *J. Chem. Theory Comput.*, 3:1004–1013, 2007.

[100] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.*, 97:10269–10280, 1993.

[101] B. H. Besler, K. M. Merz Jr., and P. A. Kollman. Atomic charges derived from semiempirical methods. *J. Comput. Chem.*, 11:431–439, 1990.

[102] P. Söderhjelm and U. Ryde. How accurate can a force field become? A polarizable multipole model combined with fragment-wise quantum-mechanical calculations. *J. Phys. Chem. A*, 113:617–627, 2009.

[103] D. W. Zhang and J. Z. H. Zhang. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein–molecule interaction energy. *J. Chem. Phys.*, 119:3599–3605, 2003.

[104] T. Schwabe. Assessing molecular dynamics simulations with solvatochromism modeling. *J. Phys. Chem. B*, 119:10693–10700, 2015.

[105] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[106] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. Interaction models for water in relation to protein hydration. In B. Pullman, editor, *Intermolecular Forces*, pages 331–342. Reidel, Dordrecht, 1981.

[107] P. Ahlström, A. Wallqvist, S. Engström, and B. Jönsson. A molecular dynamics study of polarizable water. *Mol. Phys.*, 68:563–581, 1989.

[108] P. Söderhjelm and U. Ryde. Conformational dependence of charges in protein simulations. *J. Comput. Chem.*, 30:750–760, 2009.

[109] S. Genheden, P. Söderhjelm, and U. Ryde. Transferability of conformational dependent charges from protein simulations. *Int. J. Quantum Chem.*, 112:1768–1785, 2012.

[110] P. Söderhjelm, J. Kongsted, and U. Ryde. Conformational dependence of isotropic polarizabilities. *J. Chem. Theory Comput.*, 7:1404–1414, 2011.

[111] M. Freindorf, Y. Shao, T. R. Furlani, and J. Kong. Lennard–Jones parameters for the combined QM/MM method using the B3LYP/6-31G*/AMBER potential. *J. Comput. Chem.*, 26:1270–1278, 2005.

[112] J. M. H. Olsen, N. H. List, K. Kristensen, and J. Kongsted. Accuracy of protein embedding potentials: An analysis in terms of electrostatic potentials. *J. Chem. Theory Comput.*, 11:1832–1842, 2015.

[113] P. Söderhjelm. Polarization effects in protein–ligand calculations extend farther than the actual induction energy. *Theor. Chem. Acc.*, 131:1159–1170, 2012.

[114] S. Jakobsen, K. Kristensen, and F. Jensen. Electrostatic potential of insulin: Exploring the limitations of density functional theory and force field methods. *J. Chem. Theory Comput.*, 9:3978–3985, 2013.

[115] G. Karlström, R. Lindh, P.-Å. 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. *Comput. Mater. Sci.*, 28:222–239, 2003.

[116] F. Aquilante, L. De Vico, N. Ferré, G. Ghigo, P.-Å. Malmqvist, P. Neogrády, T. B. Pedersen, M. Pitoňák, M. Reiher, B. O. Roos, L. Serrano-Andrés, M. Urban, V. Veryazov, and R. Lindh. MOLCAS 7: The next generation. *J. Comput. Chem.*, 31:224–247, 2010.

[117] A. H. Steindal, K. Ruud, L. Frediani, K. Aidas, and J. Kongsted. Excitation energies in solution: The fully polarizable QM/MM/PCM method. *J. Phys. Chem. B*, 115:3027–3037, 2011.

[118] P. Söderhjelm, C. Husberg, A. Strambi, M. Olivucci, and U. Ryde. Protein influence on electronic spectra modeled by multipoles and polarizabilities. *J. Chem. Theory Comput.*, 5:649–658, 2009.

[119] P. Söderhjelm, F. Aquilante, and U. Ryde. Calculation of protein–ligand interaction energies by a fragmentation approach combining high-level quantum chemistry with classical many-body effects. *J. Phys. Chem. B*, 113:11085–11094, 2009.

[120] D. Hršak, L. Holmegaard, A. S. Poulsen, N. H. List, J. Kongsted, M. P. Denofrio, R. Erra-Balsells, F. M. Cabrerizo, O. Christiansen, and P. R. Ogilby. Experimental and computational study of solvent effects on one-and two-photon absorption spectra of chlorinated harmines. *Phys. Chem. Chem. Phys.*, 17:12090–12099, 2015.

[121] D. Hršak, A. M. Khah, O. Christiansen, and C. Hättig. Polarizable embedded RI-CC2 method for two-photon absorption calculations. *J. Chem. Theory Comput.*, 11:3669–3678, 2015.

[122] A. H. Steindal, M. T. P. Beerepoot, M. Ringholm, N. H. List, K. Ruud, J. Kongsted, and J. M. H. Olsen. A polarizable embedding framework

for open-ended electrical response properties: Multiphoton absorption in biomolecular systems. *In preparation.*

[123] A. H. Steindal, J. M. H. Olsen, K. Ruud, L. Frediani, and J. Kongsted. A combined quantum mechanics/molecular mechanics study of the one-and two-photon absorption in the green fluorescent protein. *Phys. Chem. Chem. Phys.*, 14:5440–5451, 2012.

[124] N. H. List, J. M. H. Olsen, H. J. Aa. Jensen, A. H. Steindal, and J. Kongsted. Molecular-level insight into the spectral tuning mechanism of the DsRed chromophore. *J. Phys. Chem. Lett.*, 3:3513–3521, 2012.

[125] A. Pikulska, A. H. Steindal, M. T. P. Beerepoot, and M. Pecul. Electronic circular dichroism of fluorescent proteins: A computational study. *J. Phys. Chem. B*, 119:3377–3386, 2015.

[126] D. H. Friese, M. T. P. Beerepoot, M. Ringholm, and K. Ruud. Open-ended recursive approach for the calculation of multiphoton absorption matrix elements. *J. Chem. Theory Comput.*, 11:1129–1144, 2015.

[127] R. Di Remigio, M. T. P. Beerepoot, Y. Cornaton, M. Ringholm, A. H. Steindal, K. Ruud, and L. Frediani. Open-ended formulation of self-consistent field response theory with the polarizable continuum model for solvation. *In preparation.*

[128] K. D. Nanda and A. I. Krylov. Two-photon absorption cross sections within equation-of-motion coupled-cluster formalism using resolution-of-the-identity and Cholesky decomposition representations: Theory, implementation, and benchmarks. *J. Chem. Phys.*, 142:064118, 2015.

[129] M. J. Paterson, O. Christiansen, F. Pawłowski, P. Jørgensen, C. Hättig, T. Helgaker, and P. Sałek. Benchmarking two-photon absorption with CC3 quadratic response theory, and comparison with density-functional response theory. *J. Chem. Phys.*, 124:054322, 2006.