Computational Science and Engineering
(International Master's Program)

Technische Universität München

Master's Thesis

Adaptive QM/MM simulations of proton transfer in water with an energy based reaction coordinate

Author:	Alberto Pérez de Alba Ortíz
1st examiner:	Prof. Dr. Karsten Reuter
2nd examiner:	Prof. Dr. Thomas Huckle
Assistant advisor(s):	Dr. Harald Oberhofer, Dr. Thomas Stecher
Thesis handed in on:	October 21, 2016

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

October 21, 2016 Alberto Pérez de Alba Ortíz

Acknowledgments

I have been very fortunate to have many people helping me bring to life this thesis. I would like to give them all special thanks in these lines. First of all, I would like to thank Thomas Stecher for all his advice and encouragement. His support to do science has extended well beyond this work and, for that, I am sincerely grateful. Cheers! My thanks also go to Harald Oberhofer, for his answers always full of keen insight and his valuable guidance. Without his and other collaborators' previous developments, this thesis would not have been possible. I am more than thankful to Prof. Karsten Reuter for welcoming me in the group. I deeply appreciate his honest council and trust in my work. I also want to thank Prof. Thomas Huckle. I have very much enjoyed my meetings with him and counting with his valuable feedback. My acknowledgement also goes to Letif Mones, for an elucidating and motivating conversation during his visit to Munich.

I am very grateful and proud to be supported by Mexico, my native country. Not only this thesis, but my entire Master's degree was only made possible by CONACyT, I2T2 and the state of Nuevo León. I would specially like to show my appreciation to Ilsa Torres, who embodies the spirit of these organizations, by promoting the growth of young mexican talent. Another organization from my country, SEP, also has my gratitude for supporting me during the first year of my Master's. I also want to thank all the professors back in Mexico, whose recommendations helped me obtain these scholarships and my place at TUM.

My understanding of molecular simulations would not be the same if it wasn't for Prof. Martin Zacharias. I am very thankful for the chance of working directly with him and for his genuine interest in my growth as a young researcher. In the same way, I thank Prof. Faidon-Stelios Koutsourelakis, whose lessons have motivated my advances in Bayesian strategies.

I would like to also thank everyone at the Chair for Theoretical Chemistry. It is very enriching to work among such bright and kind people. I would like to specially mention Carlos, Kim and Bridgita, friends with whom I always enjoyed talking in between the steps of this work. In the same way, I would like to thank all the CSE community, including colleagues, coordinators and professors. Helisa and Makis, my closest friends in this city, have all my affection for the times shared during the past two years. Great conversations, long walks across the campus and well deserved dinners always kept my spirit raised.

I also want to express my gratitude to my mother, Francis Ortíz. She has always inspired me to follow my dreams, even when this means being so far way from my family. Last, but not at all least, I would like to thank my wife, Sujania Talavera, who has been hearteningly by my side at the rehearsals and premieres of every big event in this adventure. She is my greatest inspiration.

¡Gracias!

Abstract

The simulation of complex molecular systems requires vastly different time and length scales to be overcome. Multiscale modeling enables cost-effective investigations based on hierarchical multi-level descriptions, e.g., quantum mechanics/molecular mechanics (QM/MM), while advanced sampling eases the exploration of reaction paths in reasonable computing times. Care has to be taken, however, that the use of such schemes does not result in unphysical results.

In this thesis, we study the behaviour of an interesting prototype reaction, proton transfer in water, under different simulation set-ups. The nature of an excess proton in water remains a topic of debate in the literature, with both the Zundel and Eigen ions being proposed as candidate structures. In this work, we initially proceed assuming a Zundel-to-Zundel mechanism, but are ultimately unable to confirm the stability of a Zundel ion within our simulation set-up. Methodologically, the diffusive liquid medium requires the multiscale scheme to be adaptive, i.e., it needs to allow for water molecules entering and leaving the QM zone. The majority of the presented work is thus concerned with the implementation of the so-called adaptive buffered-force QM/MM (AdBF-QM/MM) method, which has previously been shown to reproduce the structure of liquid water and been applied to calculate free energy profiles of proton transfer reactions in water. We use Python's Atomic Simulation Environment to interface the FHI-aims package for density functional theory (DFT) with the LAMMPS code for molecular dynamics (MD). Taking advantage of the method's flexibility, several simulations are run to test how the precise set-up influences the reaction. To monitor proton transfer, an energy based reaction coordinate is implemented, which takes solvent effects into account. This requires the parametrization of force fields, which is achieved using Bayesian strategies. Finally, umbrella sampling is used to calculate free energy profiles of the studied reaction.

In the AdBF-QM/MM framework, we note the critical role of the QM region, as its size must be sufficient for quantum behaviour to emerge. Furthermore, the buffer size must balance the cost of the extended QM/MM calculations while preventing close interactions between the QM charge density and MM point charges. If the buffer is too small, the proton is attracted to the edge of the QM zone. We also observe the sensitivity of the reaction coordinate being severely exaggerated when the system drifts away from the intended range. Finally, our results suggest the respective stability of the Zundel and Eigen ions to be very sensitive to the precise QM/MM setup.

Contents

Outline of the thesis

Part I: Introduction and Theoretical Background

Chapter 1: Introduction

This chapter presents the motivation for the thesis, its objectives and a general overview of its structure.

Chapter 2: Theoretical Background

The fundamentals of density functional theory (DFT) and molecular dynamics (MD) are explained in this chapter. A basic distinction of quantum mechanics/molecular mechanics (QM/MM) embedding strategies is also outlined.

Part II: System and Methodology

Chapter 3: System

This chapter describes the selected model for the proton in water and the transfer reaction mechanism. The key actors of the process are identified and its importance for the methodology is highlighted.

Chapter 4: Methodology

The workhorses of the thesis are presented in this chapter. Three sections respectively elaborate on: the adaptive buffered-force QM/MM method (AdBF-QM/MM); the energy gap reaction coordinate; and the umbrella sampling (US) and free energy profile reconstruction techniques.

Part III: Simulation Protocols, Results and Conclusion

Chapter 5: Simulation Protocols

This chapter describes the corresponding settings, parameters and procedures employed during the equilibration and production simulations.

Chapter 6: Results

The outcomes of the simulations are presented in this chapter.

Chapter 7: Conclusions

This final chapter is concerned with the analysis and discussion of the obtained results, as well as with the description of an outlook for further research.

Appendix:

This appendix provides details about the Bayesian strategies applied in the definition of the force fields required by the energy gap reaction coordinate.

Part I.

Introduction and Theoretical Background

1. Introduction

1.1. Motivation

In Chemistry, molecular simulations are nowadays often employed to elucidate the properties of interesting reactions. The outcomes of such computational studies promote faster progress in science and technology, as they complement abstractly theoretical and experimental approaches. Nonetheless, this research is highly non-trivial and many obstacles must be overcome before obtaining useful insight. Two emblematic challenges can be identified when trying to capture the authentic behaviour of complex phenomena. The first one is that — typically — essential processes can only be captured through a full quantum mechanical (QM) treatment of the system. However, this is prohibitively expensive for current computational resources. Therefore, we look at hierarchical simulation schemes. In such, we perform a QM calculation only for the crucial part of the system, embedded inside a more affordable, classical molecular mechanics (MM) simulation [3, 4]. These coupling strategies involve a considerable amount of complexity, and even more so when dealing with diffusive environments, in which the adaptivity of QM/MM descriptions becomes critical [5, 7, 22, 29]. The second challenge we face is that, even when closely capturing the physics of the system, the time necessary for the actual reaction to occur might be extremely long. As a response to that, advanced — or rare event — sampling techniques have been developed [28]. In these methods, a reaction coordinate is defined as a key quantity that describes the evolution of the system from a reactant to a product state. Then, by biasing this reaction coordinate, specific stages of the reaction process can be observed and, subsequently, techniques can be applied to recover the unbiased reaction properties [16]. Evidently, the exact definition of an adequate reaction coordinate is, by itself, a challenging task [21].

The exact way in which we overcome the two challenges described above inherently impacts the outcome of our simulations. The extension of the QM treatment in hybrid QM/MM methods has been shown to affect structural properties [5], as well as potential of mean force (PMF) and free energy profiles [22, 29]. The height of energy barriers and even the location of stable states are significantly conditioned by the proceedings of the multiscale scheme. Similarly, the formulation of reaction coordinates also influences the range, shape and sensitivity of the observed free energy profiles [22]. Pertinent questions then are: How should multiscale methodologies be tuned in order to realistically reproduce chemical reactions? And how should the reaction coordinate be designed to sensibly distinguish between relevant states?

In this thesis, we engage both questions for the particular case of the proton transfer reaction in water. This reaction is not only a valid prototype for more complex systems, but also an interesting process on its own. Protonation and deprotonation require in prin-

ciple a QM treatment, but covering numerous water molecules in such detail is expensive. Furthermore, the aqueous environment allows particles to enter or leave the QM region, making adaptivity indispensable. To address these challenges, we employ the adaptive buffered-force QM/MM (AdBF-QM/MM) method [5, 22, 29], a scheme that has proven highly successful in capturing both the correct structure of bulk water and the free energy profiles of proton transfer reactions. Once the multiscale challenge is attended, we require a reaction coordinate, along which umbrella sampling biasing and free energy profile reconstruction can be applied [16, 27]. Since the effect of the surrounding solvent water during the proton transfer is considerable, geometrically motivated formulations of the reaction coordinate are not straightforward. On the other hand, energy based approaches enable a more comprehensive description of the system. In particular, we select the energy gap reaction coordinate [21], which has been previously employed for proton transfer reactions in water. The formulation of the energy gap is based on the difference between the reactant and product states potential energy functions, or force fields, which we parametrize using Bayesian strategies. Establishing such initial and final states for the proton transfer in water is not straightforward, as both the dominant structure of the proton in water [25], and the transfer reaction mechanism[13, 17] are still not fully understood. There is an ongoing discussion as to which are the stable and intermediate states of the reaction, with the Zundel ion and Eigen ion as candidates. In our work, the Zundel-Eigen-Zundel process is assumed, and considered in the implementations of the multiscale and sampling methodologies.

This thesis is structured as follows. Chapter 2 contains the essential theoretical background underpinning the independent QM and MM methods, as well as basic concepts for a non-adaptive coupling strategy; Chapter 3 is concerned with the chosen description of the proton in water and the corresponding transfer reaction; Chapter 4 focuses on the workhorses of this project, with sections dedicated to the selected adaptive QM/MM method (4.1), the energy gap reaction coordinate (4.2) and the advanced sampling technique (4.3); Chapter 5 describes the specific simulation protocols taken in each step of our investigation; Chapter 6 exhibits the results of our simulations; and Chapter 7 portrays our final conclusions, discussion and further proposed research. At the end of the thesis, Appendix A contains further details about the Bayesian calibration and validation of MM force fields for the energy gap reaction coordinate.

1.2. Objectives

The goals of this thesis project are:

- Implement the AdBF-QM/MM method [5, 22, 29] in Python's Atomic Simulation Environment (ASE) [2, 1], using the software packages FHI-aims [6, 10], for QM, and LAMMPS [24, 8], for MM.
- Formulate an energy gap reaction coordinate [21] for the proton transfer reaction mechanism in water, assuming the Zundel ion to be a stable state [25].
- Parametrize the initial and final state force fields required by the energy gap using Bayesian strategies [20, 12, 26].

- Generate a framework and protocols to simulate the proton transfer using the AdBF-QM/MM method as engine and the energy gap for monitoring and biasing.
- Perform umbrella sampling along the reaction path of the proton transfer and reconstruct free energy profiles using Gaussian process regression (GPR) [16, 27].
- Extract conclusions about the effect of the chosen multiscale scheme, and reaction coordinate, on the observed reaction mechanism. Motivate further research based on this insight.

2. Theoretical Background

2.1. Density functional theory

This section is based on textbooks by Cramer [9] and Jensen [14], as well as on the general outline of closely related work [3].

On a fundamental level, all matter can be represented by electrons and atomic nuclei. The behaviour of such particles is explained by quantum mechanics. In particular, for systems with a description that does not change in time, we employ the time-independent Schrödinger equation:

$$\hat{H} \left|\Phi_i\right\rangle = E_i \left|\Phi_i\right\rangle \tag{2.1}$$

In this eigenvalue equation, $\hat{H}$ is the Hamiltonian operator of the system and the wave function $|\Phi_i\rangle$ is a stationary state with energy E_i. We are interested in the eigenpairs of $\hat{H}$, as any state of the system can be expressed as a superposition of the stationary states. However, such problem does not have an analytical solution for systems with more than one electron. Furthermore, even for a small number of electrons, a numerical approach is extremely computationally expensive. Hence, we look at alternative methods.

Density functional theory (DFT) is a method that works based on electronic density rather than on wave functions. In order to properly illustrate it, some simplifications must be made parting from the time-dependent Schrödinger equation. First, we write the many-body Hamiltonian operator in a general full form:

$$\hat{H} = \hat{T}_{\text{el}} + \hat{V}_{\text{ext}} + \hat{V}_{\text{int}} + \hat{T}_{\text{nuc}} + \hat{V}_{\text{nuc-nuc}} \tag{2.2}$$

Where, from left to right, interactions correspond to the kinetic energy operator for electrons $\hat{T}_{\text{el}}$, the Coulomb potential between electrons and nuclei $\hat{V}_{\text{ext}}$, the Coulomb potential between electrons themselves $\hat{V}_{\text{int}}$, the kinetic energy operator for nuclei $\hat{T}_{\text{nuc}}$, and the Coulomb potential between nuclei $\hat{V}_{\text{nuc-nuc}}$. The use of the subscripts ext and int will become clear in the next paragraph.

The Hamiltonian can be simplified when considering the difference in time and mass scales between electrons and nuclei, being the latter vastly slower and more massive. The Born-Oppenheimer approximation allows us to neglect the interactions between nuclei $\hat{V}_{\text{nuc-nuc}}$ and their kinetic energy $\hat{T}_{\text{nuc}}$, as if they remained immobile. Then, while electron-electron interactions are accounted for explicitly in the internal potential $\hat{V}_{\text{int}}$, the electron-nucleus interactions are only considered as an external potential $\hat{V}_{\text{ext}}$. Furthermore, in the same term we can also include other external contributions, such as electrostatic interactions due

to embedding. This is a vital aspect for QM/MM calculations. We rewrite the electronic Hamiltonian as:

$$\hat{H} = \hat{T}_{\text{el}} + \hat{V}_{\text{ext}} + \hat{V}_{\text{int}} \tag{2.3}$$

More explicitly:

$$\hat{H} = -\frac{\hbar^2}{2m}\sum_i \nabla_i + \sum_i V_{\text{ext}}(\mathbf{r}_i) + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.4}$$

This provides a Hamiltonian that is easier to handle. However, when inserted back in the Schrödinger equation, the complete form still involves many-body wave functions.

$$\Psi_i(\{\mathbf{r}_i\}) = \Psi_i(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) \tag{2.5}$$

These wave functions are $3N$-dimensional, where N is the number of electrons in the system. The storage requirements for calculations quickly become unfeasible, as the number of grid points is elevated to the $3N$-th power. Moreover, the fermionic nature of the electrons, ruled by Pauli's exclusion principle prevents this representation to be simplified in an extensive fashion. The Thomas-Fermi method provides an option that does not treat many body wave functions, but electronic density. These two quantities can be easily related with the following equation:

$$n(\mathbf{r}) = N\int dr_2^3 ... dr_N^3 |\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N)|^2 \tag{2.6}$$

This electronic density, together with the electronic Hamiltonian, yields a total energy shown by Hohenberg and Kohn:

$$E[n] = F_{\text{LL}}[n] + \int d^3 r V_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) \tag{2.7}$$

Where the new term $F_{\text{LL}}[n]$ corresponds to the minimum electronic kinetic energy and electron-electron interaction potential, given the current density. This term is universally valid for any electron system.

$$F_{\text{LL}}[n] = \min_{\Psi\to n(\mathbf{r})} [\langle\Psi|\,\hat{T}\,|\Psi\rangle + \langle\Psi|\,\hat{V}_{\text{int}}\,|\Psi\rangle] \tag{2.8}$$

Having this expression, only a second minimization across electronic densities is required to find the minimum energy of the system. That is, the ground state. For most chemical applications, thermal fluctuations of a few hundred Kelvin can be neglected, such that the ground state is a valid description for the system. Then, we are concerned with how to solve the reformulated problem, in order to obtain the total energy and electronic density. We do so by an iterative method known as self-consistent field (SCF) illustrated in Figure 2.1.

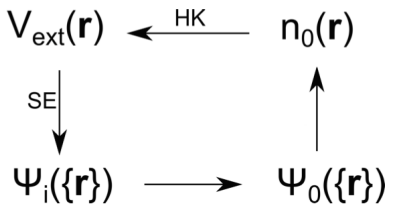

Figure 2.1.: Self-Consistent Cycle (SCF) corresponding to the first Hohenberg-Kohn theorem. Edited from [3].

We start by inserting our external potential $V_{\text{ext}}(\mathbf{r})$ in the electronic Hamiltonian, then we solve the Schrödinger equation to obtain every state of the system $\Psi_i(\{\mathbf{r}\})$, including the ground state $\Psi_0(\{\mathbf{r}\})$. This is simply related to the electronic density $n_o(\mathbf{r})$, as shown in Equation 2.6. Subsequently, the first Hohenberg-Kohn (HK) theorem, which states that the ground state electronic density $n_0(\mathbf{r})$ uniquely determines the external potential $V_{\text{ext}}(\mathbf{r})$, up to a constant. In this way, the SCF is closed.

An additional, game-changing, assumption can be done. The Hartree-Fock (HF) approximation neglects the interactions between electrons, and instead considers them as moving in an averaged effective potential V_{eff}, thus, removing the correlation between particles. When plugging this into the many-body Hamiltonian, the internal potential $V_{\text{int}}(\mathbf{r})$ is split in two terms: the first one, known as Hartree term, contains all electron-electron interactions, even spurious self-interactions; the second, known as exchange term, embodies the corrections for the self-interaction and Pauli's exclusion principle. This simplification, as applied to DFT by Kohn and Sham, provides a remarkably cheap method for many-electron systems.

The *ansatz* of Kohn-Sham (KS) DFT resorts to the HF approximation and maps the HK many-electron problem onto an auxiliary non-interacting electron system. Such connection is enabled under the key assumption that the ground state densities of both systems are equal. Before building the new SCF cycle, let us define the necessary expressions under the new KS framework.

First, the non-interacting electron density can now be simply written as:

$$n(\mathbf{r}) = \sum_{i}^{N} |\Psi_i(\mathbf{r})|^2 \tag{2.9}$$

Then, analogously to Equations 2.7 and 2.8, we express the KS total energy as:

$$E^{\text{KS}}[n] = T_{\text{s}}[n] + E_{\text{Hartree}}[n] + E_{\text{xc}}[n] + \int d^3r V_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) \tag{2.10}$$

Where the kinetic energy for non-correlated particles is defined as:

$$T_{\text{s}} = -\frac{1}{2} \sum_{i}^{N} \int d^3r |\nabla \Psi_i(\mathbf{r})|^2 \tag{2.11}$$

While the HF approximation provides us with the Hartree $E_{\text{Hartree}}[n]$ and exchange-correlation $E_{\text{xc}}[n]$ terms. The Hartree term is defined as:

$$E_{\text{Hartree}}[n] = \frac{1}{2}\int d^3r d^3r' \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \tag{2.12}$$

The exchange correlation term, on the other hand, can be related to the HK Equations 2.7 and 2.8 as:

$$E_{\text{xc}}[n] = F_{\text{LL}}[n] - (T_{\text{s}}[n] + E_{\text{Hartree}}[n]) \tag{2.13}$$

However, there exists no analytic expression for the exchange-correlation functional. More than acceptable approximations have been developed. We will explain their basic formulation toward the end of this section.

The KS cycle consist of two HK cycles, one for the original system and one for the auxiliary system, which are connected by the KS *ansatz* at their ground state densities. Then, the need to solve the interacting many-electron system is avoided.

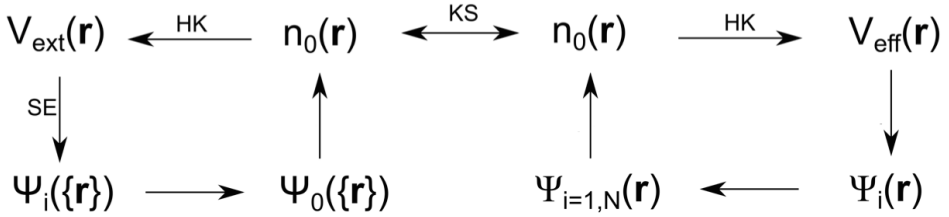

Figure 2.2.: Self-Consistent Cycle (SCF) corresponding to the first Kohn-Sham (KS) *ansatz*. Edited from [3].

In every iteration, the original system's ground state density is inserted in the auxiliary system cycle, where it uniquely determines the effective potential. In turn, the effective potential allows to solve the Schrödinger equation and obtain each of the single-particle stationary states, namely, the orbitals. Then, the ground state density can be recalculated as in Equation 2.9. Once the auxiliary system SCF is done, we can continue with the original system's cycle. In practice, the total energy and the electronic density are monitored and used as convergence criteria.

The scaling and accuracy of this method are remarkable, as its relatively straightforward implementation and parallelization. In this thesis, we shall exploit it by the means of the FHI-aims software package [6]. The features of this MPI parallelized FORTRAN based code include: convergence criteria, van der Waals corrections, etc. Moreover, for our QM/MM approach, FHI-AIMS enables electrostatic embedding by including additional interactions inside the V_{ext} term [10, 4]. This concept will be further clarified in Section 2.3. Additionally, the code offers a selection of approximations for the exchange-correlation functional $E_{xc}[n]$ [30]. In the following paragraphs, the formulation of such estimations will be explained.

An initial crude approximation of the exchange-correlation functional can based on the homogeneous electron gas. Assuming the system is piecewise homogeneous, we integrate and express the so called local density approximation (LDA):

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[n] = \int d^3r n(\mathbf{r})[\epsilon_{\mathrm{x}}^{\mathrm{homo}}([n(\mathbf{r})], \mathbf{r}) + \epsilon_{\mathrm{c}}^{\mathrm{homo}}([n(\mathbf{r})], \mathbf{r})] \tag{2.14}$$

Where, for the homogeneous electron gas, the exchange energy density is known analytically and the correlation energy density can be calculated numerically.

Inhomogeneous systems can be modeled by extending the LDA and considering the density gradient. The family of generalized-gradient approximation (GGA) functionals is founded on this idea:

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[n] = \int d^3r n(\mathbf{r})\epsilon_{\mathrm{x}}^{\mathrm{homo}} F_{\mathrm{xc}}(n, \nabla n, ...) \tag{2.15}$$

The exchange enhancement factor F_{xc} can be formulated in several ways. Among them, PBE, BLYP and rPBE are popular examples. The specific forms include information about the density gradient, up to a certain order that balances accuracy with computational cost, as well as physically motivated constraints. In general, GGAs perform better than LDAs when dealing with atoms and molecules. In this thesis, the PBE functional will be used in the DFT calculations. However, further details about its formulation will not be discussed.

Both LDAs and GGAs contain the spurious self-interaction in the Hartree term. The HF method provides an exact correction for this effect. It can be combined with DFT to generate hybrid functionals:

$$E_{\mathrm{xc}}^{\mathrm{hybrid}}[n] = E_{\mathrm{xc}}^{\mathrm{DFT}} + a(E_{\mathrm{x}}^{\mathrm{exact}} + E_{\mathrm{x}}^{\mathrm{DFT}}) \tag{2.16}$$

As a is a free parameter, and the $E_{\mathrm{xc}}^{\mathrm{DFT}}$ functional can be chosen as any LDA or GGA functional, the hybrid functionals present complete flexibility. In our work, we have used the widely known PBE0, B3LYP and HSE06. We do not dive into the respective formulations. All in all, hybrid functionals present a significant improvement, specially when studying binding energies in molecules. However, they cannot be used in metals.

The construction of new functionals, as well as the general development of DFT, are active research field nowadays. The ground-breaking KS method has settled as the standard for solving many-electron systems and it is a central actor in modern research. The richness of its possible implementations is far too wide to include in this section. The reader is referred to further literature [3, 9, 14], if looking for a deeper understanding of the exact procedures and proofs that sustain this method. The background finished here is sufficient to support the QM/MM framework.

2.2. Molecular dynamics

In many molecular processes, quantum effects can be neglected. Molecular mechanics (MM) models atomic systems from a classical perspective. Consequently, molecular dynamics (MD) provides a method to simulate such classically interacting systems as they evolve in time. The N-body problem is solved repeatedly in order to generate the trajectories of the particles as described by Newton's equation of motion. The key ingredients to perform an MD simulation include: an MM force field to calculate the potential energy of the system and its derivatives, a method to integrate the equations of motion, and a set of constraints and protocols. In the next paragraphs, while mainly focusing on the first, the fundamentals of these three aspects will be explained. Textbooks by Cramer [9], Jensen [14] and Frenkel and Smit [11] constitute the basis for this section.

An MM force field is a function that serves to calculate the potential energy of a system. The form of such a function contains terms that correspond to relevant interatomic interactions, each contributing to the total energy. These contributions come from both bonded and non-bonded interactions. Bonded, also known as covalent interactions, include bond lengths, angles and torsions, while non-bonded interactions correspond to Coulombic and van der Waals effects. A typical force field can be presented as follows:

$$E(r) = \sum_{m}^{N_m} E_{\text{Morse},m} + \sum_{b}^{N_b} E_{\text{bond},b} + \sum_{a}^{N_a} E_{\text{angle},a} + \sum_{t}^{N_t} E_{\text{torsion},t} + \sum_{j\neq k}^{N_{nb}} (E_{\text{Cou},jk} + E_{\text{vdW},jk}) \tag{2.17}$$

Where the Morse potential term corresponds to atomic bonds that can be formed or broken; the harmonic bonds term, to covalently connected atom pairs; the harmonic angles term, to the angles formed by groups of three atoms; and the harmonic torsions term, to rotations defined by groups of four longitudinally bonded atoms, in which two different planes are defined by the first and last three atoms, and the angle between the planes is measured. At the end of the equation, the unbonded terms correspond to the Coulombic (electrostatic) and van der Waals (Lennard-Jones) pair-wise interactions.

In order to keep consistency with our simulations, the specific definition for each potential energy contribution is taken from the LAMMPS molecular dynamics package [24, 8]:

$$E_{\text{Morse},m} = D_m[1 - e^{-\alpha_m(r_m - r_{0,m})}]^2 \tag{2.18}$$

$$E_{\text{bond},b} = K_b(r_b - r_{0,b})^2 \tag{2.19}$$

$$E_{\text{angle},a} = K_a(\theta_a - \theta_{0,a})^2 \tag{2.20}$$

$$E_{\text{torsion},t} = K_t(1 + \cos(n_t\phi_t))^2 \tag{2.21}$$

$$E_{\text{Cou},jk} = \frac{C}{\epsilon}\frac{q_j q_k}{r_{jk}} \tag{2.22}$$

$$E_{\text{vdW},jk} = 4\epsilon_{jk}\left[\left(\frac{\sigma_{jk}}{r_{jk}}\right)^{12} - \left(\frac{\sigma_{jk}}{r_{jk}}\right)^{6}\right] \tag{2.23}$$

In these equations, we can see that each term presents its own set of parameters: the Morse potential has an equilibrium bond distance, identified by sub-index 0, as well as two potential well related constants, D_m for the depth and α_m for the width; all of the harmonic potentials, bonds, angles and torsions, present a harmonic constant $K_{b,a,t}$, that sets the strength of the interaction; the harmonic bonds and angles have an equilibrium value, identified with sub-index 0; the harmonic torsion has an integer parameter n_t, that multiplies the torsion angle; both of the non-bonded interactions depend on the interatomic distance r_{jk}; the Coulombic contribution presents the typical form of a electrostatic pair-wise potential, with the Coulomb's constant C and the dielectric constant ϵ; and finally, the van der Waals contribution has the form of the widely used Lennard-Jones pair-wise potential, with σ_{jk} and ϵ_{jk} corresponding to the width and depth of the potential well.

Many considerations can be taken when calculating the potential energy of a system. Cut-off distances can be set in order to neglect non-bonded interactions beyond a certain points, switching can be done to make such cut-offs less abrupt, special summations can be done for periodic settings, multipole approximations and coarse graining can represent several particles as one, indirect neighbours of the same molecule can present weighted non-bonded interactions, and so on. If interested in further details, the reader is encouraged to consult the references [9, 11, 14, 24]. In this thesis, the focus will be on aspects that can be exploited either in QM/MM calculations, or in advanced sampling techniques. For example, when doing QM/MM embedding, the interactions with certain atoms can be removed from the force field and instead taken from an external reference, i.e., the DFT charge density. This complements the $\hat{V}_{\text{ext}}$ term of DFT and allows for a two-way communication between QM calculations and MM calculations. In Section 4.1, we describe how to implement this exchange as a combination of forces during the integration of the equations of motion. Furthermore, when doing rare-event sampling, MD simulations can also be biased by external potentials. Section 4.3 presents an example of this.

Once the potential energy of the system at a certain configuration is known, the forces on each atom can be calculated simply as the negative of the gradient. These forces can then be used to obtain the accelerations, velocities and new positions of the particles. The numerical integration required for this process is not trivial, as it must ensure energy conservation. Typical choices like Euler or Runge-Kutta methods perform poorly in this respect. On the other hand, symplectic integrators, like the Störmer-Verlet method, give much better results.

In general, the set-up for the simulation should be defined. This includes decisions like having fixed or periodic boundary conditions (PBC) or using explicit or implicit solvent. Furthermore, restraints or constraints can be applied to respectively limit or fix a particular degree of freedom of the system. The size of the system itself and the length of the simulation must be chosen such that the statistics of the MD run return valid insight. After taking all of these considerations, the simulations can be run while conserving certain quantities in statistical ensembles. We distinguish: the microcanonical ensemble (NVE), conserving number of particles N, volume V, and energy E; the canonical ensemble (NVT), conserving number of particles N, volume V and temperature T; and the isothermal-isobaric ensemble (NPT), conserving number of particles N, pressure P and temperature T. To this end, different thermostats and barostats are implemented.

2.3. QM/MM embedding

Even when using KS DFT, simulating large atomic systems under a QM approach is computationally expensive. Still, there are effects that can only be captured by a QM method, such as covalent bond forming/breaking or changes in charge states. For this reason, we turn to coupling techniques in which a relevant part of the system is treated quantum mechanically, while the rest is done with MM [3, 4, 5, 22, 29, 7]. In such schemes, the respective potentials for each domain are calculated according to its corresponding method. The interaction between both regions involves a higher level of complexity. Two basic strategies can be distinguished:

- Mechanical embedding: the QM region is resolved without considering the MM region around it. The interactions are then included only at the MM level.
- Electrostatic embedding: the QM region is resolved considering the atoms in the MM region as charges exerting an external potential. Therefore, the electronic density is actively affected by the environment. Correspondingly, the forces due to the atoms in the QM region are included in the MM region. In case molecules are crossing from the QM to the MM regions, the bonded interactions are treated on the MM level.

How to bring adaptivity into electrostatic embedding will be one of the main topics in Chapter 4.

Part II.

System and Methodology

3. System

3.1. The excess proton in water

In general, protonated species cannot be well represented simply by a proton addition. Water is not an exception, as hydrogen bonding must be considered to elucidate the structure and dynamics of the system [25]. Up to current research, there is no consensus on the stable state of an excess proton in water. Two opposite perspectives pose the Eigen ion, or hydronium, H_3O^+ [13] and the Zundel ion $H_5O_2^+$ [25] as main candidates. Both states occur under specific conditions of concentration and acidity. The respective configurations are depicted in Figure 3.1.

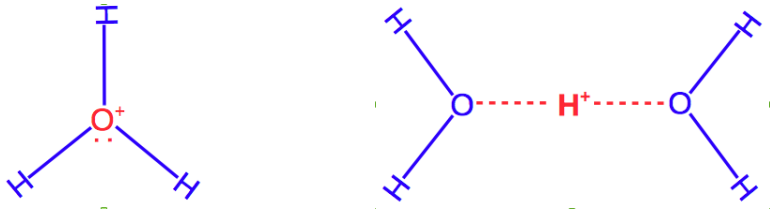

Figure 3.1.: The Eigen ion (left) and the Zundel ion (right) states of an excess proton in water

The Eigen ion provides a simple representation with a single protonated water molecule. However, from the perspective of an energy based reaction coordinate, it behaves counter-intuitively when used to identify stable configurations during the proton transfer. This is explained in Section 3.2. On the other hand, the Zundel ion, which consists of two water molecules with their oxygen atoms H-bonded to the proton, offers a better alternative. In this thesis, the Zundel ion is assumed to be the stable state of the excess proton in water. All further proceedings are based on this consideration. Accordingly, we elaborate on the structure of the Zundel. Typically, the ion is surrounded by four water molecules composing the first hydration shell, as shown in Figure 3.2. Then, we have the tetrasolvated structure $H_5O_2^+ + 4\ H_2O$.

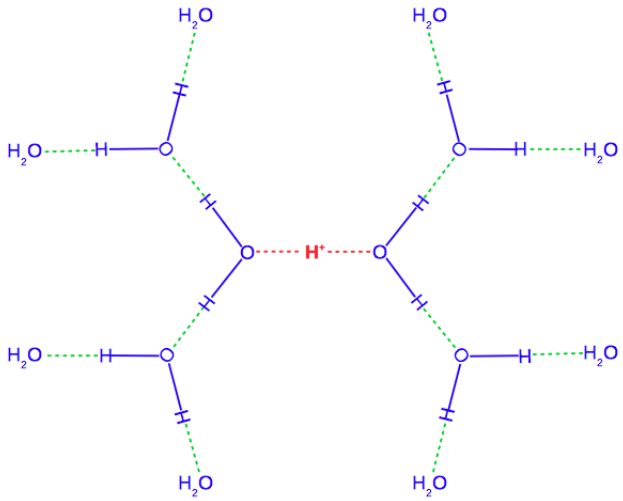

Figure 3.2.: Zundel ion in water. According to our calculations, the approximate interatomic distances are as follows: water covalent bonds (solid blue) of 1.0 Å, H-bonds (dashed green) of 1.7 Å, and Zundel ion bonds to the proton of 1.2 Å. Based on [25]. Not to scale.

This representation is still not perfect and forms only under selected conditions of water concentration and strong acids. A $H_7O_3^+ + 5\ H_2O$ representation enables more generality. However, it also increases the number of atoms involved in the protonation and de-protonation. Accounting for the quantum behaviour of the species endorses the Zundel ion as a more attractive alternative. Further advantages of it, concerning the transfer reaction, will be discussed in the next lines.

3.2. The proton transfer reaction in water

In the literature, the proton transfer has been represented as the proton traveling the distance between the two oxygen atoms of an Eigen ion and a water molecule. The initial and final states are defined as the proton being closer to either of the oxygen atoms and forming hydronium ions. Nonetheless, this scheme poses a fundamental problem. Considering non-bonded interactions, the intermediate state corresponding to the proton positioned exactly at the same distance from both oxygen atoms is stable and minimizes the energy. This problem is avoided when contemplating a Zundel ion, which corresponds to said stable structure. Then, we rather represent the transfer as the hop of the proton from an initial configuration as shown in Figure 3.2, to a final configuration as depicted in Figure 3.3. Note that the transition also involves the rearrangement of the surrounding solvent water.

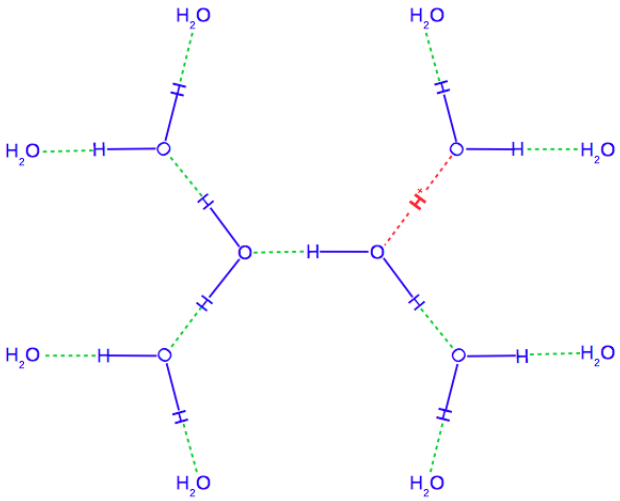

Figure 3.3.: Zundel ion in water after a proton transfer to the top-right neightbor water molecule starting from Figure 3.2. According to our calculations, the approximate interatomic distances are as follows: water covalent bonds (solid blue) of 1.0 Å, H-bonds (dashed green) of 1.7 Å, and Zundel ion bonds to the proton of 1.2 Å. Based on [25]. Not to scale.

We also wish to highlight the key actors in this transfer reaction: the two water molecules that originally form the Zundel ion, the proton and the *receptor* water molecule that becomes part of the Zundel ion after the hop. These actors are the essential part of the system for which QM behaviour should be captured. Of course, symmetry allows any of the first four neighbouring water molecules of the original Zundel ion to be a potential receptor. However, the hop can be guided. We merely require means to monitor and drive the proton transfer. This will be discussed in Sections 4.2 and 4.3, dedicated to the reaction coordinate and sampling. Another important aspect is to establish which parts of the system have a relevant quantum behaviour. We expect at least the three key actors to require a QM treatment. The first and second hydrations shells around these three components might also be of importance. We illustrate them in Figure 3.4.

It is worth noting that the Zundel ion's oxygen that is closest to the receptor water has a symmetric point of view of the transfer reaction. Standing on this oxygen, the proton hop from the initial to the final configuration is exactly symmetric to the transfer in the opposite direction. We also wish to summarize the general structure of our system, composed by the Zundel ion and the receptor water in the core, surrounded by five water molecules in the first hydration shell, and ten in the second. Section 4.1, dedicated to the adaptive QM/MM method, will further explain how to consider the quantum nature of the proton transfer in our system.

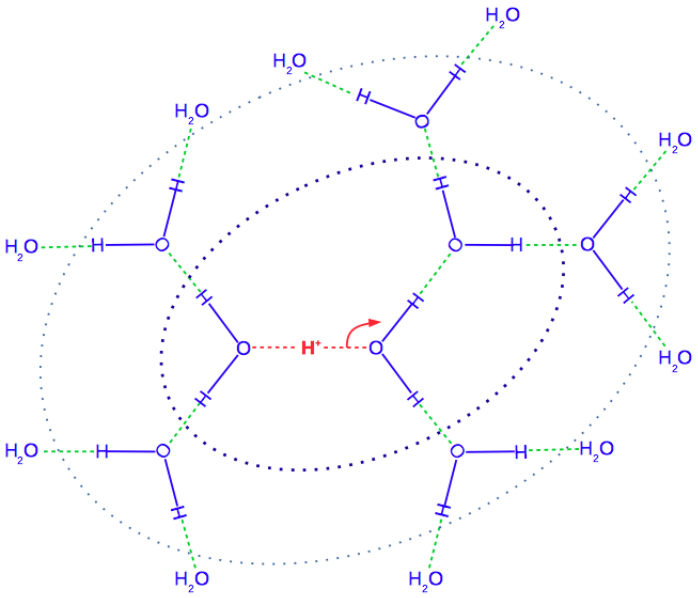

Figure 3.4.: First (between bold dotted and dotted lines) and second (beyond dotted line) hydration shells around the Zundel ion and the receptor water. According to our calculations, the approximate interatomic distances are as follows: water covalent bonds (solid blue) of 1.0 Å, H-bonds (dashed green) of 1.7 Å, and Zundel ion bonds to the proton of 1.2 Å. Based on [25]. Not to scale.

It is important to emphasize that the Zundel-to-Zundel proton transfer is taken as model for this thesis, with the Eigen ion as intermediate state [25, 17]. However, the exact reaction mechanism is not fully known. Equally valid research documents the Eigen-Zundel-Eigen proton transfer, with the hydronium as stable configuration and the Zundel ion as intermediate state [13]. It can be hypothesized that such differences depend on the particular QM method, and even more of a coupled QM/MM. In particular, considering nuclear quantum effects can favour the stability of the Zundel ion, as it enables the delocalization of the proton [13]. While keeping this Zundel-Eigen dichotomy in mind, we still opt to proceed assuming a Zundel-Eigen-Zundel transfer mechanism.

4. Methodology

4.1. Adaptive buffered-force QM/MM

As explained in Chapter 2, QM/MM embedding enables a cost-effective two-level description of a system, in which a particularly interesting region is treated as QM, while the surrounding environment is resolved with MM. This method is very powerful by itself in solid-state settings [3, 4], where each region is calculated separately, and the major challenge is posed by managing the interactions between QM and MM atoms. However, when dealing with diffusive media, further challenges appear as the system must be adaptively repartitioned to maintain focus on the relevant QM zones [5, 7, 22, 29]. Implementing such adaptivity is not trivial, as whenever a particle switches descriptions, it undergoes a change in the potential energy surface (PES). If a proper adaptive method is not applied, artefacts near the QM/MM boundary result in unphysical density variations.

Two approaches can be taken when building adaptive QM/MM methods: energy mixing [7] or force mixing [5, 22, 29]. As the name suggests, energy mixing couples QM and MM calculations in terms of energy. It allows to define a total energy for the system. However, since the chemical potential is not the same in both regions, molecules are transported to the region with lower potential, leading to the artifacts mentioned above. Efforts can be made to reconcile the energies across descriptions, but any discrepancies still lead to the same artificial forces. No general solution for this artifact of energy mixing has been yet developed. On the other hand, force mixing, while not without its problems, appears to be a more resilient alternative. In this approach, the coupling is achieved by calculating the forces of each sub-system separately, and then combining them. The chemical potential mismatch is avoided, at the cost of losing a definition for total energy and its conservation. Because the forces are not obtained from a total energy gradient, momentum conservation is also lost. Moreover, the use of non-conservative forces generates heat at the QM/MM boundary, which calls for the use of massive adaptive thermostats. This can also be complemented by constant force corrections. With these additions, a robust method is produced. Force mixing adaptive QM/MM approaches are able to correctly sample canonical distributions for molecular systems, therefore reproducing physical structures [5, 22] and free energy profiles [22, 29].

Two more aspects must be determined for the adaptive QM/MM scheme. The first is to decide if the descriptions of particles crossing the QM/MM boundary should be changed continuously or abruptly. Opting for a continuous conversion requires a transition region between QM and MM, in which forces or energies are gradually interpolated. Counterintuitively, at least in force mixing approaches, the use of a smooth transition does not improve the accuracy of the method [22]. Therefore, most implementations use abrupt transitions. The second aspect to determine has a significant effect on the quality of the method's re-

sults. It refers to the specific way in which errors near the QM/MM boundary are corrected. This is handled in different ways for energy and force mixing. In the energetic approach, the energy of the system can be defined as the sum of the QM energy, the MM energy and an interaction energy. This last term is uniquely responsible of correcting all artifacts at the boundary [22]. Adaptive schemes of this nature perform several QM/MM calculations, each with a different interface, and combine them with a weighted sum in order to correct the boundary errors. Since the energy is not an extensive property of the system, interfaces must be tried for all, or most, combinations of atoms in and out of the QM region. Force mixing, on the other hand, is done with an extensive property. This allows for the use of buffers, or extensions to the original delimited regions, as an alternative to multiple interface calculations. In the next paragraphs, we will elaborate on the particulars of the adaptive QM/MM method implemented in this thesis.

The adaptive buffered-force QM/MM (AdBF-QM/MM) method was developed by, among others, Bernstein, Várnai, Mones and Csányi in 2011 [5, 22, 29]. In categories as defined above, it is a force mixing approach with an abrupt transition and buffers. The first step in its implementation is the definition of adequate domains for the calculations. We begin by defining a core or center region, which is composed by the atoms for which it is essential to capture quantum behaviour. In the proton transfer framework described in Chapter 3, these crucial atoms would be the Zundel ion and the receptor water. This region is not dynamically adapted and it is the reference around which all other regions are defined. As it is always considered quantum mechanically, its MM force field parameters are not required. Then, for atoms that are within a distance r^{QM} from any core atom we define the QM region. This region is dynamical as any atom that enters it is considered quantum mechanically. Conceptually, we will treat the core region as included in the QM region. The atoms that are within a distance r^{buf} from any QM atom, and not included already in the QM region, are considered as the buffer region. This region is also dynamical. The MM region is composed of the buffer region and all additional atoms that are further away from the core. It is updated adaptively and treated classically. One additional feature is implemented in the definition of the adaptive regions. In order to avoid particles oscillating around a boundary to quickly change their descriptions, hysteresis is applied. The distance r_{-}^{QM}, in which a particle's QM description is turned on, is closer to the core than the distance r_{+}^{QM}, in which its QM description is turned off. This allows for a recently changed atom to maintain its description. The same principle holds for the buffer. Moreover, in order to avoid covalent bonds crossing the QM/MM boundary, only entire molecules are included or excluded from the dynamical regions. As a last remark, it is a sensible choice to set the radii of the QM and buffer regions matching the hydrations shells of the core region. Figure 4.1 shows the described regions schematically.

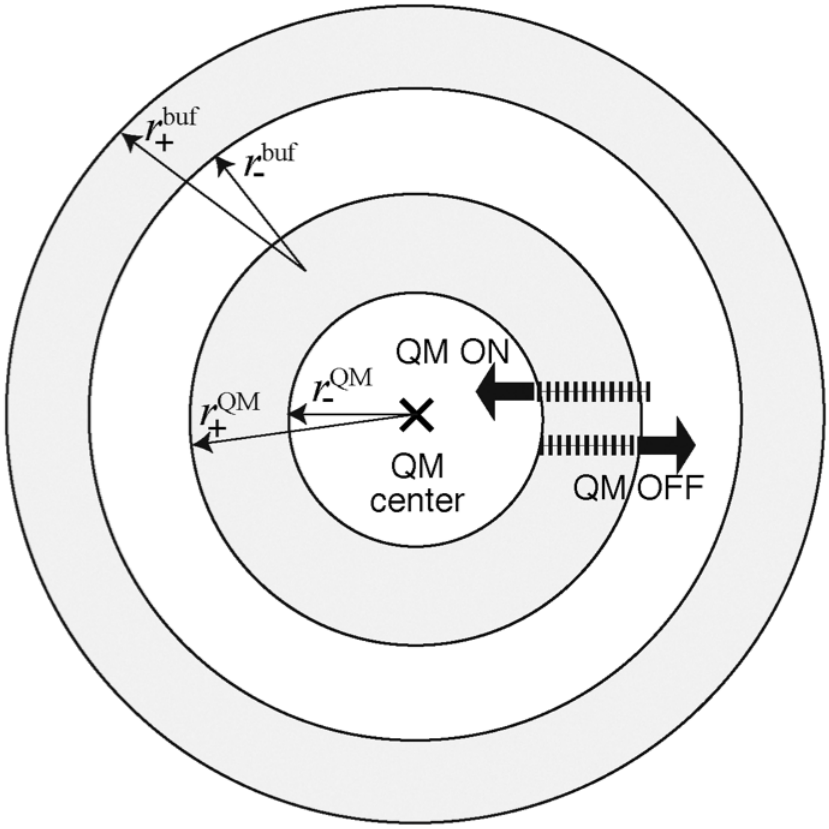

Figure 4.1.: AdBF-QM/MM regions and hysteretic radii. Reproduced from [5].

The method is implemented as follows. We start by classifying the atoms into the four regions described above. Then, we set up two completely separate and independent QM/MM electrostatic embedding calculations to extract forces. The first calculation is done by extending the QM region by the size of the buffer. The entire QM+buffer region is treated quantum mechanically, while the atoms in the rest of the MM region are included as point charges contributing to the external potential. The second calculation is done by reducing the QM region to its minimum, that is, to the core atoms. The core atoms are then treated quantum mechanically, while keeping the rest of the QM and MM atoms as point charges. In both calculations, the interactions between the atoms treated as external charges during the embedding can be resolved by a classical MM computation. However, only the MM forces in the reduced QM/MM calculation will be relevant. This is because of the way in which the forces are mixed: from the extended calculation, we take the forces for the original QM region and, from the reduced calculation, we take the forces for the original MM region. This combination of forces violates momentum conservation, and therefore a constant force correction is applied evenly to all atoms in the QM region. Figure 4.2 illustrates the flowchart for the AdBF-QM/MM.

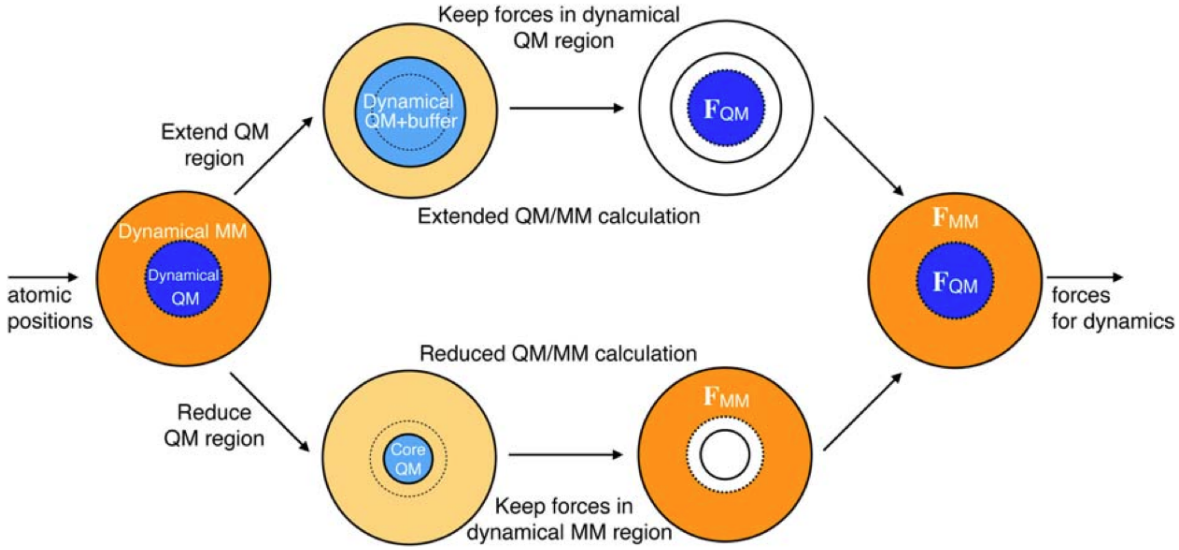

Figure 4.2.: AdBF-QM/MM flowchart. Reproduced from [22].

Once the forces are mixed, the dynamics are propagated using a massive adaptive thermostat. Current AdBF-QM/MM implementations use the adaptive Langevin thermostat, which can be understood as a Langevin thermostat coupled to a Nosé-Hoover thermostat [15]. The dynamical equations are [5, 22, 29]:

$$\dot{q} = \frac{p}{m} \tag{4.1}$$

$$\dot{p} = F(q) - (\gamma + \chi)p + \sqrt{2k_B T \gamma m}\dot{w} \tag{4.2}$$

$$\dot{\chi} = \frac{2K - nk_B T}{Q} \tag{4.3}$$

$$Q = k_B T \tau_{\mathrm{NH}}^2 \tag{4.4}$$

Where q and p are the position and momentum, m is the atomic mass, $F(q)$ is the force, $\gamma = 1/\tau_{\mathrm{L}}$ is the Langevin friction or inverse of the Langevin time constant, χ is the Nosé-Hoover degree of freedom, Q is the Nosé-Hoover fictitious mass, τ_{NH} is the Nosé-Hoover time constant, k_B is Boltzmann's constant, T is the temperature, K the kinetic energy and $\dot{w}$ is the derivative of a Wiener process.

The AdBF-QM/MM method, as described, has been successfully applied to a number of molecular systems. Published applications include: tests for the structure of bulk water, as well as water around a Cl^- solvated ion [5]; free energy calculations for the nucleophylic-substitution of methyl chloride with a chloride anion and the deprotonation of the tyrosine

side chain [29]; and calculations of potential of mean force (PMF) profiles for water autoprotolysis in the presence of zinc and dimethyl-phosphate hydrolysis [22]. The method proves not only able to capture correct structures and transition barriers, but is also highly efficient when compared against more expensive, energy based, QM/MM techniques.

Of course the performance of the AdBF-QM/MM method highly depends on the chosen regions and radii and the subsequent extension of the QM calculations. By tuning certain parameters, specific variations of the method can be identified, some of them equivalent to previously developed techniques. First, it should be understood that, while in the extended calculation the buffer acts as such for the QM region, in the reduced calculation it is the difference between the core and the QM region that acts as a buffer. Then, if one choses the QM region to be the same as the core ($r_+^{\mathrm{QM}} = 0$, $r_-^{\mathrm{QM}} = 0$) and no buffer ($r_+^{\mathrm{buf}} = 0$, $r_-^{\mathrm{buf}} = 0$), the result is a completely unbuffered calculation, i.e. adaptive conventional QM/MM (AdConv-QM/MM) [22]. Alternatively, only one of these considerations can be taken, leading to calculations that are only buffered for QM or MM forces. If the extended QM/MM calculation is avoided, the method becomes significantly cheaper, but also prone to boundary effects. Finally, if there is a reactive force field available for the core atoms, the reduced calculation can be done cheaply, entirely with MM.

In this thesis, three modifications have been applied to the original AdBF-QM/MM method. The first and most simple is that the r^{buffer} distance is defined from the positions of the core atoms rather that from the QM atoms. This makes the classification easier, while a straightforward subtraction of the r^{QM} distance serves to recover the approximate original r^{buffer} definition. The second change is concerned with the momentum conservation correction, done after the force mixing. Instead of applying the correction only in the QM atoms, we do so in every atom of the system. Moreover, the correction is not done with a constant force, but rather with a constant acceleration in order to avoid differences between lighter and heavier atoms. The third modification consists in adding van der Waals forces to the QM/MM embedding. This is done by performing an additional MM calculation with Lennard-Jones interactions exclusively. Non-bonded interactions are then completed, since the embedding already captures electrostatics. Of course, if the implemented QM/MM embedding already includes van der Waals forces, this is not necessary.

Implementations of the AdBF-QM/MM method in the CP2K quantum chemistry package and in the AMBER MD package have been published in recent years [22]. In both cases, the software packages are able to execute the entire method independently . In this thesis, we coupled two different packages. For DFT QM/MM embedding, we use FHI-aims [6, 10], an ab initio molecular simulation package developed by the Fritz-Haber Institute of the Max Planck Society. For MM calculations, we use LAMMPS [24, 8], a classical MD simulator developed by Sandia National Labs and Temple University. FHI-aims is written in Fortran95, while LAMMPS is in C++. Both codes work in parallel with message passing interface (MPI). LAMMPS can also use graphics processing units (GPUs). The interfacing between both programs and all the code development was done with Python's Atomic Simulation Environment (ASE) module [2, 1]. ASE can call FHI-aims and LAMMPS, and retrieve the corresponding results. This is done by implementing calculator objects, which are used to update the forces, kinetic and potential energy of an atoms object that represents our system. Several copies of the system are resolved by the different calculators. Then, the

returned forces are mixed, corrected and used to propagate the dynamics according to the thermostat, which can also be smoothly integrated into the ASE framework. A total of five calculators are required: two FHI-aims calculators, one for the extended and one for the reduced QM/MM embedding; two LAMMPS calculators to add the van der Waals forces in each of the embedding calculations; and one LAMMPS calculator for the interactions of the MM atoms.

One final aspect of our particular implementation should be considered. Unlike C2PK or AMBER, FHI-aims currently does not support PBC when performing QM/MM embedding. For this reason, we run our simulations in a finite box. In order to prevent the system from diffusing, we set a spherical region delimited by a wall. The sphere is centered on a particular point with symmetric significance. The wall exerts a harmonic force on any atom about to leave the spherical region. This restriction is implemented in LAMMPS for the MM atoms. It is important to remove these wall forces from the momentum conservation correction.

The AdBF-QM/MM, as implemented for this thesis, can be summarised by the following pseudo-code:

Algorithm 1 AdBF-QM/MM in Python's ASE with FHI-aims and LAMMPS

Initialize:
atoms object, core atoms indices, FHI-aims calculator object,
LAMMPS calculator object, hysteretic radii floats
for each timestep **do:**
classify atoms
according to distances and hysteretic radii
assign copies of atoms to calculators
FHI-aims extended QM/MM embedding ← QM+buffer atoms
LAMMPS LJ embedding interactions ← QM+buffer atoms
FHI-aims reduced QM/MM embedding ← core atoms
LAMMPS LJ embedding interactions ← core atoms
LAMMPS MM interactions (with wall) ← all non-core QM/MM
calculate forces
mix forces and apply on original atoms
QM atoms: extended QM/MM + LJ embedding interactions
MM atoms: reduced QM/MM + LJ embedding interactions + MM interactions
correct forces
constant acceleration for momentum conservation (no wall effect)
update atomic positions
adaptive Langevin thermostat
end for

4.2. Energy gap reaction coordinate

As mentioned in Chapter 3, in order to study the proton transfer reaction we require means to monitor and bias it. A reaction coordinate is an abstraction that allows us to express the

progress along a reaction pathway - from initial to final state - in a single coordinate [11]. In general, reaction coordinates are geometrically motivated and involve tracking key atomic positions, bond lengths, angles, torsions, etc. As using a single degree of freedom is often not enough for a good description, it is possible to perform a linear combination of reaction coordinates, or even use several of them simultaneously and explore the space spanned by them. However, finding these important degrees of freedom is not easy and there is no systematic way to do it. Furthermore, many reactions are driven by a cooperative motion of the reactants and the solvent [21]. Our proton transfer model, as exposed in Section 3.2, presents such challenges.

A very interesting and pragmatic option is to express the reaction progress in terms of energy. This idea is based on Marcus theory for electron transfer and solvent reorganization. It is possible to describe the reactant and product states with different potential energy functions, corresponding to two different bonding topologies. Then, we can construct a reaction coordinate known as energy gap, E_{gap} [21]. We do this by calculating the difference between the initial (1) and final (2) potential energy functions evaluated at any intermediate configuration of the system. The general form is then written as:

$$E_{\text{gap}} = \epsilon_1(r) - \epsilon_2(r) \tag{4.5}$$

By definition, the minimum of each potential energy function is located at the configuration for which it was formulated. Then, we can imagine a simplistic 1-D E_{gap} representation of our proton transfer reaction as the one illustrated in Figures 3.2 and 3.3:

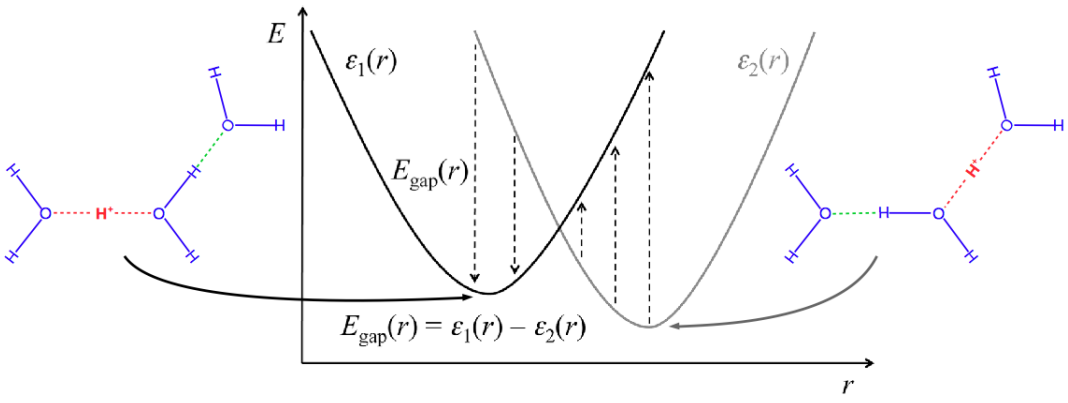

Figure 4.3.: Energy Gap reaction coodinate. Edited from [21].

We would like to make a number of observations on Figure 4.3. First, unlike what is shown in this mere exemplary depiction, both minima should have the same value given the symmetry of our system. Second, these minima should be well-separated, in order to have a selective energy gap that properly identifies the initial and final states. Third, as was said in Chapter 3, any of the water molecules in the first hydration shell may become the receptor for the proton. This simple form of the E_{gap} cannot capture such multiplicity.

To achieve that, more complete formulations have been made [21]. However, in our case we will rely on the simple form, as we will restrict the transfer to occur towards a specific water while performing the sampling. Finally, and most importantly, even if in the diagram only the Zundel and the receptor water are included, we should clarify that the rest of the surrounding water is also considered to calculate the potential energy of the system. In this last point resides the power of the energy gap [21].

In order to evaluate the energy gap, adequate potential energy functions must be inserted into the general form. For this, we look at classical MM force fields. As described in Chapter 2, force fields consist of a functional form and a set of parameters. Then, we face the two-sided challenge of designing a proper force field for our system. First, we have to capture the relevant atomic interactions contributing to the potential energy; and second, we have to correctly parametrize the selected force field terms. Educated guesses of a suitable force field form can be proposed based on similar, previously known, molecular systems. Common practice is to parametrize these models according to experimental data or simulated data from full QM calculations. Since our work is already in the context of an adaptive QM/MM simulation, it is a sensible choice to use the second data source. To perform the actual fitting, there are many techniques currently available. However, it is still critical to capture the essential interactions of the system. Having some well-parametrized contributions, but ignoring others, will not lead to physical results. In order to find an adequate model, test runs with several force fields must ultimately be performed and benchmarked against the experimental or QM reference.

In probabilistic language, the above is an inverse problem — where model parameters are to be inferred from a data set — and it provides an interesting playground for Bayesian strategies [20, 12, 26]. For a given force field form and data, Bayesian model calibration can provide not only sets of optimal parameters, but also probability distributions and associated uncertainties for their values. Complementarily, Bayesian model validation can assign relative probabilities to different models. This comparison balances the complexity and precision of the models, thus avoiding overfitting. In the Appendix A of this thesis, we elaborate on the application of these strategies to the simplification of the system described below.

Since the entire system composed by the Zundel ion, receptor water and surrounding water is far too big to handle, we simplify the problem. We decide not to include the surrounding water bonded and non-bonded parameters in the calibration, and rather let them be defined by the widely recognized fTIP3P water model, as defined in the LAMMPS documentation [8]. The same fTIP3P model is used for all Lennard-Jones parameters of the system. After these considerations, only the bonded parameters and charges for the receptor water and the Zundel ion are left to be resolved. For both molecules, DFT (PBE) QM data is used as reference. Charges are simply averaged. The receptor water bonded interactions, which comprise two harmonic bonds and one harmonic angle, are parametrized by a weighted non-linear least squares fitting. We invest all the Bayesian machinery exclusively in the Zundel ion's bonded interactions.

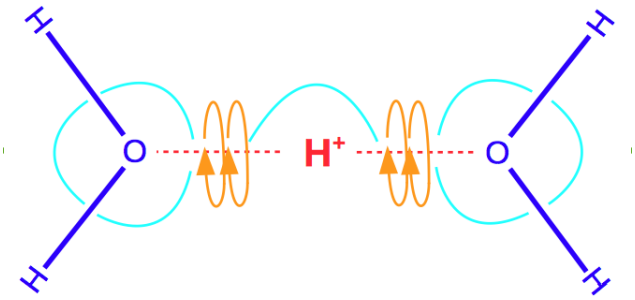

Figure 4.4.: Bonded interactions of the Zundel ion

In a first approach, we assume the Zundel ion to be structured as shown in Figure 4.4. The ion is composed of a total of six bonds, four O-H water-like bonds, plus two additional O-H^+ bonds. The nature of these interactions, Morse or harmonic, is one of the main aspects of the Bayesian analysis. Extra bonds, such as one between the oxygen atoms, are also set for comparison. The angles of the Zundel ion are accounted for as two external ones (H-O-H), four internal ones (H-O-H^+) and a central one with the proton in the middle (O-H^+-O). Four torsions are considered, one on each of the O-H^+-O-H references.

In Appendix A, Bayesian model calibration is employed to parametrize this model. Improvements are proposed based on the probability distributions of the parameters, the precision of the potential energy fitting and the structure of the Zundel ion reproduced by the MM parameters. The process is repeated to generate a broad set of models. Bayesian model validation is then used to evaluate which of the calibrated force fields is best suited to our system. The resulting force field form and parameters are included in Chapter 5.

4.3. Umbrella sampling and free energy profile reconstruction

The simulation times required for reactions to take place are often exceedingly long. Advanced sampling techniques are motivated by this challenge. In general, these rare event methods rely on biased MD simulations. That is, in driving the system into particular desired states, which can be represented by points along the reaction coordinate u [11]. Typically, the goal is to calculate the change of free energy F along u, referred to as a free energy profile. The free energy is directly related to the probability distribution of the system's states $F(u) = -k_B T \ln P(u)$. High energy intervals, or barriers, are not easily accessible and are responsible for the long simulation times. In umbrella sampling (US) [28], we divide the reaction path into several windows, each identified by a particular value of the reaction coordinate u_i. Then, we apply a u-dependent bias (Equation 4.6) to restrain the system, and sample around the given u_i value in each window (Figure 4.5).

$$w_i(u) = \frac{1}{2} k_i (u - u_i)^2 \tag{4.6}$$

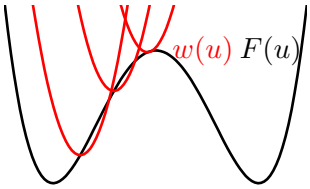

Figure 4.5.: Umbrella sampling windows (red) along a free energy profile (black). Figure by T. Stecher.

The strength of each umbrella biasing potential can be tuned by the k_i parameter. In general, areas with a negative curvature in the free energy, such as the maximum in Figure 4.5, will require a stronger bias to be effectively restrained. However, since the exact shape of the profile is not known a priori, this choice requires further analysis. Moreover, the energetic barriers that are represented by such maxima in the free energy profiles are of great importance during our study of chemical reactions. The height of the barrier provides key insight on the transition between the two adjacent stable states, for example, the initial and final configurations of our proton transfer described in Chapter 3.

Once the umbrella sampling is completed, we require methods to recover the underlying unbiased profile. A traditional approach is the weighted histogram analysis method (WHAM), which poses limitations regarding window overlapping and requires the choice of a certain bin width. A newer scheme available today is umbrella integration (UI) [16], an approach inspired by thermodynamics integration. The main idea of UI is to work with the derivative of the free energy with respect to the reaction coordinate. In each window, we assume there is a point in u for which the biasing force is exactly cancelled by the underlying free energy derivative. Then, this equilibrium point corresponds to the mode of the sampled distribution $\hat{u}_i$. We write [27]:

$$\left.\frac{\partial F(u)}{\partial u}\right|_{\hat{u}_i} = -\left.\frac{\partial w(u)}{\partial u}\right|_{\hat{u}_i} = -k_i(\hat{u}_i - u_i) \tag{4.7}$$

This estimation can be further developed. The biased free energy power series, here implicitly truncated after the first order terms, can be further expanded. In practice, however, the second order terms already introduce a significant amount of noise into the free energy. An additional note concerns statistics. Equation 4.7. is evaluated with the mode $\hat{u}_i$. Nonetheless, typically the mode is approximated by the mean $\hat{u}_i \approx \bar{u}_i$, which is easier to calculate with limited simulation times.

After obtaining the value of the free energy for each window, the data can be combined by a weighted average [16] or by more elaborate techniques, such as Gaussian process regression (GPR) [27]. This thesis will not dive into such methodologies. The provided references contain both, the first proposed version of UI [16], and a state-of-the-art implementation [27].

Part III.

Simulation Protocols, Results and Conclusions

5. Simulation Protocols

5.1. Set up and restraints

All simulations are set in a finite cubic box of $38 \times 38 \times 38$ Å. Centered inside the box, a spherical region with a radius of 18.5 Å contains all atoms. If an atom is at a distance of 1.0 Å or closer to the border of the sphere, a correcting harmonic force with a constant of 50 kcal/(molÅ^2) is applied. We refer to this restraint as a harmonic wall. Inside the sphere, we place 626 pre-equilibrated water molecules, accounting for a density similar to Reference [5]. At the center of the sphere, we position the Zundel ion and the receptor water. The Zundel ion's oxygen atom closest to the receptor water is placed exactly at the center of the spherical region. A Hookean restraint, as implemented in ASE [1], is set to keep this central atom near that position. That is, if the atom is to be displaced beyond 1 Å from the center of the sphere, a correcting harmonic force with a constant of 2.0 eV/Å^2 is applied. In the same way, Hookean restraints are set to maintain the Zundel ion and the receptor water in adequate configurations for the reaction. The O-H$^+$-O-H–O chain interatomic distances are controlled by Hookean restraints activated at 2 Å, with a constant of 2.0 eV/Å^2.

5.2. Energy gap reaction coordinate models

From the seven models calibrated and validated in Appendix A, only the fittest are employed for actual simulations. The parameters of the models are summarized in Table A.5. Models 4, 5, 6 and 7 are employed for evaluations. Only models 4 and 5 are actively used for equilibration and production runs respectively.

5.3. Equilibration runs

The equilibration runs are done exclusively with MM. The initial configuration and restraints are taken as described in Section 5.1. The potential and forces are calculated with a hybrid LAMMPS calculator, which comprises a linear combination of the initial state and the final state energy gap force fields (Equation 5.1). Specifically, we employ force field model 4, which only uses harmonic bonds and has proved the most structurally stable.

$$E = (1 - \lambda)E_{\text{initial}} + \lambda E_{\text{final}} \tag{5.1}$$

The goal of this equilibration protocol is to span the reaction path for the future sampling, and to identify the approximate energy gap values for the initial and final states

in an MM context. A total of five equilibrations runs are set, with mixing parameters λ clustered near the Zundel configuration: [0.05, 0.11, 0.21, 0.34, 0.5]. In order to optimize the equilibration, the runs are performed consecutively. The progressive equilibrations are done from 0.5 to 0.05. The 0.0 value is omitted, as this run contains a considerable amount of noise. The runs are 5 ps long, in steps of 0.5 fs. A Langevin thermostat with a friction of $\gamma = 1/\tau_{\mathrm{L}} = 0.02\ \mathrm{fs}^{-1}$ was set to maintain the system at 300 K. Equilibrations with $\lambda > 0.5$ are performed, but not reported, as they only reflect the symmetry of the system.

5.4. Unbiased production runs

All production simulations are done with our implementation of the AdBF-QM/MM method. The core atoms for all runs are the Zundel ion and the receptor water. Various sizes for the QM and buffer regions are tested. Our goal is to locate stable states and recognize reaction dynamics, as well as analyze the effects of the hysteretic radii on the system's behaviour. Additionally, the energy gap, evaluated with force field models 4, 5, 6 and 7, is monitored during the runs. This is done with the purpose of determining which energy gap formulation provides adequate sensitivity and selectivity for the relevant identified reaction states. The initial configuration for all free runs, with one exception, is taken from the equilibrated run with $\lambda = 0.5$. The restraints described in Section 5.1 still hold. Other than those restrictions, the system is let free to evolve. The time constants for the adaptive Langevin thermostat are $\tau_{\mathrm{L}} = 370$ fs and $\tau_{\mathrm{NH}} = 74$ fs [5]. The time step size is of 0.5 fs. Among the many implemented AdBF-QM/MM protocols, the following selection is presented:

Protocols / Settings	r_-^{QM}	r_+^{QM}	r_-^{buffer}	r_+^{buffer}	xc functional	equilibration mix
AdConv	0.0	0.0	0.0	0.0	PBE	0.5
AdBF0	2.0	2.6	2.0	2.6	PBE	0.5
AdBF2	2.0	2.6	4.0	4.6	PBE	0.5
AdBF3	2.0	2.6	5.0	5.6	PBE	0.5
AdBF4	2.0	2.6	6.0	6.6	PBE	0.5
AdBF3*	2.0	2.6	5.0	5.6	PBE	0.05
AdBF3-PBE0	2.0	2.6	5.0	5.6	PBE0	0.5
AdBF3-B3LYP	2.0	2.6	5.0	5.6	B3LYP	0.5
AdBF3-HSE06	2.0	2.6	5.0	5.6	HSE06	0.5

Table 5.1.: Simulation protocols selection. Each protocol consists of: the four hysteretic radii (in Å) for the AdBF-QM/MM method, the exchange correlation functional for the QM/MM DFT calculations and the equilibration mix from which the starting configuration is taken.

Table 5.1 summarizes the chosen simulation settings. The first protocol, AdConv, corresponds to a conventional unbuffered QM/MM simulation. It provides a simple approach to benchmark against. The following settings, AdBF0, AdBF2, AdBF3 and AdBF4, corre-

spond to calculations with the AdBF-QM/MM method. The QM radius is set to match the first hydration shell of the core atoms, while the buffer size is increased in each simulation. The effects of the buffer size are to be determined from these calculations. Protocol AdBF3* is a sanity check for AdBF3, as it is exactly the same, except for its starting configuration, which is taken from the initial state equilibration. The last three protocols are set to test the effects of the hybrid functionals PBE0, B3LYP and HSE06. The rest of the settings are kept as in the AdBF3 protocol. All FHI-aims calculations include a vander Waals correction based on Hirshfeld's partitioning [10].

5.5. Biased production runs

A particular energy gap formulation - model 5 - and a particular AdBF-QM/MM setting - protocol AdBF3 - are chosen to perform US. The motivation for this is explained in Chapter 6. A total of 12 windows were set for energy gap values ranging from -2.2 to 0.0 eV with steps of 0.2 eV. This increased number of windows with respect to the equilibrations was established after noticing that the free energy profile demanded more sampling than expected. The positive values of the energy gap are omitted, because of the symmetry of the reaction. Each window of the US takes the equilibrated run with the closest average energy gap value as initial configuration. The harmonic constant for all umbrellas is 0.4 eV^{-1}. Each simulation is run for at least 2.5 ps. The first 0.5 ps of each window are considered equilibration, and omitted in the GPR reconstruction of the free energy profile [27].

5.6. Computational aspects

The equilibration runs are done on a local workstation with 8 cores available. However, the LAMMPS calculators are used with a single core, as the input and output operations when coupling with ASE cause significant overhead.

Protocols AdConv and AdBF0, as well as several unmentioned test runs, are done on Arthur, the local cluster at the Chair of Theoretical Chemistry. A total of 16 cores are used for this system during FHI-aims calculations. As before, the LAMMPS operations are done with a single core.

The rest of the unbiased production runs are done on SuperMUC Phase 2 [23]. Protocols AdBF2 and AdBF3 are run using two Haswell nodes, accounting for 56 cores. Protocols AdBF3* and AdBF4 used four Haswell nodes, with 112 cores. The more expensive protocols, AdBF3-PBE0, -B3LYP, and -HSE06, use eight Haswell nodes, accounting for 224 cores. In all cases, the totality of cores is only used for the extended QM/MM embedding. The reduced QM/MM embedding is done with 8 cores, which provides optimal scaling for the smaller quantum system. Again, all LAMMPS operations use a single core. The computing times are reported in Chapter 6. The biased production runs for US windows are also done on SuperMUC Phase 2 [23] , using 112 cores. Again, the reduced QM/MM calculation is done only with 8 cores, and the MM calculations only with one core. The time limit is of 96 hours per window.

6. Results

6.1. Equilibration runs

The quality of the equilibration runs is to be evaluated. We are concerned with the structure of the system and with the spanned energy gap values. Figures 6.1 to 6.5 show the corresponding information about the runs. The left plots depict the oxygen-to-oxygen distance r_{OO}, for both the initial and the final Zundel ions in each equilibration. On the other hand, the right plots show the values of the energy gap during the run.

As expected, the O-O distances gradually evolve from the symmetric Eigen ion configuration (with approximately 2.5 Å for both distances), to the expected Zundel ion structure described in Chapter 3 (with approximately 2.4 Å for the Zundel O-O distance and 2.7 Å for the O-O distance to the receptor water). This confirms the structural correctness of force field model 4, which was used for the hybrid MM calculator. The ASE-Graphic User Interface (ASE-GUI), was employed to visualize and confirm the Eigen and Zundel ion structures. With respect to the energy gap reaction coordinate, the intermediate configuration is expectedly located at zero, while the Zundel ion states are found around ±1.8 eV. This indicates the energy gap range to be explored during the US production runs.

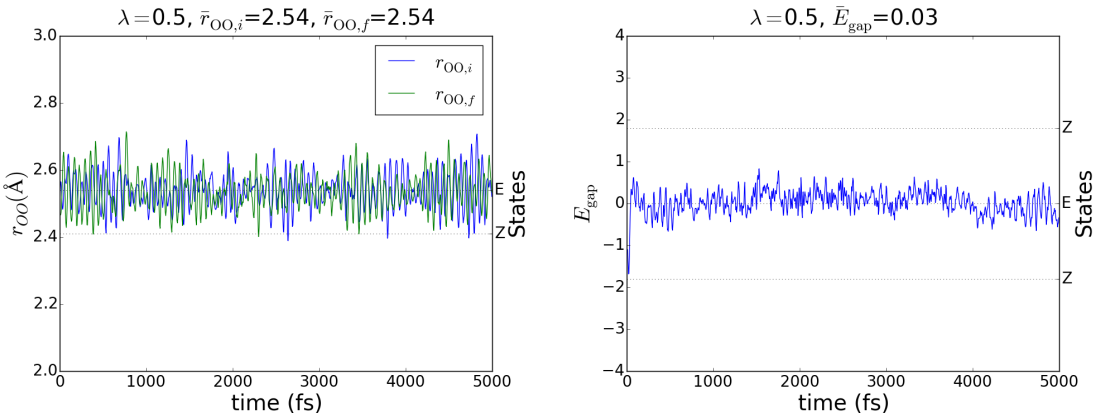

Figure 6.1.: Oxygen-to-oxygen distances r_{OO}, for both the initial and the final Zundel ions (left) and energy gap values (right) during the equilibration run with $\lambda = 0.5$

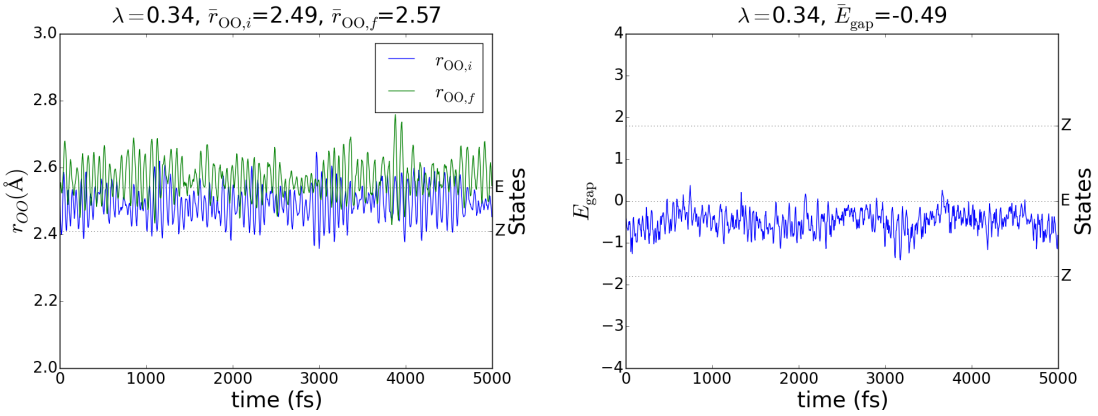

Figure 6.2.: Oxygen-to-oxygen distances r_{OO}, for both the initial and the final Zundel ions (left) and energy gap values (right) during the equilibration run with $\lambda = 0.34$

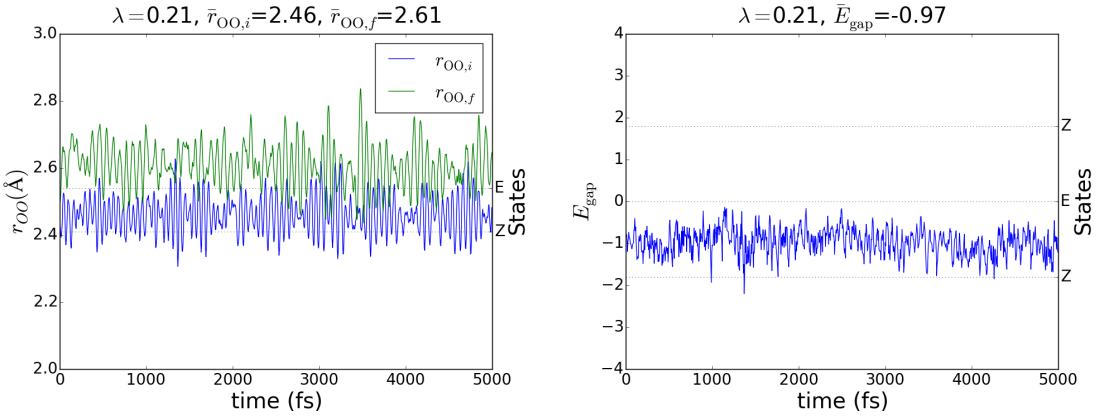

Figure 6.3.: Oxygen-to-oxygen distances r_{OO}, for both the initial and the final Zundel ions (left) and energy gap values (right) during the equilibration run with $\lambda = 0.21$

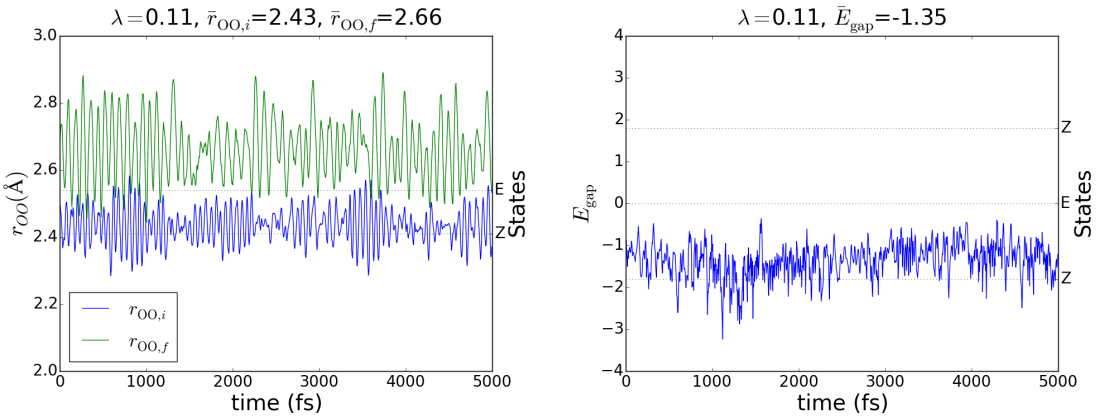

Figure 6.4.: Oxygen-to-oxygen distances r_{OO}, for both the initial and the final Zundel ions (left) and energy gap values (right) during the equilibration run with $\lambda = 0.11$

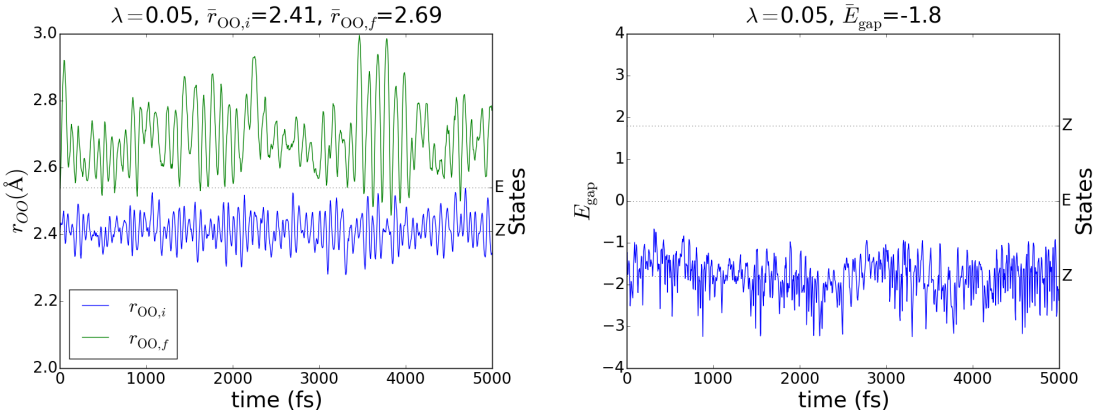

Figure 6.5.: Oxygen-to-oxygen distances r_{OO}, for both the initial and the final Zundel ions (left) and energy gap values (right) during the equilibration run with $\lambda = 0.05$

6.2. Unbiased production runs

The unbiased runs fulfil an exploratory purpose. They deliver information about the system's dynamics and stable states, before bias potentials are added. In the following Figures, we show the sampled energy gap values, both in time series plots and histograms. We also identify the Zundel and Eigen ion states with Z and E labels on the secondary axes.

Protocol AdConv

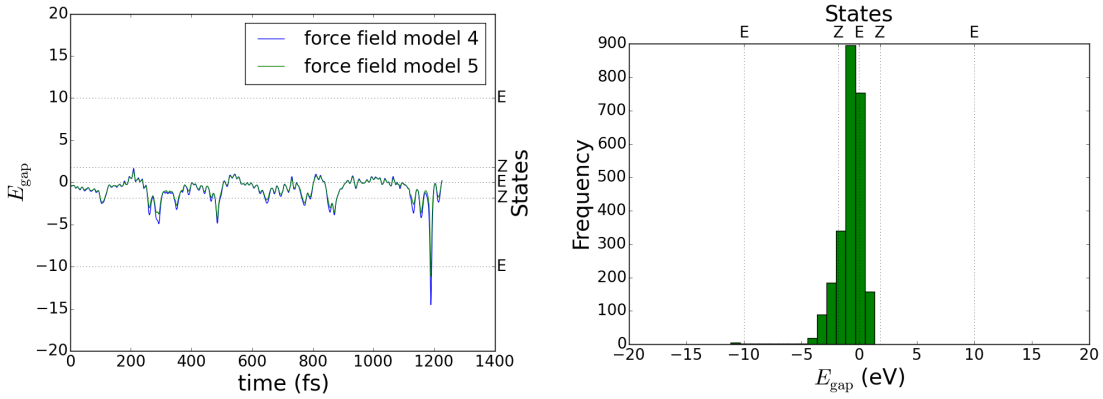

Figure 6.6.: Energy gap time series plot (left) and histogram (right) generated with protocol AdConv.

In this adaptive conventional QM/MM method, the system is visibly restricted near the Eigen ion configuration. We hypothesize that having only the core atoms as QM region, and no buffers, is the cause for this lack of mobility. Considering only this small part of the system with QM, can reasonably explain the stability of the Eigen ion, which provides symmetry. Additionally, in the histogram we observe that the system shows no preference for the Zundel ion state located near the energy gap value of -1.8 eV. Finally, it should also be noted that the energy gap formulation with force field model 5 is less sensitive than model 4, specially for peaks away from the zero. This aspect is also noticeable in the following protocols, and will be exploited to reduce the noise in the US windows.

Protocol AdBF0

This protocol, which sets the QM region to cover the first hydration shell of the core atoms and no buffer, returns interesting results. In the time series, it is observable that, after some equilibration due to the solvent effects, the system drifts far away from the initial configuration. Proton transfer occurs, but not as expected. The system quickly goes through the initial Zundel ion state defined in Chapter 3, and then continues for a second proton transfer to one of the water molecules in the first hydration shell. The final state of this transfer is not clearly identifiable as an Eigen or Zundel ion. This process reflects one of the most critical aspects of QM/MM interactions. We are convinced that the drifting of the proton toward the edge of the QM region is due to the Coulomb singularities (point charges), that represent the MM atoms during the embedding. That is, the proton is attracted by the nearby oxygen atoms outside the QM region. Fortunately, the implemented buffered method provides a solution to this problem. In the following protocols, we increase the size of the buffer, aiming to reduce the effect of the MM charges on the proton. It should also be mentioned that, having the system so far away from the originally expected configurations, pushes the limits of the parametrized energy gap reaction coordinate, which becomes considerably noisy.

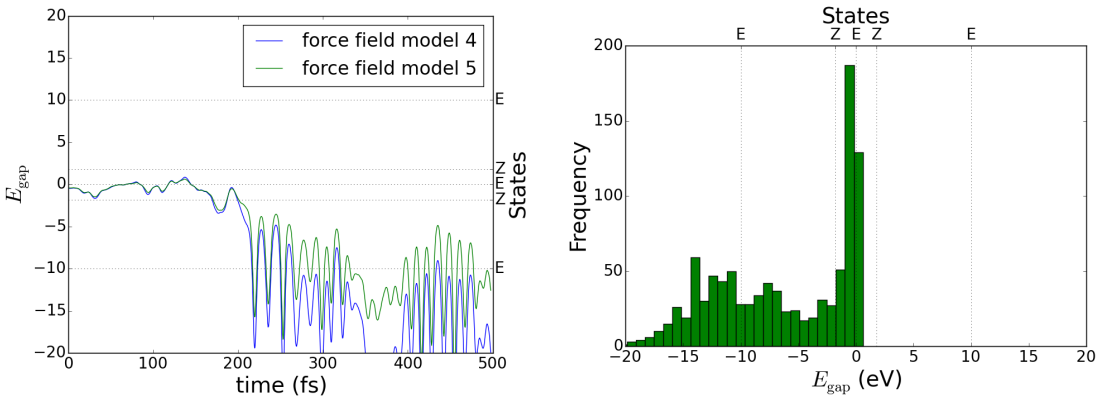

Figure 6.7.: Energy gap time series plot (left) and histogram (right) generated with protocol AdBF0.

Protocol AdBF2

In this short exploratory run, we set the QM region to match the first hydration shell, and the buffer to match the second. After a brief equilibration, the system once again drifts away to edge. However, not as fast as in the previous attempt with no buffer. The Eigen ion is still the dominant state, as seen in the histogram. In the next protocols, we proceed to increase the buffer size.

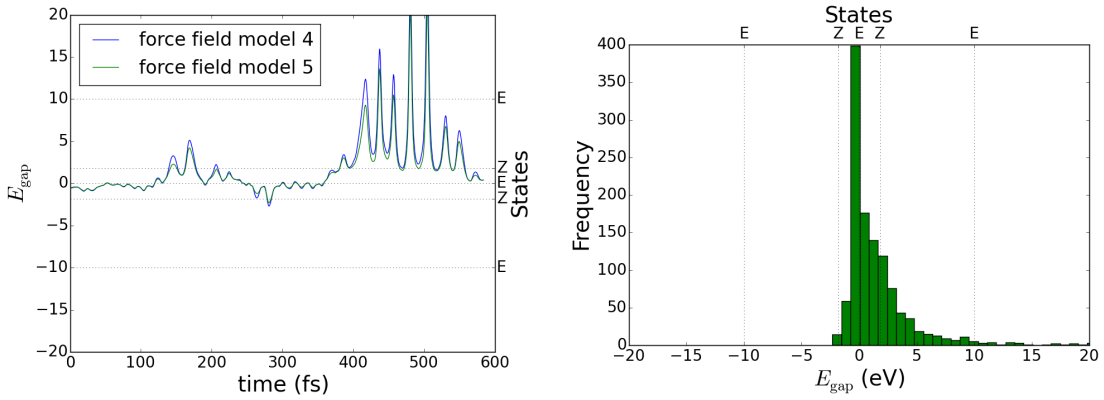

Figure 6.8.: Energy gap time series plot (left) and histogram (right) generated with protocol AdBF2.

Protocol AdBF3

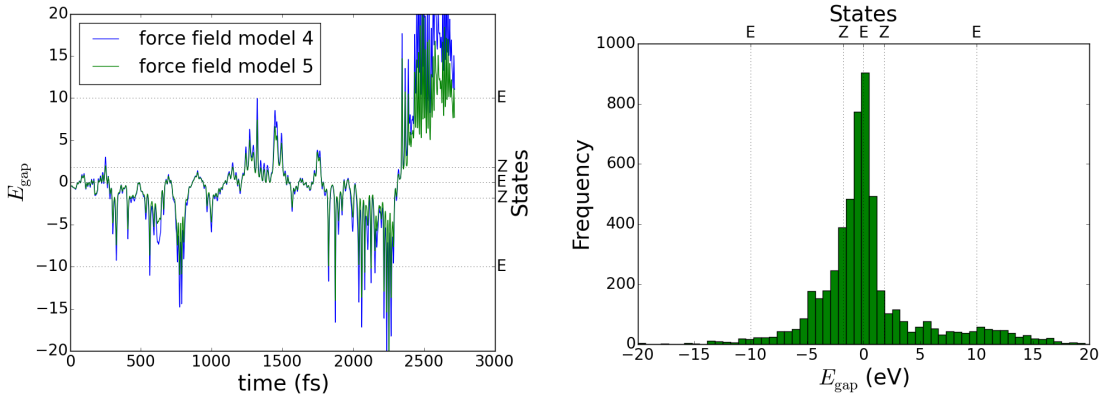

Figure 6.9.: Energy gap time series plot (left) and histogram (right) generated with protocol AdBF3.

The QM region is kept for the first hydration shell, while the buffer covers 3.0 Å on top. For approximately 2 ps, the system stays reasonably bounded, and later visits two other

states further away from the expected interval. When analyzing the dynamics, these two states were identified as Eigen ions. They formed on the core oxygen atoms on each side of the central oxygen atom. That is, one Eigen ion formed at the receptor water, and one at the opposite oxygen of the initial Zundel ion. At this point we recognize that, contrary to our expectations, the Eigen ion is the stable state of an excess proton in water under these simulation conditions. Protocol AdBF3* is set to confirm this outcome.

Protocol AdBF4

This protocol completes the buffer tests. It employs a buffer of 4.0 Å on top of the first hydration shell QM region. The simulation behaves similarly to protocols AdBF3 and, especially, AdBF3*. Stable Eigen ion configurations are now clearly identifiable in the histogram at energy gap values of 0.0 and approximately 10.0 eV.

Even larger buffer sizes are not tested for two reasons. First, in the literature, this is the maximum buffer size used for systems as or more complex than ours [22, 29]. Second, the increasing computational cost of the extended QM/MM calculation does not appear to pay off anymore, as the simulations results are not significantly affected. Of course, in order to properly affirm this, a force convergence study is required.

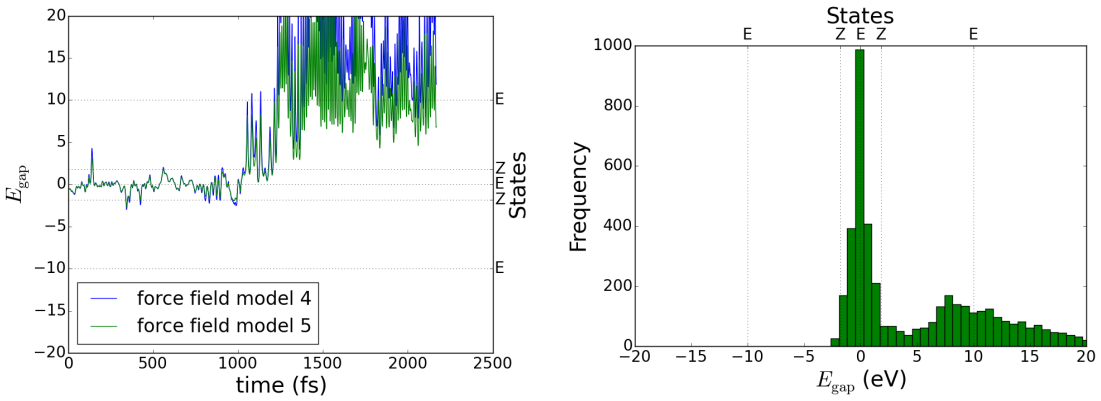

Figure 6.10.: Energy gap time series plot (left) and histogram (right) generated with protocol AdBF4.

Protocol AdBF3*

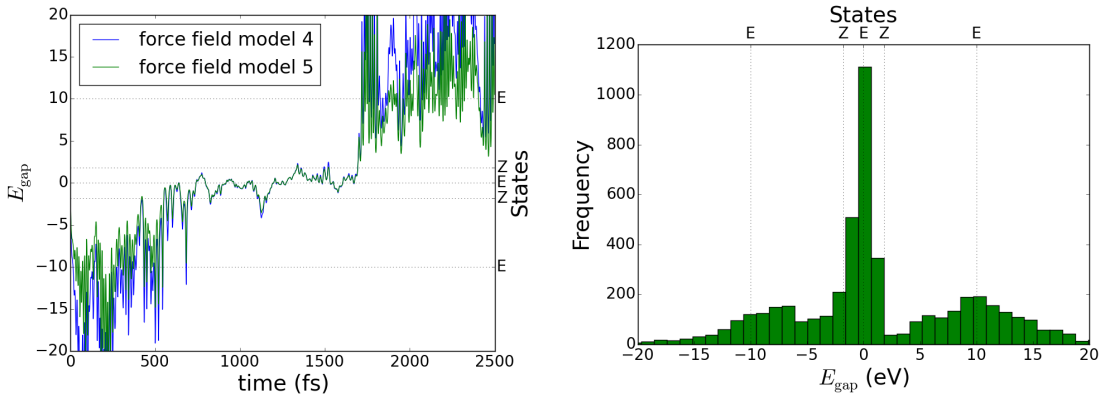

Figure 6.11.: Energy gap time series plot (left) and histogram (right) generated with protocol AdBF3*.

This protocol, equivalent to AdBF3, starts from the initial Zundel configuration, and provides very valuable insight about our system. Three Eigen ion configurations, formed on each of the core oxygen atoms, are visited (Figure 6.12). The energy gap values for such states are visible at 0.0 and approximately ± 10 eV in the histogram. The intermediate Zundel ions are not stable. The buffer size is seemingly sufficient to prevent the proton from drifting towards the edge of the QM zone, as it is never transferred to the oxygens in the first hydration shell of the core region.

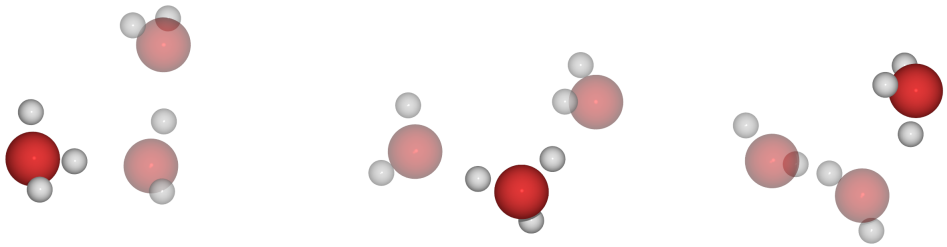

Figure 6.12.: Stable Eigen ion configurations at the three core oxygen atoms, corresponding to energy gap values of -10 eV (left), 0 eV (center) and 10 eV (right).

We also seize this protocol to verify the proper functioning of some aspects of the AdBF-QM/MM method. The first one is the adaptive Langevin thermostat. In Figure 6.13, the running average kinetic energy per atom is shown. The thermostat keeps all atoms on the same average value of 0.39 eV. The increased variance on QM and buffer atoms is simply due to the smaller number of atoms in such regions.

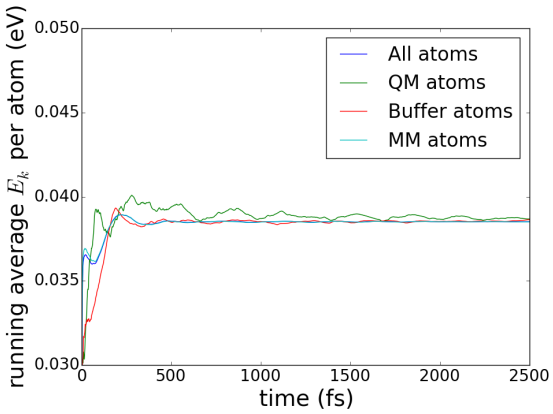

Figure 6.13.: Kinetic energy of atoms in different regions during the unbiased run with protocol AdBF3*.

The second aspect that is revised is the hysteretic classification of the atoms. Figure 6.14 shows the number of atoms in the QM and buffer regions during the run. As designed, the classification does not present quick changes.

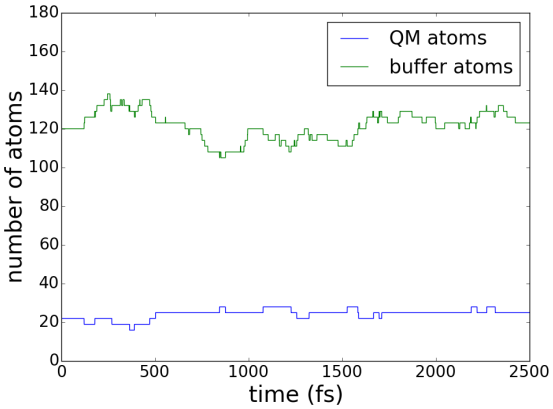

Figure 6.14.: Time series of the number of QM and buffer atoms during the unbiased run with protocol AdBF3*.

Two additional aspects that we wish to report are the orders of magnitude of both, the momentum conservation correction, and the force exerted by the wall. The forces applied for momentum conservation are of around 10^{-4} eV/Å on oxygen atoms, and 10^{-5} eV/Å on hydrogen atoms. The total force exerted by the wall is of around 10^{-2} to 10^{-1} eV/Å.

Protocol AdBF3-PBE0

After confirming that using a GGA functional for the QM DFT calculations does not return a stable Zundel ion, we evaluate hybrid functionals. This protocol is the same as AdBF3, but using the PBE0 exchange correlation functional. The result, however, is not different from the previous ones. The system does not seem to have any stable state near the Zundel ion, but rather in the vicinity of the Eigen ion. The computational cost of this hybrid functional limited the extension of our exploratory runs.

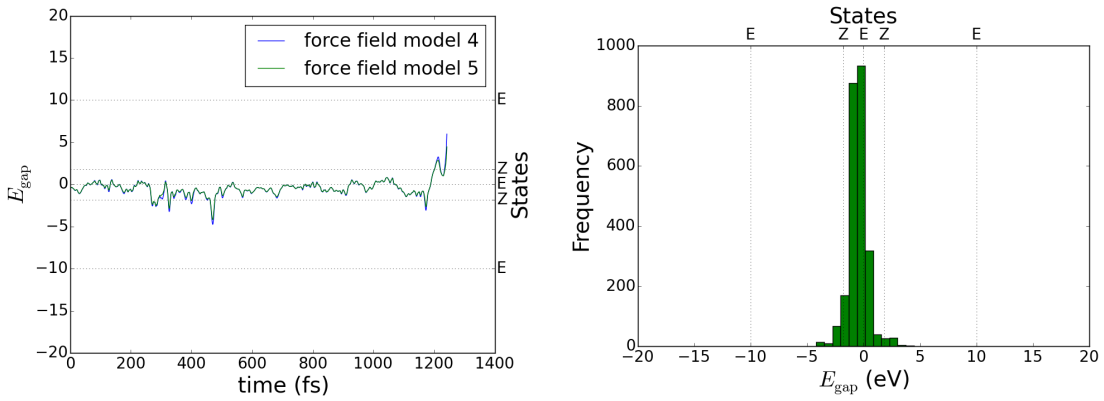

Figure 6.15.: Energy gap time series plot (left) and histogram (right) generated with protocol AdBF3-PBE0.

Protocol AdBF3-B3LYP

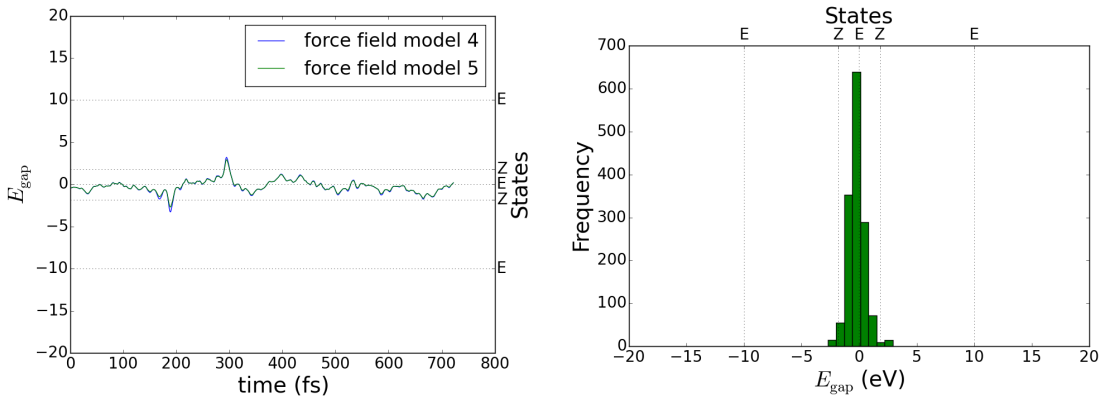

Figure 6.16.: Energy gap time series plot (left) and histogram (right) generated with protocol AdBF3-B3LYP.

As in the previous protocol, this run is designed to test the effect of a hybrid functional. B3LYP is used, and the rest of the parameters are kept as in the AdBF3 protocol. The resulting histogram still shows no stable state near the Zundel ion, but near the Eigen. As before, the length of the run is limited by the cost of the hybrid functional.

Protocol AdBF3-HSE06

In this final protocol with the hybrid functional HSE06, we observe again a preference for the Eigen ion state. The Zundel is never visited. As the run is rather short, further exploration, with more computational investment, would be required. We postpone this for further research.

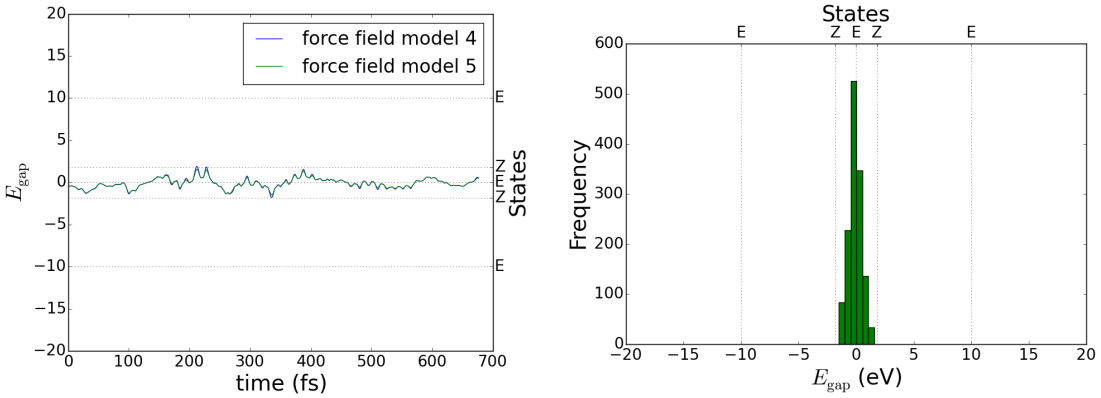

Figure 6.17.: Energy gap time series plot (left) and histogram (right) generated with protocol AdBF3-HSE06.

6.3. Biased production runs

US is performed with the protocol AdBF3, which has already been identified to generate an Eigen-Zundel-Eigen proton transfer. The UI (Figure 6.18) confirms the observations from the unbiased runs. In our current QM/MM setting, the stable state of an excess proton in water is the Eigen ion. We find the corresponding minimum free energy at the zero energy gap value. However, a maximum is not observable around -1.8 eV, where the Zundel ion configuration is located. This is certainly counterintuitive, as having the Eigen ion as minimum energy state should also imply having the Zundel as maximum. This is the reason for which the sampling was extended to -2.2 eV, but no clear maximum was detected. From the unbiased runs, the minimum free energy corresponding to the next Eigen ion should be around -10 eV. However, our hybrid calculator method for equilibration cannot produce adequate initial configurations for such states far away from the original range. Therefore, we do not set more windows closer to the next Eigen ion. Additionally, we note that the scale of the free energy changes is quite small (around 0.04 eV), which explains the quick transitions seen in the protocols above.

Umbrella integration with protocol AdBF3

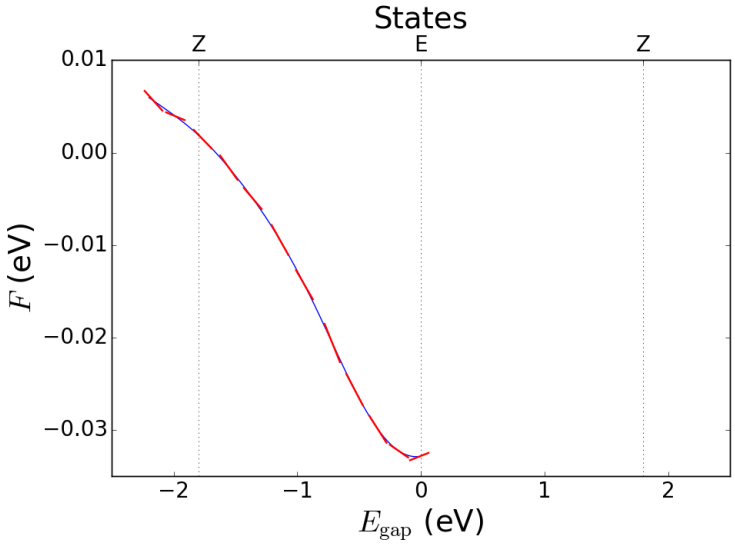

Figure 6.18.: Blue: free energy profile reconstructed after umbrella sampling and umbrella integration with protocol AdBF3. Red: sampled derivative of the free energy with respect to the energy gap reaction coordinate.

After confirming that the Eigen ion is the stable state in our QM/MM framework. We decided to perform an additional US, with a new protocol not described in Section 5. In this protocol, we set two water molecules and the proton between them as core atoms. The initial and final states are assumed to be Eigen ions formed at each of the water molecules and the intermediate state is a Zundel ion. The hysteretic radii and exchange correlation functional are set as in protocol AdBF3. The energy gap force fields for the Eigen ion were parametrized by weighted non-linear least squares fitting on DFT PBE data, considering three Morse bonds and three harmonic angles. The results, with units as in the Appendix A, are: $D_m = 72.54$, $\alpha_m = 2.67$, $r_{0,m} = 0.99$, $K_a = 36.02$, $\theta_{0,a} = 110.93$. The force fields were also used for MM equilibration, having the core atoms constrained near the center of the spherical region. Five windows were run at energy gap values of [0.0, -0.75,-1.6,-2.8,-4.95,-6.0] eV with $k_i = 0.2$ eV^{-1}. Each window was run for at least 1.5 ps, ignoring the first 0.5 ps for equilibration. The computing time limit was of 48 h. The free energy profile was reconstructed using GPR (Figure 6.19).

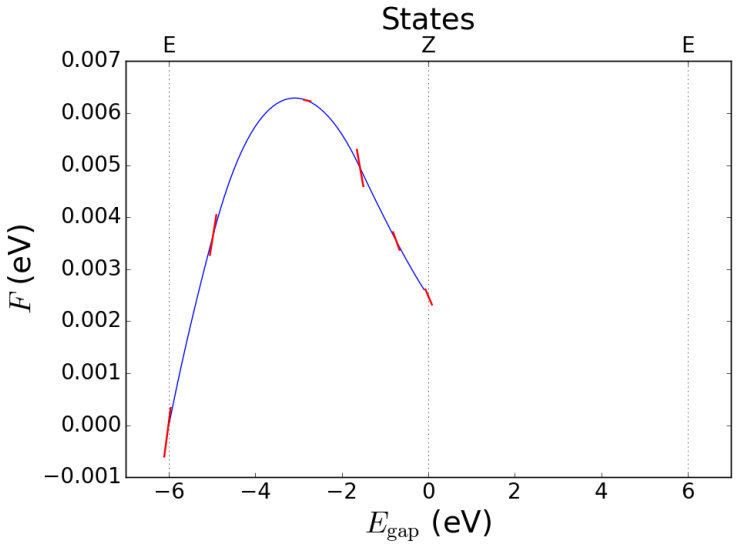

Figure 6.19.: Blue: free energy profile reconstructed after umbrella sampling and umbrella integration with new protocol assuming the Eigen ion as stable state. Red: sampled derivative of the free energy with respect to the energy gap reaction coordinate.

While this calculation might not be fully converged, it is clear that, contrary to our expectations, an Eigen-to-Eigen transfer with a barrier at the Zundel ion is not captured. Modifying the size and shape of the QM region changes the stable states of the system. Even with clear minima not detected, we observe a coexistence of both, Zundel and Eigen states, in the free energy profile. This result reinforces one of the motivations of this thesis, as it reflects how the selected methodologies affect the reaction mechanisms observed in simulations. We are now also able to grasp the difficulty of fully understanding the nature of the excess proton in water.

6.4. Computational aspects

Table 6.1 summarizes the computational resources employed for each protocol, as well as an approximate lower bound for the computing time of each time step. It should be noted that, depending on the convergence of the DFT SCF cycle, and on the number of particles entering and exiting the buffer, these time steps can vary significantly. Calculations with a bigger buffer, present larger increments during the runs, as more atoms are able to cross the QM/MM boundary.

Protocols / Resources	System	Number of cores	Run time (h)	Step time (s)
AdConv	Arthur	16	120	60
AdBF0	Arthur	16	120	90
AdBF2	Supermuc Phase 2	56	48	40
AdBF3	Supermuc Phase 2	56	192	70
AdBF4	Supermuc Phase 2	112	144	80
AdBF3*	Supermuc Phase 2	112	96	70
AdBF3-PBE0	Supermuc Phase 2	224	240	250
AdBF3-B3LYP	Supermuc Phase 2	224	144	270
AdBF3-HSE06	Supermuc Phase 2	224	192	270

Table 6.1.: Computational resources invested in each protocol. The time step times are approximated lower bounds in seconds.

7. Conclusions

7.1. Discussion

First, we dive into the multiscale modeling. The robustness of the AdBF-QM/MM method, which allows for clean and functional coding, is highly valuable. The method's portability was exploited in ASE, using FHI-aims and LAMMPS. Additionally, the several tunable parameters of the method, as the definitions of the core atoms and the hysteretic radii, allow for an important degree of flexibility when covering complex systems. This flexibility builds up, as the respective QM/MM electrostatic embedding and MM methods also have their own parameters. All in all, the AdBF-QM/MM performs resiliently and, as reported in the literature [5, 22, 29], captures structures and free energy profiles. Nonetheless, the method is not free of challenges. In our test runs, we have discovered how critical is it to adequately set the QM and buffer regions. The QM region must be large enough to enable the respective QM forces to drive the system. A too small QM zone will not allow for quantum behaviour to emerge. The shape of the region is also critical, as it should capture the key actors of the reaction, and account for relevant symmetry. Additionally, the buffer must also be chosen sensibly. It should put enough distance between the QM region and the external MM point charges during the embedding. Otherwise, the close interactions with the Coulomb singularities pull the DFT charges toward the edge of the QM region. However, this aspect should be carefully balanced, as big buffer sizes lead to increasingly expensive QM/MM calculations. For this purpose, in the Section 7.2 we propose further tests to optimize the method's regions and parameters.

Regarding the reaction coordinate, we appreciate the ability of the energy based approach to capture the effects of the surrounding solvent. While some effort is required for the force field parametrization, the resulting description of the system is much more comprehensive than a geometrically inspired formulation. Bayesian strategies provide a fit framework to calibrate and validate MM force fields based on QM DFT data. This is discussed further in Appendix A. Ultimately, force fields can be successfully designed and inserted into the energy gap formulation.The reaction coordinate can then be monitored during unbiased runs to verify its sensitivity, and eventually selected for biased simulations. One additional advantage of this approach is that the generated force fields are useful for pre-equilibration. However, the energy based reaction coordinate also has limitations. We observed how, when the system drifts away form the intended initial and final configurations, the reaction coordinate becomes highly sensitive. The energy gap is constructed to capture specific processes. Evidently, more complex energy based reaction coordinates can be formulated [21], but they still should be carefully calibrated in order to capture relevant reaction states.

Finally, concerning the investigated proton transfer in water, our understanding of the system's complexity was notably broadened. All of our proceedings were originally oriented

to a Zundel-to-Zundel proton transfer. We set both the AdBF-QM/MM regions and the reaction coordinate to match the Zundel ion as the stable state. However, during our unbiased simulations, we had to conclude that, for our set-up, the stable state of the excess proton in water was the Eigen ion. This was confirmed by the calculation of the free energy profile. Moreover, under this original setting it was not clear if the now intermediate Zundel corresponded to a maximum in the free energy profile, which compromises the symmetry of the reaction. At the end of our work, the Zundel-Eigen dichotomy remains as an open question [13, 25], inviting further investigation.

7.2. Outlook

One of the most enriching aspects of this thesis is the number of doors that it opens for subsequent work. Below, we propose some interesting studies beyond ours.

An interesting extension would be to account for nuclear quantum effects, which might conceivably favour the Zundel ion [13]. Additionally, this research could be complemented by longer tests with different exchange correlation functionals. Then, we can also think about implementing periodic boundary conditions on FHI-aims, so no wall is necessary to keep the system together. Another aspect of QM/MM that can be explored is smearing the external MM charges during the embedding, so a smaller buffer size might be sufficient. Having a polarizable MM force field might also help. Ultimately, to optimize the buffer size, a force convergence study must be performed. The size of the QM region itself can also be subject to further testing. It would also be interesting to modify the force fields calibrated for the energy gap, and substitute the reduced QM/MM embedding calculation with a full MM calculation. For this purpose, a more elaborate model parametrization would be necessary. Another development could introduce a new form of adaptiveness, by allowing the core atoms and the respective QM and buffer regions to change according to the position of the proton, thus, enabling the tracking of multiple consecutive transfers. Finally, having a full QM run of the proton transfer would be expensive, but also useful for benchmarking.

For the energy gap reaction coordinate, it would be engaging to attempt more complex fittings for the parameters. For example, explicitly including hydration shells of the core atoms can change the charges and bonded parameters of the force fields. More data would be required for these calibrations, as well as a more robust use of the Bayesian strategies. Additionally, the formulation of the coordinate itself can be changed, to account for multiple reactant and product states, enabling for more freedom in the simulations.

Finally, the Zundel-to-Zundel proton transfer should also be investigated after modifying the AdBF-QM/MM method. Fully understanding this system, would not only provide valuable experience for more complex reactions, but would also shine some light on these open questions: What is the nature of the excess proton in water? Which are the stable and intermediate states of the transfer mechanism? And how can our simulation techniques faithfully reproduce the physics of this chemical reaction?

Appendix

A. Appendix

This annex to the thesis also served as final report for the course *Bayesian Strategies for Inverse Problems*, by Prof. Faidon-Stelios Koutsourelakis from the Research Group for Continuum Mechanics. The version shown here is modified for consistency with the rest of the writing. The work is based on course material [18] and References [20, 12, 26].

In order to provide an E_{gap} reaction coordinate for the proton transfer in water described in Chapter 3, it is necessary to correctly define and parametrize a force field for the Zundel ion, the receptor water and the rest of the water. As posed at the end of Section 4.2, by using the fTIP3P model for the surrounding water and an independent fitting for the receptor, the problem can be reduced to designing a force field only for the bonded parameters of the Zundel ion.

The objectives of this appendix are:

1. Calibrate multiple models for the Zundel ion force field, each accounting for a different set of interactions.
2. Validate the calibrated models and identify the physically relevant interactions.

Our approach consists of:

- Data: using FHI-aims DFT package, run a full QM simulation of the Zundel ion to produce a trajectory of 10,000 steps of 1 fs each. Use this trajectory as set of data containing different configurations (atomic positions, bonds, angles and torsions) and their associated potential energies.
- Units: energy contributions are expressed in kcal/mol; distances in Å; angles in degrees; harmonic constants in such units that return kcal/mol when multiplied with their respective terms; the Morse parameters D_m in kcal/mol and α_m in Å^{-1}; electric charges in elementary charge e; consistently, the Coulomb's constant in $\frac{\text{kcal Å}}{\text{mol } e^2}$; and Lennard-Jones parameters ϵ_{jk} in kcal/mol and σ_{jk} in Å. Finally, the torsion integer parameter n_t and the dielectric constant ϵ have no units.
- Model Calibration: obtain maximum likelihood estimates (MLE) for the force field parameters assuming a uniform prior. Use the Laplace method to derive the posterior distributions associated with the MLE. Then, for the simplest model, obtain via Markov chain Monte Carlo (MCMC) sampling an actual posterior distribution in order to evaluate the adequacy of the Laplace approximation..
- Model Validation: use the model evidence term and Bayes factor to compare the different models parametrized by MLE and extract conclusions about the relevant interactions.

A.1. Bayesian model calibration

We engage the challenge of calibrating the bonded parameters for the Zundel ion force field. Seven models will be parametrized. We will do so via MLE. For the simplest case, we will also obtain a posterior via MCMC sampling to benchmark the MLE.

A.1.1. Formulation

We begin by inserting our current problem in the Bayes theorem:

$$\underbrace{p(\theta \mid traj, ff)}_{\text{posterior}} = \frac{\overbrace{p(traj \mid \theta, ff)}^{\text{likelihood}} \overbrace{p(\theta \mid ff)}^{\text{prior}}}{\underbrace{p(traj \mid ff)}_{\text{evidence}}} \tag{A.1}$$

Where θ is a parameter vector, $traj$ is the full QM trajectory with our configuration and potential energy data; and ff is our force field model. We are interested in the posterior distribution $p(\theta \mid traj, ff)$ of a certain parameter set θ to reproduce our trajectory data $traj$ under a certain force field form ff. This posterior probability is equal to a likelihood times a prior, normalized by a model evidence term. The likelihood quantifies how likely is our trajectory data $traj$, given the parameters θ and the force field form ff. The prior represents a preconceived idea of how should the parameters θ look like. The model evidence $p(traj \mid ff)$ acts as a normalization constant, and therefore can be neglected when trying to determine the shape of the posterior.

$$p(\theta \mid traj, ff) \propto p(traj \mid \theta, ff) p(\theta \mid ff) \tag{A.2}$$

The model evidence will regain importance when we go into model validation, and compare different force field models ff. For now, only the likelihood and the prior should be adequately defined for our problem. In order to express a likelihood, a stochastic behaviour should be introduced. As presented in literature [26], this can be done by assuming Gaussian noise in our data. Namely, to each potential energy ϵ_τ of a certain configuration r_τ in our full QM run $traj$, we add a random Gaussian distributed variable γ_τ centered at zero with standard deviation σ_τ. This returns a noisy potential energy ϵ'_τ, that enables the use of Bayesian strategies. This assumption is reasonable considering the time stepping nature of the data generator. With all aspects considered, we have:

$$\epsilon'_\tau = \epsilon(r_\tau \mid \theta) + \gamma_\tau, \qquad \gamma_\tau \sim \mathcal{N}(0, \sigma_\tau^2) \tag{A.3}$$

Inserting this Gaussian noise as likelihood in Equation A.2, we obtain:

$$p(\theta \mid traj, ff) \propto \frac{1}{(2\pi\sigma_\tau^2)^{N/2}} \exp\left[-\frac{1}{2} \sum_\tau^N \left(\frac{\epsilon'_\tau - \epsilon(r_\tau \mid \theta, ff)}{\sigma_\tau} \right)^2 \right] p(\theta \mid ff) \tag{A.4}$$

If the prior is assumed to be uniform, then finding the maximum of the posterior (maximum a posteriori or MAP) is equivalent to finding the maximum of the likelihood, hence, a maximum likelihood estimate.

A.1.2. Maximum likelihood estimate (MLE)

Taking only the likelihood term, the equation above can be easily reduced to a weighted non-linear least squares problem in which we have to minimize the argument of the exponential function:

$$p(traj \mid \theta, ff) = \frac{1}{(2\pi\sigma_\tau^2)^{N/2}} \exp\left[-\frac{1}{2} \sum_\tau^N \left(\frac{\epsilon'_\tau - \epsilon(r_\tau \mid \theta, ff)}{\sigma_\tau} \right)^2 \right] \tag{A.5}$$

This problem has been widely studied, meaning that currently there exist very stable and fast solvers. We will make use of one solver implemented in Python's Scipy library (version 0.17.1), inside the function `curve_fit()`. This same resource was used for the fitting of the receptor water.

As a last remark before diving into the fitting of the Zundel ion, one important detail should be explained. Even though the MLE is a single-point estimate, it can be connected to a Gaussian distribution by a formalism known as the Laplace method [12, 20]. First, we Taylor-expand the log-posterior:

$$\ln p(\theta \mid traj, ff) \approx \ln p(\theta_0 \mid traj, ff) - \frac{1}{2}(\theta - \theta_0)^T \Sigma_\theta^{-1} (\theta - \theta_0) \tag{A.6}$$

Where, Σ_θ is the covariance matrix of the parameters, also known as standard error matrix, and θ_0 is the vector of optimal parameters. Both terms are provided by the `curve_fit()` Python function.

$$\Sigma_\theta^{-1} = -\frac{\partial^2}{\partial\theta^2} \ln p(\theta \mid traj, ff) \bigg|_{\theta=\theta_0} \tag{A.7}$$

This allows to approximate the posterior distribution as an unnormalized distribution:

$$p(\theta \mid traj, ff) \approx p(\theta_0 \mid traj, ff) \exp\left[-\frac{1}{2}(\theta - \theta_0)^T \Sigma_\theta^{-1} (\theta - \theta_0) \right] \tag{A.8}$$

Having such a Gaussian distribution function provides an indication of the level of uncertainty associated with our MLE. We will use this notion in the immediate subsections regarding the fitting of four different models. Additionally, the Gaussian distribution facilitates the calculation of the normalizing model evidence term. This will be done in the latter sections, when performing model comparison.

A.1.3. Model 1: Morse bonds

We start with the model shown in the Figure 4.4. In a first attempt, the bonds of the water molecules are considered harmonic, while the ones to the proton are taken as Morse bonds. The angles, two external ones (H-O-H), four internal ones (H-O-H^+) and a central one (O-H^+-O), are all taken as harmonic. One torsion per hydrogen is considered along each of the four O-H^+-O-H references. All force field contributions are calculated as shown in Section 2.2. In this Appendix, Morse bonds and harmonic bonds will be distinguished in figures by red and blue coloring respectively.

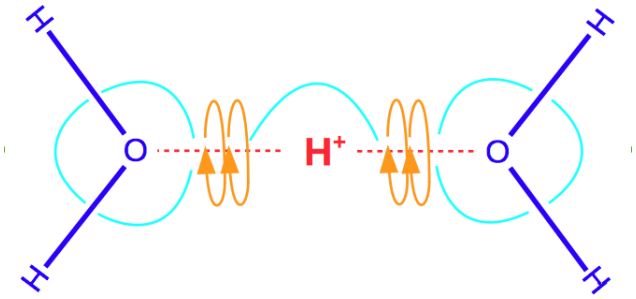

Figure A.1.: Zundel ion interactions considered in model 1. Morse bonds are depicted in red; harmonic bonds, in blue; harmonic angles, in cyan; and harmonic torsions, in orange.

It should be noted that symmetrically equivalent interactions share the same parameters. This encloses the four harmonic bonds between themselves, the two Morse bonds between themselves, the two externals angles between themselves, and so on. We use as first guesses: the average value during the full QM run for Morse and harmonic equilibrium bond lengths, as well as for harmonic equilibrium angles; the inverse of the respective data's variance times $k_BT/2$ for harmonic constants; 100 kcal/mol and 0.01 $Å^{-1}$ for Morse potential parameters D_m and α_m respectively; and 1 for the torsion parameter n_t. We introduce an additional parameter, a constant negative offset with a first guess equivalent to the minimum energy. This guarantees all energy contributions to be positive. Additionally, we bound the equilibrium parameters to be $\pm$ 10 % around the first guess. In the same fashion, we also bound all harmonic constants and the Morse D_m to be positive and < 1000 kcal/mol, as well as α_m to be also positive and < 10 Å. The torsion integer parameter is kept between 0 and 1000.

The fitting resulted in the following parameters. The $\pm$ values are defined as $2\sigma_\theta$, where σ_θ represents the diagonal terms of Σ_θ, which indicate the standard deviation of each particular parameter.

- $K_{b,\text{water}} = 526.77 \pm 9.91$
  $r_{0,b,\text{water}} = 0.98 \pm 0.0003$
- $D_{m,\text{proton}} = 399.98 \pm 1135.26$

$\alpha_{m,\text{proton}} = 0.23 \pm 0.34$
$r_{0,m,\text{proton}} = 1.17 \pm 0.008$

- $K_{a,\text{ext}} = 38.64 \pm 1.23$
  $\theta_{0,a,ext} = 108.30 \pm 0.11$
- $K_{a,\text{int}} = 15.81 \pm 0.35$
  $\theta_{0,a,\text{int}} = 116.60 \pm 0.13$
- $K_{a,\text{cen}} = 10.69 \pm 1.10$
  $\theta_{0,a,\text{cen}} = 181.65 \pm 1.57$
- $K_t = 0.45 \pm 0.02$
  $n_t = 0.99 \pm 0.01$
- $c_o = -11.76 \pm 0.13$

We recovered the following potential energy for the trajectory data:

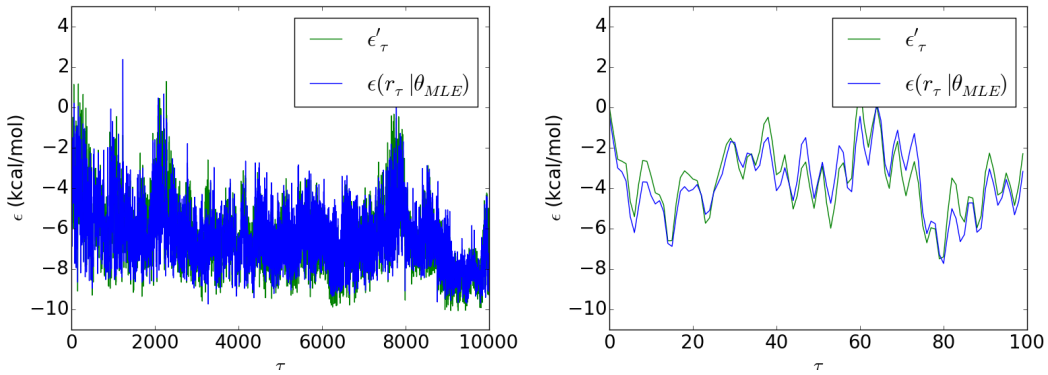

Figure A.2.: Left: Potential energies for the trajectory data given by the full QM run (green) and by the model 1 force field (blue); Right: first 100 trajectory steps of the left plot.

The root mean square error (RMSE) for the potential energy fitting is 0.75 kcal/mol, a very reasonable value. The potential energy fitting also shows a close fit. However, there are a few warning flags. First, the uncertainty for the Morse potential parameters is extremely big. Moreover, when trying to reproduce the general dynamics and structure of the zundel ion with these parameters in a full MM run, the O-O distance does not correspond to our QM reference of 2.4 Å. We hypothesize that this occurs because the derivative of the Morse potential depends on a D_m/α_m ratio. This means we are trying to fit two parameters when we only have information for one. Another reason could simply be the fact that we are simulating very close to the equilibrium bond length. Perhaps higher temperatures are needed to capture the right data. The rest of the bond lengths and angles are in accordance with the QM simulation. The harmonic constants have all physical values, similar to those of fTIP3P water.

A.1.4. Model 2: all bonds harmonic

Motivated by the results of the previous fitting, we attempted a second one. For this occasion, we have replaced the O-H^+ Morse bonds with harmonic bonds. The rest of the interactions are managed in the same way.

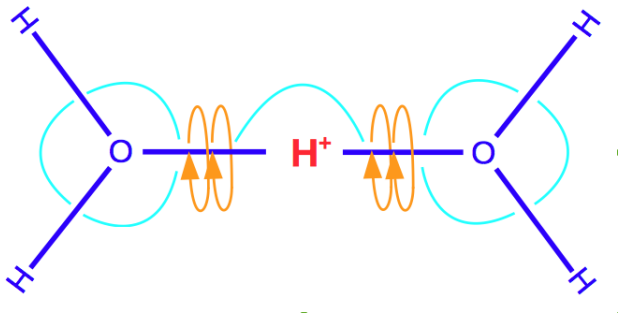

Figure A.3.: Zundel ion interactions considered in model 2. Morse bonds are depicted in red; harmonic bonds, in blue; harmonic angles, in cyan; and harmonic torsions, in orange.

We used the same policy as before for the initial guesses. The fitting resulted in the following parameters:

- $K_{b,\text{water}} = 526.46 \pm 9.90$
  $r_{0,b,\text{water}} = 0.98 \pm 0.0003$
- $K_{b,\text{proton}} = 19.98 \pm 0.86$
  $r_{0,b,\text{proton}} = 1.17 \pm 0.009$
- $K_{a,\text{ext}} = 38.61 \pm 1.23$
  $\theta_{0,a,\text{ext}} = 108.30 \pm 0.11$
- $K_{a,\text{int}} = 15.81 \pm 0.35$
  $\theta_{0,a,\text{int}} = 116.59 \pm 0.13$
- $K_{a,\text{cen}} = 10.71 \pm 1.10$
  $\theta_{0,a,\text{cen}} = 181.63 \pm 1.56$
- $K_t = 0.45 \pm 0.02$
  $n_t = 0.99 \pm 0.01$
- $c_o = -11.75 \pm 0.13$

And the following potential energy for the trajectory data:

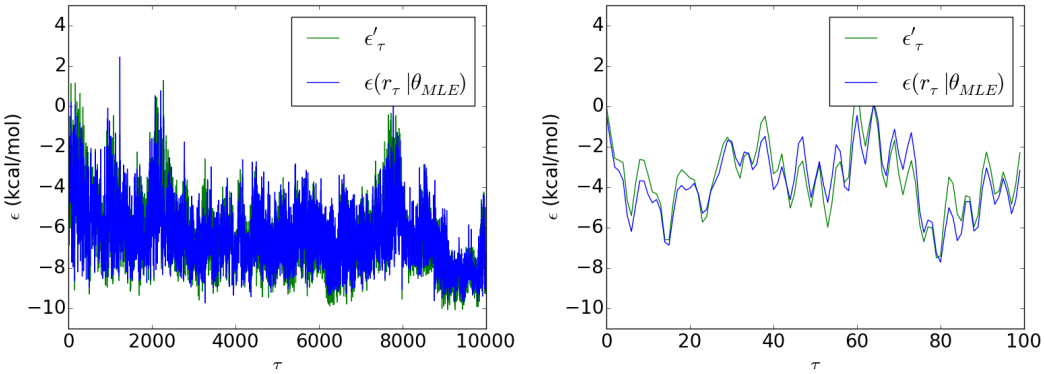

Figure A.4.: Left: Potential energies for the trajectory data given by the full QM run (green) and by the model 2 force field (blue); Right: first 100 trajectory steps of the left plot.

The RMSE of this calibration was 0.75 kcal/mol. It is safe to affirm that the change from Morse to harmonic bonds did not cause any harmful effects with respect to the energy fitting. However, when looking at the average structure generated by a full MM run with these parameters, we notice that the O-O distance is around 0.5 Å longer than expected, still not correct.

A.1.5. Model 3: extra O-O harmonic bond

The difficulty in capturing the correct O-O distance invited us to introduce an explicit bond between these two atoms. Even if this bond is not strictly physical, there are many force fields that take advantage of this sort of virtual interactions to better mimic a full QM behaviour. Then, on top of the previous model, we introduced this extra harmonic bond between the two oxygens:

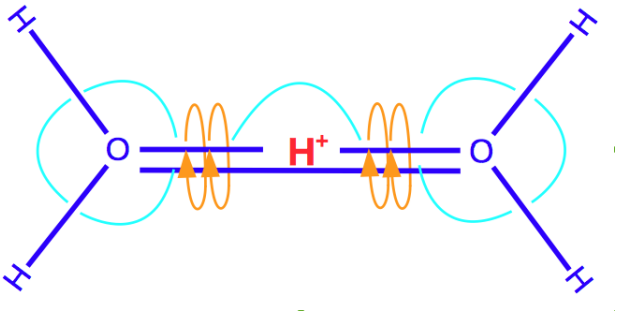

Figure A.5.: Zundel ion interactions considered in model 3. Morse bonds are depicted in red; harmonic bonds, in blue; harmonic angles, in cyan; and harmonic torsions, in orange.

Keeping the same policy for initial guesses, the fitting results in:

- $K_{b,\text{water}} = 513.82 \pm 8.70$
  $r_{0,b,\text{water}} = 0.98 \pm 0.0002$
- $K_{b,\text{proton}} = 8.72 \pm 0.87$
  $r_{0,b,\text{proton}} = 1.10 \pm 2.31$
- $K_{b,\text{oxygen}} = 70.68 \pm 2.66$
  $r_{0,b,\text{oxygen}} = 2.44 \pm 0.29$
- $K_{a,\text{ext}} = 38.29 \pm 1.08$
  $\theta_{0,a,\text{ext}} = 108.30 \pm 0.10$
- $K_{a,\text{int}} = 15.81 \pm 0.31$
  $\theta_{0,a,\text{int}} = 116.38 \pm 0.11$
- $K_{a,\text{cen}} = 10.53 \pm 12.56$
  $\theta_{0,a,\text{cen}} = 182.19 \pm 3.12$
- $K_t = 0.45 \pm 0.02$
  $n_t = 1.00 \pm 0.01$
- $c_o = -11.98 \pm 9.16$

The potential energy for the trajectory data is:

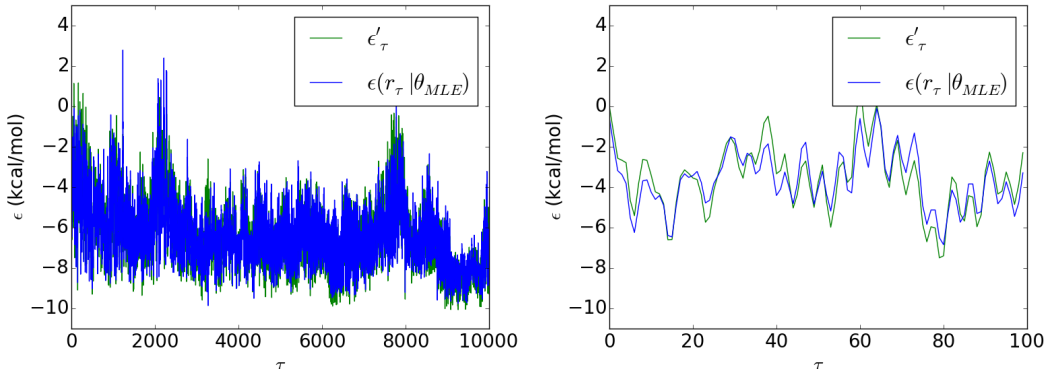

Figure A.6.: Left: Potential energies for the trajectory data given by the full QM run (green) and by the model 3 force field (blue); Right: first 100 trajectory steps of the left plot.

The RMSE of this set up, 0.66 kcal/mol, is significantly better than the ones of previous trials. There is a side effect of the extra O-O bond: an increased variance for the O-H^+ bond equilibrium length and for the central angle harmonic constant, as well as for the constant offset. We assume this comes from the redundancy that is introduced for such interactions. In spite of this, the inclusion of the O-O bond appears to pay off. The general dynamics and structure of the zundel ion are correct, except for one detail. In the QM run, the hydrogens always point outwards of the zundel ion during their rotations. However, in the MM runs reproduced with this force field, the hydrogens occasionally deviate from this orientation. This may hint the need for electrostatic repulsion. Typically, such unbonded interactions are neglected between atoms that belong to the same molecule, specially small ones like the zundel ion. However, it is possible to include some degree of interaction, for example LAMMPS allows to include weighted interactions with first, second or third neighbours by means of the `special_bonds` command [8].

A.1.6. Model 4: weighted H-to-H electrostatic interactions

We maintain the same force field for all bonded parameters. We add an electrostatic interaction only between the hydrogens on opposite sides of the zundel. Namely, they only interact if they are not bonded to the same oxygen atom. The charges for the atoms are extracted from the full QM simulation, while the dielectric constant is $\epsilon = 1$ and the Coulomb constant in adequate units is $C = 332.06 \frac{\text{kcal Å}}{\text{mol } e^2}$. This interaction will be weighted by a parameter w_{Cou}, which will be added to our fitting.

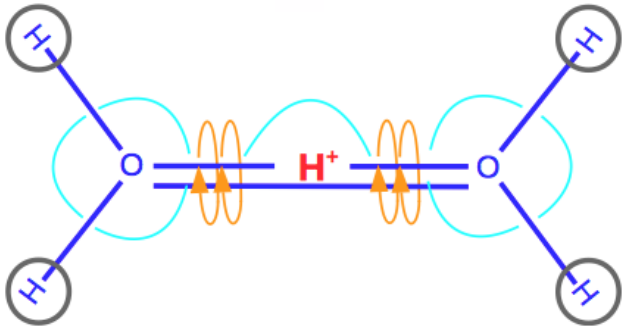

Figure A.7.: Zundel ion interactions considered in model 4. Morse bonds are depicted in red; harmonic bonds, in blue; harmonic angles, in cyan; harmonic torsions, in orange; and electrostatics, in grey.

The initial guess for the new parameter w_{Cou} will be zero (no interaction) and the value will be bounded in the [0,1] interval. The rest of the initial guesses are kept as in the previous model. The fitting resulted in:

- $K_{b,\text{water}} = 512.17 \pm 8.09$
  $r_{0,b,\text{water}} = 0.98 \pm 0.0003$
- $K_{b,\text{proton}} = 8.20 \pm 0.81$
  $r_{0,b,\text{proton}} = 1.10 \pm 2.29$
- $K_{b,\text{oxygen}} = 71.04 \pm 2.47$
  $r_{0,b,\text{oxygen}} = 2.37 \pm 0.27$
- $K_{a,\text{ext}} = 38.44 \pm 1.00$
  $\theta_{0,a,\text{ext}} = 108.53 \pm 0.09$
- $K_{a,\text{int}} = 13.77 \pm 0.30$
  $\theta_{0,a,\text{int}} = 109.74 \pm 0.43$
- $K_{a,\text{cen}} = 11.09 \pm 11.68$
  $\theta_{0,a,\text{cen}} = 181.64 \pm 2.23$
- $K_t = 0.10 \pm 0.03$
  $n_t = 1.03 \pm 0.04$
- $w_{\text{Cou}} = 0.42 \pm 0.02$

- $c_o = -45.90 \pm 6.34$

Also, we obtained the potential energy for the trajectory data:

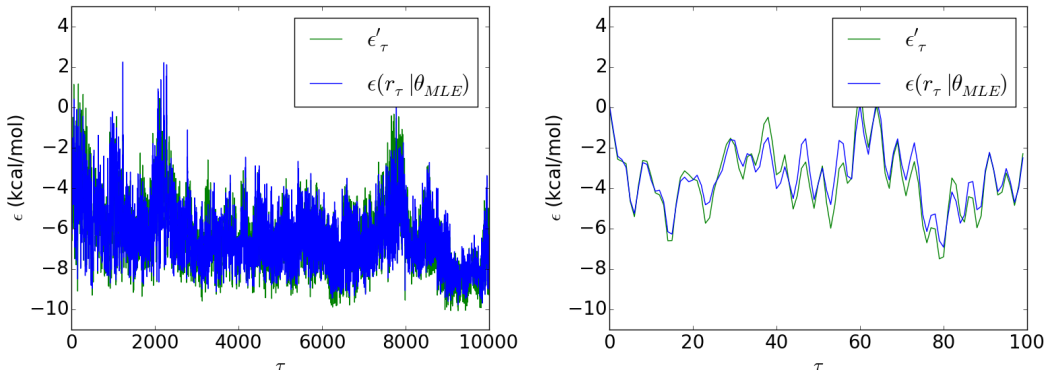

Figure A.8.: Left: Potential energies for the trajectory data given by the full QM run (green) and by the model 4 force field (blue); Right: first 100 trajectory steps of the left plot.

The RMSE of this fitting was 0.61 kcal/mol, the best one among our four calibrated models so far. We appreciate that introducing the electrostatic interactions between the hydrogens diminishes the torsion energy contribution. This hints a complementary mechanism for the hydrogen movement. We decide to neglect this interaction from now on. In general, the structure and dynamics generated with this force field fairly resemble those of the full QM run. Additionally, we see a better fitting for potential energy peaks that were not well captured with the previous models.

A.1.7. Model 5: Morse parameters for the water-like bond

During equilibration runs as described in Chapter 5, it was observed that the water-like harmonic bonds of the Zundel ion were very sensitive. Oscillations appeared whenever this bond was perturbed during the transfer. This motivated a change of this bond for a Morse type. As mentioned above, the harmonic torsion is omitted from here on, as its constant was heavily reduced in the previous model.

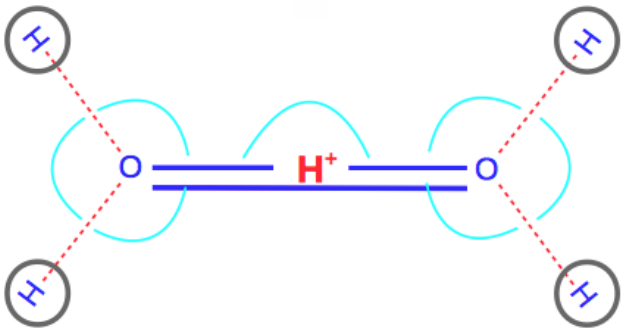

Figure A.9.: Zundel ion interactions considered in model 5. Morse bonds are depicted in red; harmonic bonds, in blue; harmonic angles, in cyan; harmonic torsions, in orange; and electrostatics, in grey.

The initial guess for the newly introduced Morse parameters is taken as in model 1. The fitting results in:

- $D_{m,\text{water}} = 80.9 \pm 24.27$
  $\alpha_{m,\text{water}} = 2.59 \pm 0.40$
  $r_{0,m,\text{water}} = 0.98 \pm 0.0004$
- $K_{b,\text{proton}} = 8.33 \pm 0.81$
  $r_{0,b,\text{proton}} = 1.10 \pm 2.24$
- $K_{b,\text{oxygen}} = 70.81 \pm 2.46$
  $r_{0,b,\text{oxygen}} = 2.36 \pm 0.27$
- $K_{a,\text{ext}} = 38.50 \pm 1.00$
  $\theta_{0,a,\text{ext}} = 108.54 \pm 0.09$
- $K_{a,\text{int}} = 13.61 \pm 0.30$
  $\theta_{0,a,\text{int}} = 108.81 \pm 0.37$
- $K_{a,\text{cen}} = 11.25 \pm 11.60$
  $\theta_{0,a,\text{cen}} = 181.21 \pm 1.80$
- $w_{\text{Cou}} = 0.47 \pm 0.02$
- $c_o = -50.13 \pm 5.95$

Also, we obtained the potential energy for the trajectory data:

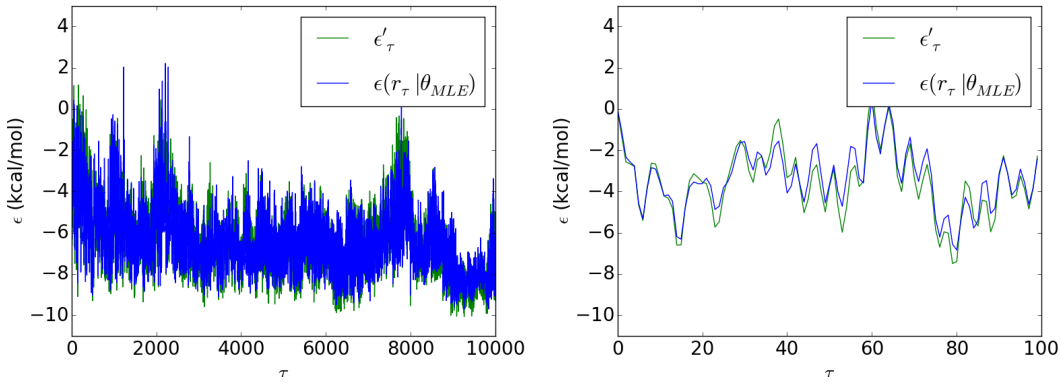

Figure A.10.: Left: Potential energies for the trajectory data given by the full QM run (green) and by the model 5 force field (blue); Right: first 100 trajectory steps of the left plot.

The RMSE of this fitting was 0.61 kcal/mol, same as model 4. The Morse parameters for the O-H bond show a reasonable degree of uncertainty. It should be remembered that the Morse potential allows for larger deviations from the equilibrium bond distance. A compromise is taken between the noise level of the energy gap and the structural stability of the Zundel ion. The following last two models are motivated by the smooth energy gap evolutions provided by Morse bonds.

A.1.8. Model 6: Morse parameters for the water-like and the proton bonds

We set up Zundel ion interactions with Morse types for both the O-H and the O-H^+ bonds.

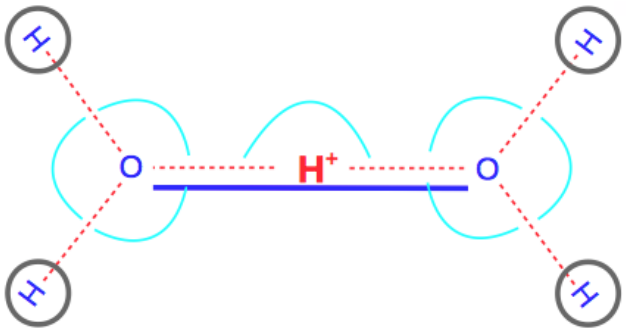

Figure A.11.: Zundel ion interactions considered in model 6. Morse bonds are depicted in red; harmonic bonds, in blue; harmonic angles, in cyan; harmonic torsions, in orange; and electrostatics, in grey.

The initial guess for the newly introduced Morse parameters is taken as in models 1 and 5. The fitting results in:

- $D_{m,\text{water}} = 81.1 \pm 22.65$
  $\alpha_{m,\text{water}} = 2.59 \pm 0.37$
  $r_{0,m,\text{water}} = 0.98 \pm 0.0004$
- $D_{m,\text{proton}} = 4.39 \pm 1.59$
  $\alpha_{m,\text{proton}} = 5.27 \pm 0.56$
  $r_{0,m,\text{proton}} = 1.10 \pm 0.01$
- $K_{b,\text{oxygen}} = 86.1 \pm 2.72$
  $r_{0,b,\text{oxygen}} = 2.42 \pm 0.02$
- $K_{a,\text{ext}} = 39.02 \pm 0.94$
  $\theta_{0,a,\text{ext}} = 108.58 \pm 0.08$
- $K_{a,\text{int}} = 13.27 \pm 0.28$
  $\theta_{0,a,\text{int}} = 108.75 \pm 0.36$
- $K_{a,\text{cen}} = 8.92 \pm 1.25$
  $\theta_{0,a,\text{cen}} = 181.68 \pm 1.45$
- $w_{\text{Cou}} = 0.47 \pm 0.01$
- $c_o = -51.24 \pm 1.43$

Also, we obtained the potential energy for the trajectory data:

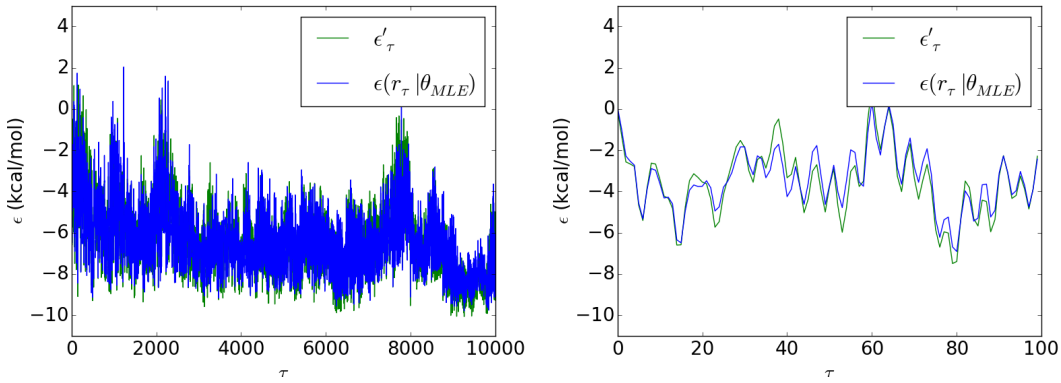

Figure A.12.: Left: Potential energies for the trajectory data given by the full QM run (green) and by the model 6 force field (blue); Right: first 100 trajectory steps of the left plot.

The RMSE of this fitting was 0.57 kcal/mol, improving that of models 4 and 5. Its noticeable that the O-O harmonic bond absorbs part of the O-H^+ interaction. The uncertainty for the central angle also improves. The certainty for the Morse parameters is fairly acceptable. Still, it should be kept in mind that Morse bonds penalize the deviation from the equilibrium distance less that harmonic bonds. Then, the atoms are allowed to drift further away.

A.1.9. Model 7: Morse parameters for all bonds

We set a Zundel ion with every bond set as a Morse type.

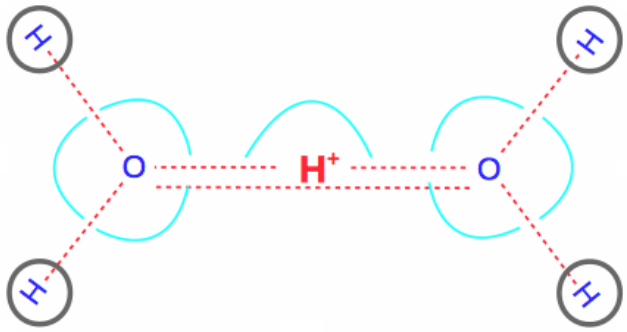

Figure A.13.: Zundel ion interactions considered in model 7. Morse bonds are depicted in red; harmonic bonds, in blue; harmonic angles, in cyan; harmonic torsions, in orange; and electrostatics, in grey.

The initial guess for the newly introduced Morse D_m and α_m parameters is taken as in models 1, 5 and 6. The fitting resulted in:

- $D_{m,\text{water}} = 84.2 \pm 23.44$
  $\alpha_{m,\text{water}} = 2.54 \pm 0.36$
  $r_{0,m,\text{water}} = 0.98 \pm 0.0004$
- $D_{m,\text{proton}} = 4.28 \pm 2.02$
  $\alpha_{m,\text{proton}} = 5.0 \pm 0.67$
  $r_{0,m,\text{proton}} = 1.10 \pm 0.02$
- $D_{m,\text{oxygen}} = 23.3 \pm 3.83$
  $\alpha_{m,\text{oxygen}} = 2.24 \pm 0.21$
  $r_{0,m,\text{oxygen}} = 2.41 \pm 0.02$
- $K_{a,\text{ext}} = 39.01 \pm 0.91$
  $\theta_{0,a,\text{ext}} = 108.62 \pm 0.08$
- $K_{a,\text{int}} = 13.41 \pm 0.28$
  $\theta_{0,a,\text{int}} = 108.89 \pm 0.34$
- $K_{a,\text{cen}} = 9.19 \pm 1.40$
  $\theta_{0,a,\text{cen}} = 181.56 \pm 1.37$
- $w_{\text{Cou}} = 0.47 \pm 0.01$
- $c_o = -50.97 \pm 1.45$

Also, we obtained the potential energy for the trajectory data:

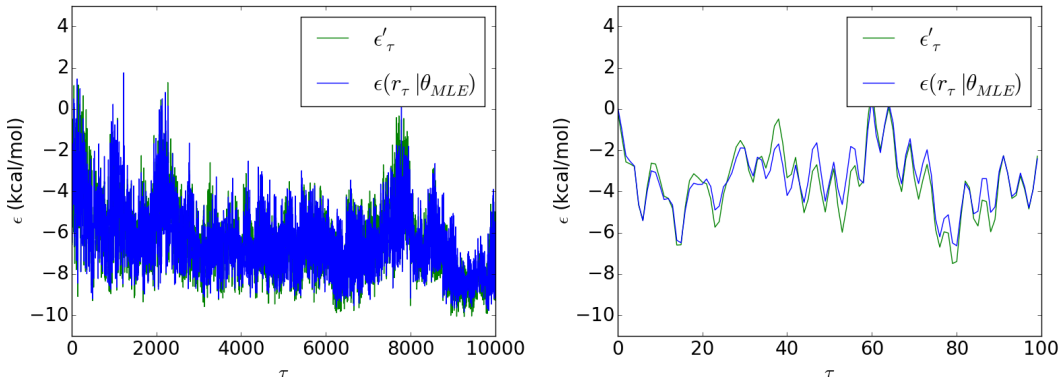

Figure A.14.: Left: Potential energies for the trajectory data given by the full QM run (green) and by the model 7 force field (blue); Right: first 100 trajectory steps of the left plot.

The RMSE of this fitting was 0.56 kcal/mol, slightly better than model 6 and the best one among all the calibrated models. The structural compromise of the Morse bonds, however, should still be considered.

A.2. Laplace approximation vs. MCMC sampling

In the previous section, seven models were calibrated by the means of MLE. As shown at the end of the MLE's formulation, the single-point estimate can be connected to a Gaussian distribution. However, the question of how to test the validity of this approximation remains. In particular, we wish to compare this Gaussian distribution to the actual shape of the posterior. Markov chain Monte Carlo methods are powerful tools to sample the posterior. In this section, we will perform MCMC sampling - in the form of a Metropolis algorithm - on the simplest of our models (model 2) and compare the obtained distribution with the one of the MLE.

First, we provide a quick reminder in the form of a Metropolis algorithm pseudo-code (Algorithm 2). Then, attending the specifics of our MCMC algorithm, we will explain the calculation of our likelihood, prior and posterior probabilities, as well as the definition of our first guess and the random steps for proposals.

Algorithm 2 MCMC Metropolis sampling of a log-posterior probability distribution

```
Initialize:
    Assign current_parameters = initial_guess
    Calculate current_loglikelihood
    Calculate current_logprior
    Calculate current_logposterior = ...
                        current_loglikelihood + current_logprior
    Define random_step
for 0 to chain_length do:
    Generate proposal_parameters = ...
                        current_parameters + random_step
    Calculate proposal_loglikelihood
    Calculate proposal_logprior
    Calculate proposal_logposterior = ...
                        proposal_loglikelihood + proposal_logprior
    Calculate acceptance_probability = ...
                        exp(proposal_logposterior- current_logposterior)
    if random_number[0,1] < acceptance_probability then:
                                                ▷ always true if acceptance > 1
        Increment accepts = accepts + 1
        Assign current_parameters= proposal_parameters
        Assign current_loglikelihood = proposal_loglikelihood
        Assign current_logprior = proposal_logprior
        Assign current_logposterior = proposal_logposterior
    end if
    Append current_parameters to sampled_parameters
end for
return normalized_histogram(sampled_parameters)
return acceptance_ratio = accepts / chain_length
```

The likelihood is calculated based on Equation A.5, with an additional consideration. In order to avoid ill-conditioned operations associated with multiplying small numbers, we take the logarithm of the likelihood:

$$\log p(traj \mid \theta, ff) = -\frac{1}{2}\sum_{\tau}^{N}\left(\frac{\epsilon'_\tau - \epsilon(r_\tau \mid \theta, ff)}{\sigma_\tau}\right)^2 - \frac{N}{2}\log 2\pi\sigma_\tau^2 \tag{A.9}$$

In order to be consistent with the MLE methodology, we consider $\sigma_\tau = 1$, as does the `curve_fit` function by default.

The prior is calculated as uniform in the vicinity of $\pm 3\sigma_\theta$ around the center of the MLE Gaussian distribution. Outside that interval it is considered as zero. In this way, we will be able to compare the shape of the posterior at the location where the MLE assigns the maximum probability. For consistency with the log-likelihood, the prior is also taken with a logarithm. Moreover, when calculating the posterior, we use a sum rather than a multiplication, as it is also a logarithm. In the same way, the acceptance ratio is calculated with a subtraction and then recovered with an exponential function.

The initial guess for the Metropolis algorithm is taken as the center of the MLE Gaussian distribution. With respect to the steps for our proposal, we consider a vector of random perturbations for the parameters. The perturbations are random numbers taken from Gaussian distributions, which satisfy the symmetric reversibility requirement of the Metropolis algorithm. After several trials, we established an optimal set of standard deviations for the steps. These are equal to $3\sigma_\theta/10$. Qualitatively, this means that our steps have a standard deviation of one tenth of the $3\sigma_\theta$ range that the prior allows us to explore in each direction.

The algorithm returned the following posterior probability shape near the MLE after one million steps with an acceptance ratio of 23.5%. Again, units are as described in the beginning of the Appendix:

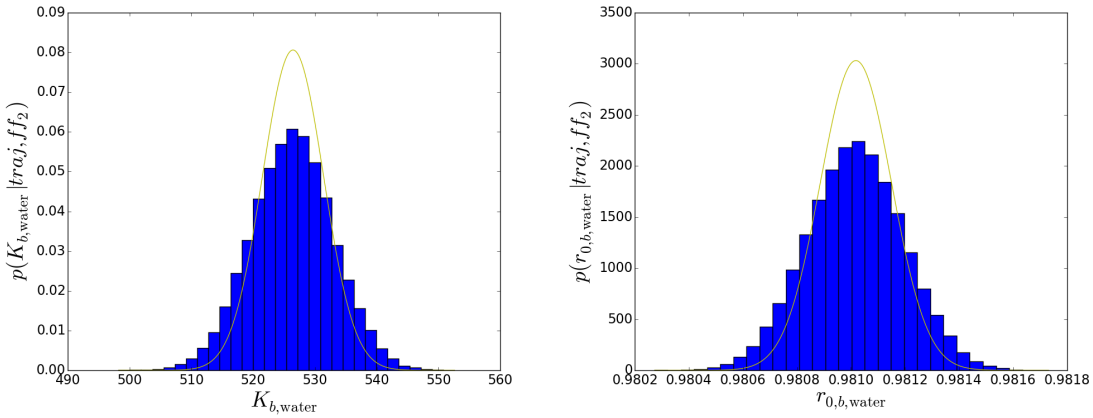

Figure A.15.: Posterior probabilities for model 2 harmonic O-H bond parameters sampled with a Metropolis algorithm (blue bars) compared to the Gaussian probability derived from the MLE using the Laplace approximation (yellow line).

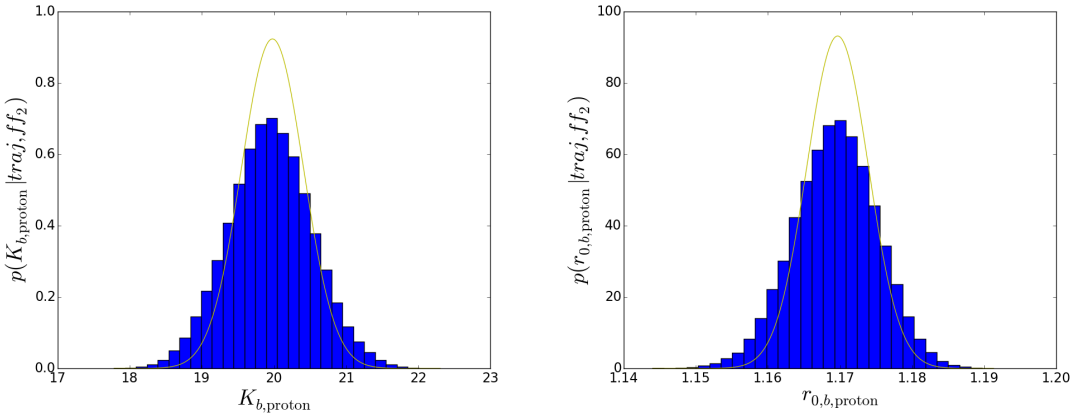

Figure A.16.: Posterior probabilities for model 2 harmonic O-H$^+$ bond parameters sampled with a Metropolis algorithm (blue bars) compared to the Gaussian probability derived from the MLE using the Laplace approximation (yellow line).

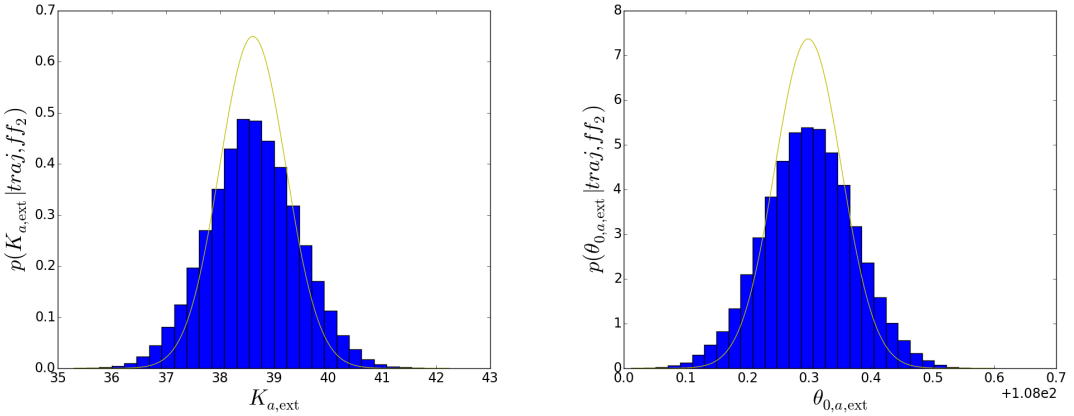

Figure A.17.: Posterior probabilities for model 2 harmonic external angle parameters sampled with a Metropolis algorithm (blue bars) compared to the Gaussian probability derived from the MLE using the Laplace approximation (yellow line).

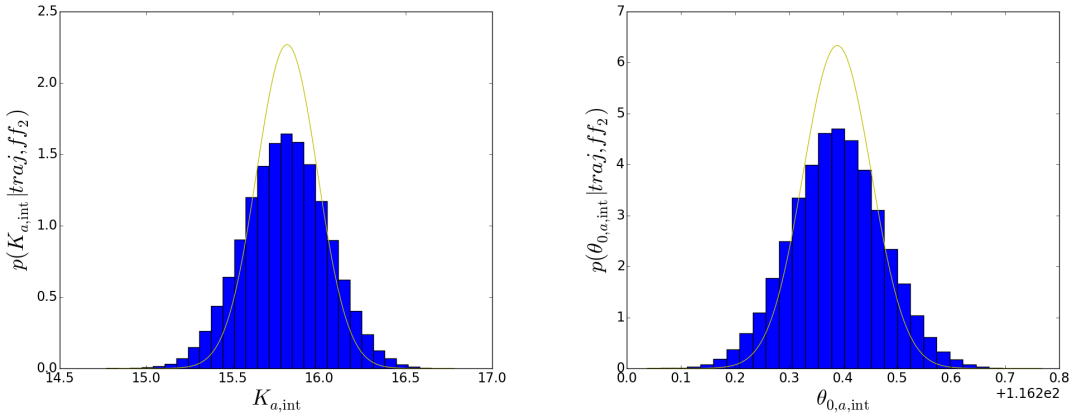

Figure A.18.: Posterior probabilities for model 2 harmonic internal angle parameters sampled with a Metropolis algorithm (blue bars) compared to the Gaussian probability derived from the MLE using the Laplace approximation (yellow line).

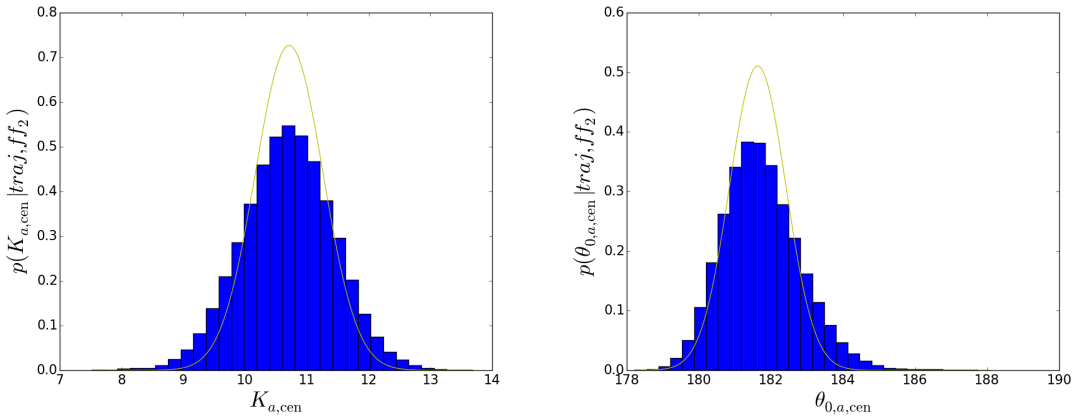

Figure A.19.: Posterior probabilities for model 2 harmonic central angle parameters sampled with a Metropolis algorithm (blue bars) compared to the Gaussian probability derived from the MLE using the Laplace approximation (yellow line).

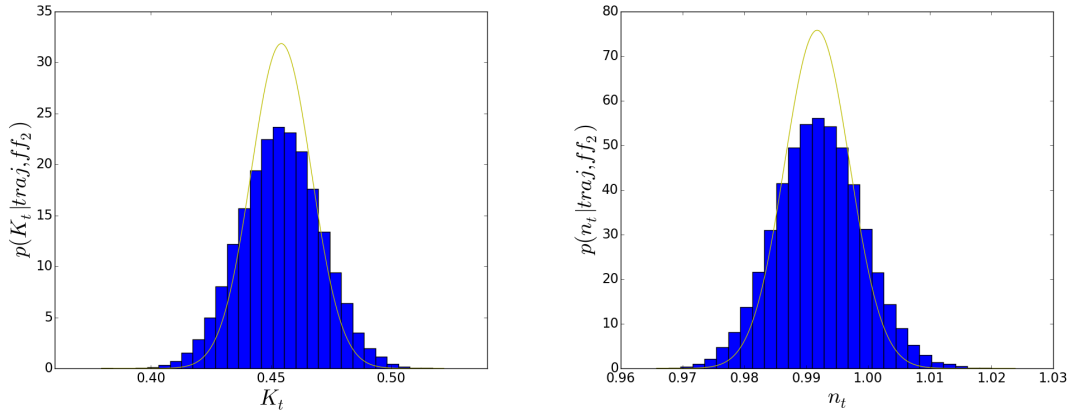

Figure A.20.: Posterior probabilities for model 2 harmonic torsion parameters sampled with a Metropolis algorithm (blue bars) compared to the Gaussian probability derived from the MLE using the Laplace approximation (yellow line).

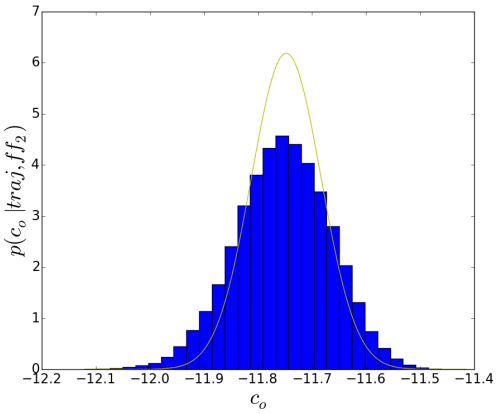

Figure A.21.: Posterior probabilities for model 2 constant offset parameter sampled with a Metropolis algorithm (blue bars) compared to the Gaussian probability derived from the MLE using the Laplace approximation (yellow line).

The Laplace approximated posterior and the MCMC sampling show a fair degree of resemblance. All of the parameters are very closely matched. The is some overestimation of the probability peak, but the accordance between the MLE and the MAP is remarkable. This result encourages us to base our model comparison on the Laplace approximated posterior probabilities.

A.3. Bayesian model comparison

A.3.1. Formulation

As mentioned during the formulation of the posterior probability for our Bayesian strategy (Equation A.1), the model evidence or normalization term is neglected during calibration. However, it is essential when performing model comparison. In the following paragraphs, we will show how to perform calculations to compare our calibrated models.

We begin by expressing a new application of the Bayes theorem, this time for the posterior probability of a given force field model to generate our trajectory data:

$$p(ff \mid traj) = \frac{p(traj \mid ff)p(ff)}{p(traj)} \tag{A.10}$$

Then, we are interested in the ratio of two of these posterior probabilities, one for model a and other for model b.

$$\frac{p(ff_a \mid traj)}{p(ff_b \mid traj)} = \frac{p(traj \mid ff_a)p(ff_a)}{p(traj \mid ff_b)p(ff_b)} \tag{A.11}$$

By the law of total probability, we can rewrite part of this ratio as the ratio between the model evidence terms of the two models. This new quantity is referred to as the Bayes Factor, while the remaining term is called the model prior:

$$\begin{aligned} \frac{p(ff_a \mid traj)}{p(ff_b \mid traj)} &= \underbrace{\frac{\int p(traj \mid \theta, ff_a)p(\theta \mid ff_a)d\theta}{\int p(traj \mid \theta, ff_b)p(\theta \mid ff_b)d\theta}}_{\text{Bayes factor}} \underbrace{\frac{p(ff_a)}{p(ff_b)}}_{\text{model prior}} \\ &= \underbrace{\frac{p(traj \mid ff_a)}{p(traj \mid ff_b)}}_{\text{Bayes factor}} \underbrace{\frac{p(ff_a)}{p(ff_b)}}_{\text{model prior}} \end{aligned} \tag{A.12}$$

In the Bayesian model comparison context, the Bayes factor indicates how much more likely is model a than model b to be the adequate for the trajectory data. The model prior provides a preconceived idea of which model is the adequate one. In the next subsections, we will obtain the terms to compare our four calibrated models. We will assume no preferred model in the model prior.

Our model comparison relies on the simplicity of our Laplace-like approximation. As explained during the Formulation subsection of the model calibration, we assume that the posterior is similar to the Gaussian distribution associated with the MLE. The normalization constant, or model evidence term, for such posterior approximation is [12, 20]:

$$p(traj \mid ff) = p(\theta_0 \mid traj, ff)\sqrt{(2\pi)^K \det(\Sigma_\theta)} \tag{A.13}$$

In order to avoid floating point errors, we take the logarithm of this expression. The

log-posterior probability is expanded in log-likelihood (Equation 12) and log-prior.

$$\log p(traj \mid ff) = \underbrace{-\frac{1}{2}\sum_{\tau}^{N}\left(\frac{\epsilon'_\tau - \epsilon(r_\tau \mid \theta_0, ff)}{\sigma_\tau}\right)^2 - \frac{N}{2}\log 2\pi\sigma_\tau^2}_{\text{log-likelihood}} + \underbrace{\log p(\theta_0)}_{\text{log-prior}} + \underbrace{\frac{K}{2}\log 2\pi + \frac{1}{2}\log\det(\Sigma_\theta)}_{\text{log normalization constant}} \tag{A.14}$$

In this equation, the likelihood maximizes the probability according to the precision of the model, while the covariance term favours broader models. In the same fashion, the prior penalizes complexity, while the first them of the normalization favours it. The balance between these contributions builds the Occam's Razor concept, as explained in the literature [20].

For simplicity and numerical stability, we can express the logarithm of the Bayes factor ratio simply as a difference $\log p(traj \mid ff_a) - \log p(traj \mid ff_b)$. Then, for a more compact representation, the models can be evaluated individually simply looking at the model evidence term. The greater the value, the better the model will result for our data when compared against the others.

A.3.2. Model 1

$$\log p(traj \mid ff_1) = \underbrace{-11992.42}_{\text{log-likelihood}} \quad \underbrace{-64.28}_{\text{log-prior}} \quad \underbrace{+12.86 - 37.52}_{\text{log norm. const.}} = -12081.36 \tag{A.15}$$

A.3.3. Model 2

$$\log p(traj \mid ff_2) = \underbrace{-11992.91}_{\text{log-likelihood}} \quad \underbrace{-65.64}_{\text{log-prior}} \quad \underbrace{+11.95 - 38.67}_{\text{log norm. const.}} = -12085.29 \tag{A.16}$$

A.3.4. Model 3

$$\log p(traj \mid ff_3) = \underbrace{-11350.52}_{\text{log-likelihood}} \quad \underbrace{-76.23}_{\text{log-prior}} \quad \underbrace{+13.78 - 41.14}_{\text{log norm. const.}} = -11454.11 \tag{A.17}$$

A.3.5. Model 4

$$\log p(traj \mid ff_4) = \underbrace{-11057.03}_{\text{log-likelihood}} \quad \underbrace{-76.23}_{\text{log-prior}} \quad \underbrace{+14.70 - 45.09}_{\text{log norm. const.}} = -11163.65 \tag{A.18}$$

A.3.6. Model 5

$$\log p(traj \mid ff_4) = \underbrace{-11036.78}_{\text{log-likelihood}} \quad \underbrace{-64.72}_{\text{log-prior}} \quad \underbrace{+13.78 - 40.53}_{\text{log norm. const.}} = -11128.25 \tag{A.19}$$

A.3.7. Model 6

$$\log p(traj \mid ff_4) = \underbrace{-10815.61}_{\text{log-likelihood}} \quad \underbrace{-63.36}_{\text{log-prior}} \quad \underbrace{+14.70 - 50.36}_{\text{log norm. const.}} = -10914.62 \tag{A.20}$$

A.3.8. Model 7

$$\log p(traj \mid ff_4) = \underbrace{-10738.27}_{\text{log-likelihood}} \quad \underbrace{-61.26}_{\text{log-prior}} \quad \underbrace{+15.62 - 54.18}_{\text{log norm. const.}} = -10838.09 \tag{A.21}$$

A.4. Outcomes

As expected from the RMSE values and the observed structure and dynamics, models 4 to 7 have the most favourable probabilities given our data. These probabilities are almost entirely determined by the likelihoods, as the differences in the complexity of the models are not enough to clearly penalize over-parametrization. However, the calibration of the parameters alone allows for parameter relevance detection. We observe the vanishing of the torsion interaction in model 4, as well as the close relationship between the O-H^+ and O-O bonds in model 3. Furthermore, the uncertainties associated with the individual parameter distributions, allow to detect the lack of adequate data for the Morse bond of model 1, as well as the effect of having redundant parameters or cooperative interactions.

In this thesis, models 4 and 5 are be actively employed. Model 4, because of its full harmonic bonds, presents greater structural stability. We take advantage of this to perform equilibration runs in water solvent. For production runs, model 5 is employed. Its Morse O-H bond returns smoother energy gap evolutions. Models 6 and 7 are not used, as we decide to maintain part of our original physical motivation. Of course, these force fields, and new ones, could also be employed, as we motivate in the proposed outlook of Chapter 7. Published research has shown that different energy gap parametrizations return valid and, when normalized, consistent free energy profiles [21].

At the termination of this Appendix we appreciate the power of probabilistic formulations. Bayesian model calibration and comparison are able to support and quantify our beliefs about the models. This provides valuable assistance and information when formulating the force fields. While doing a simple weighted non-linear squares fitting would have sufficed to provide a set of parameters, Bayesian strategies brought to our attention further knowledge about the Physics of the system, which ultimately led to more comprehensive results.

A.5. Summary

Parameters / Model	1	2	3	4	5	6	7
$K_{b,\text{water}}$	526.77	526.46	513.82	512.17	—	—	—
$r_{0,b,\text{water}}$	0.98	0.98	0.98	0.98	—	—	—
$D_{m,\text{water}}$	—	—	—	—	80.9	81.1	84.2
$\alpha_{m,\text{water}}$	—	—	—	—	2.59	2.59	2.54
$r_{0,m,\text{water}}$	—	—	—	—	0.98	0.98	0.98
$K_{b,\text{proton}}$	—	19.98	8.72	8.20	8.33	—	—
$r_{0,b,\text{proton}}$	—	1.17	1.10	1.10	1.10	—	—
$D_{m,\text{proton}}$	399.98	—	—	—	—	4.39	4.28
$\alpha_{m,\text{proton}}$	0.23	—	—	—	—	5.27	5.0
$r_{0,m,\text{proton}}$	1.17	—	—	—	—	1.10	1.10
$K_{b,\text{oxygen}}$	—	—	70.68	71.04	70.81	86.1	—
$r_{0,b,\text{oxygen}}$	—	—	2.44	2.37	2.36	2.42	—
$D_{m,\text{oxygen}}$	—	—	—	—	—	—	23.3
$\alpha_{m,\text{oxygen}}$	—	—	—	—	—	—	2.24
$r_{0,m,\text{oxygen}}$	—	—	—	—	—	—	2.41
$K_{a,\text{ext}}$	38.64	38.61	38.29	38.44	38.50	39.02	39.01
$\theta_{0,a,\text{ext}}$	108.30	108.30	108.30	108.53	108.54	108.58	108.62
$K_{a,\text{int}}$	15.81	15.81	15.81	13.77	13.61	13.27	13.41
$\theta_{0,a,\text{int}}$	116.60	116.59	116.38	109.74	108.81	108.75	108.89
$K_{a,\text{cen}}$	10.69	10.71	10.53	11.09	11.25	8.92	9.19
$\theta_{0,a,\text{cen}}$	181.65	181.63	182.19	181.64	181.21	181.68	181.56
K_t	0.45	0.45	0.45	—	—	—	—
n_t	0.99	0.99	1.00	—	—	—	—
w_{Cou}	—	—	—	0.42	0.47	0.47	0.47
c_o	−11.76	−11.75	−11.98	−45.90	−50.13	−51.24	−50.97
RMSE	0.75	0.75	0.66	0.61	0.61	0.57	0.56
log model evidence	−12081	−12085	−11454	−11164	−11128	−10915	−10838

Table A.1.: Parameters, RMSE and log model evidence terms for all calibrated models. Units are as described in the beginning of the Appendix.

Bibliography

[1] ASE-developers. ASE documentation. `https://wiki.fysik.dtu.dk/ase/`, 2016.

[2] S. R. Bahn and K. W. Jacobsen. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.*, 4:56–66, 2002.

[3] D. Berger. *Solid State QM/MM Embedding for a First-Principles Description of Catalytic Processes*. PhD thesis, Technische Universität München, 2014.

[4] D. Berger, A. Logsdail, H. Oberhofer, M. Farrow, R. Catlow, P. Sherwood, A. Sokol, V. Blum, and K. Reuter. Embedded-Cluster Calculations in a Numeric Atomic Orbital Density-Functional Theory Framework. *J. Chem. Phys.*, 141:024105, 2014.

[5] N. Bernstein, C. Várnai, I. Solt, S. A. Winfield, M. C. Payne, I. Simon, M. Fuxreiter, and G. Csányi. QM/MM simulation of liquid water with an adaptive quantum region. *Phys. Chem. Chem. Phys.*, 14:646–656, 2012.

[6] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals . *Comp. Phys. Comm.*, 180:2175–2196, 2009.

[7] R. Bulo, B. Ensing, J. Sikkema, and L. Visscher. Toward a Practical Method for Adaptive QM/MM Simulations. *J. Chem. Theory Comput.*, 5:2212–2221, 2009.

[8] Sandia Corporation. LAMMPS WWW Site. `http://lammps.sandia.gov`, 2016.

[9] C. J. Cramer. *Essentials of Computational Chemistry. Theories and Models*. John Wiley & Sons, Ltd, 2nd edition, 2004.

[10] FHI-aims team. *Fritz Haber Institute ab initio molecular simulations: FHI-aims. All-Electron Electronic Structure Theory with Numeric Atom-Centered Basis Functions*. Fritz-Haber-Institut der Max-Planck-Gesellschaft, 2015.

[11] D. Frenkel and B. Smit. *Understanding Molecular Simulation. From Algorithms to Applications*. Academic Press, 2002.

[12] Z. Ghahramani. Bayesian methods for machine learning. In *International Conference on Machine Learning Tutorial*, 2004.

[13] A. A. Hassanali, J. Cuny, V. Verdolino, and M. Parrinello. Aqueous solutions: state of the art in ab initio molecular dynamics. *Phil. Trans R. Soc. A*, 372:20120482, 2014.

[14] F. Jensen. *Introduction to Computational Chemistry*. John Wiley & Sons, Ltd, 2nd edition, 2007.

[15] A. Jones and B. Leimkuhler. Adaptive stochastic methods for sampling driven molecular systems. *J. Chem. Phys.*, 135:084125, 2011.

[16] J. Kästner and W. Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a nivel analysis method: "Umbrella Integration". *J. Chem. Phys.*, 123:144104, 2005.

[17] P. V. Komarov, P. G. Khalatur, and A. R. Khokhlov. Large-scale atomistic and quantum-mechanical simulations of a Nafion membrane: Morphology, proton solvation and charge transport. *Beilstein J. Nanotechnol.*, 4:567–587, 2013.

[18] F.S. Koutsourelakis. Bayesian Strategies for Inverse Problems. Course Material.

[19] Leslie Lamport. *LaTeX : A Documentation Preparation System User's Guide and Reference Manual*. Addison-Wesley Professional, 1994.

[20] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2005.

[21] L. Mones and G. Csányi. Topologically Invariant Reaction Coordinates for Simulating Multistate Chemical Reactions. *J. Phys. Chem. B*, 116:14876–14885, 2012.

[22] L. Mones, A. Jones, A. W. Goetz, T. Laino, R. C. Walker, B. Leimkuhler, G. Csányi, and N. Bernstein. The Adaptive Buffered Force QM/MM Method in the CP2K and AMBER Software Packages. *J. Comput. Chem.*, 36:633–648, 2015.

[23] Leibniz Super Computing Centre of the Bavarian Academy of Science and Humanities. SuperMUC Petascale System. `https://www.lrz.de/services/compute/supermuc/`, 2016.

[24] S. Plimpton. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comp. Phys.*, 117:1–19, 1995.

[25] C. Reed. Myths about the Proton. The Nature of H^+ in Condensed Media. *Acc. Chem. Res.*, 46(11):2567–2575, 2013.

[26] F. Rizzi, H. N. Najm, B. J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, and O. M. Knio. Uncertainty Quantification in MD Simulations. Part II: Bayesian Inference of Force-field Parameters. *Multiscale Model. Simul.*, 10(4):1460–1492, 2012.

[27] T. Stecher, N. Bernstein, and G. Csányi. Free Energy Sorface Reconstruction from Umbrella Samples Using Gaussian Process Regression. *J. Chem. Theory Comput.*, 10:4079–4097, 2014.

[28] G. M. Torrie and J. P. Valleau. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett.*, 28:578–581, 1974.

[29] C. Várnai, N. Bernstein, L. Mones, and G. Csányi. Test of an Adaptive QM/MM Calculation on Free Energy Profiles of Chemical Reactions in Solution. *J. Phys. Chem. B*, 117:12202–12211, 2013.

[30] Ren X, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler. Resolution-of-identity approach to Hartree-Fock, hybrid density functionals, RPA, MP2, and GW with numeric atom-centered orbital basis functions. *New J. Phys.*, 14:053020, 2012.