HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Development and Application of Hybrid Quantum Mechanical/Molecular Mechanical Methods with an Emphasis on the Implementation of a Fully Polarizable Model

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Eliot Boulanger
aus Liège

Mülheim an der Ruhr/Düsseldorf 2014

Aus dem Institut für Theoretische Chemie und Computerchemie der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Walter Thiel

Koreferent: Prof. Dr. Christel M. Marian

Tag der mündlichen Prüfung: 03/12/2014

Hiermit versichere ich, die hier vorgelegte Arbeit eigenständig und ohne unerlaubte Hilfe angefertigt zu haben. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner Institution eingereicht. Ich habe keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den

(Eliot Boulanger)

Acknowledgement

Foremost, I would like to express my sincere gratitude to my advisor Prof. Walter Thiel for the continuous support of my Ph.D. research, for his patience and his advice. I am grateful for the freedom and the trust he has given me, and for procuring an ideal research environment that allowed me to develop both professionally and personally.

I am grateful to Prof. Christel Marian for accepting to co-advise this thesis and for agreeing to review it. I thank every member of the jury for accepting to read this thesis and to allow its defense.

This thesis would not have been possible if it wasn't for the support of my loving wife, Amélie. I thank her from the bottom of my heart for allowing me to pursue the path of PhD studentship abroad, joining me in Germany and creating a new life with our son Basile. I would not exchange what we have created for anything in the world and never will.

I would like to thank all my colleagues, collaborators, and friends that have contributed to the success of this thesis in a professional and/or social manner. More specifically, I thank Claudia Loerbroks for coping with me in our office on a daily basis, for our fruitful exchanges and collaborations as well as for our moments of fun. I am grateful to Iakov Polyak for our joint projects and great coffee discussions, it was all fun. I thank Mario Barbatti for giving me the opportunity to join an exciting collaboration, for helpful advice during my whole PhD time, and for morning coffees. I thank all other group members for helping me in one way or another to get around this challenge called PhD thesis. Also, thanks to all my Belgian friends and family with whom I could keep in touch despite the distance and who continuously supported me in whatever I undertook.

My gratitude also goes to Alexander MacKerell and Pedro Lopes for allowing me to visit their group in Baltimore and for helping me with the implementation of the Drude oscillator force field into ChemShell.

List of papers included in this thesis

(1) Solvent boundary potentials for hybrid QM/MM computations using classical Drude oscillators: a fully polarizable model.

Eliot Boulanger and Walter Thiel, *J. Chem. Theory Comput.* **2012**, 8, 4527-4538.

Designed and implemented the method, ran and analyzed all computations, and wrote the draft of the paper.

(2) Quantum mechanics/molecular mechanics dual Hamiltonian free energy perturbation.

Iakov Polyak, Tobias Benighaus, Eliot Boulanger, and Walter Thiel*, J. Chem. Phys.* **2013**, 139, 064105.

Participated in analyzing the results, in designing the tests, and in writing the draft of the paper.

(3) Photochemical Steps in the Prebiotic Synthesis of Purine Precursors from HCN.

Eliot Boulanger, Anakuthil Anoop, Dana Nachtigallova, Walter Thiel, and Mario Barbatti, *Angew. Chem. Int. Ed.* **2013**, 52, 8000-8003.

Ran and analyzed the QM/MM computations and participated in the discussion of the proposed mechanism.

(4) *A microiterative intrinsic reaction coordinate method for large QM/MM systems.*

Iakov Polyak, Eliot Boulanger, Kakali Sen, and Walter Thiel, *Phys. Chem. Chem. Phys.* **2013**, 15, 14188-14195.

Wrote part of the code and designed the test computations.

(5) *Towards QM/MM Simulation of Enzymatic Reactions with the Drude Oscillator Polarizable Force Field.*

Eliot Boulanger and Walter Thiel, *J. Chem. Theory Comput.* **2014**, 10, 1795-1809.

Implemented the method, ran and analyzed all computations, and wrote the draft of the paper.

à Basile

Summary

This thesis presents work on the hybrid quantum mechanical/molecular mechanical (QM/MM) method, both on development and applications. The main focus was on developing a polarizable embedding scheme using the Drude Oscillator (DO) polarizable force field as MM component, in combination with any QM method. An efficient procedure was implemented to obtain the proper polarization state of the QM and MM parts of the system simultaneously. Further improvements could be achieved by coupling this approach with solvent boundary potentials (BPs) making use of an implicit representation of the distant solvent environment through a polarizable dielectric continuum, which reduces the number of degrees of freedom substantially. The QM/MM-DO/BP implementation covers the generalized solvent boundary potential (GSBP) for molecular dynamics simulations and the solvated macromolecule boundary potential (SMBP) for geometry optimizations. These approaches account for long-range electrostatic interactions in a fully polarizable three-layer QM/MM-DO/BP framework.

Making use of our new code and the recently published polarizable version of the CHARMM force field for proteins, we performed the first QM/MM-DO study of enzymatic reactions with polarizable embedding. This involved resolving several technical issues, with regard to the convergence behavior in systems with many polarizable interacting MM atoms and the treatment of polarization at the QM/MM boundary when cutting a covalent bond. We validated the consistency of our QM/MM-DO model for several small test systems through comparisons with full QM results. The QM/MM-DO computations on the enzymatic reactions in chorismate mutase and *p*-hydroxybenzoate hydroxylase showed polarization effects on the potential energy barriers of the order of 5 to 20%.

We participated in the development of an intrinsic reaction coordinate (IRC) method capable of tackling large QM/MM systems by using a microiterative approach, in which the IRC treatment is applied to a subset of atoms and the remainder of the environment is relaxed by geometry optimization at every step. The method was shown to work well for suitably chosen IRC subsets. We also participated in the development of a QM/MM free energy method that combines efficient low-level sampling with infrequent high-level energy evaluations in a Dual Hamiltonian Free Energy Perturbation (DH-FEP) approach, the merits of which were demonstrated both for small test systems and for two enzymatic reactions.

On the application side, we performed a standard QM/MM study on the Baeyer-Villiger reaction catalyzed by phenylacetone monooxygenase, with emphasis on the role of the active-site residues. We explored their possible configurations, identified the most relevant of these residues, and addressed the role of an extra water molecule in the active site. We also carried out a less conventional QM/MM application by computing the energy dissipation in aqueous solution of a hot ground state obtained after relaxation from an electronically excited state of a HCN tetramer, in the context of a theoretical study that aimed at establishing a photochemical pathway for the prebiotic synthesis of purines.

Zusammenfassung

Die vorliegende Doktorarbeit befasst sich mit der weiteren Entwicklung und mit Anwendungen der quantenmechanischen/molekülmechanischen (QM/MM) Methode. Der Schwerpunkt lag auf der Entwicklung eines polarisierbaren Einbettungs-Schemas unter Verwendung eines auf Drude Oszillatoren (DO) basierenden Kraftfeldes als MM Komponente, in Kombination mit beliebigen QM Methoden. Dabei wurde ein effizientes Verfahren implementiert, mit dem die Polarisation in den QM und MM Regionen gleichzeitig konvergiert werden kann. Weitere Verbesserungen konnten erreicht werden durch Kopplung dieses Ansatzes mit einem Lösungsmittel-Grenzpotential (boundary potential, BP), wobei die entfernte Solvens-Umgebung implizit durch ein polarisierbares Dielektrikum dargestellt wird, was die Zahl der Freiheitsgrade deutlich verringert. Die QM/MM-DO/BP Implementierung umfasst zwei verschiedene Versionen von Lösungsmittel-Grenzpotentialen, zum Einsatz in molekulardynamischen Simulationen (GSBP) und in Geometrieoptimierungen (SMBP). Diese Methoden erlauben die Einbeziehung von langreichweitigen elektrostatischen Wechselwirkungen im Rahmen eines vollständig polarisierbaren dreilagigen QM/MM-DO/BP Ansatzes.

Unter Verwendung des neuen Codes und der vor kurzem publizierten polarisierbaren CHARMM Kraftfelds für Proteine konnten wir die erste QM/MM-DO Studie von enzymatischen Reaktionen mit polisierbarer Einbettung durchführen. Dabei mussten zuerst einige technische Probleme gelöst werden, im Hinblick auf das Konvergenzverhalten in Systemen mit einer großen Zahl von polarisierbaren MM Atomen und der Behandlung der Polarisation an der QM/MM Grenze beim Schneiden kovalenter Bindungen. Die Konsistenz unseres QM/MM-DO Modells wurde an einigen kleinen Testsystemen durch Vergleiche mit QM Rechnungen validiert. QM/MM-DO Rechnungen an den enzymatischen Reaktionen in Chorismat-Mutase und *p*-Hydroxybenzoat- Hydroxylase ergaben Polarisationseffekte auf die Barrieren im Bereich von 5 bis 20%.

Weiterhin waren wir an der Implementierung von intrinsischen Reaktionskoordinaten (IRC) für große QM/MM Systeme beteiligt, unter Verwendung eines mikroiterativen Ansatzes, bei dem die IRC-Rechnung auf eine Untergruppe vom Atomen beschränkt ist und die Umgebung in jedem Schritt durch eine Geometrieoptimierung relaxiert wird. Für geeignet gewählte Untergruppen funktioniert diese IRC-Methode gut. Wir waren auch an der Entwicklung einer Methode zur QM/MM Berechnung freie Energien beteiligt, die ein effizientes „low-level" Sampling mit periodischen „high-level" Energieberechnungen kombiniert (Dual Hamiltonian Free Energy Perturbation, DH-FEP). Die Vorzüge der DH-FEP Methode konnten für kleine Testsysteme und für zwei enzymatische Reaktionen gezeigt werden.

Im Bereich der Anwendungen wurde eine Standard-QM/MM-Studie zur Baeyer-Villiger-Reaktion in Phenylaceton-Monooxygenase durchgeführt, mit einem Fokus auf der Rolle der Aminosäuren im aktiven Zentrum. Untersucht wurden deren mögliche Konfigurationen und Relevanz sowie die Funktion eines zusätzlichen Wassermoleküls. Eine weniger konventionelle QM/MM Anwendung betraf die Energiedissipation eines HCN Tetramers in Wasser im heißen Grundzustand nach Relaxation aus dem angeregten Zustand, im Rahmen einer theoretischen Studie zu möglichen Reaktionspfaden bei der prebiotischen Synthese von Purinen.

Table of Contents

1. Introduction

1.1 A Story of Scale

The concept of scale is inherent to the world of science. Ever since the scientific revolution in the late 19th century, the field has been subdivided into disciplines that study events happening at different scales. For instance, biology studies systems ranging from the nanometer scale to the kilometer scale, going from enzymology to ecology. At larger scale earth science takes over and then astrophysics. Chemistry, the subject of this thesis, takes place around the nanometer scale which corresponds to the size of a big molecule with the chemical bond lengths being about one order of magnitude smaller. Zooming in, one reaches the worlds of first quantum physics and then particle physics. Nowadays these separations have become obsolete, and interdisciplinary research takes a central role on the scientific scene. Indeed, a quick look at the list of recent Nobel laureates will convince anyone that there is a real tendency of escaping traditional scientific disciplines and moving towards hybrid approaches.

The work presented in this thesis follows this trend as it applies mathematical models and computational techniques to implement physical laws in order to simulate chemical reactions possibly relevant to biology. More precisely, we will focus on the implementation, improvement, and application of hybrid methods combining quantum mechanics (QM) and molecular mechanics (MM) approaches for simulating enzymatic reactions.[1]

In this introduction, we will first describe QM and MM methods separately, and then their combination in the QM/MM framework. The second section will address QM/MM method development. The main focus is on the implementation of polarizable embedding using the Drude oscillator polarizable force field and its combination with a solvent boundary potential, which results in a fully polarizable three-layer model.[2,3] Other investigated topics include the implementation of QM/MM intrinsic reaction coordinates[4] and the development of a QM/MM dual Hamiltonian free-energy method.[5] The third section will cover two examples of QM/MM applications, namely the formation of the Criegee intermediate in the enzymatic Baeyer-Villiger reaction catalyzed by phenylacetone monooxygenase (PAMO) and the energy dissipation in aqueous solution of the hot ground state of trans-2,3-diaminomaleonitrile (trans-DAMN) after relaxation from an electronically excited state.[6]

1.2 Resolution, Tools, Methods, and Size

To avoid misunderstandings later on, some notational aspects are addressed in this section. The aim is to clarify the meaning of terms used in this thesis (without claiming rigorous definitions).

When it comes to theory and simulation, the concept of "scale" implies another key aspect: the "resolution". For instance, to study a chemical reaction occurring at the molecular level, it is necessary to take into account the electronic redistribution that occurs at a smaller scale. The scale at which a given process takes place and the one necessary for its explanation can thus be different. Normally, the required resolution has a scale of at least one order of magnitude below that of the phenomenon being studied.

Before advancing further, it is important to describe the difference between what will be called in this thesis a tool and a method. We consider as "tools" numerical procedures used either to get static data such as an optimum geometry or to sample over dynamical data to extract properties such as the temperature. Typical tools are thus geometry optimization and molecular dynamics (MD) techniques. We consider as "methods" mathematical formulations based on physical laws that allow us to calculate potential energy surfaces and other physical properties. QM/MM is a hybrid method that combines QM and MM methods. We also use the term "level of theory" or simply "level" as synonym of method.

Another important notion is "size". Whether a molecule or a system is small, big, or huge strongly depends on the method used to describe it and the tools one is willing to apply. It thus depends on the property of interest and the type of explanation one wants to obtain. Molecules considered here as small may appear huge for someone computing high-resolution spectra. In the same way, large macromolecules studied in this thesis will have a negligible size for someone in the field of system biology. The notion of size can become rather ambiguous when considering hybrid methods such as QM/MM that can be subjected to all kinds of different tools. This is why we will, in the following sections, concentrate on providing an appreciation of what a method can do for some given resolution, scale, and tools, rather than discussing the underlying theoretical aspects in great detail.

1.3 Atomic Resolution: Molecular Mechanics

Molecular mechanics employs classical potentials to study chemical or biochemical systems.[7,8] It is based on Newtonian physics and aims at studying large systems such as solvated macromolecules with high efficiency. The method used to compute the potential energy of a system is called a force field. Normally, it uses atomic resolution and neglects any electronic degrees of freedom. Force fields are commonly applied in combination with molecular dynamics or Monte Carlo simulations, which are sampling tools designed to extract data out of long simulations using statistical techniques.

Force fields are purely empirical. Their classical nature does not allow for the description of electronic events such as chemical reactions. They are parameterized using experimental and/or theoretical data. They have mainly been developed by the biophysics community. A standard force field such as CHARMM,[9,10] the one mainly used in this thesis, considers every atom as a point object and includes two types of interactions: bonded and non-bonded. Bonded interactions are usually represented by harmonic potentials for bond stretching and bond angle deformations, and by periodic potentials for torsions, in order to achieve a realistic description of geometries, vibrational spectra, and dynamical properties. Non-bonded interactions consist of two parts, namely the electrostatic interactions between the point charges located at each atomic position (as a rough model of the electronic distribution) and the Lennard-Jones terms representing the attractive and repulsive van der Waals (vdW) interactions. In its simplest form, the potential energy (U) function for a CHARMM-like force field can be written as a sum of sums of these terms.

$$U = \sum_{bonds} k_r(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\varphi(1 + cos(n\varphi - \delta)) + \sum_{impropers} k_\omega(\omega - \omega_0)^2 + \sum_{nonbonded} \varepsilon_{ij}\left[\left(\frac{r_0}{r_{ij}}\right)^{12} - 2\left(\frac{r_0}{r_{ij}}\right)^{6}\right] + \sum_{nonbonded} \frac{q_i q_j}{r_{ij}}$$

In this equation, the force constants k are force field parameters, and the subscript 0 refers to a standard value that needs to be parameterized or is directly taken from literature. The letter r denotes distances (between atoms i and j), θ is an angle, φ is a dihedral angle, and ω is an

improper angle for an out-of-plane deformation. The first sum over non-bonded terms represents the Lennard-Jones interactions; the parameters ε_{ij} are related to the depth of the vdW energy minimum. The second sum describes the Coulomb electrostatic interactions between the point charges (q) at atoms i and j. Other terms can be added to increase the accuracy of a given force field, but they are generally seen as corrections and not as key components during force field parameterization.

1.4 Electronic Resolution: Quantum Mechanics

Working with atomic resolution is usually not enough to tackle chemical problems, which requires inclusion of electronic effects. For this purpose, quantum mechanics has to be used and the Schrödinger equation has to be solved.[11]

QM calculations can be very challenging, especially for large molecules with many electrons. A hierarchy of approximate QM methods has been developed. Depending on the target accuracy, the size of the considered system, and the tolerable computational effort, different methods can be applied. It is not the purpose of this thesis to cover the whole ensemble of available QM methods, nor to detail their development and derivation. Numerous books and reviews cover these aspects far more accurately than what can possibly be done here. Instead, we compile a general overview over the different classes of available QM methods, focusing on accuracy and computational efficiency as well as their usefulness in current QM/MM simulations of ground-state properties.

There are three main families of QM methods which have been used in this thesis to compute the properties of the electronic ground state of molecules. The so-called *ab initio* methods are entirely based on first principles. In increasing order of accuracy and complexity as well of computational effort, the most prominent ones are the Hartree-Fock method (HF), second-order Møller-Plesset perturbation theory (MP2),[12] and coupled cluster theory.[13] HF calculations are rarely used nowadays, because they have been largely superseded by density functional theory which generally provides higher accuracy at similar cost. MP2 calculations (normally in combination with medium-size basis sets and the resolution-of-identity approximation) offer a good compromise between accuracy and cost; MP2 is thus well suited for computing fairly accurate ground-state properties of a reasonably large systems, also in a QM/MM framework. Coupled cluster calculations with single, double, and perturbative triple excitations (CCSD(T)) and large basis sets are currently the "gold standard" of theoretical chemistry for ground states. They have been used in some QM/MM simulations, but they are generally too expensive for

regular use. It should be noted that there is ongoing work aimed at drastically reducing the cost of such computations, which should facilitate the use of CCSD(T) in future QM/MM studies.[14,15]

Nowadays, the most popular method when it comes to chemistry, and therefore for QM/MM, is density functional theory (DFT), which is based on a one-to-one relationship between the ground-state electronic density and the nuclear geometry of a molecule.[16,17] DFT can be seen, in an intuitive manner, as a clever way to separate what you can handle exactly in the electronic Hamiltonian from what you cannot (using a formulation in terms of the electron density). The unknown parts are lumped together in the exchange-correlation functional, which covers exchange and electron correlation effects (as well as kinetic energy corrections in the Kohn-Sham framework). The accuracy of DFT computations thus essentially depends on the chosen exchange-correlation functional. Standard choices, in order of increasing accuracy, involve the local density approximation (LDA), the generalized gradient approximation (GGA), and hybrid functionals (with partial inclusion of HF exchange). For chemical purposes, LDA functionals are generally not accurate enough, GGA functionals are known to give good geometries and reasonable energies, and hybrid functionals usually show the best performance with regard to energies (when used with reasonably large basis sets). Several other types of functional have been developed, but their description is beyond the scope of this thesis.

The third family of QM methods used in this thesis are semiempirical approaches based on the modified neglect of differential overlap and the use of a minimal valence basis.[18] They rely on a careful parameterization and are extremely fast. Their accuracy can also be quite good, but this has to be validated on a system-by-system basis, by comparing their results with those obtained from higher-level methods. If applicable, semiempirical methods are extremely useful for QM/MM simulations, especially for molecular dynamics and free-energy calculations.

1.5 Multiscale Modelling: Quantum Mechanics/Molecular Mechanics

Hybrid QM/MM approaches aim at combining the best of both worlds by performing accurate QM computations of the electronically relevant part of the system and simulating the rest using efficient MM methods.[1,19-24] They can therefore be used only if the properties of interest are well localized in the QM part of the system. This approach is particularly suitable for simulating enzymatic reactions, which take place rather locally (in the active site) in a structured environment (the enzyme) that is difficult to represent in an implicit manner. It was originally proposed by Warshel and Levitt in their seminal paper in 1976.[25] This work earned them,

together with Martin Karplus, the 2013 Nobel Prize in chemistry. In the following sections, we will describe some aspects relevant to this technique.[26-28]

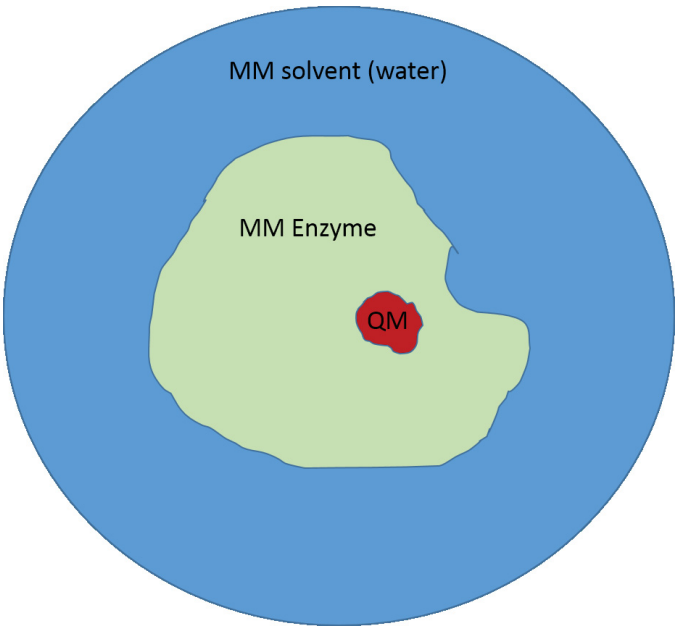


Figure 1: Partitioning of a typical QM/MM system for an enzymatic reaction. The QM region is part of the enzyme and is shown in red. The rest of the enzyme is defined at MM level and is shown here in green. The rest of the MM region, the blue sphere, is the solvent (generally water).

1.5.1 Energy computation scheme

Figure 1 represents the basic partitioning of the system for a QM/MM computation on an enzyme. The energy for a setup of this kind can be determined using two different strategies, from a subtractive or an additive scheme.[1]

In the subtractive scheme, one computes the energy of the whole system using the MM force field, then removes the part associated with the QM region and replaces it by the QM energy of the QM region. It is a simple interpolation scheme, which does not directly compute the QM/MM interactions but approximates them at the MM level. It is easy to implement, but often lacks accuracy. It can be especially problematic if there are no suitable force field parameters for the QM region of the system.

The additive scheme is generally more accurate. There are separate QM calculations for the active site (QM region) and MM calculations for the environment (MM region), and the QM/MM interactions are treated explicitly using a particular embedding scheme (see next subsection 1.5.2 below). Particular care is required when the QM/MM boundary cuts through

a covalent bond (see subsection 1.5.3). The total QM/MM energy is obtained as the sum of the QM energy, the MM energy, and the QM/MM interaction energy. The additive scheme is the only one used in this thesis.

1.5.2 Embedding

The embedding defines how the QM and MM regions are coupled through non-bonded interactions,[29] i.e., vdW and electrostatic terms. The simplest choice is mechanical embedding, which computes all these terms at the MM level (as in the subtractive scheme).

Nowadays, the standard technique in QM/MM studies is electronic embedding. It still treats the vdW interactions at the MM level, but determines the electrostatic QM/MM interactions by including the MM point charges into the QM computation (as additional terms in the one-electron part of the Hamiltonian). This allows the QM wavefunction to polarize under the influence of the MM point charges. The corresponding gradient contributions at the MM atoms can be easily computed by evaluating the electrostatic field due to the QM region at their positions. This technique has the advantage of explicitly computing the electrostatic QM/MM interaction energy at the QM level, while at the same time avoiding any in-depth parameterization of the force field for the QM region. Standard vdW parameters are normally sufficient for computing the vdW part of the QM/MM interaction energy.

A polarizable embedding is required when using a polarizable force field in the QM/MM framework. This creates complications since both the QM and MM regions are now polarizable, which calls for a self-consistent treatment. This is one of the main topics of this thesis and will thus not be covered here, but later in much detail (see section 2.1).

1.5.3 Covalent Bond Crossing at the QM/MM Boundary

When defining the QM and MM regions in a given system, it is often impossible not to have a covalent bond being cut at the QM/MM boundary.[1,30] This is generally the case for enzymes in which one or more active-site residues are involved in the reaction. The problems arising from such cuts have received much attention from several groups over the years, and several remedies have been suggested. There is still not a universal and accurate method to treat these issues, and hence such cuts need to be handled with care. Here we outline the main approaches to deal with the problem and list a few rules of thumb.

There are three main strategies that have been considered. The link atom schemes satisfy the valence at the cut by adding a hydrogen atom at the frontier QM atom. Other techniques use a specially parametrized boundary atom or a pseudopotential for this purpose. Finally, the valence can also be satisfied by introducing frozen orbitals that replace the cut bond.

In this thesis, we use the link atom approach in combination with a charge shift scheme, which transfers the charge on the frontier MM atom in the cut bond to its closest MM neighbor. Additionally, point charges are added to maintain the dipole moment of the MM group next to the QM/MM boundary (in order to minimize electrostatic perturbations). This scheme is not perfect, but has performed well in many previous QM/MM applications.

Admittedly, there will be inevitable errors when using the link atom approach with the charge shift scheme. To keep them as small as possible, one should try to follow a set of rules:

- The cut bond should be as distant as possible from the active part of the QM region.
- Atoms involved in a chemical reaction should not be close to the frontier bond being cut. They should be at least three bonds away.
- It is advisable to cut less polar bonds, for example a simple "C-C" bond.
- One should avoid cutting through a MM charge group since this can create a local artificial charge in the vicinity of the QM region.

1.5.4 Boundary Conditions and Long-range Interactions

Systems used in QM/MM calculations are finite in size. Compared to what is found in nature, this is an approximation. Traditionally, there are two ways to compensate: periodic boundary conditions[31] and solvent boundary potentials.[32]

Periodic boundary conditions are the method of choice for MM computations.[8] In this approach, the simulation box is repeated in all three directions periodically for an infinite number of times. A molecule leaving the system on one side will therefore appear again on the opposite side. The presence of the repeated images allows the inclusion of long-range electrostatic effects. The approach is only valid if the simulation box itself is large enough to represent its own environment accurately enough. It has been implemented at the QM/MM level by Laino et al.[33] I have done some initial (incomplete) work on a ChemShell implementation, which is not presented in this thesis. This work is continued by Tatiana Vasilevskaya.

Solvent boundary potentials (BP) make use of a polarizable dielectric continuum (PDC) outside of the atomistically defined QM/MM system. The applied PDC approach is similar to the continuum solvation models used in pure QM computations. QM/MM/BP treatments also capture long-range electrostatic effects. They may well be the method of choice for solvated enzymes as they work with a relatively small number of atoms (compared to periodic boundary conditions). In this thesis, we employ the generalized solvent boundary potential (GSBP) originally developed in the group of Benoit Roux[34,35] and the solvated macromolecule boundary potential (SMBP) developed in our group.[36,37] The GSBP is designed for highly efficient MD simulations; at the QM/MM level, it has only been coupled with semiempirical QM methods. SMBP is less efficient but can be used together with any QM method. It has been designed for geometry optimization. These two methods will be covered in detail in this thesis when we discuss their combination with the Drude oscillator polarizable force field (section 2.1).

2. Improving and Developing QM/MM Methods

The main purpose of this thesis was to further develop the QM/MM methodology to increase its accuracy and efficiency. The key improvement is the implementation of a polarizable embedding using the Drude oscillator force field, both in a standard QM/MM framework and in a three-layer scheme employing boundary potentials (section 2.1). Two other development projects cover tools used in QM/MM computations, namely the implementation of intrinsic reaction coordinates (IRC) for large QM/MM systems (section 2.2) and of a dual Hamiltonian approach to compute QM/MM free energies with a high-level QM method while sampling with a lower-level method (section 2.3).

2.1 Polarizable Embedding

Already at an early stage of QM/MM development in the 1990s, polarizable embedding was proposed as a more accurate strategy than standard electrostatic embedding.[25,29] Applications of this approach have been impeded for a long time by the lack of fully parametrized polarizable force fields,[38,39] which have become available only recently for proteins.[40,41]

In the following chapters, we will first describe the different kinds of polarizable force field. We will then focus on the Drude Oscillator (DO) model which was used in this thesis.[42] We continue with technical aspects addressing in some detail the combination of the QM/MM-DO methods with a solvent boundary potential to improve the computational efficiency while including long-range electrostatic effects.[3] Finally, we present pilot applications of the QM/MM-DO model to enzymatic reactions.[2]

2.1.1 Polarizable Force Fields

There are two main techniques to simulate electronic polarization at the MM level. One makes use of an explicit induced dipole description,[41,43-46] while the other applies a charge equilibration procedure.[47-49] In this subsection we will focus on the explicit induced dipole description, which underlies the MM method used in this thesis: the Drude Oscillator force field.[42,50-52]

In this approach, an induced dipole moment (μ_i) is added at all (or some) MM atoms *i* with position x_i. It is obtained from the electric field (*E*) at x_i and the polarizability (α_i) which is a parameter of the force field:

$$\mu_i = \alpha_i E(x_i)$$

The electric field at x_i can be split into a static part coming from permanent atomic point charges (E^0) and a dynamic part arising from the other induced dipoles of the system. Denoting by T_{ij} the interaction tensor elements between the induced dipoles μ_i and μ_j, the electric field can be written as:[39]

$$E(x_i) = E^0(x_i) - \sum_{j \neq i} T_{ij}\mu_i$$

In additive (non-polarizable) force fields, the non-bonded interactions between atoms involved in bonded interactions (bond, angle, and dihedral terms) are normally omitted. By contrast, it is necessary to include them in polarizable force fields in order to obtain the proper polarization state of the molecule. Given the fact that such atoms are close in space, a screening function needs to be introduced to avoid overpolarization. Such a function was initially proposed by Thole.[53,54] It contains parameters that need to be optimized when parameterizing the force field. Generally, the screening function can be included as a prefactor (γ_{ij}) that depends on the positions of atoms *i* and *j*. Therefore, the induced dipole moments can be obtained by solving a linear system of equations.[2,39]

$$(\boldsymbol{\alpha}^{-1}\boldsymbol{\gamma}\boldsymbol{T})\boldsymbol{\mu} = \boldsymbol{E}^0$$

From left to right, this equation contains a diagonal matrix with the inverse of the atomic polarizabilities, the screening function tensor (elements γ_{ij}), the dipole-dipole interaction tensor (elements T_{ij}), the vector with the Cartesian components of the induced dipoles, and the vector with the Cartesian components of the static electric field at each polarizable atom. The dimension of all these objects is 3*N*, with *N* being the number of polarizable atoms. This system of equations can be easily solved by standard linear algebra methods when *N* is reasonably small. However, force fields are designed to tackle very large systems, for which this straightforward approach becomes very costly (scaling with N^3), and one then normally resorts to iterative solvers to obtain the proper polarization of each center.

2.1.2 Drude Oscillator Force Field

A straightforward way to represent an induced dipole at an MM atom is to mimic it by two point charges of same magnitude but opposite sign (Figure 2). This strategy leads to the shell,[55-57] charge-on-spring,[58-62] and Drude oscillator (DO)[40,42,50-52,63,64] models. In all these models the two point charges (q) are linked by a spring. The first charge is fixed at the nucleus of the polarizable atom while the second one is mobile. Polarization arises from the competition between the forces acting on the mobile charge, which are due to the spring and the electrostatic interactions with the environment. The optimum position (d) of the mobile charge (Drude particle) is obtained when reaching equilibrium:

$$\frac{\partial(U_{spring} + U_{elec})}{\partial d} = 0$$

The potential energy U_{spring} of the harmonic spring is evaluated using a force constant (k_d) that is in the DO case defined in terms of the polarizability of the corresponding atom:

$$k_d = \frac{q^2}{\alpha}$$

In the chosen DO approach, k_d is always fixed to 1000 kcal mol^{-1}Å^{-2} to maintain a small d value (to keep the point-dipole approximation valid) and to avoid introducing additional parameters. To take DOs into account, the electrostatic part of the MM potential function has to be extended by the following terms.

$$E_{MM}^{elec} = \sum_i \sum_{j>i} \frac{q_i q_j}{r_{ij}} + \sum_i \sum_{j'} \left(\frac{q_i q_{j'}}{r_{ij'}} - \frac{q_i q_{j'}}{r_{ij}} \right) + \sum_{i'} \sum_{j'>i'} \left(\frac{q_{i'} q_{j'}}{r_{i'j'}} + \frac{q_{i'} q_{j'}}{r_{ij}} \right) - \sum_{i'} \sum_{j'} \left(\frac{q_{i'} q_{j'}}{r_{ij'}} \right) + \frac{1}{2} \sum_{i'} k_{d,i'} d_{i'}^2$$

Here, indices *i* and *j* run over MM atoms, *i'* and *j'* denote DO terms (involving Drude particles), and *r* is the corresponding distance. The 1-2 and 1-3 interactions between Drude particles are screened by applying the Thole function, which is represented by the following expression:

$$1 - \left(\frac{\mu_{ij}}{2} + 1 \right) exp(-\mu_{ij})$$

where $\mu_{ij} = r_{ij}t_{ij}$, r_{ij} being the inter-Drude particle distance and t_{ij} is the Thole parameter:

$$t_{ij} = \left(\frac{t_i + t_j}{\sqrt[6]{\alpha_i \alpha_j}}\right)$$

Here, t_i and t_j are the Thole parameters coming from the force field parameterization[64] of atoms *i* and *j*. The interactions involving Drude particles are illustrated in Figure 2 (without showing the standard interactions between the MM atoms).

The chosen force field[40] also includes an explicit representation of lone pairs to better describe the charge distribution around heteroatoms (grey balls in Figure 2). This provides a local reference frame to define an anisotropic polarizability for DOs on heteroatoms. The lone pairs are rigidly bonded to their hosting atom, and their positions are evaluated at every step with the use of internal coordinates. The forces acting on the lone pairs are distributed across the hosting atom and its neighbors in a way that conserves the total force and torque. Therefore, they are not included as additional degrees of freedom in MD simulations or geometry optimizations.

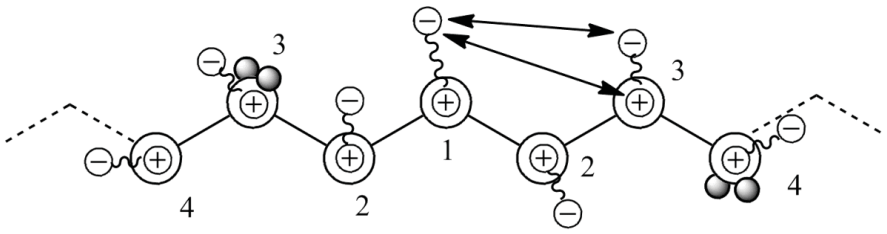


Figure 2: A molecule represented by the Drude oscillator model. Atoms are represented as spheres containing the fixed positive charge of the DO model. The Drude particles are bonded to the corresponding atom by a spring. Thole-type screenings (arrows) are applied in the DO model only for 1-2 and 1-3 interactions (see atom numbering). Lone pairs on heteroatoms are represented by grey balls.

2.1.3 Drude Oscillators in a QM/MM Framework

The implementation of the DO model within a QM/MM framework was first discussed for GROMOS (COS model)[65] and then for CHARMM with a preliminary version of the force field which did not include Thole-type interactions or lone pairs.[66] In both cases, the Drude oscillators were included in the QM computation by modifying the one-electron terms in the Fock matrix, in complete analogy to the classical MM point charges. The Drude particles give

rise to extra one-electron terms, while the compensating charge at the nucleus of the polarizable atom is taken into account by adjusting the corresponding atomic charge.

The electrostatic potential needed to obtain the polarization of the Drude centers is evaluated from the electric field (*E*) at the position of the Drude particle. It is split into three parts: QM, MM, and DO contributions. The force (*F*) on a Drude particle *i'* by a given component of the electric field can be evaluated in the following manner:

$$F_{d_{i\prime}} = \frac{\alpha_{i\prime}}{q_{i\prime}}\left(E_{i\prime}^{MM} + E_{i\prime}^{QM} + E_{i\prime}^{DO}\right)$$

As the contributions $E_{i\prime}^{MM}$, $E_{i\prime}^{QM}$, and $E_{i\prime}^{DO}$ to the electric field are interdependent, a self-consistent field approach is needed as shown in Figure 3. In our implementation,[3] we first evaluate $E_{i\prime}^{QM}$ for a set of fixed DO positions, which are then updated in an iterative scheme through an MM inner cycle. In this cycle, $E_{i\prime}^{MM}$ and $E_{i\prime}^{DO}$ are computed for the given geometry, and the DO positions are updated using the associated forces. This inner cycle is iterated until the DO positions are converged in the field of the given QM wave function (as judged by their maximum and average displacement from one step to another). Thereafter the convergence of the QM energy is checked. If not converged, the process is iterated by recalculating $E_{i\prime}^{QM}$ and going again through the inner cycle, until full overall convergence is achieved for both the DO positions and the QM total energy.

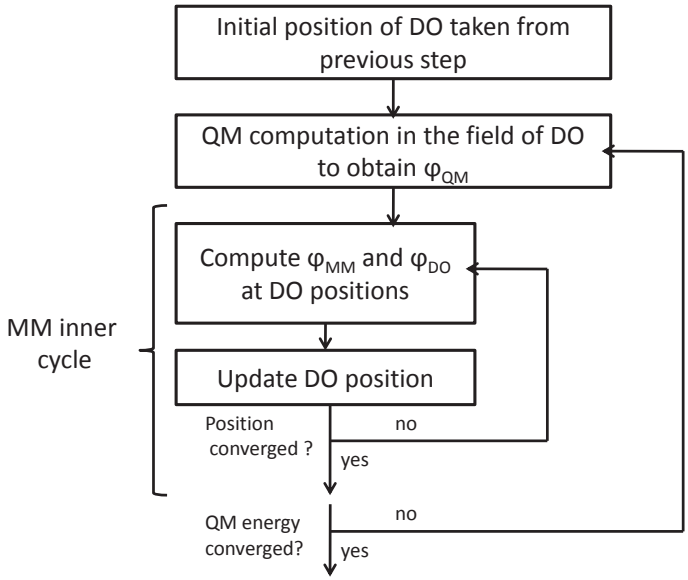


Figure 3: Flowchart of the dual SCF approach for determining the DO positions and the MM polarization in a QM/MM framework. The outer SCF procedure converges the QM wave function, while the inner one converges the DO positions in the field of each other and of the MM atoms.

We have adopted this basic method and have chosen not to refine the implementation in order to avoid modifying any QM code. This choice is motivated by the strategy to be compatible with any QM program that is interfaced to ChemShell.[67] Also, for MD simulations, Drude oscillators can be included as extra degrees of freedom, and an extended Lagrangian scheme[68] or a predictor-corrector approach[69,70] can be used to propagate them in time together with the atomic degrees of freedom. We have implemented the extended Lagrangian approach in a developmental version of ChemShell but do not cover this work here in detail. More information can be found in the associated paper.[3]

2.1.4 Three-Layer QM/MM-DO/Boundary Potential Approach

This section covers the paper: "Solvent boundary potentials for hybrid QM/MM computations using classical Drude oscillators: a fully polarizable model", which is reprinted in the annex of this thesis.

Including polarization in a self-consistent manner can be computationally very expensive. The computation time sharply increases with the number of Drude polarizable centers and hence with system size. In order to speed up the computations we have decided to investigate the use of solvent boundary potentials to decrease the number of explicit point charges included in the QM/MM-DO computation. This three-layer approach has the further advantage of including

long-range electrostatic effects by making use of a polarizable dielectric continuum to implicitly represent the solvent far away from the reactive center.[37]

2.1.4.1 Solvent Boundary Potentials

The purpose of a boundary potential is to simulate the electrostatic effects of a virtually infinite implicit outer region on a finite explicit inner region,[32] for example a QM/MM system.[35] In this thesis, we have followed the formalism used by Roux and coworkers who introduced two such approaches at the MM level, namely the standard solvent boundary potential[32] and its generalized version (GSBP).[34] They showed that, if only the degrees of freedom of the inner region are relevant to the computation of some property, this property can be calculated on the surface of its potential of mean force (PMF) obtained by integrating out the degrees of freedom of the outer region. For a system of *N* atoms, the first *n* of them being in the inner region and the *n*+1 to *N* remaining ones being in the outer region with coordinates R_o:

$$e^{-\beta W_i} = \frac{1}{C}\int dR_o d(n+1) \dots dN e^{-\beta U}$$

By picking the appropriate normalization constant (C), it can be shown that the PMF is equal to the reversible work (ΔW) necessary to assemble the inner region inside the outer region. This work can be split into contributions from the potential energy (U) of the inner region and the free energies arising from conformational restrictions (ΔW_{cr}), nonpolar interactions (ΔW_{np}), and electrostatic interactions (ΔW_{elec}).

$$\Delta W = U + \Delta W_{cr} + \Delta W_{np} + \Delta W_{elec}$$

If the outer region is chosen to be representative of the average outer conformation, one can further approximate the PMF by assuming that ΔW_{cr} and ΔW_{np} will both remain constant throughout the process studied in the inner region, so that they can be ignored when computing changes of inner-region properties, for example relative energies along a reaction path.

The generalized solvent boundary potential (GSBP) is designed for MD simulations.[34] It evaluates the inner-outer electrostatic interactions in two different fashions. The non-solvent (e.g. enzyme) part of the outer region is treated atomistically, with explicit calculation of standard Coulomb interactions. The solvent part of the outer region is described by a polarizable dielectric continuum (PDC), and the interactions with the inner region are evaluated

by solving the linearized Poisson-Boltzmann equation for a given dielectric constant of the solvent. The central idea of GSBP is to precompute this interaction with the use of a Green's function which can be projected onto a set of basis functions.[34] This precomputation is expensive but needs to be done at the beginning of the computation only once and for all, which makes this technique particularly attractive for long MD simulations requiring a large number of steps.

The current QM/MM implementation of GSBP is restricted to semiempirical QM methods.[35] It employs Mulliken charges to represent the QM part of the system in the GSBP computation. As the BP contribution is evaluated at every step of the iterative QM SCF procedure, the QM code has to be modified.

Geometry optimizations generally require much less steps than long MD simulations, and hence the GSBP strategy with its significant initial overhead is no longer advantageous. This motivated the development of the solvated macromolecule boundary potential (SMBP), in which the BP interactions are not precomputed but calculated on-the-fly.[36,37] This not only enhances the computational efficiency of geometry optimizations, but also avoids any modifications of the QM codes used for QM/MM. This makes it fully compatible with the ChemShell philosophy. Details of this method can be found in the attached paper[3] or in the initial publication.[36]

2.1.4.2 QM/MM-DO/GSBP

When performing polarizable-embedding QM/MM simulations with GSBP, we use an extended Lagrangian approach to propagate the Drude particle in time.[50,68] In this scheme, the Drude particles provide additional degrees of freedom to the system and can thus be considered in the same way as any other MM point charges. This makes the QM/MM-DO/GSBP combination easy for the inner region. For the outer region we make the approximation of a frozen polarization state when precomputing the GSBP terms. It is therefore necessary to run a single-point calculation to obtain the positions of the Drude particles and lone pairs from a standard QM/MM-DO computation. Once this is done, the precomputation of the Green's matrix can be carried out in a similar manner as for standard QM/MM/GSBP computations.[35]

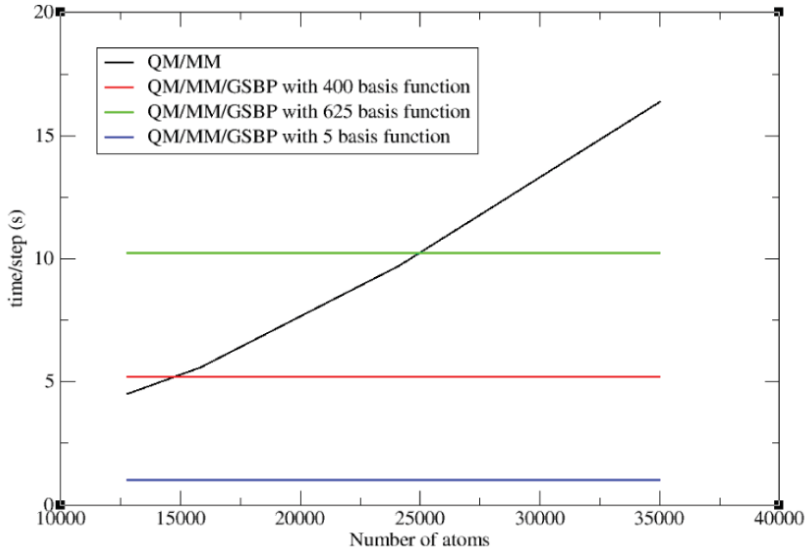


Figure 4: Computation time per MD step vs. number of atoms in the system. The black line refers to a standard QM/MM computation, and the green, red, and blue lines to QM/MM/GSBP computations with different basis sets.

We assessed the efficiency of the QM/MM-DO/GSBP method for a test system consisting of a zwitterionic glycine solvated in a ball of water molecules. We progressively increased the size of the water ball and ran MD simulations to determine the average computation time per step. The inner region of the system encompassed 903 water molecules having their hydrogen atom within 18 Å from the center. Water molecules between 14 and 18 Å were frozen in order to avoid diffusion of water out of the inner region. Figure 4 compares the computation times from standard QM/MM-DO calculations (black line) to those from QM/MM-DO/GSBP calculations employing basis sets of different size for the projection of the GSBP Green's matrix. With the default value of 400 basis functions, the three-layer QM/MM-DO/GSBP scheme becomes advantageous at a system size of 14500 atoms. This threshold is similar to the one obtained in analogous QM/MM/GSBP simulations with an additive force field. For larger systems that are commonly used in QM/MM studies, three-layer hybrid approaches with GSBP thus afford appreciable gains in efficiency in MD simulations, without loss of accuracy.

2.1.4.3 QM/MM-DO/SMBP

When combining the QM, MM-DO, and SMBP models, there are three self-consistent procedures that need to be solved simultaneously. According to the ChemShell philosophy, this should be done in a transferable manner, without modifications to the underlying QM codes. In our implementation, the QM-SCF treatment is thus incorporated as is into the algorithm. On the other hand, the DO-SCF and SMBP-SCF procedures are performed concomitantly to

reduce the computational effort. We have designed an algorithm that minimizes the number of QM-SCF calculations, which are normally the most expensive part of the evaluation of the potential. As shown in Figure 5, data for the DO and SMBP treatments are extracted from the same QM calculation, and parts of SMBP potential are evaluated on-the-fly for the next step. The QM energy is adopted as overall convergence criterion, because it is the most relevant quantity in typical applications (i.e., studying chemical reactions). Details concerning the individual steps of the flowchart in Figure 5 can be found in the associated paper.[3]

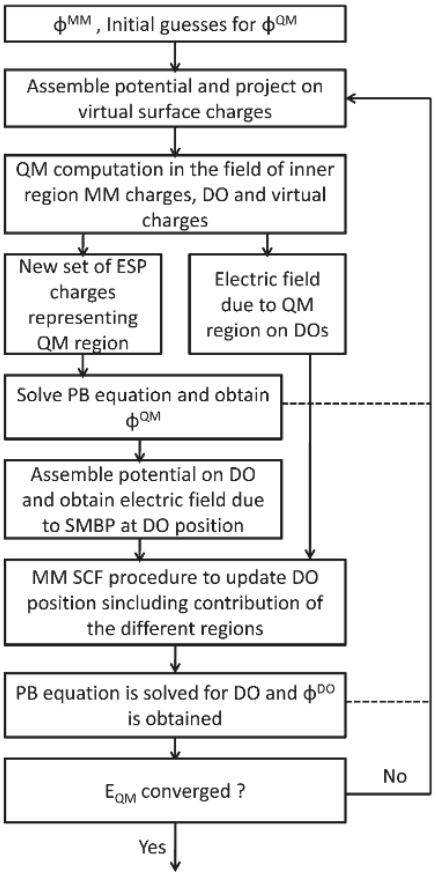


Figure 5: SCF procedure used to converge all three component of the QM/MM/SMBP approach. See text for details.

To assess the numerical validity of our implementation we used the same test system (glycine in a ball of water) as in the GSBP case. We checked the mean average deviation and the maximum deviation of the computed gradient when switching from QM/MM-DO to QM/MM-DO/SMBP. We confirmed that, if the inner-outer region boundary was far enough from the QM region, the differences in the computed gradient for the relevant part of the system were far below the convergence criteria generally used for geometry optimization. On the other hand, this numerical precision was lost when we tried to further simplify the algorithm shown in

Figure 5 (e.g., by neglecting some time-consuming parts of the coupled SCF treatments). Hence, the efficiency cannot be further improved without compromising the precision of the method.

QM	basis	QM/MM	QM/MM/SMBP	% saved
AM1		122	66	46
BLYP	SVP	505	262	48
BLYP	TZVPP	1355	726	47
B3LYP	SVP	613	362	41
B3LYP	TZVPP	1622	1021	33

Table 1: Computation time per geometry optimization step for QM/MM and QM/MM/SMBP calculations using different QM methods and different basis sets. Timings were obtained on 2.93 GHz Intel Xeon X5670 machines with 12 GB of memory.

As already noted, the SMBP approach can be used with any QM method, and the speedups will thus depend on the chosen QM method. Table 1 compiles average computation times per optimization step and the percentages saved in QM/MM-DO/SMBP calculations using different semiempirical and DFT methods with different basis sets (compared to standard QM/MM-DO). The test system was the same as before (glycine in water). The standard QM/MM-DO setup included 10 QM atoms (glycine) and 21260 MM point charges, which was reduced to 4515 point charges in the QM/MM-DO/SMBP case (without loss in accuracy). Evidently, there is a significant improvement for all tested QM methods, and hence our objective of speeding up such computations is achieved.

2.1.5 QM/MM-DO Computations for Enzymatic Reactions

This section covers the paper "Toward QM/MM Simulation of Enzymatic Reactions with the Drude Oscillator Polarizable Force Field", which is reprinted in the annex to this thesis.

Once DO parameter for proteins became available,[40] we could investigate enzymatic reactions with the polarizable QM/MM-DO method. Since this was the first time that such a study could be performed with a well-parameterized polarizable MM-DO force field, we did not use any boundary potential in order to be able to identify and analyze each energy contribution at the QM/MM level.

2.1.5.1 Technical Aspects of Computations on Enzymatic Reactions

Several problems were encountered when treating enzymes within the QM/MM-DO framework. First, in contrast to our previously used glycine-in-water test system, Thole-type interactions had to be taken into account, which caused oscillations in the QM/MM-DO SCF iterations and prevented convergence for larger systems. This problem was overcome by a successive over-relaxation approach.[29] In this technique, the positions of the Drude particles from the previous iteration are kept in memory, and their new positions are linearly interpolated between the previous and the newly predicted ones (thus introducing some damping in the DO iterations). In some cases this was not sufficient, and so we had to design a more complicated strategy, which utilizes a problem-adapted partitioning of the system and an analytical solution for the polarization state of the subsystems via Cholesky factorization of the underlying linear system of equations. This technique is described in detail and assessed in the corresponding paper.[2]

A second crucial aspect that had never been covered before was the treatment of cuts through covalent bonds at the QM/MM boundary when a polarizable force field is employed.[1,30] We tackled this problem by retaining the charge shift scheme for the standard MM point charges and trying different approaches for the Drude oscillators. The most accurate model turned out to be the complete deletion of the Drude oscillators at the bond being cut. Any of the tested more elaborate alternatives (e.g., transfer of polarizability, charge, or Thole parameter to adjacent polarizable atoms) significantly decreased the accuracy of the QM/MM results for a standard test system comprised of n-butanol and a sodium cation in 100 different positions (with cuts being applied through C-C bonds in n-butanol). For our preferred model, the deviations of the QM/MM-DO results for the deprotonation enthalpy and proton affinity of n-butanol from the full QM reference results were similar to those obtained for a standard QM/MM treatment with an additive force field. The simplest being the best, we adopted the deletion of the Drude oscillators in bonds being cut at the QM/MM-DO boundary as our standard procedure.

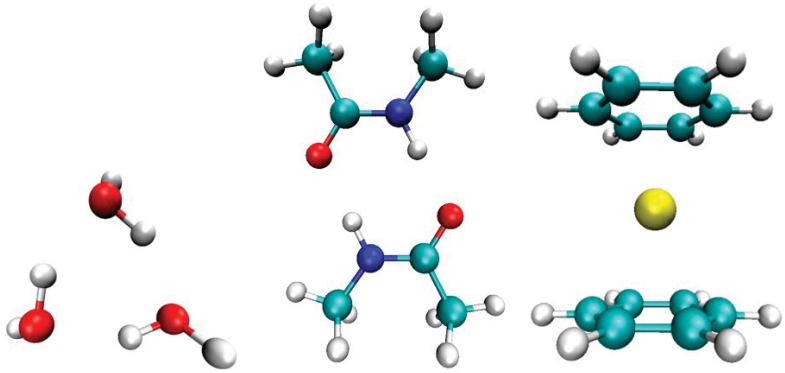


Figure 6: Test systems used for validating QM/MM-DO compatibility. From left to right: the cyclic water trimer, the cis-NMA dimer, and the bis(benzene)sodium sandwich complex.

Before running QM/MM-DO simulations of enzymatic reactions we chose three complexes to test QM/MM-DO compatibility (Figure 6): the most stable water trimer, the cis-N-methylamide (cis-NMA) dimer, and the bis(benzene)sodium sandwich complex. For each of these complexes, we determined the optimum geometries and binding energies at the QM/MM and QM/MM-DO levels and compared the results to those from full QM reference calculations. We checked a variety of commonly used QM methods covering semiempirical methods, pure and hybrid DFT functionals as well as RI-MP2 methods (basis sets: SVP, TZVP, and TZVPP basis set). It turned out that the QM/MM-DO results are fairly sensitive to the choice of QM method (and basis set), significantly more so than the QM/MM results with an additive force field. The RI-MP2 method with moderate basis sets had been employed in the parameterization of the currently adopted MM-DO force field, and it may thus not be surprising that this RI-MP2 approach offered the best compromise in terms of QM/MM-DO compatibility.

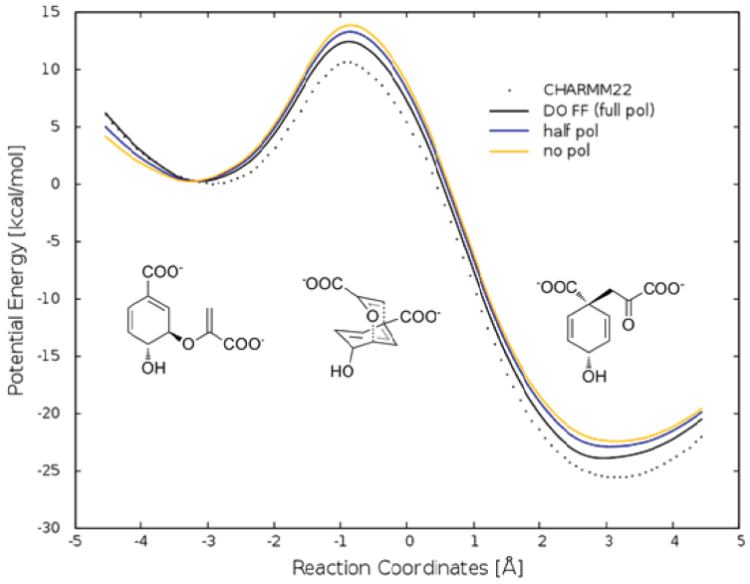


Figure 7: Potential energy profile for the Claisen rearrangement of chorismate to prephenate catalyzed by Chorismate Mutase. The dotted line shows the standard QM/MM energy profile obtained with the CHARMM additive force field. For each structure along the path, single-point QM/MM-DO calculations were performed with full polarization (black line) and with polarization scaled down by a factor 2 (blue line) and switched off totally (orange line).

We first investigated Chorismate Mutase (CM).[71] This enzyme catalyzes the Claisen rearrangement of chorismate to prephenate (Figure 7). It is often chosen to test QM/MM methods as there is no covalent bond crossing the QM/MM boundary so that problems with cuts through this bond are avoided.[72-74] We used five independent snapshots from a previous study[5] on CM free-energy calculations (see section 3.3). For each snapshot we performed single-point calculations at the previously optimized structures along the reaction path obtained at the standard QM/MM level using RI-MP2/SVP as the QM method and the CHARMM27 additive force field[9] for the MM part. Using a set of scripts to set up and to carry out the corresponding QM/MM-DO single-point calculations in an automatic fashion, we recomputed the energy profile with MM polarization included. To assess the importance of MM polarization in this enzymatic reaction we switched it off in a second set of calculations, and we also ran a third set including only half of the MM polarization on each DO center. The results for each snapshot are given in Table 2, and the potential energy profiles for one selected snapshot are shown in Figure 7 for the different types of computation. We find that switching off the MM-DO polarization influences the results notably: the computed barriers are lowered by 5 to 15% which is a rather small change that should however not be neglected when aiming for accurate results.

snapshot	$\Delta\Delta E^{\ddagger}$		$\Delta\Delta E$	
1	−1.3	(−0.29)	−1.21	(−0.41)
2	−1.25	(−0.51)	0.23	(0.09)
3	−1.42	(−0.62)	−2.01	(−0.73)
4	−0.52	(−0.13)	−0.46	(0.02)
5	−1.15	(−0.51)	−2.99	(−1.16)

Table 2: Changes in the potential energy barrier ($\Delta\Delta E^{\ddagger}$) and the reaction energy ($\Delta\Delta E$) for CM when switching on MM polarization. Values in parenthesis are obtained by switching on only half of the MM polarization.

For the sake of completeness, we investigated another enzyme, p-hydroxybenzoate hydroxylase (PHBH), in an analogous manner using again the QM/MM setup from a previous study in our group.[37] For the reaction catalyzed by PHBH, the influence of MM polarization is slightly larger, with changes in the computed barriers of up to 20% (Table 3). Based on these initial studies we conclude that MM polarization may affect the computed QM/MM barriers notably, on the order of 5-20%. A more complete assessment would of course require geometry optimizations at the QM/MM-DO level and computation of free energy profiles using a sampling technique.

snapshot	$\Delta\Delta E^{\ddagger}$		$\Delta\Delta E$	
1	2.1	(1.7)	1.7	(1.2)
3	0.8	(0.6)	2.8	(1.9)
4	0.8	(0.6)	0.5	(0.3)
5	2.2	(1.3)	2.7	(1.7)

Table 3: Changes in the potential energy barrier ($\Delta\Delta E^{\ddagger}$) and the reaction energy ($\Delta\Delta E$) for PHBH when switching on MM polarization. Values in parenthesis are obtained by switching on only half of the MM polarization.

2.2 Internal Reaction Coordinate for Large QM/MM Systems

This section summarizes the paper: "A microiterative intrinsic reaction coordinate method for large QM/MM systems", which is reprinted in the annex of this thesis.

Intrinsic reaction coordinate (IRC) methods are an invaluable tool for the QM investigation of chemical reactions. They are used to verify the character of a transition state and to ensure that it directly connects reactants and the desired products. In QM-only studies of reactions of small molecules, it is a standard practice to run such IRC calculations. At the QM/MM level, an appropriately adapted IRC method has been missing in the toolbox.[4]

The IRC is defined as the steepest-descent pathway in mass-weighted coordinates starting from the transition state and ending in a local minimum on the potential energy surface.[75] Several algorithms exist to follow this path in discrete steps. The straightforward Euler method only requires gradients but lacks accuracy. The local quadratic approximation (LQA) gives more precise results by employing information from the Hessian. The predictor-corrector approach is an advanced method that corrects the predicted steps using the stored information about previous steps on the path being followed.[76,77]

QM/MM simulations normally cope with very large systems, and a huge number of degrees of freedom would thus need to be included in a standard IRC computation. This would quickly become impractical, especially if one wants to include Hessian information in the process. We have implemented a microiterative procedure, in which only a small subset of relevant atoms (normally QM atoms) is included into the IRC computation itself, whereas the rest of the system is relaxed by geometry optimization after each IRC step. This follows the philosophy of the microiterative transition state search.[78] In our implementation, all the IRC algorithms mentioned above are available, and we use a Hessian update method to avoid recomputing it at each step of the IRC.[79]

We validated our implementation for several test systems including the enzymatic reactions catalyzed by CM and PHBH (see above). We find that our microiterative IRC technique is capable of handling large QM/MM systems efficiently and with good accuracy. We recommend the use of the LQA method with a relatively small step size to ensure that the outer part of the system can properly relax at every step. It is of course important to carefully define the inner region included in the standard IRC treatment, which should encompass every atom involved in the reaction.

2.3 Dual Hamiltonian Free Energy Perturbation

This section summarizes the paper "Quantum mechanics/molecular mechanics dual Hamiltonian free energy perturbation", which is reprinted in the annex of this thesis.

Free energy evaluation techniques are standard tools in MM and QM/MM computations. For chemical reactions they allow for computing their rates and the thermodynamics properties of the species involved. In the canonical ensemble with partition function Z, the Helmholtz free energy can be expressed as:[80]

$$A = -k_B T \ln(Z)$$

where k_B is the Boltzmann constant and T is the temperature. For any system of significant size, Z has to be determined numerically, by using sampling techniques such as molecular dynamics[8] or Monte Carlo.[81] The main methods to extract free energies out of such simulations are umbrella sampling,[82] thermodynamic integration,[83] and free energy perturbation (FEP).[80] Each of them has been adapted to QM/MM simulations (for more detailed information see the introduction of the attached paper[5]).

Free energy computations necessitate extensive sampling, e.g. by means of long MD simulations that require a large number of successive QM/MM computations. This often rules out the direct use of accurate first-principles QM methods, and one then often resorts to semiempirical QM methods for this purpose. There are some methods available that use thermodynamic cycles to derive higher-level free energies from lower-level (semiempirical or force field) simulations. We propose an FEP-based method of this kind. In the FEP treatment of a chemical reaction, one starts from a precomputed potential energy profile with optimized geometries along the reaction path, which are then split into discrete windows. For each of these windows, one then computes, through MD simulations, the energy necessary to bring the system to the next window by perturbing the reaction coordinate.

In our scheme, a standard QM/MM MD simulation is run using a semiempirical QM method, and the QM/MM energy is evaluated with a higher-level QM method every n steps by perturbing the reaction coordinate to the next window. We found that n=15 is a good compromise between accuracy and efficiency. The use of two distinct QM methods leads to the name "dual Hamiltonian free energy perturbation (DH-FEP)". Our scheme differs from previously proposed approaches by not separating the QM and MM degrees of freedom. Instead we sample along all QM degrees of freedom with the exception of the reaction coordinate, whereas previously proposed methods sample only along MM degrees of freedom.[84]

We first validated the DH-FEP method by using an analytical two-dimensional energy surface. Good results compared to standard thermodynamic integration were obtained when the overlap of the two surfaces in the region of sampling was sufficient. Hence, the geometries given by both QM methods should be similar along the reaction path, especially for the reaction coordinate. We applied our method to the Claisen rearrangement catalyzed by chorismate mutase (see subsection 2.1.5.2) using OM3 and SCC-DFTB as low-level methods and RI-

MP2/SVP as high-level method. The reaction coordinate was initially defined as the difference of the distances of the forming and breaking bonds. With this choice, we had to combine calculations with both semiempirical methods (for different parts of the path) to produce a surface overlapping well enough with the MP2 surface. The key issue turned out to be the match of the coordinates of the atoms directly involved in the reaction (as obtained from the lower-level and higher-level methods). Constraining the two key distances of the forming and breaking bonds in the DH-FEP MD simulations gave fully satisfactory results regardless whether OM3 or SCC-DFTB was chosen as lower-level method (without the need of constructing a hybrid surface). We therefore recommend to apply the DH-FEP method first by using a single appropriate reaction coordinate, and to constrain further relevant reactive degrees of freedom in problematic cases. These additional degrees of freedom can be identified, for instance, by using the microiterative IRC tool described in the previous chapter.

3. Examples of QM/MM Simulations

In this section, two applications of QM/MM methods are reported. The first one is a "standard" QM/MM study of reactions catalyzed by the enzyme phenylacetone monooxygenase (PAMO), in collaboration with Prof. Manfred Reetz (previously at the Max Planck Institute, now at Marburg University). It is a continuation of previous QM/MM work in our group by Iakov Polyak on cyclohexanone monooxygenase (CHMO).[85] The current status of this project is described in some detail here (section 3.1), because it has not yet been published. The PAMO study will be completed by another group member, Yiying Zheng.

The second project concerns the prebiotic synthesis of purines nucleobases. In this case, QM/MM was employed in a less conventional way to model the energy dissipation process after relaxation of an electronically excited state producing a "hot" ground state. This work was performed in a collaboration led by Mario Barbatti.[6] It is summarized in section 3.2.

3.1 Phenylacetone Monooxygenase

3.1.1 Introduction

Phenylacetone monooxygenase (PAMO) is a Baeyer-Villiger monooxygenase (BVMO)[86] that exhibits good thermal stability and performs well in a variety of solvents.[87,88] It catalyzes the oxidation of phenylacetone (PHAC) to benzyl acetate using NADPH as an electron donor and molecular oxygen as oxidative reactant. NADPH first reduces the enzyme-bound FAD cofactor to $FADH^-$ which reacts with molecular oxygen yielding a C4a-peroxyflavin intermediate. This species reacts with PHAC and generates the product, presumably via a Criegee intermediate (Figure 8).[89] As in the previous QM/MM study on CHMO,[85] we focus on this part of the catalytic cycle. Here we only treat the first step leading to the Criegee intermediate, and not the second migration step (Figure 8). A direct reaction pathway to the product bypassing the Criegee intermediate could not be found.

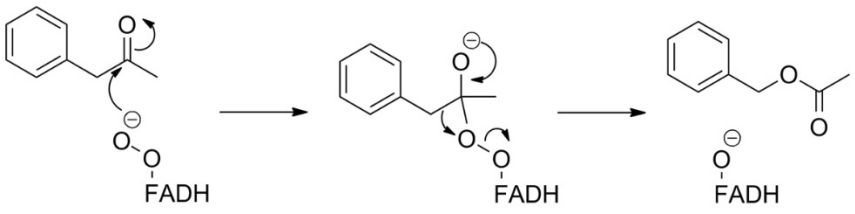


Figure 8: Mechanism of the formation of the Criegee intermediate and the product in WT PAMO.

The crystal structure of PAMO was taken from the work of Orru *et al*. (pdb: 2YLT).[90] It includes the protein, the two cofactors (NADPH and FADH), and an inhibitor with a structure similar to PHAC. The C4a-peroxyflavin starting structure was generated by manually adding an O_2 moiety to FADH in the same manner as in the CHMO study.[85] The inhibitor in the crystal structure was replaced by PHAC, again as before.[85] Force field parameters for the substrate were adapted from related compounds. The QM/MM setup followed standard procedures.[85]

The binding pocket in PAMO is very similar to the one of the previously studied CHMO.[85] By analogy, we first considered the same QM region. It included PHAC, the peroxyflavin part of $FADOO^-$, the region of $NAPD^+$ close to the reacting species, and the residue ARG337. Figure 9 shows the QM region from the generated starting structure.

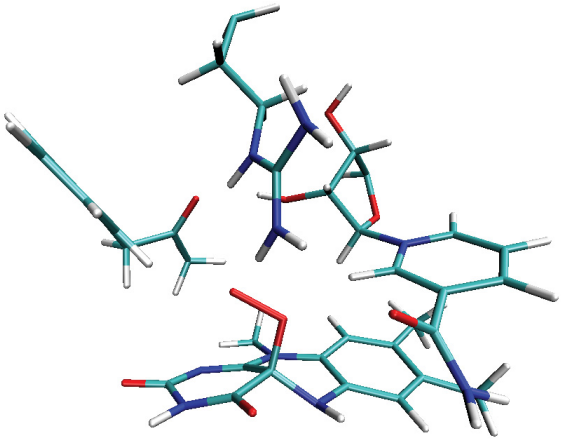


Figure 9: Starting structure showing the atoms in the initially chosen QM region.

3.1.2 Orientation of the $NADP^+$ Nicotinamide Moiety

A closer look at the active site (Figure 9) reveals one key difference from the previous setup,[85] namely the orientation of the nicotinamide moiety in the $NADP^+$ cofactor. In the case of CHMO, one of the hydrogen atoms of the carboxamide was found to form a hydrogen bond with the proximal oxygen atom of the peroxy group of $FADHOO^-$, in line with experimental evidence for the stabilization of peroxyflavin by $NADP^+$. Here, the amide is oriented into the opposite direction having the oxygen atom pointing towards ARG337. It engages in a moderate hydrogen bond with the closest hydrogen atom of this residue, at a donor-acceptor distance of 2.97 Å. This conformation is not specific to our QM/MM setup since it is also found in the crystal structure (with FADH and no peroxy group).[90] In the latter, it could be an artifact

induced by the crystallization and thus not representative of the real experimental (or natural) conditions. Also, at room temperature, there might be a facile rotation of the carboxamide group giving access to both possible orientations.

These issues were studied by performing a MD simulation at the MM level. As we cannot run simulations of more than a few nanoseconds, we cannot expect to observe enough rotations during this time to obtain quantitative results. Therefore we applied an enhanced sampling technique. We ran a metadynamics[91] simulation that allowed us to determine the free energy profile along one particular degree of freedom, namely the dihedral angle between the nitrogen and the carbon of the amide, the carbon of the ring to which it is bonded, and an adjacent carbon atom. During this simulation all other degrees of freedom were freely sampled. In the metadynamics procedure, the selected dihedral angle explored all possible conformations, and the free energy profile was built taking into account the artificial forces that had to be introduced. The simulation was run in the NPT ensemble at 300 K and 1 atm for 12 ns (shorter runs of 3 and 6 ns were not long enough to offer adequate sampling).

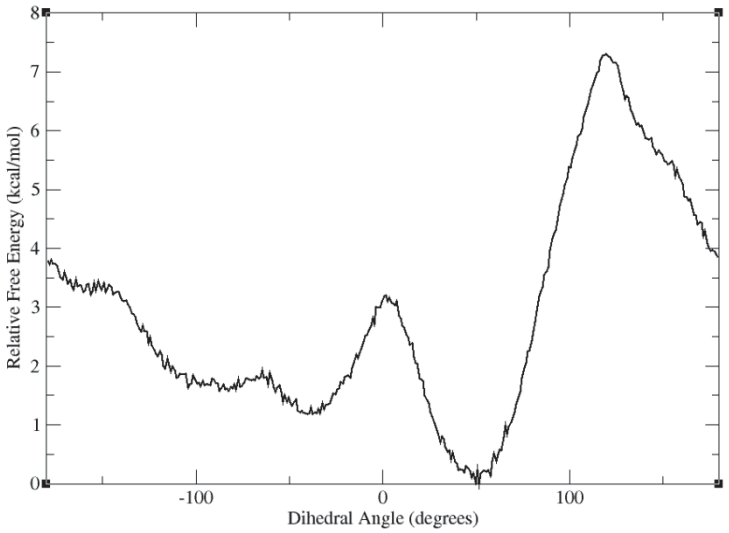


Figure 10: Free energy of rotation of the carboxamide moiety of the nicotinamide part of the NADP+ cofactor.

In the computed free energy profile (Figure 10), the minimum at around 50° corresponds to the arrangement observed in the crystal structure. The second minimum at -50° has the same general orientation but with the oxygen atom pointing down this time. The barrier between the two minima is quite low so that switching from one to the other is possible. The structures with dihedral angles around -160° benefit from hydrogen bonding of the carboxamide with the peroxy group of $FADHOO^-$ (analogous to the CHMO case, see above), but they are almost 4

kcal/mol above the overall minimum. While they are accessible at room temperature, the minimum conformation should clearly be more populated.

3.1.3 Preliminary QM/MM Computations

Several QM/MM potential energy scans were run from the initial structure to the Criegee intermediate. We chose an analogous QM region as in the previous CHMO study (with PHAC as substrate, around 100 QM atoms), and we adopted the same definitions of the reaction coordinate and the active region as before.[85] For QM regions of this size, DFT is the QM method of choice. We tested several exchange-correlation functionals and two different basis sets (BP86,[92,93] BLYP,[94] B3LYP,[95] PBE,[96] PBE0[97]; SVP,[98] TZVP[99]). It turned out that the B3LYP/TZVP level was needed to provide an acceptable accuracy (as previously in the case of CHMO[85]). The computed barrier was around 7-8 kcal/mol, with the Criegee intermediate occupying a shallow minimum about 6 kcal/mol above the reactant (again quite similar to what had been observed for CHMO[85]). Note that every DFT method used showed the presence of the Criegee intermediate, and no reaction coordinate leading directly to the product could be found.

However, these preliminary scans also indicated that the choice of the QM region may be questionable. The energy profiles were not entirely smooth and, more importantly, drastic changes in MM energy were sometimes observed along the reaction coordinate. This suggests that some more MM residues should be included in the QM region to treat all the relevant parts of the system at the QM level on an equal footing.

3.1.4 A Water Molecule Stabilizing the Criegee Intermediate

Visual analysis revealed a water molecule in the direct vicinity of the phenylacetone substrate. The ketone part of the substrate is stabilized by three hydrogen bonds: from $NADP^+$, from ARG337, and from this water molecule, as depicted in Figure 11.

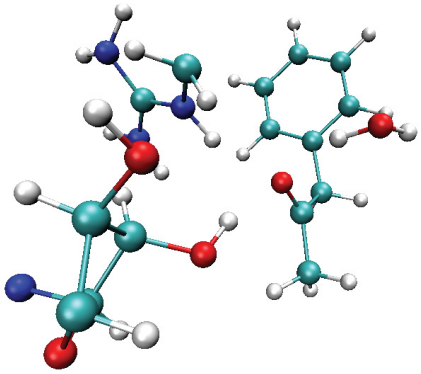


Figure 11: H-bonds to the ketone of PHAC in PAMO, involving ARG337, $NADP^+$ and a water molecule.

Visual inspection also indicates which amino acids are involved in the stabilization of this extra water molecule (Figure 12). It accepts a hydrogen bond from TYR502 and donates two hydrogen bonds to PHAC and to the peptide bond between ALA442 and LEU443. It is important to note that the latter are part of a loop (440-443) which has been in the focus of mutation studies to extend the scope of the reaction, either by deleting some or all amino acids (441-443) or by mutating PRO440 to make the chain more flexible and allow some kinds of rearrangement.[87]

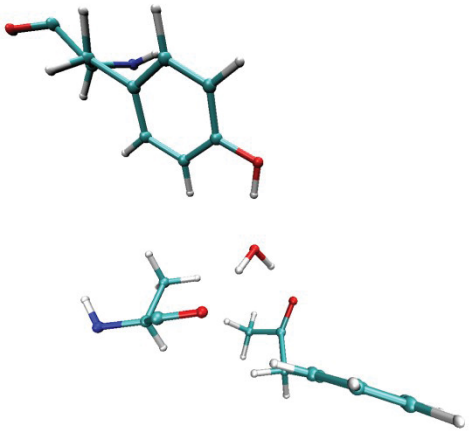


Figure 12: Environment of the water molecule stabilizing the substrate. It is H-bonded to ALA442 and TYR502.

The water molecule (Figure 12) is not present in the crystal structure and has also been missing in the CHMO study.[85] In the QM/MM setup for PAMO, it appeared very soon during equilibration of the system (after ca. 10-20 ps) and stayed there for the rest of the initial classical MD simulation (ca. 10 ns). However, using this computational approach, it is impossible to guarantee that this water molecule will always be present and remain near the active site in

such a big system. To check the likeliness for this to happen, we performed MD simulations with locally enhanced sampling[100] at the MM level. This ensures a more extensive sampling of the relevant part of the system, the water molecule in our case, and more importantly increases the rate of possible transitions to other conformations. The water molecule again stayed close to its initial position during the whole simulation time of 10 ns. Therefore we assume that it is present when phenylacetone is in the active site. Its absence in the crystal structure might be due to the fact that it does not contain phenylacetone in the active site, but a different inhibitor.

The scan of the first step of the reaction was rerun at the B3LYP/TZVP level including this water in the QM region. This modification did not change the reaction path qualitatively but lowered the barrier and gave a smoother curve (Figure 13, left). Comparing again different QM methods still suggests that the use of B3LYP/TZVP is required (Figure 13, right).

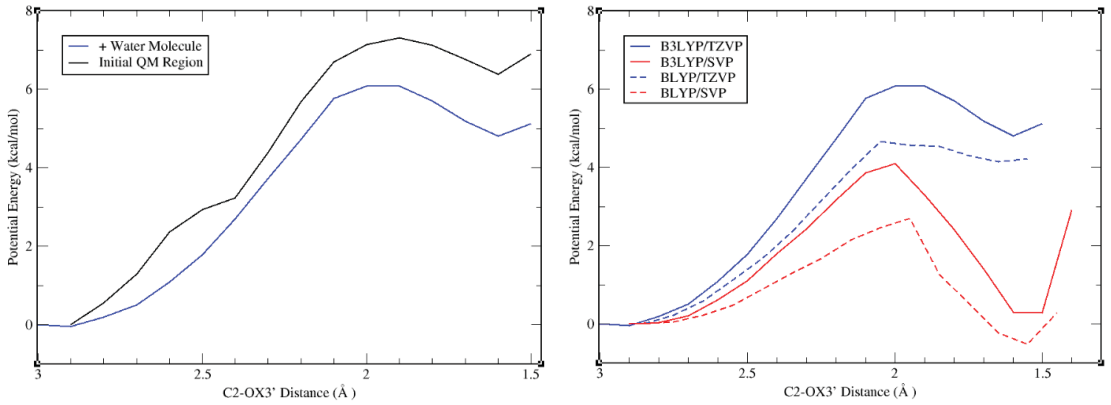


Figure 13: Potential energy scan for the formation of the Criegee intermediate. On the left: comparison of initial QM region and the QM region including the water molecule at B3LYP/TZVP level. On the right: comparison of different QM methods for the new QM region.

To check whether the Criegee intermediate might react with the extra QM water molecule, we ran potential energy scans at the B3LYP/TZVP level with this QM region and at the OM3[101,102] level with an even larger QM region including the residues stabilizing the QM water molecule. In both cases, the energy kept going up as the proton of the QM water molecule approached the oxygen atom of the Criegee intermediate.

3.1.5 Another Key Residue: ASP66

Visual inspection of the optimized structures along the reaction path and an assessment of the electrostatic effects of the active-site environment at the OM3/MM level indicated that the ASP66 residue may affect the formation of the Criegee intermediate by interacting with ARG337. The importance of this residue had already been highlighted in the paper reporting the PAMO crystal structure,[90] but without giving a clear interpretation of its role.

Figure 14 shows the effects of switching off the MM charges of residues (left) or water molecules (right) in the active part of the MM region on the energy of the Criegee intermediate relative to the reactant complex. The semiempirical OM3 Hamiltonian[101,102] was used to optimize the geometries of both species followed by single-point B3LYP/TZVP energy calculations. Figure 14 shows the differences between the relative energies of the Criegee intermediate with the MM charges being switched off and being retained (standard values) as obtained at the QM and QM/MM (tot) level; the QM region consisted of the oxidized flavin, phenylacetone, and the interacting part of $NADP^+$. This analysis identifies the MM residues that have a significant electrostatic influence on the reaction and should thus be included in the QM region. In the present case, ASP66 and one further water molecule (2099) were included on this basis in the following QM/MM calculations. In this QM region, the ASP66 residue is present in its deprotonated form and thus possesses a formal negative charge so that the QM region is neutral overall.

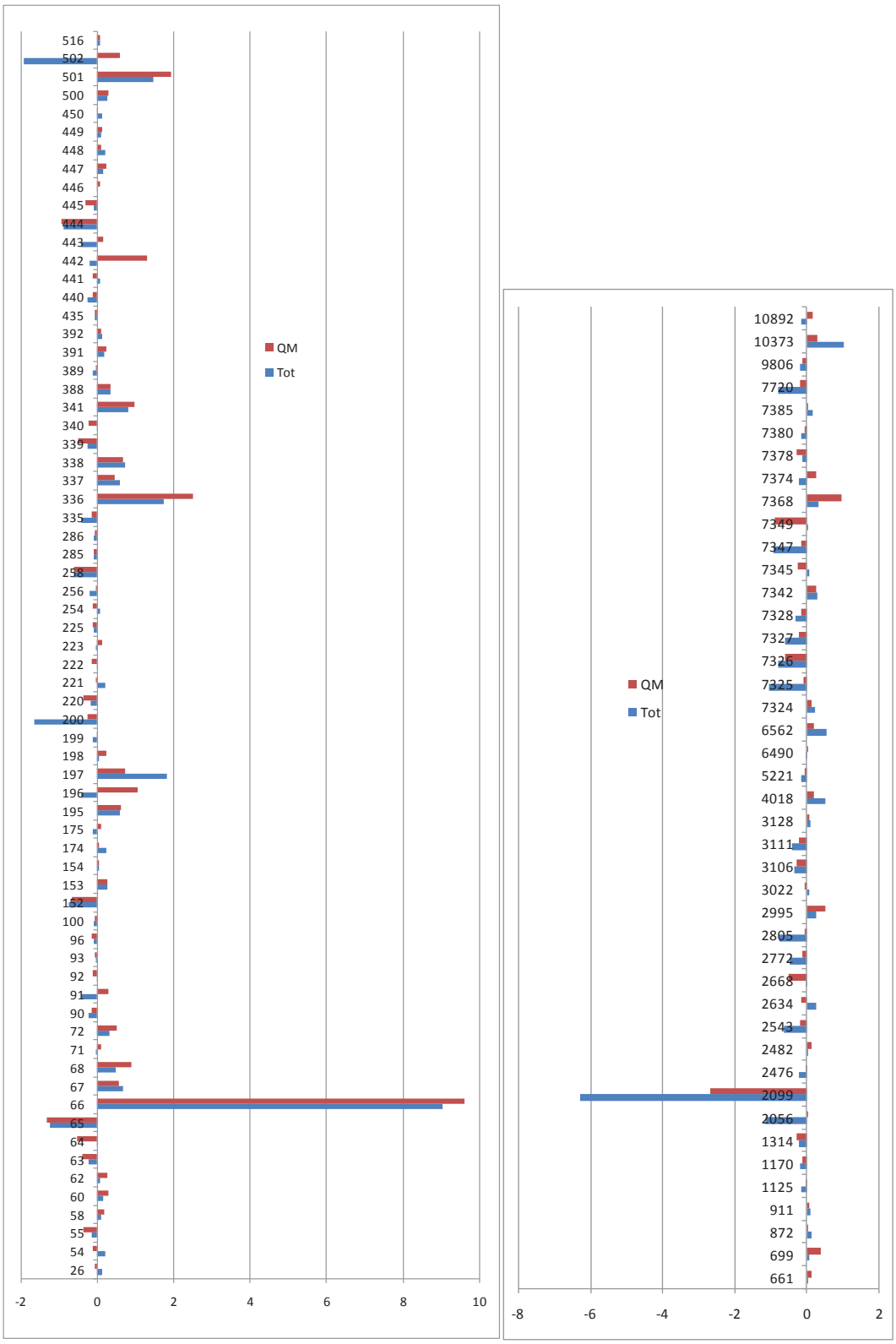


Figure 14: Residue analysis for the formation of the Criegee intermediate in PAMO. Electrostatic effects of PAMO residues (left) and water molecules (right) in kcal/mol. See text for details.

3.1.6 Formation of the Criegee Intermediate

After deciding to include ASP66 in the QM region, we examined three QM/MM partitions (Figure 15): "small", adding just ASP66; "medium", adding ASP66 and ILE67; "large", adding ASP66, ILE67, and CYS65. We also considered ILE67 and CYS65 as they are hydrogen bonded to the flavin moiety. Scans were run with B3LYP/TZVP as QM method, without and with D2 dispersion corrections.[103]

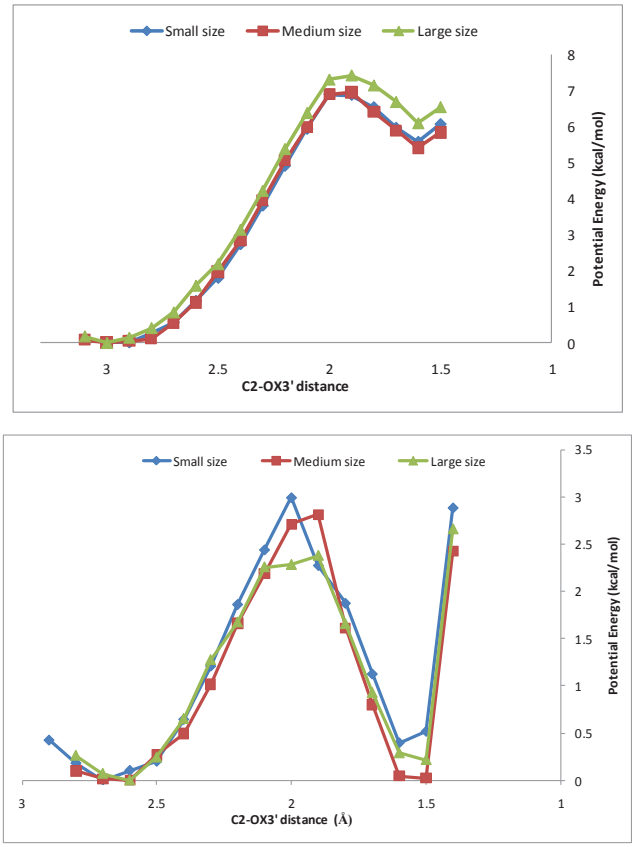


Figure 15: QM(B3LYP/TZVP)/MM energy profiles without (left) and with (right) D2 dispersion corrections for formation of the Criegee intermediate using different QM regions including Asp66 (see text).

The resulting profiles are fairly similar for the three chosen QM regions. They are smoother when D2 dispersion corrections are not applied. Here, as in previous scans, it proved to be technically difficult or impossible to precisely locate the transition state by an unconstrained transition state search. As in the CHMO case,[85] the following mechanistic reasoning is thus based on the computed energy profiles.

3.1.7 Suggested Mechanism and Relevant Residues

In this section, our aim is to give a detailed interpretation of the role of different active-site residues in PAMO. As shown previously,[85] ARG337 plays a crucial role in catalysis by Baeyer-Villiger monooxygenases. Hence we first focus on this residue and its environment. Figure 16 shows the active site of PAMO in a 2D representation, with the peroxyflavin just partially included for the sake of clarity.

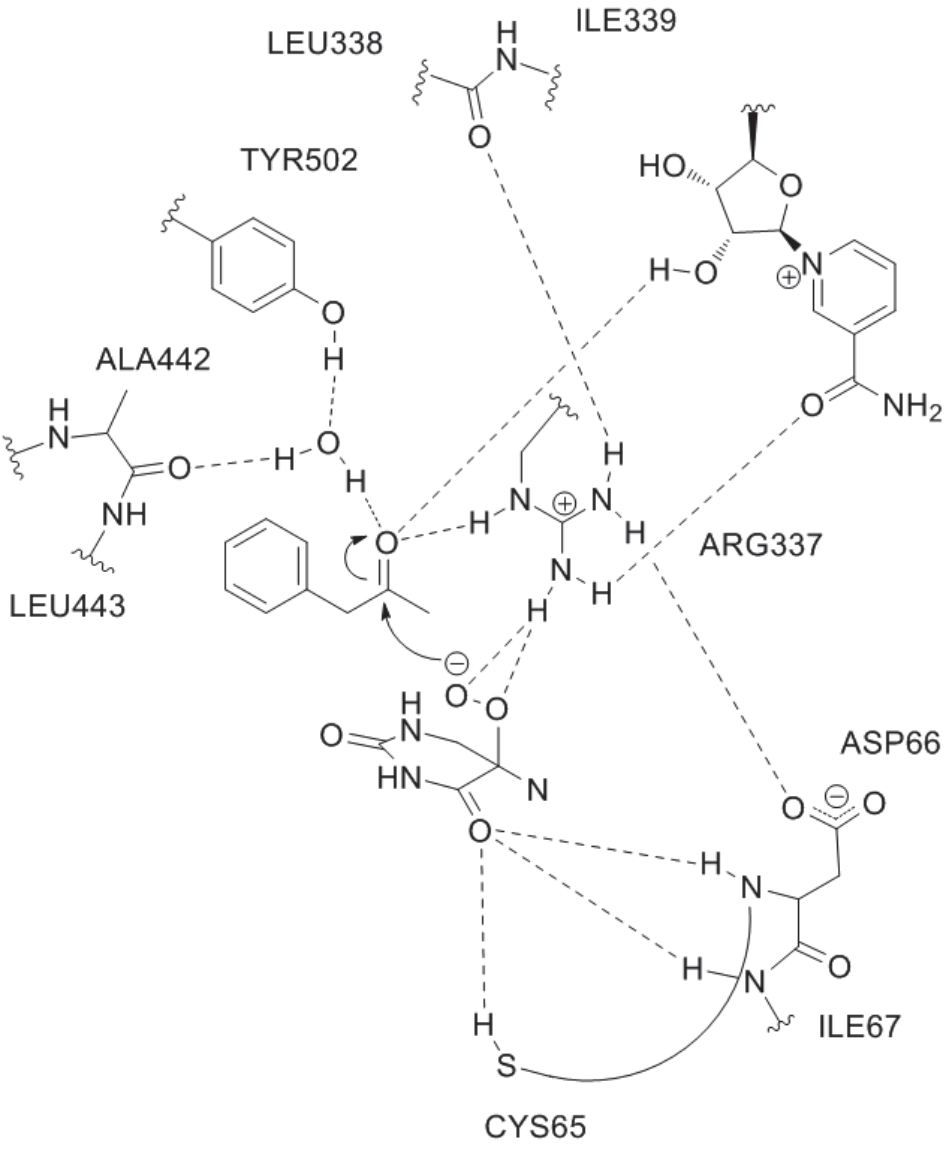


Figure 16: 2D representation of the residues and substrates in the active site of PAMO. For clarity, only a small part of the peroxyflavin moiety is represented. Hydrogen bonds are indicated by dashed lines.

Wild-type PAMO does not catalyze the Baeyer-Villiger reaction of substrates lacking the phenyl ring of phenylacetone. We supposed that π-cation interactions of the π-conjugated substrate with the guanidinium moiety of the ARG337 residue could be the cause for this

observation. To investigate this aspect, we ran a series of high-level QM computations in the gas phase on the guanidinium-toluene complex. In these tests, we replaced PHAC by toluene and ARG337 by its guanidinum moiety to directly assess the interactions of the π-conjugated ring with the cation. We adopted the optimized geometries obtained from the QM/MM scan in the protein for both the substrate and the residue. We added hydrogen atoms to satisfy the valence of these compounds and optimized their positions while keeping the rest of the molecule frozen. We performed single-point energy evaluations at each geometry along the scan at the MP2/aug-cc-pVTZ level, which is expected to model π-cation interactions faithfully. We also computed the energies of isolated toluene and guanidinium cation with the same method. Figure 17 shows the results from these computations. Apparently (see left side of Figure 17), when toluene and guanidinium approach each other in PAMO-derived geometries, the energy of isolated toluene does not change much, while the energies of isolated guanidinium and of the complex are lowered significantly in the transition state region (by ca. 2 kcal/mol) and still somewhat in the Criegee intermediate (by ca. 0.5 kcal/mol). The π-cation interaction energy along the scan is obtained as the difference of the energies of the complex and its two constituents. Evidently (see right side of Figure 17), it remains rather small throughout the scan (up to 0.3 kcal/mol) and even shows a drop when approaching the conformation corresponding to the transition state. We note that accounting for basis set superposition errors may change these results. Furthermore, the scan reported here did include the extra QM water molecule in the QM region, but not the ASP66 residue that has later been shown to be important (see above).

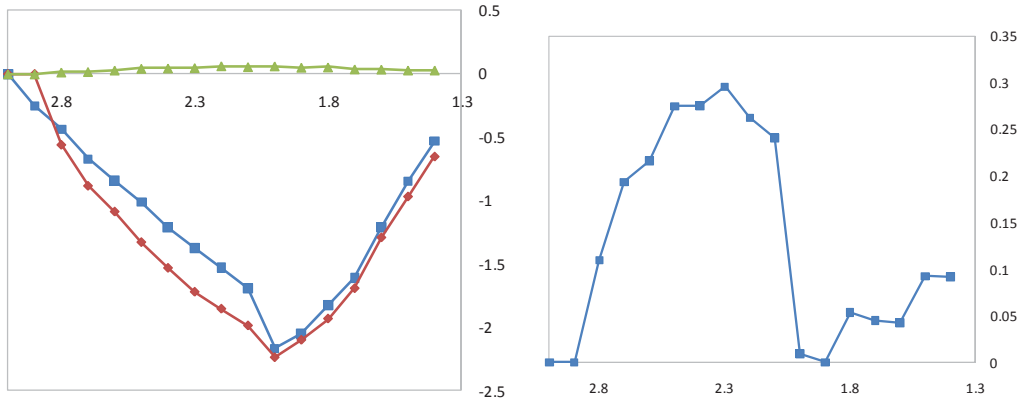


Figure 17: MP2/aug-cc-pVTZ energy profiles of the toluene-guanidinium complex at geometries taken from QM/MM optimizations of PAMO (see text). Left: energies of the complex (blue), toluene (green) and guanidinium (red). Right: complexation energy (difference between complex and constituents).

The toluene-guanidinium test calculations thus do not support a major role for substrate-ARG337 π-cation interactions in PAMO. They show, however, that the guanidinium part of ARG337 adopts a more favorable (more stable) conformation around the transition state leading to the Criegee intermediate. For further analysis, we now monitor the hydrogen bonds between ARG377 and its environment.

The guanidinium part of ARG337 can form hydrogen bonds with the substrate (PHAC), the distal (O87) and proximal (O86) oxygen atoms of the peroxyflavin moiety, ASP66, LEU338 (oxygen atom in the 338-339 peptide bond), and $NADP^+$. We monitored the corresponding hydrogen bond distances in a QM/MM scan obtained with the largest QM region considered presently (see above). The results are plotted in Figure 18.

Along the reaction path, ARG337 does not follow the "migrating oxygen": the ARG337-O87 distance increases from 1.68 to 2.58 Å when going from the reactant complex to the Criegee intermediate. The major part of this change takes place in the transition state region, at values of 2.0-1.9 Å for the reaction coordinate. At the same stage, the ARG336-O96 distance shrinks by ca. 0.1 Å, thus providing some compensation. Another prominent feature in Figure 18 is the continuous decrease of the hydrogen bond distance between ARG337 and the ketone moiety of PHAC along the reaction path. Due to geometrical constraints, it is inevitable that the substrate approaches ARG337, which leads to enhanced ARG337-PHAC hydrogen bonding and increasing stabilization along the reaction path towards the Criegee intermediate.

The hydrogen bonding effects of the other residues are less pronounced. We summarize them in Figure 19 by highlighting the most stabilizing influence of each hydrogen bond involving ARG337. It happens that every one of these hydrogen bonds favors the reaction. The hydrogen bond to FADO serves as a switch, those to PHAC and LEU338 mainly stabilize the Criegee intermediate, and those to NADP+ and ASP66 lower the transition state energy. All these effects are rather smaller individually, but they add up and thus result in an important role of ARG337 in facilitating the formation of the Criegee intermediate.

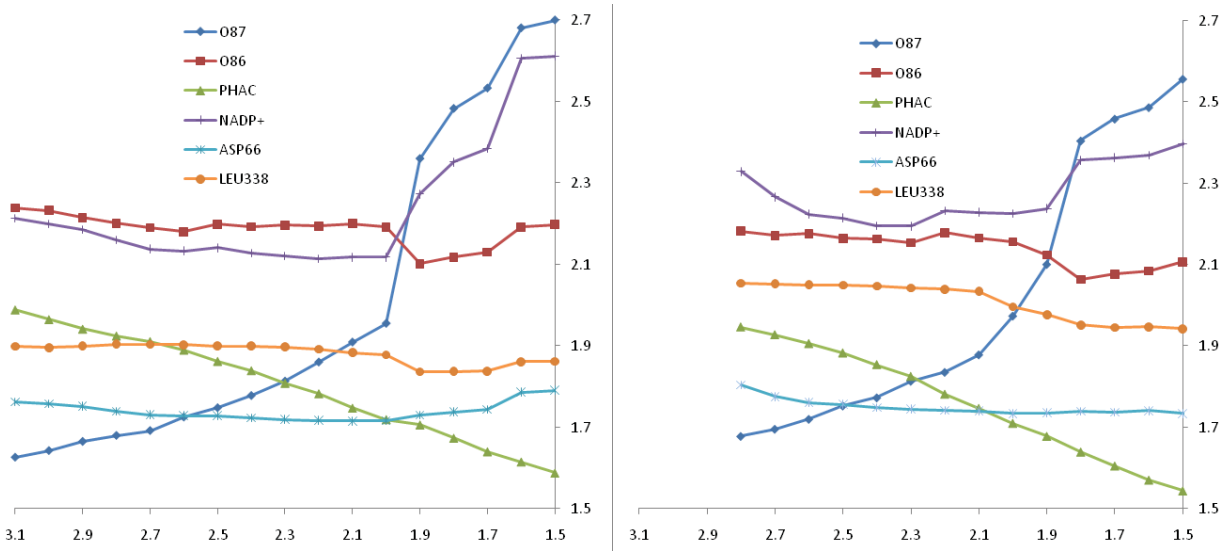


Figure 18: Hydrogen bond distances of ARG337 with the substrate and different active-site residues in PAMO along the reaction path to the Criegee intermediate, from QM(B3LYP/TZVP)/MM optimizations without (left) and with (right) D2 dispersion corrections (see text)

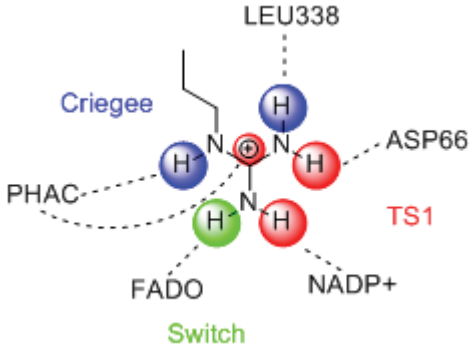


Figure 19: Effects of the different hydrogen bonds of ARG337. The most stabilizing influence of each hydrogen bond is indicated by a color code. See text for details.

3.2 Prebiotic Synthesis of Purines

This section summarizes the paper "Photochemical steps in the prebiotic synthesis of purine precursors from HCN", which is reprinted in the annex of this thesis.

Since its discovery in 1996 by Ferris and Orgel,[104] the oligomerization of four HCN molecules to cis-2,3-diaminomaleonitrile (cis-DAMN) followed by the rearrangement to 4-amino-1H-imidazole-5-carbonitrile (AICN) is considered as one of the most probable routes for the prebiotic synthesis of purines nucleobases and nucleotides.[105,106] Despite years of investigation, the mechanism is still unknown. We tackled the problem by using theoretical and computational methods. All proposed intermediates found in the literature[107-109] were

considered as well as some others that we could imagine. The investigated mechanism is depicted in Figure 20.

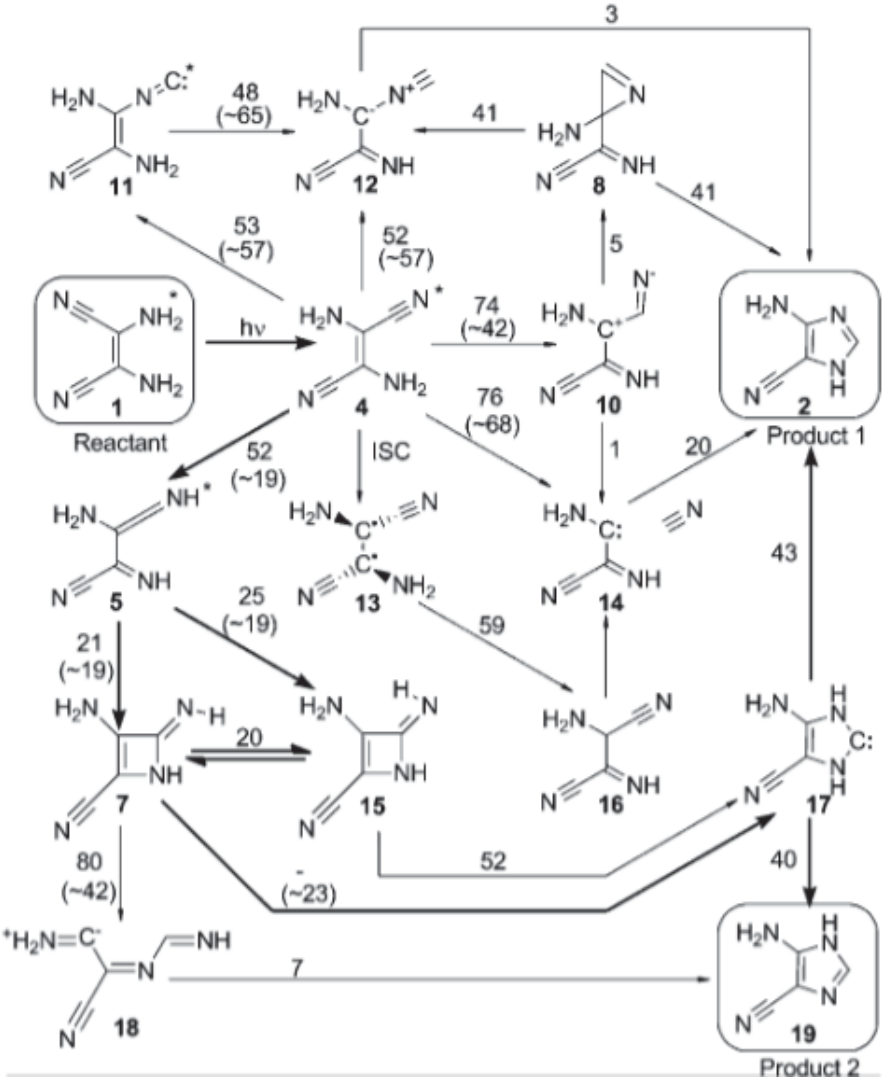


Figure 20: Intermediates considered in this study. Computed barriers in kcal/mol are given for each of the steps (arrows). Molecules with an asterisk can be electronically excited, and excited-state reaction energies are given in parenthesis. The determined pathway is highlighted by bold arrows.

As this oligomerization is known to be robust and to happen in any solvent as long as there is a high concentration of the HCN monomer, we studied the reactions in the gas phase. We used DFT as the QM method: B3LYP[95] for ground-state and CAM-B3LYP[110] for excited-state species, in combination with the aug-cc-pVTZ basis set[111] in both cases.

In terms of the kinetics, only one pathway emerged as possible. It is highlighted in Figure 20 with bold arrows. As suggested before,[112] it starts with a photoisomerization of the cis-DAMN molecule (**1**) into trans-DAMN (**4**) through a twisted conical intersection and without an energy

barrier. This isomer is subsequently excited, followed by a hydrogen transfer with a barrier of 19 kcal/mol leading to 2-amino-3-iminoacrylimidoyl cyanide (AIAC, **5** in Figure 20). Thereafter, an azetene intermediate (**7** or **15**) can be formed in the ground or excited state, which can then to rearrange to **17** or **18**. This latter step requires energy barriers which are too high in the ground state, and we thus suggest that both steps will occur in the excited state and go through the N-heterocyclic carbene **17**. In the presence of water or other protic solvents, this will lead to AICN. The potential energy profile of this pathway and the different excitations that are involved are summarized in Figure 21.

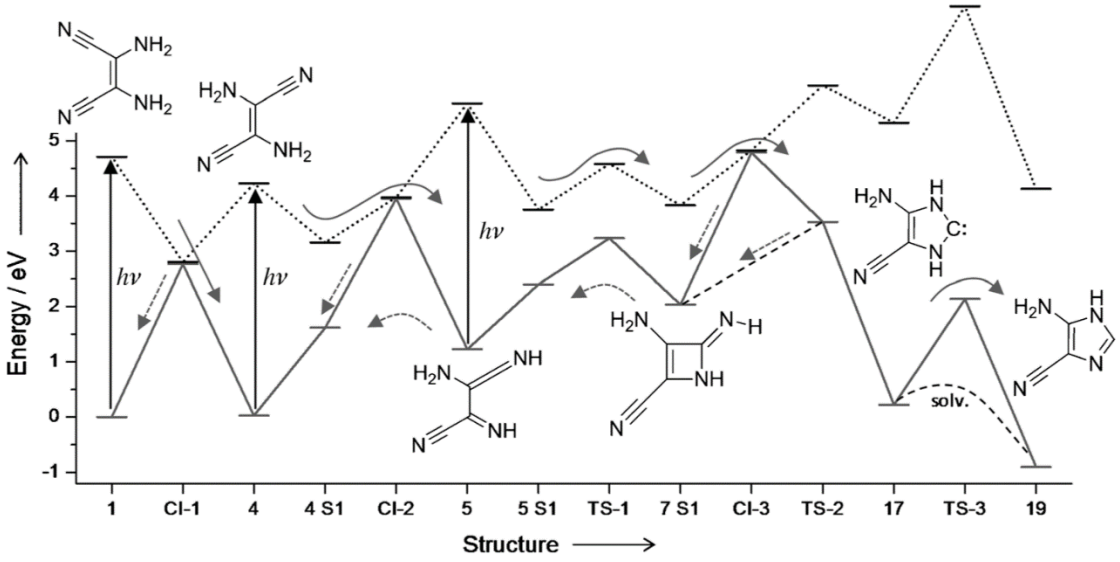


Figure 21: Free energy profile of the proposed mechanism with the individual intermediates and transition states. Excited-state steps are indicated by dashed lines. The reaction pathway is marked by arrows.

As the suggested mechanism involves several excitations and internal conversions, other possibilities in terms of reactivity are conceivable. In particular, upon relaxation from the excited to the ground state, the excess energy goes into the vibrational modes, and the molecule is formed in a so-called "hot" ground state, which might undergo reactions that are normally inaccessible under standard condition of temperature and pressure. For example, when considering the possible transformations of trans-DAMN (**4**) in Figure 20, compounds **10** to **13** could be formed by hot ground-state reactions despite high energy barriers. This could, of course, only happen if the molecule stays in this hot ground state long enough for one of these reactions to occur. In solution the excess energy could be transferred to the solvent in a competing process. This energy dissipation is still an open issue, and we decided to study it for the case of trans-DAMN (**4**) to provide evidence for or against its relevance in the proposed mechanism.

Our aim was to determine the time for which trans-DAMN will stay in the hot ground state while being solvated in water. For this purpose, water needs to be represented explicitly in order to take into account the different modes (vibration, rotation, translation) that may be involved in energy dissipation. Since the use of a QM-only approach is not feasible computationally for a system including enough water to represent many solvation shells, we performed QM/MM molecular dynamics simulations to tackle this problem.

The structure of trans-DAMN embedded in a first solvation shell of nine water molecules was first optimized at the B3LYP/6-31G** level of theory.[95] This cluster was then solvated in a 100x100x100 Å^3 box of MM water molecules described by the TIP3P water model.[113] MM parameters for trans-DAMN were assembled from those provided in the CGenFF force field.[114] The system was progressively heated to 300 K, and a 1 ns MD run was performed in the NPT ensemble with the micro-solvated trans-DAMN molecule frozen in the center of the box in order to maintain its initial geometry. From the final structure of this MM MD simulation, we extracted the embedded cluster and all water molecules with an oxygen atom within 48 Å of the center of the system. This structure served as the starting point of QM/MM MD simulations.

For the QM/MM computations, we used the semiempirical OM2[115] method to represent trans-DAMN and the nine surrounding water molecules. The system was equilibrated again for 500 ps with a time step of 1 fs to establish a temperature of 300 K. Thereafter, we performed two simulations in the NVE ensemble. The first one was started with the velocities acquired by each atom at the end of the heating process, while the second one employed modified velocities on the atoms of trans-DAMN to include the excess photoenergy transferred to its vibrational modes. This was done by conserving the direction of the initial velocities for each atom of trans-DAMN while modifying their norms to incorporate an excess of 4 eV of photoenergy plus the ground-state zero-point vibrational energy of 2.25 eV. Four different trajectories (with different energy distributions along the modes) were run and they all showed the same tendency.

The ratio between the kinetic energies during the two MD simulations is shown in the upper panel of Figure 22. Evidently, the energy dissipation occurs extremely fast and the thermal equilibrium is reached after only 2 ps. Two distinct dissipation steps can be seen: the first one is an extremely fast transfer of about one third of the total excess energy to the neighboring water molecules occurring on the same time scale as a few N-H stretching oscillations; the

second one takes somewhat longer and dissipates the remaining two thirds of the excess photoenergy within about 2 ps.

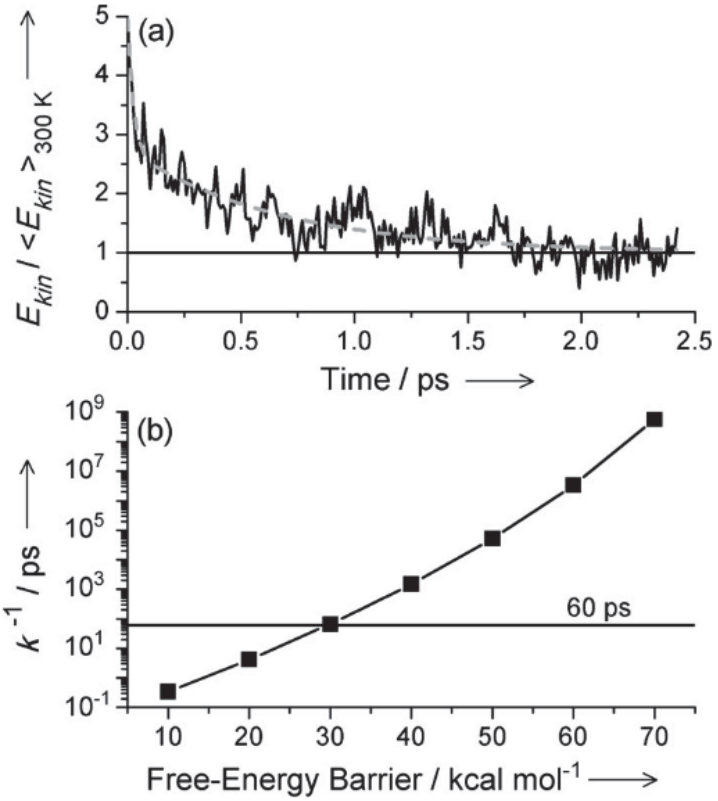


Figure 22: a) Ratio of the kinetic energy of trans-DAMN in a MD simulation of the hot ground state and a standard MD run at 300K (see text); b) inverse of the unimolecular reaction rate as a function of the free-energy barrier.

To determine if this fast energy dissipation allows for a competing hot ground-state reaction to occur, we focus on the first step of the dissipation taking place within 0.2 ps. From an evaluation of the available experimental data[116] we can conclude that there are on average 300 excitation and relaxation processes preceding the reaction, which translates into a 60 ps timescale for the reaction to occur. Given this value and the computed excess energy of 4 eV, we can use the Rice-Ramsperger-Kassel-Marcus (RRKM)[117-119] approach to estimate the unimolecular rate constant *k*(E). The density and the number of states were determined by the Beyer-Swinehart direct count method[120] on the basis of the computed harmonic frequencies for the reactant and the relevant transition state. By estimating *k*(E) for several values of the free energy barrier we find that a barrier of at most 30 kcal/mol can be overcome under these conditions (see Figure 22b). This value is significantly smaller than the computed barriers for all alternative ground-state reactions, which are all above 50 kcal/mol. Therefore we can rule out the possibility that hot ground-state reactions invalidate the proposed mechanism, on the

basis of QM/MM MD simulations that allow us to estimate the time required for energy dissipation in water.

4. Conclusion

In this thesis we have focused on QM/MM methods. Development work was done both on QM/MM methodology and QM/MM-related tools. The main project was the development and validation of a polarizable embedding scheme for the Drude oscillator force field. We combined this polarizable force field with boundary potentials for efficiency and inclusion of long-range electrostatic effects, and we validated the performance of the resulting three-layer scheme for an enzymatic reaction. We carried out the first fully polarized QM/MM simulations with a well parameterized polarizable force field. We contributed to the development of a microiterative scheme for intrinsic reaction coordinate computations for large QM/MM systems and of a dual Hamiltonian free energy perturbation QM/MM method that combines high-level energy evaluations with low-level MD simulations.

We performed one standard application to study the enzymatic reaction catalyzed by phenylacetone monooxygenase and another less conventional application to assess the energy dissipation in solution of a hot ground state after relaxation from an electronically excited state, on a topic relevant to prebiotic chemistry.

Other ongoing work not covered in this thesis addresses periodic boundary conditions and adaptive partitionings for QM/MM systems (with Tatiana Vasilevskaya), a metadynamics study of cellulose conformations with the aim of explaining its hydrolysis (with Claudia Loerbroks), and three-layer QM/MM/coarse-grained force field approaches (with Pandian Sokkar).

References

(1) Senn, H. M.; Thiel, W. *Angewandte Chemie International Edition* **2009**, *48*, 1198.

(2) Boulanger, E.; Thiel, W. *Journal of Chemical Theory and Computation* **2014**, *10*, 1795.

(3) Boulanger, E.; Thiel, W. *Journal of Chemical Theory and Computation* **2012**, *8*, 4527.

(4) Polyak, I.; Boulanger, E.; Sen, K.; Thiel, W. *Physical Chemistry Chemical Physics* **2013**, *15*, 14188.

(5) Polyak, I.; Benighaus, T.; Boulanger, E.; Thiel, W. *The Journal of Chemical Physics* **2013**, *139*, 064105.

(6) Boulanger, E.; Anoop, A.; Nachtigallova, D.; Thiel, W.; Barbatti, M. *Angewandte Chemie International Edition* **2013**, *52*, 8000.

(7) Ponder, J. W.; Case, D. A. *Advances in Protein Chemistry* **2003**, *66*, 27.

(8) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*; Academic press: San Diego, 2001; Vol. 1.

(9) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R.; Evanseck, J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.; Mattos, C.; Michnik, S.; Ngo, T.; Nguyen, D.; Prodhom, B.; Reiher, W.; Roux, B.; Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *The Journal of Physical Chemistry B* **1998**, *102*, 3586.

(10) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *Journal of Computational Chemistry* **1983**, *4*, 187.

(11) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Courier Dover Publications: New York, 2012.

(12) Møller, C.; Plesset, M. S. *Physical Review* **1934**, *46*, 618.

(13) Purvis III, G. D.; Bartlett, R. J. *The Journal of Chemical Physics* **1982**, *76*, 1910.

(14) Riplinger, C.; Neese, F. *The Journal of Chemical Physics* **2013**, *138*, 034106.

(15) Krause, C.; Werner, H.-J. *Physical Chemistry Chemical Physics* **2012**, *14*, 7591.

(16) Hohenberg, P.; Kohn, W. *Physical review* **1964**, *136*, B864.

(17) Kohn, W.; Sham, L. J. *Physical Review* **1965**, *140*, A1133.

(18) Dewar, M. J.; Thiel, W. *Journal of the American Chemical Society* **1977**, *99*, 4899.

(19) Senn, H. M.; Thiel, W. *Current Opinion in Chemical Biology* **2007**, *11*, 182.

(20) Warshel, A. *Annual Review of Biophysics and Biomolecular Structure* **2003**, *32*, 425.

(21) Schaefer, P.; Riccardi, D.; Cui, Q. *The Journal of Chemical Physics* **2005**, *123*, 014905.

(22) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H. *The Journal of Physical Chemistry B* **2006**, *110*, 6458.

(23) Friesner, R. A.; Guallar, V. *Annual Review of Physical Chemistry* **2005**, *56*, 389.

(24) Lin, H.; Truhlar, D. G. *Theoretical Chemistry Accounts* **2007**, *117*, 185.

(25) Warshel, A.; Levitt, M. *Journal of Molecular Biology* **1976**, *103*, 227.

(26) Warshel, A. *Angewandte Chemie International Edition* **2014**, *DOI: 10.1002/anie.201403689*.

(27) Levitt, M. *Angewandte Chemie International Edition* **2014**, *DOI: 10.1002/anie.201403691*.

(28) Karplus, M. *Angewandte Chemie International Edition* **2014**, *DOI: 10.1002/anie.201403924*.

(29) Bakowies, D.; Thiel, W. *The Journal of Physical Chemistry* **1996**, *100*, 10580.

(30) Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. *The Journal of Physical Chemistry A* **2000**, *104*, 1720.

(31) Ewald, P. P. *Annalen der Physik* **1921**, *369*, 253.

(32) Beglov, D.; Roux, B. *The Journal of chemical physics* **1994**, *100*, 9050.

(33) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *Journal of Chemical Theory and Computation* **2006**, *2*, 1370.
(34) Im, W.; Berneche, S.; Roux, B. *The Journal of Chemical Physics* **2001**, *114*, 2924.
(35) Benighaus, T.; Thiel, W. *Journal of Chemical Theory and Computation* **2008**, *4*, 1600.
(36) Benighaus, T.; Thiel, W. *Journal of Chemical Theory and Computation* **2009**, *5*, 3114.
(37) Benighaus, T.; Thiel, W. *Journal of Chemical Theory and Computation* **2010**, *7*, 238.
(38) Warshel, A.; Kato, M.; Pisliakov, A. V. *Journal of Chemical Theory and Computation* **2007**, *3*, 2034.
(39) Antila, H. S.; Salonen, E. In *Biomolecular Simulations*; Springer: Heidelberg, 2013, p 215.
(40) Lopes, P. E.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell Jr, A. D. *Journal of Chemical Theory and Computation* **2013**, *9*, 5430.
(41) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. *Journal of Chemical Theory and Computation* **2013**, *9*, 4046.
(42) Lamoureux, G.; MacKerell Jr, A. D.; Roux, B. *The Journal of Chemical Physics* **2003**, *119*, 5185.
(43) Kaminski, G. A.; Jorgensen, W. L. *The Journal of Physical Chemistry B* **1998**, *102*, 1787.
(44) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *The Journal of Physical Chemistry A* **2004**, *108*, 621.
(45) Ren, P.; Ponder, J. W. *The Journal of Physical Chemistry B* **2003**, *107*, 5933.
(46) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A. *The Journal of Physical Chemistry B* **2010**, *114*, 2549.
(47) Rick, S. W.; Stuart, S. J.; Bader, J. S.; Berne, B. *Studies in Physical and Theoretical Chemistry* **1995**, *83*, 31.
(48) Patel, S.; Brooks, C. L. *Journal of Computational Chemistry* **2004**, *25*, 1.
(49) Patel, S.; Mackerell, A. D.; Brooks, C. L. *Journal of Computational Chemistry* **2004**, *25*, 1504.
(50) Lamoureux, G.; Roux, B. *The Journal of Chemical Physics* **2003**, *119*, 3025.
(51) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. *Journal of Chemical Theory and Computation* **2005**, *1*, 153.
(52) Jiang, W.; Hardy, D. J.; Phillips, J. C.; MacKerell Jr, A. D.; Schulten, K.; Roux, B. *The Journal of Physical Chemistry Letters* **2010**, *2*, 87.
(53) Thole, B. T. *Chemical Physics* **1981**, *59*, 341.
(54) Van Duijnen, P. T.; Swart, M. *The Journal of Physical Chemistry A* **1998**, *102*, 2399.
(55) Nüsslein, V.; Schröder, U. *Physica Status Solidi (b)* **1967**, *21*, 309.
(56) Schröder, U. *Solid State Communications* **1993**, *88*, 1049.
(57) De Leeuw, N.; Parker, S. *Physical Review B* **1998**, *58*, 13901.
(58) Yu, H.; Hansson, T.; van Gunsteren, W. F. *The Journal of Chemical Physics* **2003**, *118*, 221.
(59) Yu, H.; van Gunsteren, W. F. *Computer Physics Communications* **2005**, *172*, 69.
(60) Geerke, D. P.; van Gunsteren, W. F. *The Journal of Physical Chemistry B* **2007**, *111*, 6425.
(61) Geerke, D. P.; van Gunsteren, W. F. *Journal of Chemical Theory and Computation* **2007**, *3*, 2128.
(62) Straatsma, T.; McCammon, J. *Molecular Simulation* **1990**, *5*, 181.
(63) Baker, C. M.; Anisimov, V. M.; MacKerell Jr, A. D. *The Journal of Physical Chemistry B* **2010**, *115*, 580.
(64) Baker, C. M.; MacKerell Jr, A. D. *Journal of Molecular Modeling* **2010**, *16*, 567.
(65) Geerke, D. P.; Thiel, S.; Thiel, W.; van Gunsteren, W. F. *Journal of Chemical Theory and Computation* **2007**, *3*, 1499.
(66) Lu, Z.; Zhang, Y. *Journal of Chemical Theory and Computation* **2008**, *4*, 1237.

(67) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *Journal of Molecular Structure: THEOCHEM* **2003**, *632*, 1.

(68) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Molecular Physics* **1996**, *87*, 1117.

(69) Kolafa, J. *Journal of Computational Chemistry* **2004**, *25*, 335.

(70) Kolafa, J. *The Journal of Chemical Physics* **2005**, *122*, 164105.

(71) Kast, P.; Asif-Ullah, M.; Hilvert, D. *Tetrahedron Letters* **1996**, *37*, 2691.

(72) Martí, S.; Moliner, V.; Tuñón, I.; Williams, I. H. *Organic & Biomolecular Chemistry* **2003**, *1*, 483.

(73) Illingworth, C.; Parkes, K.; Snell, C.; Marti, S.; Moliner, V.; Reynolds, C. *Molecular Physics* **2008**, *106*, 1511.

(74) Claeyssens, F.; Ranaghan, K. E.; Lawan, N.; Macrae, S. J.; Manby, F. R.; Harvey, J. N.; Mulholland, A. J. *Organic & Biomolecular Chemistry* **2011**, *9*, 1578.

(75) Ishida, K.; Morokuma, K.; Komornicki, A. *The Journal of Chemical Physics* **1977**, *66*, 2153.

(76) Hratchian, H. P.; Schlegel, H. B. *The Journal of Chemical Physics* **2004**, *120*, 9918.

(77) Hratchian, H.; Schlegel, H. *Journal of Chemical Theory and Computation* **2005**, *1*, 61.

(78) Turner, A.; Williams, I. *Physical Chemistry Chemical Physics* **1999**, *1*, 1323.

(79) Bofill, J. M. *Journal of Computational Chemistry* **1994**, *15*, 1.

(80) Zwanzig, R. W. *The Journal of Chemical Physics* **1954**, *22*, 1420.

(81) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *The Journal of Chemical Physics* **1953**, *21*, 1087.

(82) Torrie, G. M.; Valleau, J. P. *Chemical Physics Letters* **1974**, *28*, 578.

(83) Kirkwood, J. G. *The Journal of Chemical Physics* **1935**, *3*, 300.

(84) Zhang, Y.; Liu, H.; Yang, W. *The Journal of Chemical Physics* **2000**, *112*, 3483.

(85) Polyak, I.; Reetz, M. T.; Thiel, W. *Journal of the American Chemical Society* **2012**, *134*, 2732.

(86) Leisch, H.; Morley, K.; Lau, P. C. *Chemical Reviews* **2011**, *111*, 4165.

(87) Reetz, M. T.; Wu, S. *Journal of the American Chemical Society* **2009**, *131*, 15424.

(88) Reetz, M. T. *Journal of the American Chemical Society* **2013**, *135*, 12480.

(89) Yachnin, B. J.; Sprules, T.; McEvoy, M. B.; Lau, P. C.; Berghuis, A. M. *Journal of the American Chemical Society* **2012**, *134*, 7788.

(90) Orru, R.; Dudek, H. M.; Martinoli, C.; Pazmiño, D. E. T.; Royant, A.; Weik, M.; Fraaije, M. W.; Mattevi, A. *Journal of Biological Chemistry* **2011**, *286*, 29284.

(91) Bussi, G.; Laio, A.; Parrinello, M. *Physical Review Letters* **2006**, *96*, 090601.

(92) Becke, A. D. *Physical Review A* **1988**, *38*, 3098.

(93) Perdew, J. P. *Physical Review B* **1986**, *33*, 8822.

(94) Lee, C.; Yang, W.; Parr, R. G. *Physical Review B* **1988**, *37*, 785.

(95) Becke, A. D. *The Journal of Chemical Physics* **1993**, *98*, 5648.

(96) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Physical Review Letters* **1996**, *77*, 3865.

(97) Adamo, C.; Barone, V. *The Journal of Chemical Physics* **1999**, *110*, 6158.

(98) Schäfer, A.; Horn, H.; Ahlrichs, R. *The Journal of Chemical Physics* **1992**, *97*, 2571.

(99) Schäfer, A.; Huber, C.; Ahlrichs, R. *The Journal of Chemical Physics* **1994**, *100*, 5829.

(100) Roitberg, A.; Elber, R. *The Journal of Chemical Physics* **1991**, *95*, 9277.

(101) Scholten, M. Ph.D. thesis, Universität Düsseldorf, 2003.

(102) Otte, N.; Scholten, M.; Thiel, W. *The Journal of Physical Chemistry A* **2007**, *111*, 5751.

(103) Grimme, S. *Journal of Computational Chemistry* **2006**, *27*, 1787.

(104) Ferris, J. P.; Orgel, L. *Journal of the American Chemical Society* **1966**, *88*, 1074.

(105) Barks, H. L.; Buckley, R.; Grieves, G. A.; Di Mauro, E.; Hud, N. V.; Orlando, T. M. *ChemBioChem* **2010**, *11*, 1240.

(106) Al-Azmi, A.; Elassar, A.-Z. A.; Booth, B. L. *Tetrahedron* **2003**, *59*, 2749.

(107) Bigot, B.; Roux, D. *The Journal of Organic Chemistry* **1981**, *46*, 2872.

(108) Becker, R. S.; Kolc, J.; Rotham, W. *Journal of the American Chemical Society* **1973**, *95*, 1269.

(109) Yamada, Y.; Nagashima, N.; Iwashita, Y.; Nakamura, A.; Kumashiro, I. *Tetrahedron Letters* **1968**, *9*, 4529.

(110) Yanai, T.; Tew, D. P.; Handy, N. C. *Chemical Physics Letters* **2004**, *393*, 51.

(111) Dunning Jr, T. H. *The Journal of Chemical Physics* **1989**, *90*, 1007.

(112) Ferris, J. P.; Orgel, L. *Journal of the American Chemical Society* **1965**, *87*, 4976.

(113) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of Chemical Physics* **1983**, *79*, 926.

(114) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I. *Journal of Computational Chemistry* **2010**, *31*, 671.

(115) Weber, W.; Thiel, W. *Theoretical Chemistry Accounts* **2000**, *103*, 495.

(116) Koch, T. H.; Rodehorst, R. M. *Journal of the American Chemical Society* **1974**, *96*, 6707.

(117) Baercor, T.; Mayerfn, P. M. *Journal of the American Society for Mass Spectrometry* **1997**, *8*, 103.

(118) Rice, O. K.; Ramsperger, H. C. *Journal of the American Chemical Society* **1927**, *49*, 1617.

(119) Kassel, L. S. *The Journal of Physical Chemistry* **1928**, *32*, 225.

(120) Beyer, T.; Swinehart, D. *Communications of the ACM* **1973**, *16*, 379.

Solvent boundary potentials for hybrid QM/MM computations using classical Drude oscillators: a fully polarizable model.

Eliot Boulanger and Walter Thiel

J. Chem. Theory Comput. **2012**, 8, 4527-4538.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Solvent Boundary Potentials for Hybrid QM/MM Computations Using Classical Drude Oscillators: A Fully Polarizable Model

Eliot Boulanger and Walter Thiel*

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

ABSTRACT: Accurate quantum mechanical/molecular mechanical (QM/MM) treatments should account for MM polarization and properly include long-range electrostatic interactions. We report on a development that covers both these aspects. Our approach combines the classical Drude oscillator (DO) model for the electronic polarizability of the MM atoms with the generalized solvent boundary Potential (GSBP) and the solvated macromolecule boundary potential (SMBP). These boundary potentials (BP) are designed to capture the long-range effects of the outer region of a large system on its interior. They employ a finite difference approximation to the Poisson−Boltzmann equation for computing electrostatic interactions and take into account outer-region bulk solvent through a polarizable dielectric continuum (PDC). This approach thus leads to fully polarizable three-layer QM/MM-DO/BP methods. As the mutual responses of each of the subsystems have to be taken into account, we propose efficient schemes to converge the polarization of each layer simultaneously. For molecular dynamics (MD) simulations using GSBP, this is achieved by considering the MM polarizable model as a dynamical degree of freedom, and hence contributions from the boundary potential can be evaluated for a frozen state of polarization at every time step. For geometry optimizations using SMBP, we propose a dual self-consistent field approach for relaxing the Drude oscillators to their ideal positions and converging the QM wave function with the proper boundary potential. The chosen coupling schemes are evaluated with a test system consisting of a glycine molecule in a water ball. Both boundary potentials are capable of properly reproducing the gradients at the inner-region atoms and the Drude oscillators. We show that the effect of the Drude oscillators must be included in all terms of the boundary potentials to obtain accurate results and that the use of a high dielectric constant for the PDC does not lead to a polarization catastrophe of the DO models. Optimum values for some key parameters are discussed. We also address the efficiency of these approaches compared to standard QM/MM-DO calculations without BP. In the SMBP case, computation times can be reduced by around 40% for each step of a geometry optimization, with some variation depending on the chosen QM method. In the GSBP case, the computational advantages of using the boundary potential increase with system size and with the number of MD steps.

1. INTRODUCTION

Hybrid quantum mechanical/molecular mechanical (QM/MM) methods have become reliable tools for studying chemical reactions in large biomolecules.[1−7] Already the first such study[8] considered the embedding of a QM subsystem in a polarizable MM environment to be important for the proper description of enzymes. Since then, polarizable force fields (PFFs) have undergone much development and are now approaching maturity. They are being used increasingly in biomolecular simulations, and highly optimized PFFs for such applications are expected to be available soon.[9,10]

There are several ways to simulate the polarizability of MM atoms.[9−11] These include induced dipoles (or multipoles),[12−14] fluctuating charges,[15−19] and classical Drude oscillators (DOs).[20−26] The DO approach is also called the shell model[27−29] or charge-on-spring (COS) model.[11,30−33] In our early work, we represented MM polarization by induced dipoles at the MM atoms.[34] More recently, we have adopted the DO (COS) model, in view of its inherent simplicity and its widespread use in PFF development.[23,35−38] The model consists of a mobile charge linked to a polarizable MM atom by a spring; a charge of the same magnitude and opposite sign is added at the nucleus of this atom, so that these two virtual charges form a dipole. The mobile charge moves in response to the electrostatic interactions with the environment, thus simulating MM polarizability. We have included the DO model in the QM/MM ChemShell software using the GROMOS COS force field.[39] Other interfacing methods have been discussed for the CHARMM DO force field.[40,41]

In a QM/MM framework, the influence of MM polarization should be especially important for processes that involve charged or very polar species and significant charge relocation. In these cases, long-range electrostatic interactions are also expected to play a prominent role.[42] This calls for the development of treatments that cover both these aspects in a balanced manner.

Long-range electrostatic interactions can be taken into account in classical MM simulations using several well established techniques. At the QM/MM level, two such techniques have been implemented by different groups, namely periodic boundary conditions and boundary potentials. Periodic boundary conditions have been applied using Ewald summation.[43−45] This approach will require huge unit cells in biomolecular work (containing the large nonperiodic biomolecule and a solvent environment of sufficient size) and may thus be less practical at the QM/MM level. The alternative boundary potential approach circumvents this problem by considering only a restricted number of atoms explicitly and

Received: August 16, 2012
Published: October 12, 2012

representing the distant environment by a continuum model. To be more specific, it splits the system into an explicit inner region (including the QM subsystem as well as the adjacent MM part of the macromolecule) and an implicit outer region (consisting of the distant MM atoms of the macromolecule and the bulk solvent). The long-range electrostatic effects of the outer region are captured by a boundary potential that represents the influence of the discrete MM charges in this region and of the bulk solvent treated as a polarizable dielectric continuum (PDC). This method is well suited for describing localized process such as those commonly studied with QM/MM methods.

The generalized solvent boundary potential (GSBP)[46−48] and the solvated macromolecule boundary potential (SMBP)[49,50] allow the use of irregularly shaped dielectric boundaries between the macromolecule and bulk solvent. This feature is important in QM/MM computations of enzymes as the protein and the surrounding water possess very different dielectric constants.[51] GSBP was originally designed for use in molecular dynamics (MD) simulations. At the QM/MM level, it has up to now only been interfaced with semiempirical methods. SMBP was developed for geometry optimizations with any kind of QM method and can thus be used to compute QM/MM potential energy surfaces (PESs) also with high-level QM methods. GSBP and SMBP complement each other in QM/MM free energy perturbation (FEP) calculations of free energy differences.[52,53] These can be used to estimate the entropic contribution of the environment by sampling over the MM degrees of freedom.

Polarizable force fields have already been interfaced with PDC models, with results that are promising in terms of accuracy,[54−59] and excited-state properties of small molecules have been studied by embedding the solute molecule (QM) in a few explicit polarizable solvent molecules (PFF) and a bulk solvent (PDC).[56,60] In our treatment, we combine a QM part that is intrinsically polarizable, an explicit MM region described by a PFF, and a solvent represented by a PDC, which leads to a fully polarizable three-layer model. We also note that boundary potentials have the general advantage of being quite efficient compared with full QM/MM treatments, since they consider only a small part of the system explicitly. This reduced size of the explicit MM region can be particularly beneficial when using PFFs which normally require some kind of iterative scheme to determine the proper MM polarization.

The purpose of this paper is to present combination schemes for Drude oscillators with GSBP and SMBP in order to obtain a fully polarizable three-layer QM/MM-DO/BP model at a reasonable computational cost.

2. THEORY

The system under study is separated into different spatial regions. As usual in QM/MM methods, there is a central QM region surrounded by an MM region. The latter is further partitioned into an explicit inner part (treated atomistically at the MM level) and an implicit outer part (represented by the boundary potential). All these subsystems interact with each other, and the total energy is given by the following additive expression:

$$E_{\mathrm{tot}} = E_{\mathrm{QM}} + E_{\mathrm{MM}} + E_{\mathrm{BP}} + E_{\mathrm{QM/MM}} + E_{\mathrm{QM/BP}} + E_{\mathrm{MM/BP}} \tag{1}$$

In our implementation in the ChemShell package,[61,62] E_{QM} is the energy of the QM part obtained with any available QM method, and E_{MM} is the energy obtained from any MM force field function interfaced with the program. E_{BP} denotes the energy contribution from the boundary potential. The three last terms arise from the interactions between the subsystems. In the following, we describe the interactions that occur when using the polarizable DO force field for the MM part and the GSBP or SMBP for the boundary potential.

2.1. Drude Oscillators in a QM/MM Framework. The DO model aims at simulating the electronic polarizability at the MM level. It represents the induced dipole at every polarizable atom by two charges of the same magnitude (q) and opposite sign linked by a harmonic spring. The first charge is located at the nucleus of the atom, while the second one is mobile. Polarization arises from the competition between the forces acting on the mobile charge, which are due to the spring and the electrostatic interactions with the environment. The optimum position (d) of the mobile charge (Drude particle) is obtained by requiring that these two forces compensate each other.

$$\frac{\partial(U_{\mathrm{spring}} + U_{\mathrm{elec}})}{\partial d} = 0 \tag{2}$$

The electrostatic potential energy U_{elec} is obtained by summing over all point-charge interactions applying Coulomb's law. The potential energy U_{spring} of the harmonic spring is evaluated using a force constant (k_d) that is generally defined in terms of the polarizability (α) of the corresponding atom:

$$k_d = \frac{q^2}{\alpha} \tag{3}$$

In the CHARMM force field, k_d is always fixed to 1000 kcal mol^{-1} Å^{-2} in order to maintain a small d value and to keep the point dipole approximation valid.[26] The implementation of the DO terms within a standard force field involves only modifications in the electrostatic part of the MM potential energy function.

$$E_{\mathrm{MM}}^{\mathrm{elec}} = \sum_i \sum_{j>i} \frac{q_i q_j}{r_{ij}} + \sum_i \sum_{j'} \left(\frac{q_i q_{j'}}{r_{ij'}} - \frac{q_i q_{j'}}{r_{ij}} \right) + \sum_{i'} \sum_{j'>i'} \left(\frac{q_{i'} q_{j'}}{r_{i'j'}} + \frac{q_{i'} q_{j'}}{r_{ij}} \right) - \sum_{i'} \sum_{j'} \left(\frac{q_{i'} q_{j'}}{r_{ij'}} \right) + \frac{1}{2} \sum_{i'} k_{d,i'} d_{i'}^2 \tag{4}$$

Here, indices i and j run over atoms, the prime denotes a DO term, and r is the distance between the two corresponding centers. In the CHARMM force field, several other terms have been included in an attempt to properly simulate the electronic distribution and the dipole response of a molecule.[63] The 1−2 and 1−3 interactions between Drude particles (at atoms located one or two covalent bonds away from each other) are screened by applying the Thole function:[64]

$$1 - \left(1 + \frac{T_{ij}}{2}\right) \exp(-T_{ij}) \tag{5}$$

dx.doi.org/10.1021/ct300722e | J. Chem. Theory Comput. 2012, 8, 4527−4538

$$T_{ij} = t\left(\frac{r_{ij}}{\sqrt[6]{\alpha_i \alpha_j}}\right) \tag{6}$$

where t is the Thole parameter. Interactions at larger distances (with three or more bonds in between) are evaluated without any screening. Also, the charge of a heteroatom (without the DO contribution) can be split and partly located at one or two nearby positions (fixed in terms of internal coordinates) to represent lone pairs. The simultaneous presence of such lone pair charges and Drude particles may allow the simulation of anisotropic polarizability.[65] Another correction has been introduced to avoid the so-called polarization catastrophe, which is caused by strong Coulombic interactions at close distances that may lead to an excessive drift of the Drude particle.[11] The associated anharmonic hyperpolarizability damping term is taken into account only if d is higher than a predefined limit (typically 0.2 Å).[24]

The implementation of the DO model within a QM/MM framework has been discussed for GROMOS (COS model)[39] and more recently also for CHARMM.[40] In both cases, the Drude oscillators are included in the QM computation by modifying the one-electron terms in the Fock matrix, in complete analogy to the classical MM point charges. The Drude particle gives rise to extra one-electron terms, while the compensating charge at the nucleus of the polarizable atom is taken into account by adjusting the corresponding atomic charge.

To determine the position of the Drude particle, and thus obtain the induced dipole, the left-hand side of eq 2 must be minimized. The required electrostatic potential is evaluated from the electric field (ϕ) at the position of the Drude particle, which is composed of QM, MM, and DO contributions. The force (F) exerted on Drude particle i' by a given component of the electric field can be written as

$$F_{d_{i'}} = \frac{\alpha_{i'}}{q_{i'}}(\varphi_{i'}^{\mathrm{MM}} + \varphi_{i'}^{\mathrm{QM}} + \varphi_{i'}^{\mathrm{DO}}) \tag{7}$$

The contributions $\varphi_{i'}^{\mathrm{MM}}$, $\varphi_{i'}^{\mathrm{QM}}$ and $\varphi_{i'}^{\mathrm{DO}}$ to the electric field are interdependent, and their computation thus calls for a dual self-consistent-field (SCF) approach[39] as indicated in Figure 1. We followed the same implementation as in our previous work[39]

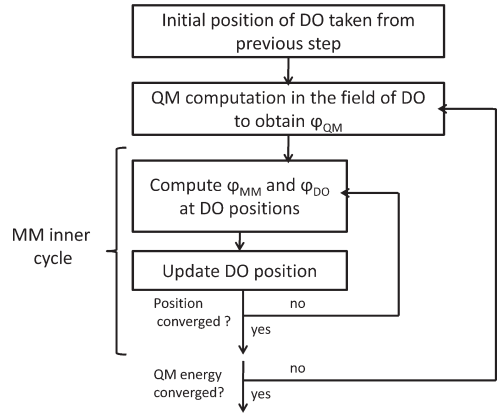


Figure 1. Dual SCF approach for determining the DO positions in a QM/MM framework. The outer SCF procedure converges the QM wave function, while the inner one converges the DO positions in the field of each other and of the MM atoms. See text for details.

but introduced an additional option: when using the CHARMM force field, the electric field is computed at the position of the Drude particle (to conform to CHARMM conventions), whereas it is calculated at the corresponding atomic position for the GROMOS COS model.[32] In our computational approach, we first evaluate $\varphi_{i'}^{\mathrm{QM}}$ for a set of fixed DO positions, which are then updated in an iterative scheme through an MM inner cycle. In this cycle, $\varphi_{i'}^{\mathrm{MM}}$ and $\varphi_{i'}^{\mathrm{DO}}$ are computed for the given geometry, and the DO positions are updated using the forces from eq 7. This inner cycle is iterated until the DO positions are converged in the field of the given QM wave function (as judged by their maximum and average changes). Thereafter, the convergence of the QM energy is checked. If this is not the case, the process is restarted at the first step by recalculating $\varphi_{i'}^{\mathrm{QM}}$ and going again through the inner cycle, until full overall convergence is achieved.

A microiterative scheme has also been proposed, with one update of the DO positions in each step of the QM SCF procedure.[40] This approach is computationally more efficient but necessitates the modification of the QM program and thus cannot be applied directly for any QM code.

The iterative relaxation of the Drude particle to its minimum energy during each step of an optimization or a sampling procedure is accurate but can become very expensive if there are a large number of polarizable atoms in the system. Therefore, this scheme may no longer be practical for long MD runs.

One key advantage of Drude oscillators is that they can be treated as dynamical degrees of freedom. Their direct inclusion in a standard MD scheme is problematic, however, since the oscillations of the Drude particles would have very high frequencies and would thus require very small integration time steps. This problem can be overcome by the use of extended Lagrangian dynamics.[20] In this scheme, the overall motion of the atoms and the relative motion within the atom−DO pairs are separated and propagated in a coupled manner as follows. A small mass (m_{D}) is taken from the polarizable atom and assigned to the Drude particle. The polarizable site is propagated in the overall MD scheme using the center of mass (R_i) of the atom−DO pair and the total mass of the atom (m_i). The relative motion in the atom−DO pair is propagated using an extended mass $m_i' = m_{\mathrm{D}}(1 - m_{\mathrm{D}}/m_i)$.[20] In the NVT ensemble, this leads to the so-called cold DO model.[20] The dynamics of the system is controlled by a thermostat of the desired temperature (T) while the relative motion within the DOs is frozen at temperature $T_* = 1$ K to avoid high-frequency oscillations. At each MD step, the DO positions are not fully converged, but it is assumed that, during the sampling process, they will oscillate around their respective minima. The equations of motions, using two Nosé−Hoover thermostats, are

$$m_i\ddot{R}_i = F_{R,i} - m_i\dot{R}_i\dot{\eta} \tag{8.a}$$

$$m_i'\ddot{d}_i = F_{d,i} - m_i'\dot{d}_i\dot{\eta}_* \tag{8.b}$$

$$Q\ddot{\eta} = \sum_j m_j\dot{R}_j^2 - N_{\mathrm{f}}\,k_{\mathrm{B}}T \tag{8.c}$$

$$Q_*\ddot{\eta}_* = \sum_j m_j'\dot{d}_j^2 - 3N_{\mathrm{D}}\,k_{\mathrm{B}}T_* \tag{8.d}$$

The indices i and j run over all atoms. The variables associated with the thermostats are the inertia factor Q and the friction

dx.doi.org/10.1021/ct300722e | J. Chem. Theory Comput. 2012, 8, 4527−4538

coefficients $\dot{\eta}$ that are obtained by solving eqs 8.c and 8.d. The subscript "$*$" refers to the thermostat for the relative motion of the atom−DO pair. N_f is the total number of degrees of freedom excluding constrained components and DOs, and N_D is the number of DOs. Note that if an atom is not polarizable, m_i' is zero and $R_i = r_i$. This extended system can be propagated in time using a velocity Verlet scheme.[66,67] We have implemented this approach for the NVT ensemble. Analogous MD runs can also be performed with the NPT ensemble using a barostat for the extended system,[20] but this variant was not implemented here because the use of BPs constrains the system to a constant volume (see below). We also did not consider other methods that propagate classical representations of electronic polarization using always stable estimator-corrector algorithms.[40,68,69]

The implementation of this integration scheme for QM/MM molecular dynamics necessitates the computation of the forces that act on the different particles. The contributions to these forces from the QM and MM regions are additive:

$$F_{R,i} = -\frac{\partial(U^{QM} + U^{MM})}{\partial r_i} - \frac{\partial(U^{QM} + U^{MM})}{\partial r_{i'}} \tag{9.a}$$

$$F_{d,i} = \left(\frac{m_D}{m_i}\right)\frac{\partial(U^{QM} + U^{MM})}{\partial r_i} - \left(1 - \frac{m_D}{m_i}\right) \times \frac{\partial(U^{QM} + U^{MM})}{\partial r_{i'}} \tag{9.b}$$

The QM contributions are computed only once per MD step using the fully converged QM SCF wave function for the given configuration of MM atoms and DOs. Note that forces are computed with the charges of each polarizable center reduced by the DO partial charge $q_{i'}$ (see above). QM atoms are propagated classically by treating them as dynamical degrees of freedom in the equation of motion of the general system; they are coupled to the same thermostat.

2.2. Boundary Potentials. A boundary potential simulates the electrostatic influence of an implicit infinite outer region on the explicit inner region of interest. Figure 2 illustrates this separation for a solvated protein in a QM/MM framework.

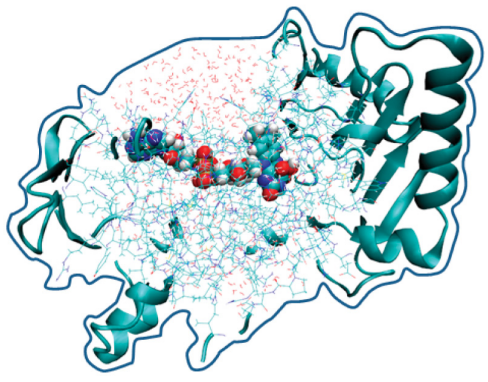


Figure 2. Schematic representation of the separation of regions in the GSBP and SMBP. The atoms in the QM region atoms are represented by their van der Waals radii. The MM atoms from the inner region are pictured explicitly by lines. The outer region of the macromolecule is symbolized by the ribbons. The region outside of the blue boundary line corresponds to the bulk solvent simulated by a PDC.

We use R to denote the generic coordinates of a macromolecule surrounded by N solvent molecules. The inner region consists of the inner part of the macromolecule (R_i) as well as n inner solvent molecules, while the outer region includes the outer part of the macromolecule (R_o) and the remaining $N - n$ solvent molecules. Statistical observations are assumed to depend only on the degrees of freedom of the inner region. They can thus be computed on the surface of its potential of mean force (PMF) by integrating out the degrees of freedom of the outer region:

$$e^{-\beta W(R_i,1,\ldots,n)} = \frac{1}{C}\int dR_0\, d(n+1)\ldots dN\, e^{-\beta U(R,1,\ldots N)} \tag{10}$$

Note that only configurations for which outer region atoms do not overlap with the inner region are considered here. By picking an appropriate normalization constant (C), Beglov and Roux demonstrated that the PMF is equivalent to the reversible work necessary to assemble the inner region inside the outer region.[70] They proposed to proceed stepwise, considering separately the different contributions to this assembly:

$$\Delta W = U + \Delta W_{cr} + \Delta W_{np} + \Delta W_{elec} \tag{11}$$

U is the potential energy of the isolated inner region, and the three following terms are the free energy contributions arising from configurational restrictions, nonpolar interactions, and electrostatic interactions, respectively. This approach is only valid if the configuration of the atoms in the outer region can be considered representative of the average of all possible configurations. It will thus be particularly suitable for studying processes localized in the center of the inner region, while its accuracy will decrease in the vicinity of the inner−outer boundary. It is commonly assumed that the configurational restrictions and nonpolar interactions will remain constant for a given system and that one may thus focus on the electrostatic contributions to the boundary potential. The following subsections describe two approaches to their determination.

2.2.1. Generalized Solvent Boundary Potential. The GSBP aims at approximating the electrostatic contribution to the PMF in a scheme suitable for MD simulations. There are two parts, arising from the direct Coulomb interaction of inner-region charges with the outer-region charges of the macromolecule and with the outer-region solvent molecules described by a PDC. The latter term can be expressed as the interaction of the inner-region point charges of the macromolecule (q_A) with the reaction field potential ϕ_{rf} at their position (r_A).

$$\Delta W_{elec}^{solv} = \frac{1}{2}\sum_A q_A \phi_{rf}(r_A) \tag{12}$$

The reaction field potential is defined as the difference of the electrostatic potentials in solution and in vacuum. It can be obtained by solving the linearized Poisson−Boltzmann (PB) equation for both situations using the corresponding dielectric constants (ε).

$$\nabla[\varepsilon(r)\nabla\phi(r)] - \bar{\kappa}^2(r)\phi(r) = -4\pi\rho(r) \tag{13}$$

Here, $\rho(r)$ is the charge density and $\bar{\kappa}(r)$ is the modified Debye−Hückel screening factor. A straightforward implementation would require solving the PB equation for every configuration, which would quickly become too expensive for typical MD runs. To overcome this problem, ΔW_{elec}^{solv} is

dx.doi.org/10.1021/ct300722e | J. Chem. Theory Comput. 2012, 8, 4527−4538

separated by splitting the charge distribution into an inner and outer part.

$$\Delta W_{\mathrm{elec}}^{\mathrm{solv}} = \Delta W_{\mathrm{elec}}^{\mathrm{outer-outer}} + \Delta W_{\mathrm{elec}}^{\mathrm{inner-outer}} + \Delta W_{\mathrm{elec}}^{\mathrm{inner-inner}} \tag{14}$$

The first term represents the interaction of the outer charge distribution with its self-induced reaction field. It is constant during the sampling and can thus be neglected or computed once and for all. The inner−outer contribution to the solvation free energy can be combined with the calculation of the inner-outer Coulomb interactions in an efficient scheme using the electrostatic potential of the outer region in solution ($\phi_{\mathrm{s}}^{\mathrm{outer}}(r)$).

$$\begin{aligned}\Delta W_{\mathrm{elec}}^{\mathrm{inner-outer}} + U_{\mathrm{elec}}^{\mathrm{inner-outer}} &= \sum_{A\in\mathrm{inner}} q_A \phi_{\mathrm{rf}}^{\mathrm{outer}}(r_A) + U_{\mathrm{elec}}^{\mathrm{inner-outer}} \\ &= \sum_{A\in\mathrm{inner}} q_A \phi_{\mathrm{s}}^{\mathrm{outer}}(r_A)\end{aligned} \tag{15}$$

Since the outer region is in a frozen configuration, its potential is constant during the simulation and can be calculated and stored once and for all, giving rise to a significant decrease of on-the-fly computational costs. The only terms remaining are thus the inner−inner contributions. An analytical solution for this part is provided by the Green's function (G_{rf}) that describes the inner-region reaction field potential.

$$\phi_{\mathrm{rf}}^{\mathrm{inner}}(r) = \int \mathrm{d}r' \rho_i(r')\, G_{\mathrm{rf}}(r, r') \tag{16}$$

This formulation allows the projection of the inner charge distribution and of the Green's function onto the same set of basis functions $\{b_n\}$ with associated generalized multipole moments Q_n. The solvation free energy can be expressed as the matrix product of the reaction field matrix (M_{mn}) with these multipole moments, which yields the final expression for the GSBP.

$$\Delta W_{\mathrm{elec}}^{\mathrm{GSBP}} = \sum_{A\in\mathrm{inner}} \phi_{\mathrm{s}}^{\mathrm{outer}}(r_A) + \frac{1}{2}\sum_{mn} Q_m M_{mn} Q_n \tag{17}$$

The matrix M_{mn} can be computed once and for all at the beginning of the simulation. This requires solving the PB equation repeatedly, depending on the size of the basis set used.

In a QM/MM framework, the inner region also includes the QM part of the system. Its contribution is taken into account separately by splitting the $\Delta W_{\mathrm{elec}}^{\mathrm{GSBP}}$ expression into QM and MM parts.

$$\begin{aligned}\Delta W_{\mathrm{elec}}^{\mathrm{GSBP}} = &\sum_{A\in\mathrm{inner,MM}} \phi_{\mathrm{s}}^{\mathrm{outer}}(r_A) + \int \mathrm{d}r\, \rho^{\mathrm{QM}}(r)\, \phi_{\mathrm{s}}^{\mathrm{outer}}(r) \\ &+ \frac{1}{2}\sum_{mn} Q_m^{\mathrm{QM}} M_{mn} Q_n^{\mathrm{QM}} + \sum_{mn} Q_m^{\mathrm{QM}} M_{mn} Q_n^{\mathrm{MM}} \\ &+ \frac{1}{2}\sum_{mn} Q_m^{\mathrm{MM}} M_{mn} Q_n^{\mathrm{MM}}\end{aligned} \tag{18}$$

In previous GSBP implementations, the continuous QM charge distribution was represented by Mulliken charges.[46,48] This necessitates changes in the QM code when implementing the GSBP scheme.

2.2.2. Solvated Macromolecule Boundary Potential. The SMBP is a solvent boundary potential designed for geometry optimizations with any kind of QM method. It relies on the same approximations as the GSBP by using the same decomposition into an inner and outer region. However, as geometry optimizations require much fewer steps than MD runs, the PB equation is now solved at each step. Compared with GSBP, this saves the initial effort of computing the reaction field matrix (i.e., solving the PB equation typically 800 times for common basis sets). To allow the use of the SMBP with any QM/MM Hamiltonian, the interactions with the reaction field potential are computed separately for the QM and MM regions.

$$\Delta W_{\mathrm{elec}}^{\mathrm{SMBP}} = \int \mathrm{d}r\, \rho_{\mathrm{QM}}(r)\phi_{\mathrm{tot}}^{\mathrm{QM}}(r) + \int \mathrm{d}r\, \rho_{\mathrm{MM}}(r)\, \phi_{\mathrm{tot}}^{\mathrm{MM}}(r) \tag{19}$$

Here, $\phi_{\mathrm{tot}}^{\mathrm{QM}}(r)$ and $\phi_{\mathrm{tot}}^{\mathrm{MM}}(r)$ are the effective potentials experienced by the QM and MM regions, respectively.

$$\phi_{\mathrm{tot}}^{\mathrm{QM}}(r) = \phi_{\mathrm{s}}^{\mathrm{outer}}(r) + \phi_{\mathrm{rf}}^{\mathrm{inner-MM}}(r) + \frac{1}{2}\phi_{\mathrm{rf}}^{\mathrm{QM}}(r) \tag{20.a}$$

$$\phi_{\mathrm{tot}}^{\mathrm{MM}}(r) = \phi_{\mathrm{s}}^{\mathrm{outer}}(r) + \frac{1}{2}\phi_{\mathrm{rf}}^{\mathrm{inner-MM}}(r) \tag{20.b}$$

The reaction field potentials depend of the instantaneous configuration of the inner region and must thus be updated in every optimization step. For nonpolarizable MM point charges, the term $\phi_{\mathrm{rf}}^{\mathrm{inner-MM}}(r)$ can be computed by solving the PB equation once. On the other hand, $\phi_{\mathrm{rf}}^{\mathrm{QM}}(r)$ depends on the polarizable QM density, and hence a self-consistent reaction field procedure is needed to determine both. This involves the following steps: (1) With an initial guess for the QM charges, compute $\phi_{\mathrm{rf}}^{\mathrm{QM}}(r)$. (2) Assemble $\phi_{\mathrm{tot}}^{\mathrm{QM}}(r)$ and project it on a set of virtual charges distributed on a sphere around the inner region. (3) Evaluate the QM wave function in the field of these virtual charges and the inner-region MM point charges. (4) Determine ESP charges that represent the QM charge distribution well enough to generate a realistic electric field. (5) Loop over steps 2−4 until convergence is reached, i.e., until the QM reaction field potential changes from one iteration to the next one by less than a predefined criterion.

2.3. QM/MM-Based Combination of Polarizable Force Fields with Boundary Potentials. The use of a polarizable force field for the explicitly treated inner MM region in a QM/MM/BP setup leads to a three-layer approach which accounts for polarization effects in all layers. This introduces additional interdependences which will be described in the following.

To include Drude oscillators in the MM layer of a QM/MM/BP treatment, eq 11 needs to be supplemented with an additional term (ΔW_{pol}) that describes the free energy necessary to switch on the polarizability of the polarizable MM atoms in the field of the boundary potential. This term is expressed in different forms in the GSBP and SMBP formalisms.

2.3.1. Combination with the GSBP. Since GSBP is designed for MD simulations, the extended Lagrangian approach appears as the method of choice to treat the equations of motion. In this approach, the positions of the Drude particles are not relaxed to their energy minima at every step, and hence there is no need to apply an iterative method that would converge both the induced explicit polarization and the boundary potential simultaneously. Therefore, the DOs can be handled in the GSBP treatment just like fixed classical point charges. As usual, their contribution can be separated into inner−inner and

dx.doi.org/10.1021/ct300722e | J. Chem. Theory Comput. 2012, 8, 4527−4538

inner−outer terms. To account for the inner−inner terms, additional MM point charges (Q_m^{pol}) are introduced into the formalism and are projected on the same basis set as the other charges. The expression for ΔW_{elec}^{GSBP}, eq 17, is thus extended by adding the following terms.

$$\frac{1}{2}\sum_{mn} Q_m^{pol} M_{mn} Q_n^{pol} + \sum_{mn} Q_m^{pol} M_{mn} Q_n^{MM} + \sum_{mn} Q_m^{pol} M_{mn} Q_n^{QM} \quad (21)$$

The treatment of the inner-outer contribution depends of the description of the outer region. Two assumptions can be made. The first one is to describe the entire outer region by the PDC model, with different dielectric constants for the macromolecular part and the bulk solvent. In this case, the system is fully polarizable, and there is no need to include the DO model in the outer region or to make any further modification to the GSBP expression. This option differs, of course, from the standard GSBP implementation in QM/MM methods (see section 2.2). The second option is to assume that the polarization of the MM atoms of the macromolecule in the outer region remains the same during the MD run. This is clearly compatible with the basic GSBP assumption of neglecting the outer-region thermal fluctuations and hence keeping the outer-region MM atoms fixed during the MD simulation. The outer-region DO positions are thus determined by an initial single-point computation on the full system and are then kept fixed. The error arising from having a constant outer-region polarization is expected to be small. The second option allows us to solve the PB equation for the macromolecule with a dielectric constant of 1 and to use the previously introduced approximations that lead to an appreciable gain in efficiency.[46] In practice, the second option is implemented by adding Drude particles during the computation of ϕ_s^{outer} in the same manner as the outer-region MM point charges and by correcting the latter for the polarizable atoms by the DO counter charges.

To propagate the dynamical degrees of freedom, the force contributions from the GSBP have to be included in eqs 9.a and 9.b. Before solving the finite-difference PB equation, the charges describing the electrostatics of the system are projected onto a grid using B-splines. Thus, the force acting on any MM point charge will depend on its position on the grid. It is obtained by taking the first derivative of ΔW_{elec}^{GSBP}.

$$\frac{\partial \Delta W_{elec}^{GSBP}}{\partial r_i} = \frac{\partial \phi_s^{outer}}{\partial r_i} q_i + q_i \sum_{mn} \left[\frac{\partial b_n(r_i)}{\partial r_i}\right] M_{mn} \times [Q_{mn}^{QM} + Q_{mn}^{MM} + Q_{mn}^{pol}] \quad (22)$$

2.3.2. Combination with the SMBP. Inclusion of Drude oscillators into the SMBP formalism gives rise to additional contributions ($\Delta W_{elec,DO}^{SMBP}$) to the solvation free energy, accounting for the charge density $\rho_{DO}(r)$ that experiences the field $\phi_{tot}^{DO}(r)$.

$$\Delta W_{elec,DO}^{SMBP} = \int \mathrm{d}r\, \rho_{DO}(r)\, \phi_{tot}^{DO}(r) \quad (23.a)$$

In addition, the reaction field $\phi_{rf}^{DO}(r)$ will contribute to the boundary potential

$$\phi_{tot}^{DO}(r) = \phi_s^{outer}(r) + \frac{1}{2}\phi_{rf}^{DO}(r) + \phi_{rf}^{inner-MM}(r) \quad (23.b)$$

Likewise, eqs 20.a and 20.b are modified by adding $\phi_{rf}^{DO}(r)$ to take into account the DO contributions. The DO positions are updated by including the electric field contribution from the SMBP on the Drude particles in the iterative procedure. The electric field is determined at these positions by interpolation from the nearest point on the grid depending of the order of the B-spline used in the representation. The proper polarization of the inner region is obtained by solving eq 2 for every polarizable center in the field of the whole system.

$$\varphi_{tot}^{DO} = \varphi_{MM}^{DO} + \varphi_{QM}^{DO} + \varphi_{DO}^{DO} + \varphi_{SMBP}^{DO} \quad (24.a)$$

$$\varphi_{SMBP}^{DO}(r) = \phi_s^{outer}(r) + \phi_{rf}^{inner-MM}(r) + \phi_{rf}^{QM}(r) + \phi_{rf}^{DO}(r) \quad (24.b)$$

The treatment must take into account the interdependences between the different subsystems, since the QM wave function, the PFF, and the boundary potential are all polarizable and depend on each other. A sequential combination of the dual self-consistent-field procedures used to update DO positions and to determine the SMBP would lead to a very expensive approach, and therefore another technique needs to be considered.

For a QM/MM computation with a QM region of significant size modeled by an accurate method, the QM calculation is the bottleneck in terms of computational time. Therefore, in any iterative process, the number of QM calculations has to be kept as small as possible. We thus propose an iterative scheme that performs only one QM calculation for updating both the boundary potential and the DO polarization, as shown in Figure 3. Before starting the iterative procedure, the constant contribution to the boundary potential from the nonpolarizable outer-region MM part is computed. In the first step, the reaction field potential $\phi_{rf}^{QM}(r)$ is evaluated, and the total potential acting on the QM region is projected on the virtual charges. Thereafter, a QM computation is carried out to evaluate the wave function in the field of the inner MM region, the DO charges, and the virtual charges representing the SMBP. A new set of ESP charges is determined to represent the QM region in the SMBP, and the QM electric field is computed at the position of the Drude particles. Using the ESP charges, ϕ_{rf}^{QM} is computed by solving the PB equation, and the electric field arising from the boundary potential is evaluated at the DO positions. With all external contributions to the DO electric field being known at this point, an SCF procedure is performed to update the DO positions in the field of each other and of the environment. These new DO positions are then used to compute their contribution to the SMBP. The convergence criterion of this iterative procedure is the change in the QM energy from one step to another, which has been found to be the quantity that converges most slowly. The default criterion is the same as that for the QM computation itself (typically on the order of 10^{-7} atomic units). For the inner MM cycle, we have adopted the same convergence criteria as before (see section 3).

The procedure outlined above allows full convergence of the different parts. Compared with full QM/MM-DO calculations without boundary potentials, it is efficient because the inner SCF cycles for optimizing the DO positions are now restricted to a small number of inner-region DOs, and the QM computations need to include only rather few external MM point charges. Also, the overall process normally requires less QM calculations than when the full system is represented explicitly.

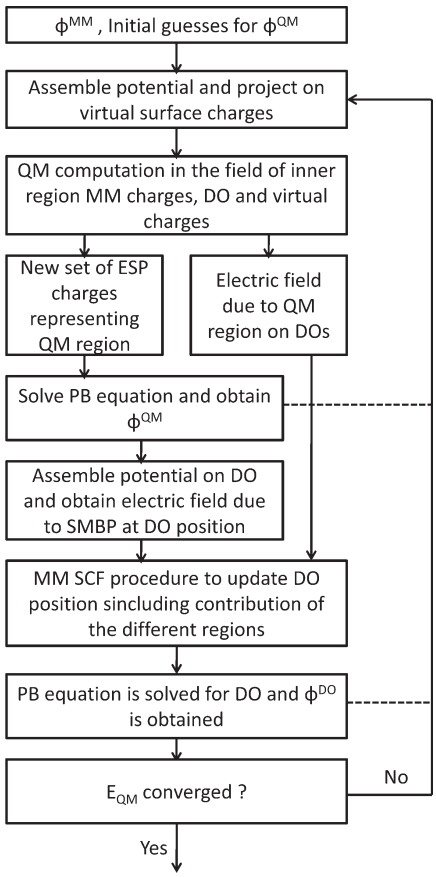


Figure 3. SCF procedure used for the update of DO positions, QM wave function, and SMBP contribution. The MM SCF procedure is the same as before (Figure 1). See text for details.

3. IMPLEMENTATION DETAILS

The methods described in section 2 have been implemented in the modular package ChemShell.[62] The code for Drude oscillators was implemented in a stand-alone module independent of the program used for MM force field evaluation. It is compatible with the CHARMM DO and GROMOS COS formalisms and can be used with any interfaced QM code. The code for lone pairs is available in a separate module and can be used together with any MM force field (polarizable or not).[23] The interaction energies involving lone pairs are computed first, and the associated gradient is assigned to the atom carrying the lone pair and its two nearest neighbors in a manner that conserves its total value and the torque (without generating any additional degrees of freedom). The propagation of cold Drude oscillators in an extended Lagrangian scheme has been implemented into the ChemShell MD module following the previously described implementation,[20] and the settle algorithm[71] was included for constraining water molecules.

The previously implemented PB equation solver[46] has been modified to handle the DO model. We have kept the approximations introduced to increase its efficiency as well as the rigid partitioning between inner and outer solvent molecules (i.e., not allowing for a dynamical and flexible separation). We use a spherical boundary both for GSBP and SMBP. When including higher-order terms such as the polarizability in the MM description, it is common practice to project the charges of the system on a discrete grid using higher-order functions.[57] For our models, fourth-order instead of third-order B-splines did not offer any improvement in terms of accuracy, and we thus adopted the latter. Similarly, the description of the dielectric boundary[57] did not improve significantly when using up to seventh-order polynomial switching functions instead of a direct approach, so we kept the latter.

In the current tests (see below), the QM calculations were performed using the MNDO2004[72] and Turbomole 6.3 programs.[73] DL_POLY was used for the additive part of the MM computations with the CHARMM force field. The HDCLOpt optimizer was employed for geometry optimizations using hybrid delocalized internal coordinates.[74]

4. ASSESSMENT

The present assessment consists of two parts. First, the accuracy of the model is discussed in terms of its ability to reproduce the gradient and the proper polarization of the inner region both for GSBP and SMBP. Second, the efficiency is evaluated by comparing computation times with and without a boundary potential.

4.1. Accuracy. Our standard test system for accuracy checks was a glycine molecule in its zwitterionic form solvated in a water ball of radius 30 Å, which has already been used in one of our previous studies.[49] Its high flexibility and polarity make it a challenging test case. The system was first thermalized using the standard nonpolarizable CHARMM force field and TIP3P water molecules. Five independent snapshots from an equilibrated MD run were investigated. These configurations were used in the QM/MM calculations without any further QM/MM-based equilibration. Since there was no significant difference between the five sets of QM/MM test results obtained, we present data only for one of the snapshots.

The glycine molecule is the QM part of the system, and the water ball is centered on its C_α carbon atom. The MM region includes 4252 water molecules, bringing the total number of atoms to 12 766. The 903 water molecules with the oxygen atom less than 18 Å away from the central carbon atom were considered as part of the inner region (together with the QM region). The 473 water molecules with their oxygen atom located in a buffer region between 14 and 18 Å from the center were frozen and represented explicitly. The other inner-region molecules were free to move. The AM1[75] semiempirical QM Hamiltonian was used for glycine, and the water molecules were represented using the SWM4-NDP PFF.[76] The relevant parameters are given in Figure 4. All inner-region DOs were considered to be active, even those assigned to a frozen explicit atom, to retain the full polarizability of the inner region.

For direct comparisons between the full Coulomb electrostatic interactions and the SMBP model, the outer-region dielectric constant was set to 1 (vacuum). Drude oscillators were included at outer-region atoms in fixed positions and were not allowed to reorganize themselves later on. Their positions were obtained by running an initial single-point computation on the full system without using a boundary potential. This way the accuracy of the SMBP could be evaluated by comparing QM/MM/SMBP results to those obtained for the entire QM/MM system. The PB equation was solved using a focusing procedure described previously[46]—first on a coarse-grained grid covering the full system and then with the use of a finer grid for the inner region, with the previously optimized spacings of 1.25 and 0.6 Å for the outer and inner grid,

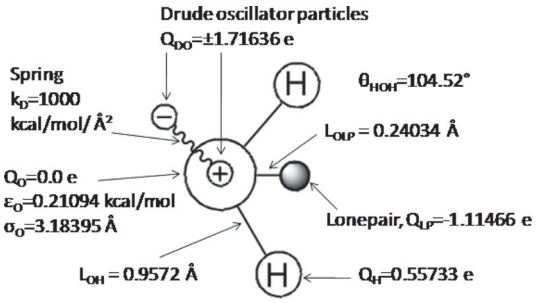


Figure 4. Schematic representation and parameters of the SWM4-NDP water model. The Drude oscillator is represented by a negative and a positive charge linked by a spring. The lone pair is shown in gray. Note that the oxygen atom has no charge by itself and is the only entity with van der Waals parameters. Geometric parameters (except for the oxygen-lone pair distance) are the same as for the TIP3P water model.

respectively. To project the SMBP potential acting on the QM region, 89 virtual surface charges were employed. Note that different definitions of the boundary potential will be used in the following tests.

Figure 5 shows (a) the mean absolute deviation (MAD) of gradient components (x,y,z) for all atoms located in the inner

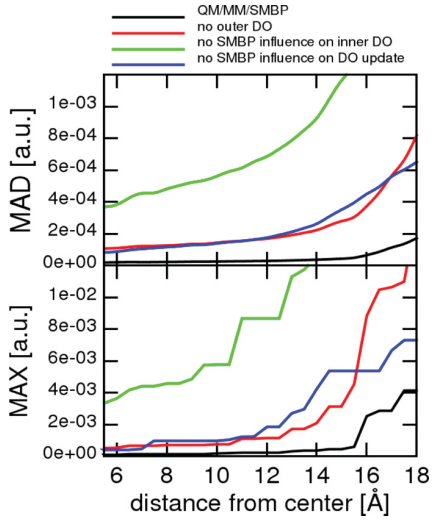


Figure 5. Deviations of the gradient components on atoms located in the inner region with a radius of 18 Å, compared with the full QM/MM results. The top panel shows the mean absolute deviation (MAD) as a function of the distance from the central carbon atom of glycine. The corresponding maximum absolute deviation (MAX) is plotted in the bottom panel. Different approximations are examined (see text).

region and (b) their maximum absolute deviation (MAX). The data are plotted as a function of the distance to the central carbon atom of glycine. The black curves correspond to the QM/MM/SMBP method as discussed in section 2. For QM/MM geometry optimizations, the convergence criteria used on gradients are typically on the order of 5×10^{-4} au. Therefore, the results obtained for this test case show that the SMBP gives a good approximation of the electrostatic interaction with the outer region. Indeed, for an active region that encompasses any molecule within 14 Å of the center, the maximum absolute deviation from QM/MM results is less than the standard convergence criterion, and the mean absolute deviation is even 1 order of magnitude smaller (i.e., only 3.6×10^{-5} au). In the buffer region between 14 and 18 Å, the gradients deviate more and more from the standard QM/MM values, which supports the convention to keep this region frozen in SMBP studies.[49] The gradients in the QM region show almost no difference, and localized processes in this region should thus be well simulated. The results in Figure 5 do not differ significantly from previously reported SMBP results obtained with identical parameters and the standard fixed-charge CHARMM force field.[49] Since the accuracy of the calculated QM/MM/SMBP gradients is essentially the same with and without DO contributions, it is not governed by the DO treatment but rather by the intrinsic errors arising from the finite-difference solution of the PB equation.[49]

How important are the DO contributions to the gradient? Do we need the full doubly iterative SCF procedure to obtain proper gradients? We address these questions by separately considering the influence of the Drude oscillators in the outer region (frozen) and in the inner region (active). We assess the necessity of including DO in the SMBP expression by considering three cases. The red curves in Figure 5 show the effects of removing the DO contributions from the outer region, which are not represented by the boundary potential. Note that for every DO, the $-q$ charge located at the atomic position is removed as well. Compared to QM/MM/SMBP, the deviations (MAD and MAX) from the full QM/MM reference gradients not only are significantly higher but also increase faster as the vicinity of the boundary is approached. This implies that the DO contributions from the outer region may have a significant influence even on localized inner-region processes, which is captured by our QM/MM/SMBP approach. The green curves show the effects of neglecting the influence of the SMBP terms on the inner-sphere Drude oscillators, both with regard to the update of the DO positions during the SCF process and their final contribution to the gradient. The deviations (MAD and MAX) are prohibitively large, so that this approximation is not to be used. The blue curves in Figure 5 show the deviations (MAD and MAX) that arise when neglecting the electric field from the SMBP during the SCF process of updating the DO positions, while including it during the final gradient computation. These deviations are smaller than those obtained upon total neglect of SMBP effects (green curves), but they are still too large to be tolerated. We conclude from these computational experiments that the full SCF procedure should be applied to ensure the needed accuracy.

We now investigate the influence of long-range electrostatic interactions on the polarization of the Drude oscillators, as indicated by the self-energy of polarization (i.e., the last term of eq 4) which is proportional to the DO polarization. Figure 6 shows the variation of this self-energy with the dielectric constant of the PDC that describes the outer-region solvent in the SMBP. The total polarization of the 903 inner-region Drude oscillators quickly increases for dielectric constants up to 20 and then levels off, with the self-energy slowly converging to a value of about 0.012 au. This corresponds for each Drude oscillator to a maximum variation of d by ca. 10^{-3} Å and a maximum increase of DO polarization by ca. 10−20%. The convergence of DO polarization in the absence of any damping terms suggests that the present SMBP treatment does not lead to the so-called "polarization catastrophe" for high dielectric constants, so that no special measures need to be taken in this regard.

dx.doi.org/10.1021/ct300722e | J. Chem. Theory Comput. 2012, 8, 4527−4538

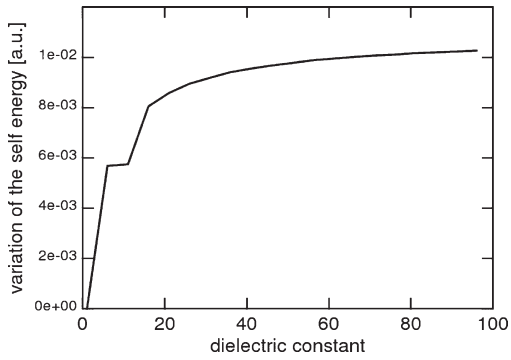


Figure 6. Total self-energy of polarization of the 903 Drude oscillators located in the inner region of the system, as a function of the dielectric constant used in the SMBP to represent the outer solvent through a PDC.

We next check in more detail how the self-energy of polarization evolves with respect to the distance from the center of the sphere (i.e., the central carbon atom of glycine). For this purpose, we again use a reference system consisting of a 30 Å sphere of water including the zwitterionic glycine molecule (12 766 atoms in total), and we represent the bulk solvent beyond this sphere in the SMBP by a PDC with a dielectric constant of 80. We compare the results obtained for this reference system with those computed for our standard system, i.e., an 18 Å sphere of water including glycine (2719 atoms in total) with bulk solvent treated as a PDC with different dielectric constants. To assess the accuracy of the results for the truncated system compared with the reference system, we consider the mean average percentage of deviation (MAD) and maximum percentage of deviation (MAX) of the self-energy of polarization of the Drude oscillators. These quantities are plotted in Figure 7 for several choices of the dielectric constant ε in the truncated system. For ε = 80, the MAD value is close to zero in the inner part of the active region and remains below 5% throughout the active region (for distances R up to 14 Å), while the MAX value rises to 23% at the active/frozen boundary (at R = 14 Å). This confirms again that the SMBP provides a reasonably accurate description of the inner active region of the truncated system. The situation is less favorable when neglecting the bulk solvent in the truncated system: for ε = 1 (vacuum), the MAD (MAX) value rises from ca. 12% (35%) in the inner active region to ca. 15% (80%) at the active/frozen boundary. When using truncated systems, the full SMBP approach (with a PDC treatment of bulk solvent) thus provides a clear improvement compared with a treatment that neglects the bulk solvent, as far as the polarization of the MM region is concerned. We also note that the distant-dependent self-energies converge very quickly with increasing dielectric constant (being very similar for ε = 15 and ε = 80, see Figure 7), in analogy to the fast convergence of the total self-energy (Figure 6).

We have evaluated the ability of the GSBP to properly reproduce the inner−outer electrostatic interactions in a similar way as for the SMBP. Initial positions for the Drude particles and the lone pairs were obtained from a fully converged iterative QM/MM computation on the complete system. The gradients on the Drude particles necessary for proper propagation of the dynamics were evaluated both at the full QM/MM and the QM/MM/GSBP level. Using the same computational parameters as in the SMBP case, the gradients on the mobile inner-region Drude particles (within the 14 Å active region) show a maximum absolute deviation of 2.7 × 10^{-4} au and a mean absolute deviation of 7.3 × 10^{-5} au from the full QM/MM reference data. These deviations are slightly higher than those obtained for the SMBP but in the same range as in previous GSBP validations,[46] in which this accuracy has been considered good enough for molecular dynamics simulations.

Going from a fixed-charge to a polarizable force field in the QM/MM/GSBP treatment may be expected to make the representation of the electrostatic interactions more demanding. To compute the inner region−inner region interactions, the GSBP approach employs a projection of the associated Green's function onto a basis set, which is also used to represent the corresponding multipole moments. We checked the convergence of the GSBP energy with increasing basis set size for the same test system as before. Figure 8 shows the

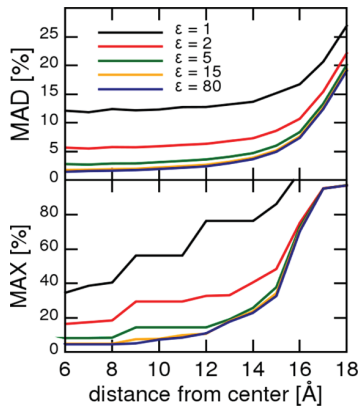


Figure 7. Mean average percentage of deviation (MAD) and maximum percentage of deviation (MAX) of the self-energy of polarization of Drude oscillators (truncated vs reference system, see text) plotted as a function of the distance from the central carbon atom of glycine. In the truncated system, different dielectric constants are used in the SMBP to represent the outer solvent through a PDC.

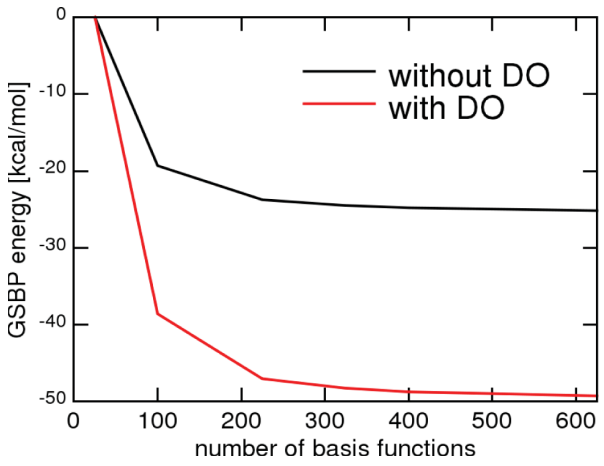


Figure 8. Variation of the GSBP energy with the size of the basis set used to project the inner−inner potential (relative to the value obtained with 25 basis functions). The dielectric constant of the outer region was fixed to 80.

dx.doi.org/10.1021/ct300722e | J. Chem. Theory Comput. 2012, 8, 4527−4538

variation of the GSBP energy for the two limiting cases, with the Drude oscillators being fully included or fully neglected in both the inner and the outer region. In the former case (red curve), the GSBP energy converges not quite as fast as in the latter case (black curve), indicating the need for a larger basis when including polarization. However, the previously recommended expansion up to l = 20 (400 functions)[46] is seen to capture the main part of the GSBP energy: an extension to l = 25 provides an additional energy gain of 0.52 kcal/mol (with DO) vs 0.37 kcal/mol (without DO). Hence, only a rather slight increase in the order of the expansion is needed in the DO case to ensure the same accuracy as before.

4.2. Efficiency. When using boundary potentials, explicit atomistic simulations are restricted to the inner region. For systems that are big enough, one may thus expect an appreciable reduction of the computational effort compared with a full QM/MM calculation.

To evaluate the efficiency of the SMBP, we use our test system (glycine in water). It consists of 10 QM atoms and 21 260 point charges in the MM region (taking into account lone pairs and Drude oscillators). The SMBP treatment employed the previously adopted parameters[49] and an 18 Å inner region containing the 10 QM atoms and a total of 4515 MM point charges. Table 1 lists the computation times and

Table 1. Average Time (s) for One Geometry Optimization Step at the Full QM/MM Level and the QM/MM/SMBP Level with the Associated Savings[a]

QM	basis	QM/MM	QM/MM/SMBP	% saved
AM1		122	66	46
BLYP	SVP	505	262	48
BLYP	TZVPP	1355	726	47
B3LYP	SVP	613	362	41
B3LYP	TZVPP	1622	1021	33

[a]Computations were run on 2.93 GHz Intel Xeon X5670 machines with 12 GB of memory. See text for further details.

associated savings obtained on average for one geometry optimization step for an active region encompassing glycine and all water molecules with their oxygen atom within 14 Å of the center. For the sake of consistency, interactions within the fixed outer region were neglected in both cases. The convergence criteria for the QM energy were 10^{-7} eV for the AM1[75] Hamiltonian and 10^{-7} au for the DFT computations using the BLYP[77,78] and B3LYP[79] exchange-correlation functionals with the SVP[80] and TZVPP[81] basis sets. The DO positions were considered converged if the maximum absolute deviation from one step to another was below 10^{-5} Bohr and the mean absolute deviation was below 2×10^{-5} Bohr. Computations were run on 2.93 GHz Intel Xeon X5670 machines with 12 GB of memory. Averages were taken over the 100 last steps of the geometry optimization (110 steps overall). The computation time for the optimizer was included and assumed to be the same in both cases. We obtain appreciable savings ranging from 33% to 48% depending on the chosen QM method. These savings arise from different contributions. In both cases, the use of the SMBP causes a strong reduction of the number of explicitly treated point charges that is even more pronounced for the polarizable force field (DO-PFF), which represents each water molecule by five point charges (rather than three without DO). We also observe that the implicit representation of the outer region in the SMBP treatment leads to faster convergence both in the overall and the inner SCF procedure.

To evaluate the efficiency of the GSBP, the glycine molecule was solvated in several water balls of different sizes. Computation times were determined for MD simulations using the extended Lagrangian approach. To ensure proper propagation, we used multiple time steps, i.e., 1 fs for the atomic motions and 1/30 fs for the thermostat.[20] An artificial mass of 0.4 au was assigned to each DO. Dynamics were run for 100 steps at 300 K to determine average times. The definitions of the active and inner regions were maintained for every system size. Figure 9 compares the average time per MD step

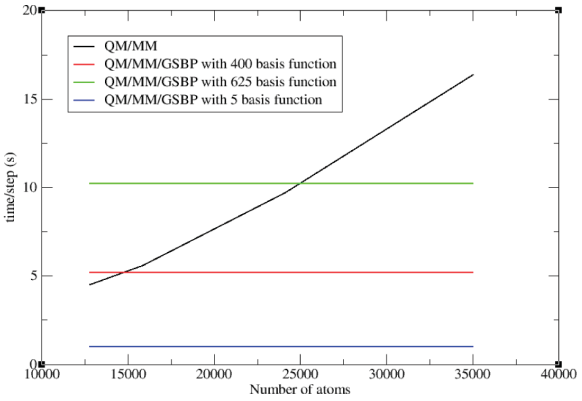


Figure 9. Computation time per MD step as a function of the total number of atoms. The black line shows the times for the full QM/MM treatment. In color: QM/MM/GSBP times for different numbers of basis functions used to project the inner−inner potential.

for QM/MM and QM/MM/GSBP MD simulations, as a function of the total number of atoms (not including the virtual DO and lone pair charges). The black line indicates the linear increase of the QM/MM computation time with system size. The colored lines show the QM/MM/GSBP computation times, which are essentially independent of system size, but depend strongly on the size of the basis set used for the required projections. The crossing points specify the system size, beyond which the QM/MM/GSBP treatment becomes more efficient than the standard QM/MM approach. For the previously recommended basis set (l = 20, 400 basis functions), the crossing occurs at a system size of about 14 500 atoms when using a polarized force field (DO-PFF), compared with about 12 500 atoms for fixed-charge force fields.[46] In QM/MM studies of enzymes, the system size is often in the range of 20 000−40 000 atoms so that the use of the GSBP provides significant savings in both cases.

5. CONCLUSION

In this article, we have combined two types of boundary potentials (SMBP and GSBP) with a QM/MM treatment, in which the MM part is described by a polarizable force field formulated in terms of Drude oscillators. For both boundary potentials, this leads to a fully polarizable three-layer QM/MM/BP model. The boundary potentials account for long-range electrostatic interactions, since they simulate the outer-region solvent by a polarizable dielectric continuum. In the case of the GSBP, the effects of the outer region are represented by a reaction field matrix, which is obtained once and for all at the beginning of a MD simulation. In the case of the SMBP, they

are computed on-the-fly during geometry optimization, in a manner that allows the use of any kind of QM method. The Drude oscillators simulate the polarization of the explicit inner MM region by two point charges of opposite sign linked by a spring, thus forming a dipole. One of the two DO charges is located at the polarizable atom, while the position of the other one is optimized in the field of the environment, with contributions from the QM and MM atoms, the boundary potential, and the other Drude oscillators. Likewise, DO terms are included in the evaluation of the QM wave function and contribute to the boundary potential arising from the inner region. Hence, the polarization effects in the three layers of our model are interdependent and coupled with each other. In the GSBP case, this coupling was treated by using an extended Lagrangian scheme, which propagates the DO positions as dynamical degrees of freedom so that they can be considered as fixed at every MD step, thus allowing us to handle the DO charges as any other classical MM point charge. In the SMBP case, the polarizations in the three layers were fully converged through an SCF procedure designed to minimize the number of QM evaluations to increase efficiency.

The accuracy of the two combination schemes was checked for a test system consisting of a glycine molecule in a water ball. Both schemes reproduce the gradients from corresponding full QM/MM calculations quite well. A proper SMBP representation of the inner−outer potential energy term requires the inclusion of frozen outer-region Drude oscillators. On the other hand, it also seems advisable to include the DO contributions during the SCF procedure to determine the SMBP. The influence of long-range electrostatic interactions on the Drude model is found to be significant, but there is no evidence for a "polarization catastrophe" when using a high dielectric constant to describe the outer-region solvent. The effects of the SMBP on the DO positions typically amount to less than 10^{-3} Å and thus to less than 10% of typical polarization effects. In the GSBP case, the results depend on the size of the basis set used to project the potential. When using Drude oscillators, maintaining the desired accuracy may require a slight extension of the multipole expansions used to describe the inner−inner interactions. QM/MM/SMBP computations become faster than the corresponding QM/MM computations beyond a certain system size, which depends on the chosen QM method. In the GSBP case, the efficiency also depends on the size of the basis set used for projection. In both cases, appreciable savings can be realized for large systems.

The presented three-layer models are particularly suitable for accurate ab initio free energy calculations of localized processes in macromolecules. Such studies require suitable polarizable force fields for proteins and other biomolecules.

AUTHOR INFORMATION

Corresponding Author

*E-mail: thiel@mpi-muelheim.mpg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors are grateful to Tobias Benighaus and Yan Zhang for supporting work and helpful discussions.

REFERENCES

(1) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198.
(2) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H. *J. Phys. Chem. B* **2006**, *110*, 6458.
(3) Hu, H.; Yang, W. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573.
(4) Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389.
(5) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186.
(6) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185.
(7) Warshel, A. *Annu. Rev. Biophys.* **2003**, *32*, 425.
(8) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
(9) Warshel, A.; Kato, M.; Pisliakov, A. V. *J. Chem. Theory Comput.* **2007**, *3*, 2034.
(10) Lopes, P. E. M.; Roux, B.; MacKerell, A. D. *Theor. Chem. Acc.* **2009**, *124*, 11.
(11) Yu, H.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69.
(12) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515.
(13) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.
(14) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.
(15) Rappe, A. K.; Goddard, W. A., III. *J. Phys. Chem.* **1991**, *95*, 3358.
(16) Rick, S. W.; Stuart, S. J.; Bader, J. S.; Berne, B. *J. Mol. Liq.* **1995**, *65*, 31.
(17) Stuart, S. J.; Berne, B. *J. Phys. Chem.* **1996**, *100*, 11934.
(18) Patel, S.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1.
(19) Patel, S.; MacKerell, A. D., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1504.
(20) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.
(21) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.
(22) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 1120.
(23) Jiang, W.; Hardy, D. J.; Phillips, J. C.; MacKerell, A. D., Jr; Schulten, K.; Roux, B. *J. Phys. Chem. Lett.* **2011**, *2*, 87.
(24) Yu, H.; Whitfield, T. W.; Harder, E.; Lamoureux, G.; Vorobyov, I.; Anisimov, V. M.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 774.
(25) Lopes, P. E. M.; Zhu, X.; Lau, A.; Roux, B.; MacKerell, A. D., Jr. *Biophys. J.* **2011**, *100*, 612.
(26) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2005**, *1*, 153.
(27) Nüsslein, V.; Schröder, U. *Phys. Status Solidi B* **1967**, *21*, 309.
(28) Schröder, U. *Solid State Commun.* **1993**, *88*, 1049.
(29) de Leeuw, N. H.; Parker, S. C. *Phys. Rev. B* **1998**, *58*, 13901.
(30) Yu, H. B.; Hansson, T.; van Gunsteren, W. *J. Chem. Phys.* **2003**, *118*, 221.
(31) Straatsma, T. P.; McCammon, J. A. *Mol. Simul.* **1990**, *5*, 181.
(32) Geerke, D. P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2007**, *3*, 2128.
(33) Geerke, D. P.; van Gunsteren, W. F. *J. Phys. Chem. B* **2007**, *111*, 6425.
(34) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580.
(35) Baker, C. M.; Anisimov, V. M.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2011**, *115*, 580.
(36) Vosmeer, C. R.; Rustenburg, A. S.; Rice, J. E.; Horn, H. W.; Swope, W. C.; Geerke, D. P. *J. Chem. Theory Comput.* **2012**, *8*, 3839.
(37) Luo, Y.; Jiang, W.; Yu, H.; MacKerell, A. D., Jr.; Roux, B. *Faraday Discuss.* **2012**, DOI: 10.1039/C2FD20068F. http://pubs.rsc.org/en/content/articlepdf/2012/fd/c2fd20068f (accessed Sept. 5, **2012**).
(38) Kunz, A.-P. E.; Allison, J. R.; Geerke, D. P.; Horta, B. A. C.; Hünenberger, P. H.; Riniker, S.; Schmid, N.; van Gunsteren, W. F. *J. Comput. Chem.* **2012**, *33*, 340.
(39) Geerke, D. P.; Thiel, S.; Thiel, W.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2007**, *3*, 1499.
(40) Lu, Z.; Zhang, Y. *J. Chem. Theory Comput.* **2008**, *4*, 1237.

(41) Rowley, C. N.; Roux, B. *J. Chem. Theory Comput.* **2012**, *8*, 3526.
(42) Meier, K.; Thiel, W.; van Gunsteren, W. F. *J. Comput. Chem.* **2011**.
(43) Nam, K.; Gao, J.; Darrin, M. *J. Chem. Theory Comput.* **2005**, *1*, 2.
(44) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2006**, *2*, 1370.
(45) Gao, J.; Alhambra, C. *J. Chem. Phys.* **1997**, *107*, 1212.
(46) Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2008**, *4*, 1600.
(47) Im, W.; Berneche, S.; Roux, B. *J. Chem. Phys.* **2001**, *114*, 2924.
(48) Schaefer, P.; Riccardi, D.; Cui, Q. *J. Chem. Phys.* **2005**, *123*, 014905.
(49) Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2009**, *5*, 3114.
(50) Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2011**, *7*, 238−249.
(51) Gilson, M. K.; Honig, B. H. *Biopolymers* **1986**, *25*, 2097.
(52) Zhang, Y.; Liu, H.; Yang, W. *J. Chem. Phys.* **2000**, *112*, 3483.
(53) Kästner, J.; Senn, H. M.; Thiel, S.; Otte, N.; Thiel, W. *J. Chem. Theory Comput.* **2006**, *2*, 452.
(54) Jon, R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694.
(55) Li, H.; Gordon, M. S. *J. Chem. Phys.* **2007**, *126*, 124112.
(56) Steindal, A. H.; Ruud, K.; Frediani, L.; Aidas, K.; Kongsted, J. *J. Phys. Chem. B* **2011**, *115*, 3027−3037.
(57) Schnieders, M. J.; Baker, N. A.; Ren, P.; Ponder, J. W. *J. Chem. Phys.* **2007**, *126*, 124114.
(58) Lipparini, F.; Barone, V. *J. Chem. Theory Comput.* **2011**, *7*, 3711.
(59) Schwabe, T.; Olsen, J. M.; Sneskov, K.; Kongsted, J.; Christiansen, O. *J. Chem. Theory Comput.* **2012**, *7*, 2209.
(60) Mennucci, B. Personal communication, July 2012.
(61) ChemShell. www.chemshell.org (accessed Aug 14, 2012).
(62) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *J. Mol. Struct.: THEOCHEM* **2003**, *632*, 1.
(63) Zhu, X.; Lopes, P. E. M.; MacKerell, A. D., Jr. *WIREs Comput. Mol. Sci.* **2012**, *2*, 167.
(64) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341.
(65) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587.
(66) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117.
(67) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637.
(68) Kolafa, J. *J. Comput. Chem.* **2004**, *25*, 335.
(69) Kolafa, J. *J. Chem. Phys.* **2005**, *122*, 164105.
(70) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050.
(71) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952.
(72) Thiel, W. *MNDO program*; Max-Planck-Institut für Kohlenforschung: Mülheim an der Ruhr, Germany, 2004.
(73) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.
(74) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177.
(75) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(76) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. *Chem. Phys. Lett.* **2006**, *418*, 245.
(77) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
(78) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
(79) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(80) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.
(81) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

NOTE ADDED AFTER ASAP PUBLICATION

This article was published ASAP on October 24, 2012. The second line in equation 4 has been modified. The correct version was published on October 30, 2012.

Quantum mechanics/molecular mechanics dual Hamiltonian free energy perturbation.

Iakov Polyak, Tobias Benighaus, Eliot Boulanger,
and Walter Thiel

J. Chem. Phys. **2013**, 139, 064105.

THE JOURNAL OF CHEMICAL PHYSICS **139**, 064105 (2013)

Quantum mechanics/molecular mechanics dual Hamiltonian free energy perturbation

Iakov Polyak, Tobias Benighaus,[a] Eliot Boulanger, and Walter Thiel[b]
Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D–45470 Mülheim an der Ruhr, Germany

(Received 6 May 2013; accepted 18 July 2013; published online 9 August 2013)

The dual Hamiltonian free energy perturbation (DH-FEP) method is designed for accurate and efficient evaluation of the free energy profile of chemical reactions in quantum mechanical/molecular mechanical (QM/MM) calculations. In contrast to existing QM/MM FEP variants, the QM region is not kept frozen during sampling, but all degrees of freedom except for the reaction coordinate are sampled. In the DH-FEP scheme, the sampling is done by semiempirical QM/MM molecular dynamics (MD), while the perturbation energy differences are evaluated from high-level QM/MM single-point calculations at regular intervals, skipping a pre-defined number of MD sampling steps. After validating our method using an analytic model potential with an exactly known solution, we report a QM/MM DH-FEP study of the enzymatic reaction catalyzed by chorismate mutase. We suggest guidelines for QM/MM DH-FEP calculations and default values for the required computational parameters. In the case of chorismate mutase, we apply the DH-FEP approach in combination with a single one-dimensional reaction coordinate and with a two-dimensional collective coordinate (two individual distances), with superior results for the latter choice. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4817402]

I. INTRODUCTION

Free energy is a key thermodynamic quantity to characterize chemical processes. It governs the relative stability of different species and the rate of chemical reactions. Knowledge of the potential energy of the system along the reaction coordinate (RC) is not sufficient to determine the reaction rate because of the entropic contributions to the free energy. In systems that obey classical statistical mechanics, one needs information about all accessible configurations of the system through the partition function to calculate the free energy exactly. The Helmholtz free energy is given by

$$A = -\frac{1}{\beta}\ln(Z), \tag{1}$$

where Z is the canonical ensemble partition function of the system and $\beta = \frac{1}{k_B T}$ is available from the Boltzmann constant k_B and the temperature T. Free energy differences can be expressed in terms of ensemble averages that can be approximately evaluated with the use of sampling techniques, such as molecular dynamics (MD) or Monte Carlo (MC) simulations.[1]

There are several well-established procedures to calculate the free energy, e.g., umbrella sampling,[2] thermodynamic integration,[3] and free energy perturbation (FEP).[4] For example, FEP can be used to determine the free energy difference between a perturbed and an unperturbed state of the system, which are described by two different Hamiltonians, through the sampling of the potential energy difference between them.

Regardless of the chosen procedure, the configurational phase space needs to be sampled extensively to obtain accurate free energies. This will become computationally demanding when going to ever larger systems and to ever more accurate and time-consuming methods for computing the potential energy during the sampling. Nowadays, classical force fields are widely used to describe thermodynamic properties of large biomolecular systems. If electronic effects are important, e.g., as in chemical reactions, one can apply hybrid quantum mechanical/molecular mechanical (QM/MM) methods,[5] in which the electronically relevant part of the system is treated quantum-mechanically, while the remainder is described by a classical force field. QM calculations require significantly more computational time than MM calculations, and therefore extensive sampling of large systems is demanding at the QM/MM level, especially when using first-principles QM methods. Due to this limitation, there have been many efforts[6–26] to develop QM/MM free energy methods, which aim at avoiding direct sampling at high levels of theory while still giving an accurate estimate of the free energy changes during the reaction.

A powerful approach, initially proposed and developed by Warshel and co-workers,[6–12] and also employed in a modified form by Ryde and co-workers,[13,14,27] makes use of thermodynamical cycles; an initial estimate of the free energy is determined by sampling with some approximate reference Hamiltonian and then corrected by evaluating via FEP the free energy change when going from the approximate reference Hamiltonian to the target QM/MM Hamiltonian. In some of these studies,[12,27] the reference potential has been generated using semiempirical QM methods. Another approach[22–24] is to accelerate the sampling of configurational phase space by using auxiliary MC simulations performed with an

[a]Permanent address: Lanxess Deutschland GmbH, 51369 Leverkusen, Germany.
[b]Electronic mail: thiel@mpi-muelheim.mpg.de

approximate Hamiltonian; the resulting final MC structures are subjected to MC update tests, which are based on the phase space overlap of the two Hamiltonians, thus significantly increasing the rate of the overall convergence. The two approaches have also been combined.[25]

There are also QM/MM free energy calculations that conduct a direct sampling of the whole phase space of the full QM/MM system on a single potential surface using umbrella sampling,[28] thermodynamic integration,[29] or umbrella integration.[30] Such calculations usually employ efficient semiempirical methods as QM component and trajectories of less than 100 ps (sufficient to obtain converged results in the investigated enzymatic systems according to standard statistical tests[29]). In a recent study,[27] the use of semiempirical QM/MM sampling for evaluating the entropic contributions was however considered questionable, because the phase space showed only weak overlap with the one derived from higher-level methods. In the dual-level approach of Tuñon and co-workers,[31,32] higher-level single-point calculations are employed to determine correction terms for the semiempirical QM/MM energy and gradient as a continuous function of a distinguished reaction coordinate, and free energy calculations are then done on the resulting surface using umbrella sampling.

The QM/MM-FE technique developed by Yang *et al.*[26] is based on the FEP method and targets an especially efficient QM/MM sampling. In this approach, the reaction path is divided into windows, and in each of them the geometry of the QM region is obtained by a restrained QM/MM optimization. This geometry is then kept fixed during the sampling which is performed only for the MM region, with the QM atoms being represented by partial charges (derived by an ESP fit of the electrostatic potential). The perturbations are defined by the exchange of the two subsequent geometries of the QM region. This procedure offers an inexpensive way to directly obtain the free energy profile of a reaction at the QM/MM level since the sampling of the MM region is effectively done at the MM level. The conceptual drawback of this approach is the lack of sampling in the QM region, and hence the entropic QM contribution can only be evaluated at the stationary points within the rigid-rotor harmonic-oscillator approximation of statistical thermodynamics.

A more general formulation of the QM/MM-FE approach proposed by Rod and Ryde[13,14] and named QTCP (quantum-mechanical thermodynamic-cycle perturbation) uses the FEP method both for evaluating the MM $\rightarrow$ MM perturbation along the reaction coordinate and for estimating the vertical MM $\rightarrow$ QM/MM free energy differences in a thermodynamical cycle.

In this paper, we present a modified version of the QM/MM-FE method, in which the phase space of the QM region is freely sampled, except for the RC which is the subject of the perturbation. The sampling of the QM region combines MD simulations at the efficient semiempirical QM/MM level with first-principles QM/MM energy evaluations (using *ab initio* or density functional QM methods). We therefore call this approach Dual Hamiltonian Free Energy Perturbation (DH-FEP). In Secs. II–III, we first describe the method and its implementation. Thereafter we validate it for two test systems: a two-dimensional analytic model potential and the enzymatic reaction catalyzed by chorismate mutase.

II. METHOD

A. QM/MM-FE

According to Zwanzig,[4] the free energy difference between a perturbed (2) and an unperturbed (1) state can be expressed as

$$\Delta A = A_2 - A_1 = -\frac{1}{\beta}\ln\int P_1(\mathbf{r})\exp\{-\beta[E_2(\mathbf{r}) - E_1(\mathbf{r})]\}\mathrm{d}\mathbf{r}, \tag{2}$$

where $E(\mathbf{r})$ is the potential energy and $P_1(\mathbf{r})$ is the probability of finding the unperturbed system in the configuration $\mathbf{r}$. For a QM/MM Hamiltonian, the energy is decomposed into three parts and therefore we have

$$\Delta A = -\frac{1}{\beta}\ln\int P_1(\mathbf{r}) \times \exp\{-\beta[\Delta E_{QM} + \Delta E_{QM-MM} + \Delta E_{MM}]\}\mathrm{d}\mathbf{r}. \tag{3}$$

In the QM/MM-FE method introduced by Yang and co-workers,[26] the perturbation is defined as the exchange of two neighboring QM structures that result from restrained optimizations of points along the reaction path. The underlying assumption is that the QM and MM degrees of freedom (DOFs) can be treated separately and that the sampling needs to be done only over the MM DOFs, whereas the contributions to the free energy arising from the fluctuations of the QM region around its "optimum reaction path" are assumed to be constant along the RC. The expression for the free energy difference between "windows" A and B along the RC is[26]

$$\Delta A(R_c) = \Delta E_{QM}(\mathbf{r}_{QM}^{min}) - \frac{1}{\beta}\ln\int P(R_c^A)\exp\{-\beta[E_{QM/MM}(\mathbf{r}_{QM}^{min}(R_c^B)) - E_{QM/MM}(\mathbf{r}_{QM}^{min}(R_c^A))]\}\mathrm{d}\mathbf{r}_{MM}. \tag{4}$$

Corrections for zero-point vibrational energies and entropic contributions are only included at the stationary points using the rigid-rotor harmonic-oscillator approximation,

$$\Delta A_{QM} - \Delta E_{QM} = \Delta E_{QM}^{ZPE} + \Delta U_{QM}^{th} - T\Delta S_{QM}. \tag{5}$$

The QM/MM-FE method outlined above involves two major assumptions. The first one is conceptual, namely not to sample the QM region, which causes a truncation of the accessible configurational space and may thus lead to an underestimation of the entropic contributions. The second and less critical one arises from the implementation: the representation of the QM atoms by ESP charges to allow for an efficient sampling (technically at the MM level).

B. Dual Hamiltonian free energy perturbation

In our approach, we do not separate the QM and MM DOFs but define the perturbation in terms of a pre-determined RC ξ on the potential energy surface. The RC is split into

discrete windows, each having a specific ξ_i value assigned, so that ξ_i and ξ_{i+1} are two constraints defining two neighboring windows along the RC,

$$\Delta E_{pert}^{\xi_i \to \xi_{i+1}} = E(\mathbf{r}', \xi_{i+1}) - E(\mathbf{r}', \xi_i), \tag{6}$$

where $\mathbf{r}'$ represents any configuration that fulfills the constraint ξ_i. We thus have a constrained Hamiltonian and can write the free energy along the RC as

$$A(\xi_i) = -\frac{1}{\beta} \ln \int \exp\{-\beta E(\mathbf{r}', \xi_i)\} \mathrm{d}\mathbf{r}'. \tag{7}$$

In standard notation,[4] the free energy perturbation between two subsequent points is given by

$$\Delta A^{\xi_i \to \xi_{i+1}} = -\frac{1}{\beta} \ln \int P_i(\mathbf{r}', \xi_i) \times \exp\{-\beta[E(\mathbf{r}', \xi_{i+1}) - E(\mathbf{r}', \xi_i)]\} \mathrm{d}\mathbf{r}'. \tag{8}$$

In practice, the integration is replaced by a discrete sum over MD steps. In the limit of complete sampling over all $\mathbf{r}'$ we obtain

$$\Delta A^{\xi_i \to \xi_{i+1}} = -\frac{1}{\beta} \ln \left[\frac{1}{N} \sum_{i=1}^{N} \exp\{ -\beta \Delta E_{pert}^{\xi_i \to \xi_{i+1}} \} \right]. \tag{9}$$

Applying this approach directly in combination with high-level QM methods would be expensive. Therefore we look for an approximation that will make our computations efficient. The integration step size in the MD simulation is usually chosen rather small to ensure a stable and accurate propagation of the system. Two consecutive points are thus rather close in geometry and ΔE_{pert} does not vary much, i.e., the step size is ideal for the MD run, but not for sampling ΔE_{pert} efficiently. Therefore we adopt a procedure, in which ΔE_{pert} is computed regularly only after skipping a pre-determined number of steps; this also decreases the correlation between subsequent configurations. The intermediate MD steps are disregarded during the computation of the free energy, which is thus determined from a limited number of configurations. This allows us to introduce the next approximation: ΔE_{pert} is evaluated with a computationally demanding high-level QM method at the selected steps (which is affordable because of the relatively small number of such calculations), while the sampling is performed at the semiempirical QM/MM level. Denoting the low-level and high-level Hamiltonian by *Ham1* and *Ham2*, Eq. (8) can then be reformulated accordingly,

$$\Delta A^{\xi_i \to \xi_{i+1}} = -\frac{1}{\beta} \ln \int P_i^{Ham1}(\mathbf{r}', \xi_i) \times \exp\{-\beta[E^{Ham2}(\mathbf{r}', \xi_{i+1}) - E^{Ham2}(\mathbf{r}', \xi_i)]\} \mathrm{d}\mathbf{r}'. \tag{10}$$

Using a cumulant expansion,[33] the free energy difference can be expressed as a function of the central moments of the energy difference distribution,

$$\Delta A = \langle \Delta E \rangle - \frac{\beta}{2} \sigma^2 + O(\beta)^2. \tag{11}$$

We use this expansion to overcome the problem of possible random occurrences of low ΔE_{pert} values in the trajectory, which may adversely affect the direct exponential average. In practice, we neglect all higher-order terms (as in Ref. 33), and the free energy difference is calculated as a sum of the average value and the variance of the energy difference distribution.

In actual applications, the reaction path is obtained from a sequence of restrained optimizations for suitably defined "windows," each one with a given RC value ξ_i. A semiempirical QM/MM MD simulation is then performed for each window with the constrained RC value ξ_i. Every x number of steps, the RC is perturbed to ξ_{i+1} and ΔE_{pert} is evaluated using a high-level QM Hamiltonian; note that the system is always propagated at RC $= \xi_i$. The ΔE_{pert} value obtained is then tested for equilibration as described in Ref. 29 by ensuring that there is no trend in the coarse-grained average and variance, and by checking the distribution for normality and lack of correlation. If the test for trend reveals non-stationarity of ΔE_{pert} or its variance, some MD steps from the beginning of simulation (and rarely from the end) are dropped until the resulting data becomes stationary. If the above analysis results in less than 400 equilibrated data points, further sampling is performed for the given window. The free energy difference between RC values ξ_i and ξ_{i+1} and the related confidence interval are then calculated based on the cumulant expansion.[33] Finally, the free energy profile of the reaction is obtained by summing up all the free energy differences between adjacent windows.

So far our development has been in terms of a one-dimensional RC (e.g., an internal coordinate or a linear combination of internal coordinates) that gives rise to a single constraint ξ_i. However, our formalism, in particular Eq. (10), remains valid when using a more general collective coordinate, for example a collection of several (N) independent internal coordinates $\{d_j(i)\}$ that are individually and simultaneously constrained during the sampling. A typical case is a one-dimensional RC defined as a linear combination of two distances, where the corresponding collective coordinate is composed of these two distances ($N = 2$). The use of a collective coordinate may lead to improved results, when the individual constraints are chosen appropriately and reflect the most relevant changes during the reaction.

DH-FEP is related to the several existing methods[11,12,27,31,32] in the sense that it uses a reference potential in order to perform efficient sampling, while obtaining the free energy difference at a higher theory level. It differs from previously proposed dual-level free energy methods in that we do not evaluate and apply high-level perturbation corrections after the low-level sampling is finished,[11,12,27] nor do we perform a semiempirical QM/MM sampling with a pre-calculated first-principles correction function along the reaction path.[31,32] Instead, our goal is to approximate an accurate high-level QM/MM sampling by using efficient semiempirical QM/MM MD simulations and directly evaluating first-principles QM/MM perturbation energies at a relatively small number of selected MD steps. We thus avoid a perturbation treatment in the method space, based on the assumption that there is a sufficient overlap in the phase space of the low-level and high-level methods used. Both our approach and the methods based on a thermodynamic cycle may suffer from a possibly weak

overlap of the two underlying phase spaces. In Ref. 12 this problem is approached by refining the reference potential, while we try to tackle it by finding a suitable semiempirical method that will represent the high-level QM method phase space well and/or by using an appropriate collective reaction coordinate (see Sec. IV). DH-FEP thus shares some basic strategic ideas with the MM based importance function method of Iftimie *et al.*,[22,23] which uses a classical MM potential to guide a first-principles MC simulation, but the computational framework is of course entirely different in these two approaches.

The convergence of QM/MM free energy perturbations based on semiempirical QM/MM simulations has recently been studied by Heimdal and Ryde.[27] The main distinction from our approach is the use of a thermodynamic cycle to account for the differences between low-level and high-level QM/MM methods via FEP. Within this framework, the so-called QTCP-free calculations are conceptually similar to our approach in the sense that only the reaction coordinate is kept fixed (rather than the whole QM system); it is not specified, however, whether the atoms involved in the reaction coordinate are fixed to their initial Cartesian coordinates or whether a constraint is applied, which impedes direct comparisons. In the QTCP-free calculations,[27] the error bars for the perturbation along the reaction coordinate are rather small (as in our approach, see below), whereas those for the perturbation in the method space (which have no counterpart in our approach) are quite large.

III. COMPUTATIONAL DETAILS

The DH-FEP method was implemented in a developmental version of the Chemshell package.[34] Constraints were imposed using the SHAKE procedure,[35] which was extended to include the difference of two distances between four different atoms. When evaluating ΔE_{pert} during the MD simulations, the SHAKE procedure is applied twice for the four atoms involved in the reaction coordinate, first to satisfy the constraint on the unperturbed system (RC value ξ_i), and then to satisfy the constraint on the perturbed system (RC value ξ_{i+1}); thereafter the potential energy is computed for the two resulting structures. For the calculation of free energy differences and the statistical validation of sampled data, we used a supplementary program written by Kästner for our original QM/MM-FE implementation.[33]

In the QM/MM calculations, we employed the following codes: MNDO2005[36] for the semiempirical QM methods OM3 (orthogonalization model 3)[37,38] and SCC-DFTB (self-consistent-charge density functional tight binding),[39] TURBOMOLE 6.3[40] for the *ab initio* QM method RI-MP2 (resolution-of-identity Møller-Plesset second-order perturbation theory),[41,42] and DL-POLY[43] for the CHARMM22 force field.[44]

IV. ASSESSMENT

We assess our method using two examples. The first one involves an analytic potential function that allows an exact evaluation of the free energy and can thus be used to validate our ansatz for calculating free energy differences along a pre-defined RC. We use two potential functions that differ slightly from each other, one of which is used for sampling and the other one for evaluating the perturbation energy differences, in order to test the importance of configurational phase space overlap between the two potentials. The second example addresses the evaluation of the activation free energy in the enzymatic reaction catalyzed by chorismate mutase: here we examine the performance of our method for a chemically meaningful QM/MM system and compare the results to experimental data.

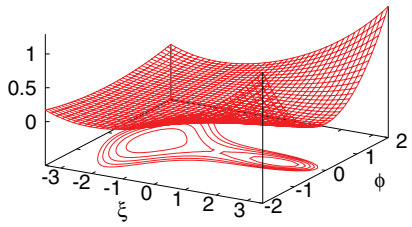


FIG. 1. 3D plot and contour plot of the analytic potential $E_1(\xi, \phi)$ with a contour spacing of 0.005. All values in atomic units.

A. Analytical model potential

For numerical validation of our method, we use a two-dimensional model potential taken from Ref. 30, for which the free energy can be computed analytically: $E_1(\xi, \phi) = f(\xi) + k(\xi)\phi^2$ with $f(\xi) = b - c\xi^2 + (c^2/4b)\xi^4$ and $k(\xi) = k_{min} + 2db/c + \sqrt{(8d^2b)/c}\xi + d\xi^2$. The 3D plot and a contour plot of the potential are shown in Fig. 1. The RC is represented by ξ while ϕ is an additional degree of freedom, along which the surface will be sampled to compute the free energy; on the RC, we always have $\phi = 0$. This model potential has two minima with $E_1 = 0$, which have different surroundings and thus differ in free energy (lower at the minimum with a broader potential because of higher entropic contributions). The free energy along the RC can be evaluated analytically as $A_1(\xi) = f(\xi) + \ln(k(\xi))/2\beta + \text{const}$.

We used the same parameters as in our previous work.[30] In atomic units, the barrier is chosen to be $b = 0.01$, the minima are placed at $\xi_{min} = \pm 2$ by assigning $c = 0.005$, while the width of E_1 in the direction of ϕ is defined by setting $d = 0.01$ and $k_{min} = 0.01$.

Constrained Metropolis MC simulations[45] were carried out on this model potential in the NVT ensemble at a temperature of 298.15 K. The path from $\xi = -3$ to $\xi = 3$ was split into windows separated by a width of 0.05. Fifty thousand MC trial steps with a maximum step size of 0.05 were performed for each window along the RC, with each new run starting at $\phi = 0$ and the RC being constrained to ξ_i. At each step, both ϕ and ξ were shifted in a random direction. If the step was accepted, the ξ value was replaced first with ξ_i and then with ξ_{i+1}, and the energies at both points were evaluated. Thereafter the next step was performed starting from ξ_i. The free energy difference was calculated from the direct exponential average of all sampled ΔE_{pert} values.

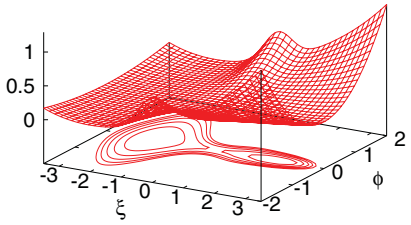


FIG. 2. 3D plot and contour plot of the analytic potential $E_{2a}(\xi, \phi)$ with a contour spacing of 0.005. All values in atomic units.

The resulting activation and reaction free energies are in excellent agreement with the analytic results. Compared with the analytic values of 28.250 and 3.512 kJ/mol for the activation and reaction free energies, the errors were as small as 0.043 and 0.049 kJ/mol, respectively, which clearly validates the FEP ansatz for calculating free energy differences along the reaction coordinate. With this justification in hand, we now test the approximation of using two different potentials for sampling and for evaluating ΔE_{pert} at the sampled geometries.

For this purpose, we constructed two new model potentials that differ in the transition state region but are the same at both minima. This choice is motivated by the intended QM/MM applications, where we expect low-level QM methods to mimic high-level QM methods more closely near the minima than near the transition states.

We first introduced into E_1 a term that depends in a Gaussian fashion on ξ and quadratically on ϕ, being zero at $\phi = 0$. In the resulting function $E_{2a}(\xi, \phi) = f(\xi) + k(\xi)\phi^2 + a \exp\{-\xi^2/(2s)\}\phi^2$, we chose the parameters as $a = 0.1$ and $s = 0.2$ (see Fig. 2).

Next we shifted the zero of the new term along the ϕ axis, thus slightly changing the minimum energy path in the region of the transition state. The new function was $E_{2b}(\xi, \phi) = f(\xi) + k(\xi)\phi^2 + a \exp\{-\xi^2/(2s)\}(\phi + \Delta)^2$ (see Fig. 3). We confirmed that MC calculations of the reaction and activation free energies for these modified potentials were as accurate as before (for the $E_1(\xi, \phi)$ potential, see above) when the energy differences were evaluated with the same potential that was used for sampling.

We then ran MC simulations with the same parameters as before, with the sampling done on the $E_{2a}(\xi, \phi)$ potential and the evaluation of ΔE_{pert} done on the $E_{2b}(\xi, \phi)$ potential. As expected, the results deteriorate with increasing values of the shift parameter Δ that governs the deviation from the $E_{2a}(\xi, \phi)$ sampling potential. In the sequence $\Delta = 0.05, 0.1,$ and 0.2, the error in the activation free energy rises from 0.45 kJ/mol via 2.33 kJ/mol to 9.74 kJ/mol. Due to the deliberate choice of the shape of the potentials (see above), the error in the reaction free energy grows much more slowly, from 0.00 kJ/mol via 0.17 kJ/mol to 0.92 kJ/mol, respectively.

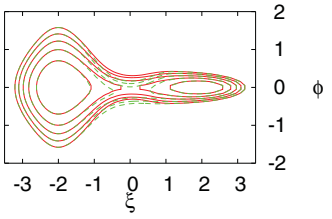


FIG. 3. Contour plot of the analytic potentials $E_{2a}(\xi, \phi)$ (solid lines) and $E_{2b}(\xi, \phi)$ (dashed lines) with a contour spacing of 0.005 and $\Delta = 0.2$. All values in atomic units.

The drastic rise of the error in the activation free energy confirms the importance of having sufficient overlap between the configurational phase space accessible on the two surfaces. At the transition state, the ϕ values that can be sampled on the $E_{2a}(\xi, \phi)$ potential range from -0.2 to 0.2 due to the steep rise of energy along the ϕ axis. Therefore, as soon as the transition state on the $E_{2b}(\xi, \phi)$ potential is moved close to the border of the ϕ values accessible at the $E_{2a}(\xi, \phi)$ level, we no longer sample the correct configurational space, and hence the computed activation free energy can no longer be trusted.

The DH-FEP method is thus clearly sensitive to the degree of the overlap of configurational phase space between the two potentials that are used for sampling and for evaluating ΔE_{pert} at the sampled geometries. Therefore the geometrical correspondence of the two potentials along the RC must be carefully checked prior to free energy calculations.

B. Chorismate Mutase

As second example we chose a "real-life" QM/MM system and calculated the activation free energy of the Claisen rearrangement of chorismate to prephenate, catalyzed by the *Bacillus subtilis* Chorismate Mutase (BsCM) enzyme. This reaction is a key step on the shikimate pathway of the aromatic amino acid synthesis in plants, fungi, and bacteria. It has been intensively investigated theoretically.[46] One peculiar trait of this system is the lack of covalent bonds between the substrate and the protein environment during the whole reaction, making it a rather convenient model for testing QM/MM methods. Experimentally, the entropic contribution to the activation free energy has been determined[47] to be $T\Delta S = -11.4 \pm 1.5$ kJ/mol at $T = 300$ K, which may serve as a reference value for assessing the results from QM/MM free energy calculations. In our present work on BsCM, we first focus on technical issues relevant to the proposed DH-FEP approach: we test the number of steps that may be skipped between two subsequent ΔE_{pert} evaluations, as well as the overall number of MD steps needed to obtain converged results, and we address the problem of configurational phase space overlap between the two potentials and how this affects the results.

In the QM/MM calculations, we treated the substrate (24 atoms) at the QM level (OM3, SCC-DFTB, RI-MP2/SVP) and the rest of the system comprising the protein and the solvent shell (13421 atoms in total) with the CHARMM22 force field.[44] The initial preparation of the system has been described elsewhere.[48] The first MD snapshot from the previous study[48] was subjected to further MD sampling using CHARMM33b1,[49,50] and six independent new snapshots were randomly chosen from this MD run.

Following standard conventions, we first defined the RC as the difference between the lengths of the breaking C–O and the forming C–C bond (see Fig. 4). Potential energy profiles

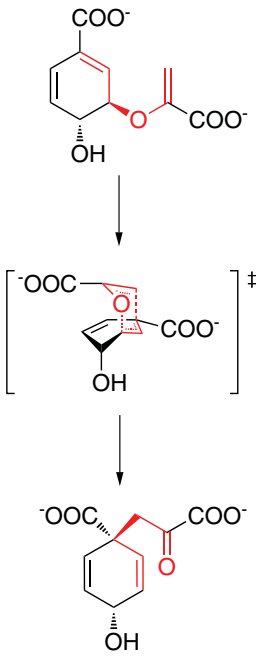


FIG. 4. Claisen rearrangement of chorismate to prephenate in chorismate mutase. The two parallel red dashed lines in the transition state indicate the forming and the breaking bonds. The difference between the corresponding distances is the reaction coordinate.

were calculated at all applied QM/MM levels (see above) for all the snapshots, via a series of restrained optimizations with the RC being sequentially changed from −2.4 Å to 2.4 Å in steps of 0.05 Å. For some of the snapshots, the reaction pathways were calculated several times in forward and backward direction until any unevenness was removed from the potential energy profile. Subsequent transition state optimizations and intrinsic reaction coordinate computations confirmed that the chosen RC is perfectly adequate and a valid reference to perform the FEP calculations.

In DH-FEP applications, two parameters need to be set, namely the number of steps skipped between two subsequent perturbations (x) and the total number of ΔE_{pert} evaluations to be performed. We have tested these options in a single MD run for an arbitrarily chosen window (at RC = −1.15 Å). The system was heated up to 300 K in steps of 10 K during 3 ps and then equilibrated for 20 ps, before the sampling was performed for 25 ps with ΔE_{pert} evaluated at every step. In this and all further MD calculations, the step size was 1 fs. All MD simulations were run in the canonial ensemble using the Nosé-Hoover chain thermostat[51,52] with a chain length of 4 and a characteristic time for the first thermostat of 0.02 ps. We used OM3/CHARMM both for sampling and for evaluating ΔE_{pert}. Results for different values of x with the number of ΔE_{pert} evaluations fixed to 1000 are shown in Fig. 5. We depict both the direct exponential average of all ΔE_{pert} evaluations taken (dashed) and the values obtained via cumulant expansion from a reduced number of ΔE_{pert} values (solid) selected after applying the statistical test on the lack of trend. The error bars shown in Fig. 5 refer to the latter; they were evaluated according to Ref. 33. The two sets of data do not deviate significantly, reflecting the lack of trend in most

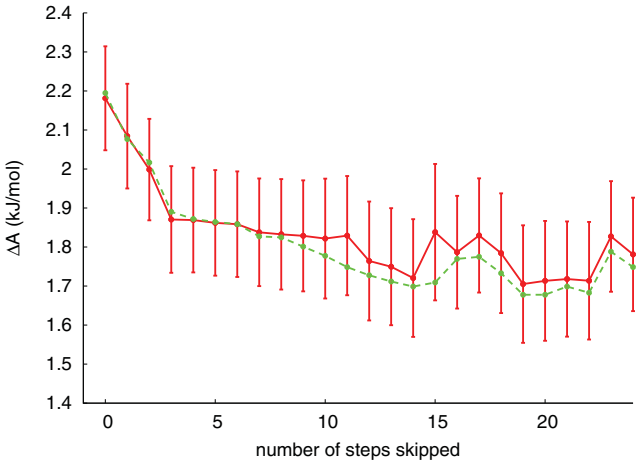


FIG. 5. Free energy difference between two windows calculated with a different number of steps skipped between the ΔE_{pert} evaluations, with the overall number of these evaluations fixed to 1000. The values in red (solid line) were obtained after subjecting the data to statistical tests for lack of trend and decorrelation. The values in green (dashed line) were obtained from direct exponential averaging of all data points. Data were taken starting from the end of a 25 ps OM3/CHARMM MD sampling run of one of the windows along the CM reaction profile (see text for further details).

datasets. ΔA converges with increasing x, showing that the decreasing dependency between subsequent ΔE_{pert} calculations improves the quality of the sampling. For x between 0 and 4, the free energy is clearly not converged, while values above 10 seem to be a reasonable choice. In this study, we adopted $x = 14$ (i.e., we evaluate ΔE_{pert} at every 15th step) since ΔA fluctuates around some average value for higher x. In an additional test, we have confirmed that this remains true up to $x = 149$, i.e., when extending the time between ΔE_{pert} evaluations up to 150 fs (see Fig. S1a of the supplementary material[53]).

Concerning the second option, Fig. 6 shows the variation of ΔA against the overall number of steps taken, with a fixed

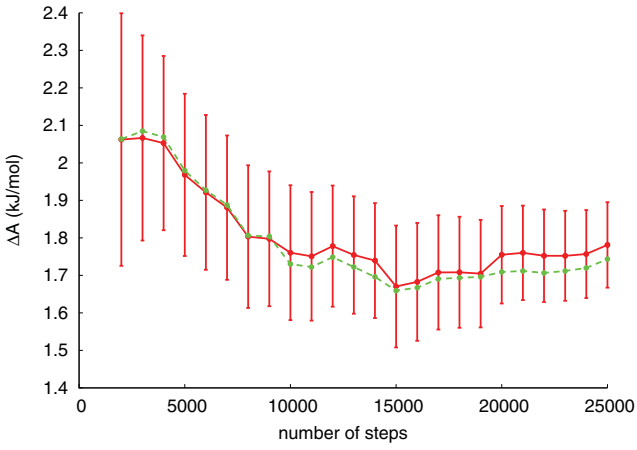


FIG. 6. Free energy difference between two windows calculated with 14 steps skipped between two ΔE_{pert} evaluations, with the overall number of these evaluations being varied. The values in red (solid line) were obtained after subjecting the data to statistical tests for lack of trend and decorrelation. The values in green (dashed line) were obtained from direct exponential averaging of all data points. Data were taken during a 25 ps OM3/CHARMM MD sampling run of one of the windows along the CM reaction profile (see text for further details).

value of $x = 14$. ΔA seems to converge after MD sampling times of around 10 ps. As expected, the error bar for ΔA decreases with increasing sampling time, i.e., with the number of ΔE_{pert} evaluations performed. This kind of convergence is confirmed by further test calculations with sampling times up to 105 ps (see Fig. S1b of the supplementary material[53]). In the following, we normally limit ourselves to 10 ps of sampling and use $x = 14$ throughout. We note in this context that more extensive sampling will often be hardly affordable in practice when the ΔE_{pert} evaluations are carried out with a high-level QM method.

Next we assess the accuracy of our method by comparing it to the well-established thermodynamic integration (TI) method. Both the sampling and ΔE_{pert} evaluations were performed at the OM3/CHARMM level. Since we focus on the activation free energy, we only considered the first 50 windows from the energy profile, covering the reactant minimum and transition state areas. The MD calculations were done for four snapshots in the following way: in every window, the system was first heated up to 300 K in steps of 10 K during 3 ps, then equilibrated for 25 ps, and finally sampled for 15 ps, with ΔE_{pert} being computed at every 15th step.

The results from the OM3/CHARMM FEP runs were in good agreement with those from TI calculations performed for the same snapshots with the same MD parameters: for all four snapshots tested, the activation free energies agreed to within 0.8 kJ/mol, which is of the same order as the error estimate[29] of 1.0 kJ/mol for the TI values with the currently adopted setup. The computed activation free energies $\Delta A^{\ddagger}$ for the four snapshots range between 66.5 and 71.5 kJ/mol, hence the snapshot-dependent fluctuations are significantly larger than the uncertainties in the TI and FEP calculations (both run on the same single potential surface). We have also tested the convergence of the OM3/CHARMM FEP results for one particular snapshot with regard to the MD sampling time in the FEP procedure: when prolonging the sampling time per window from 15 to 105 ps, the resulting free energy profiles remain virtually identical (see Fig. S2 of the supplementary material[53]), the activation free energies agree to within 0.2 kJ/mol, and the associated uncertainties decrease from 0.7 to 0.3 kJ/mol.

We now test the central DH-FEP approximation, namely the use of two different QM Hamiltonians in the QM/CHARMM calculations: OM3 or SCC-DFTB for sampling and MP2/SVP for evaluating ΔE_{pert}. As shown in Subsection IV A for the analytic model potential, reasonably accurate DH-FEP results can be expected only if the two QM methods yield reasonably similar geometries along the RC. To check this crucial DH-FEP issue, we define two criteria of geometrical correspondence: first, the interatomic distances entering the expression for the RC, and second, the root-mean-square deviation (RMSD) between the geometries of the whole QM region along the RC.

For a given value of RC defined as the difference of the distances in the forming C–O and the breaking C–C bond, restrained QM/MM optimizations (as well as constrained QM/MM dynamics) with two different QM methods will give different individual C–O and C–C distances, and therefore comparison of these distances can be a straightforward way to examine the geometrical correspondence of the two QM methods. Fig. 7 shows that the optimized distances from OM3/CHARMM nearly coincide with those from QM(MP2/SVP)/CHARMM up to RC = −1.4 Å, but start to deviate thereafter, with the difference growing up to 0.4 Å at RC = 0. The optimized C–O and C–C distances from SCC-DFTB/CHARMM show the opposite behavior: they differ from the QM(MP2/SVP)/CHARMM distances somewhat up to RC = −1.1 Å, but then follow them closely up to RC = 0 except for the region of RC = {−0.6 Å, −0.2 Å}. Concerning the RMSD values for the optimized QM regions relative to the QM(MP2/SVP)/CHARMM geometries along the RC: they vary from 0.04 to 0.06 Å for SCC-DFTB (being lowest in the region of RC = {−1.2 Å, 0.0 Å}) while they range from 0.06 to 0.09 Å for OM3 (being lowest for RC = {−1.15 Å, −0.7 Å}).

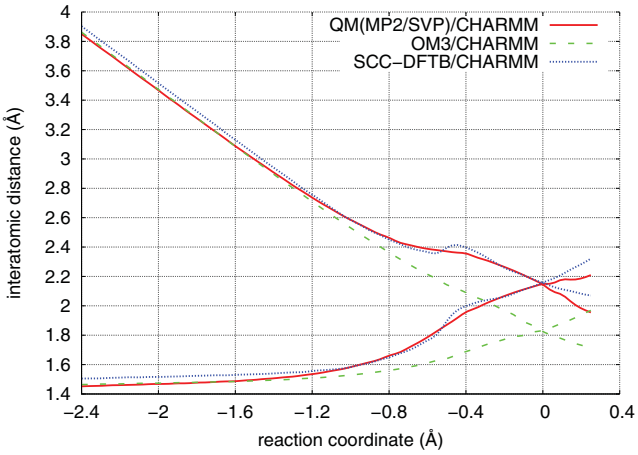


FIG. 7. Optimized C–C and C–O distances along the RC for the three different QM methods.

Going beyond geometry considerations, we performed a series of QM(MP2/SVP)/MM single-point energy calculations at the optimized OM3/MM and SCC-DFTB/MM geometries along the RC (see Fig. 8). None of the resulting two curves was exactly matching the QM(MP2/SVP)/MM

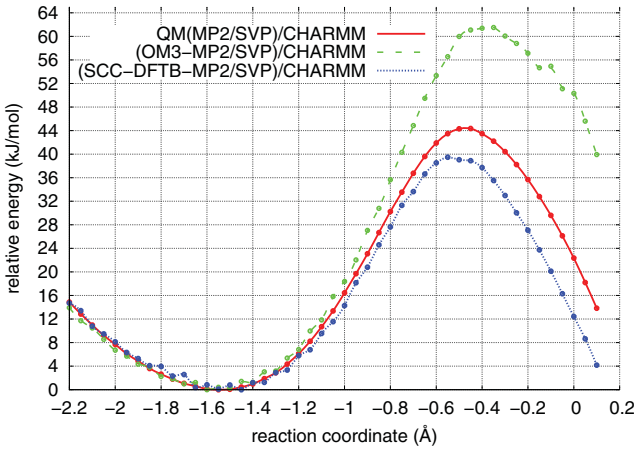


FIG. 8. Potential energy profile computed at the QM(MP2/SVP)/CHARMM level and QM(MP2/SVP)/CHARMM single-point energies at the optimized OM3 and SCC-DFTB structures along the reaction path.

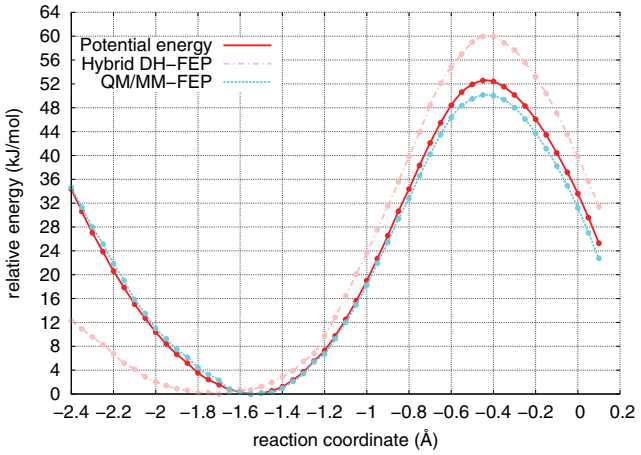


FIG. 9. Potential energy, DH-FEP, and QM/MM-FE profiles obtained for snapshot 6. The potential energy was computed at the QM(MP2/SVP)/CHARMM theory level. The DH-FEP profile was determined with a hybrid approach, in which the first part of the reaction path was sampled with OM3/CHARMM, and the second part with SCC-DFTB/CHARMM, while ΔE_{pert} was evaluated with QM(MP2/SVP)/CHARMM. The conventional QM/MM-FE profile was computed at the QM(MP2/SVP)/CHARMM level.

energy profile, but the relative energies computed at the SCC-DFTB/MM geometries were clearly closer to the QM(MP2/SVP)/MM reference values.

In an overall assessment of the QM/MM geometries for BsCM, SCC-DFTB thus seems superior to OM3 in reproducing the MP2-based results, and hence it should be a good choice for performing the sampling in QM/MM DH-FEP calculations. However, the corresponding QM(MP2/SVP//SCC-DFTB)/MM DH-FEP results (see Fig. S3 of the supplementary material[53]) were unsatisfactory: the DH-FEP free energy profile started rising much too fast at an early stage of the reaction close to the reactant state, and the activation free energy was too high compared with the QM(MP2/SVP)/MM reference value. Moreover, these calculations failed to reproduce the entropic contribution to the activation free energy that is known experimentally (see above). By contrast, the QM(MP2/SVP//OM3)/MM DH-FEP free energy profile was found to "behave" very well close to the reactant equilibrium, but to become quite different in shape from the QM(MP2/SVP)/MM reference curve closer to the TS, as expected from the geometry correspondence tests (see above).

Given the fact that neither OM3 nor SCC-DFTB provides sufficiently accurate QM/MM geometries along the whole RC, we decided to test a hybrid approach, running the MD sampling for the first part of the reaction (RC = {−2.4 Å, −1.25 Å}) with OM3/MM and using SCC-DFTB/MM for the second part (RC = {−1.25 Å, 0.1 Å}). To limit the computational effort for the MP2-based evaluation of ΔE_{pert}, the MD procedure was slightly changed: the heating was done in steps of 5 K during 6 ps, thereafter the system was equilibrated for 15 ps and sampled for 10 ps. We thus performed 1333 MP2/CHARMM calculations per window. The hybrid approach (dashed-dotted curve in Fig. 9) gave satisfactory results: the difference between $\Delta E^{\ddagger}$ and $\Delta A^{\ddagger}$ ranged from −2.0 to −18.0 kJ/mol for the individual snapshots, with an average value of −10.3 kJ/mol and a confidence interval for the barrier of about 1 kJ/mol. Taking into account the difference ΔE_{QM}^{ZPE} between the zero-point vibrational energies of TS and reactant (−4.2 kJ/mol for each snapshot in harmonic approximation) and assuming the thermal corrections ΔU^{th} to be negligible, we arrive at an average $T\Delta S^{\ddagger}$ value of −14.5 kJ/mol, which is close to the experimental result of −11.4 ± 1.5 kJ/mol.[47] It is obvious from Table I that the $\Delta A^{\ddagger}$ value fluctuates much less from snapshot to snapshot than the $\Delta E^{\ddagger}$ value, implying that the sampling was adequate. The fluctuations in the entropic contributions ($\Delta E^{\ddagger} - \Delta A^{\ddagger}$) thus mainly arise from the differences in the energy barriers for the individual snapshots.

The error estimates given in Table I account only for statistical fluctuations and incomplete sampling during the MD runs. They do not include errors caused by an insufficient overlap of the two underlying configurational spaces, as we do not apply an explicit reweighting of the semiempirical surface via FEP, as done, e.g., in Refs. 12 and 27. In the latter work, the errors associated with the perturbations along the reaction coordinate were fairly small (as in our case), while those associated with the perturbations in the method space (avoided in our approach) were rather large, thus raising general concerns about using semiempirical methods to provide the reference potential. We note that there was no attempt in Ref. 27 to evaluate the configurational space overlap between the chosen semiempirical and higher-level QM method prior to performing MD simulations, or to go beyond standard MNDO-type semiempirical methods. Doing so may enhance the quality of the reference potential in such dual-level free energy calculations.

For comparison, we also performed conventional QM/MM-FE calculations[26,29] for snapshot 6 (see Fig. 9). As expected from the lack of sampling in the QM region, the entropic contribution is underestimated: the free energy profile basically follows the potential energy profile, and the TS is even slightly lower, suggesting an entropic contribution with the wrong sign. Following the conventional procedure,[26,29] the entropic contribution for the QM region can be evaluated at the stationary points using the rigid-rotor

TABLE I. Free energy and potential energy barriers and entropic contributions to the barrier of the BsCM-catalyzed reaction for the six snapshots considered. All values in kJ/mol.

	Snapshots						Average	Exp.[47]
Snapshot number	1	2	3	4	5	6		
$\Delta A^{\ddagger}$	57.1±0.7	59.2±0.7	62.4±0.9	56.7±0.7	62.1±0.8	60.0±0.7	59.6±0.75	64.4
$\Delta E^{\ddagger}$	47.5	41.9	44.4	49.2	60.1	52.6	49.3	
$\Delta E^{\ddagger} - \Delta A^{\ddagger}$	− 9.6±0.7	− 17.3±0.7	− 18.0±0.9	− 7.5±0.7	− 2.0±0.8	− 7.4±0.7	− 10.3±0.75	− 11.4±1.5

harmonic-oscillator approximation; this gives a $T\Delta S^{\ddagger}_{QM}$ contribution of 2.5 kJ/mol, which is clearly too small to get close to the experimental value of the entropic contribution (see above). This example confirms that the degrees of freedom in the QM region should also be sampled to obtain a realistic entropic contribution to activation free energies in chemical reactions.

Our results with the hybrid approach indicate that the DH-FEP approach can provide free energies that closely mimic those from high-level QM/MM approaches, if the low-level QM/MM approach used for sampling yields realistic geometries along the RC (close to the high-level QM/MM geometries). However, such close matching of low-level and high-level geometries, e.g., from semiempirical and *ab initio* QM/MM calculations, may not always be achievable, as presently demonstrated for OM3 or SCC-DFTB versus MP2/SVP. In such cases, we can generalize the DH-FEP strategy by using more than one constraint, based on the observation that it is crucial to match the decisive geometrical variables entering the RC. In the case of BsCM, instead of only constraining the RC (i.e., the difference between the distances of the forming C–O and the breaking C–C bond), we now constrain the individual C–O and C–C distances to their reference values from QM(MP2/SVP)/MM restrained optimizations. This choice removes two DOFs of the QM region from sampling (rather than one DOF as before) and may thus entail the risk to underestimate the entropic contributions. This disadvantage is expected to be outweighed by the advantage of sampling a more appropriate configurational phase space, with better coverage of the region that is important in the high-level treatment.

We checked the performance of this collective coordinate approach by running DH-FEP calculations for snapshot 6, constraining both relevant C–O and C–C distances separately and using either OM3/CHARMM or SCC-DFTB/CHARMM for sampling throughout the whole reaction (see Fig. 10).

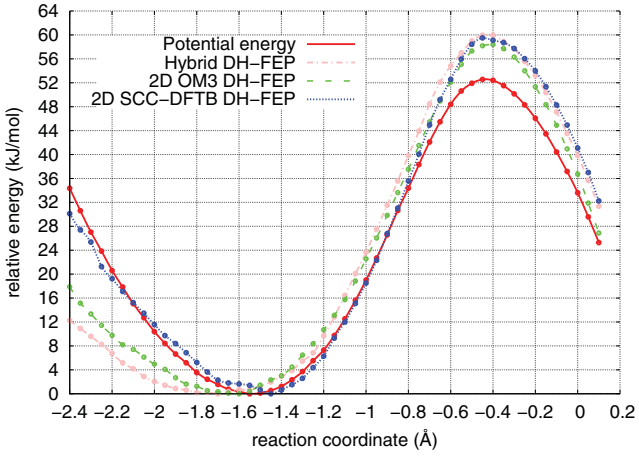


FIG. 10. Potential energy profile from QM(MP2/SVP)/CHARMM calculations and three free energy profiles computed for snapshot 6. Hybrid DH-FEP profile, sampling with OM3/CHARMM and SCC-DFTB/CHARMM for the first and second part of the reaction path, respectively (see text); 2D DH-FEP profiles, evaluated with the use of a two-dimensional collective coordinate, sampling with OM3/CHARMM and with SCC-DFTB/CHARMM. ΔE_{pert} obtained from QM(MP2/SVP)/CHARMM single-point calculations (see text).

Both DH-FEP calculations gave similar free energy profiles and reproduced the $\Delta A^{\ddagger}$ values that had previously been obtained with the hybrid DH-FEP approach. The use of a collective coordinate (here composed of the two relevant interatomic distances) in the DH-FEP calculations thus helps to overcome the limitations associated with the use of a single one-dimensional RC.[32]

The DH-FEP treatment may thus be improved by the judicious choice of an appropriate collective coordinate, thereby replacing the single constraint on the RC with two (or more) constraints on suitably chosen DOFs. This allows for successful applications even when there are appreciable differences between the low-level and high-level geometries along the reaction path. Obviously, a careful analysis of these differences is essential for identifying the DOFs that should enter the collective coordinate and be constrained in the DH-FEP calculations. Compared with the conventional QM/MM-FE procedure,[26,29] the DH-FEP approach, regardless of whether used with a single or a collective reaction coordinate, is expected to give a better estimate of the entropic contributions to the free energy profile, because of the explicit sampling of most of the QM region.

V. CONCLUSION

We have presented the DH-FEP method for evaluating free energies differences in large QM/MM systems. Compared with the conventional QM/MM-FE approach,[26,29] our method samples not only the MM region but also the QM region, i.e., the full configurational space except for the reaction coordinate. For the sake of computational efficiency, we introduced the approximation to use a less expensive low-level QM/MM method for sampling, while the perturbation energy differences ΔE_{pert} are evaluated through higher-level single-point QM/MM calculations performed at regular intervals, after skipping a pre-determined number of MD sampling steps. We examined the performance of our method using two test systems, namely a two-dimensional analytic model potential and a prototypical enzymatic reaction, the chorismate-to-prephenate conversion catalyzed by the BsCM enzyme.

Our implementation of the FEP approach was validated using the same potential for sampling and for evaluating ΔE_{pert} (i.e., a single Hamiltonian approach). The FEP results were shown to accurately reproduce the exact solutions for an analytic model potential and the activation free energy of the BsCM reaction obtained from standard thermodynamic integration.

In the numerical tests of the dual Hamiltonian approximation for the analytic model potential, the computed free energies were found to be quite sensitive to the overlap of the two surfaces in the region accessible to the sampling, thus calling for a careful analysis of the geometrical correspondence between the low-level and high-level methods chosen for DH-FEP calculations.

In the QM/MM tests for the enzymatic BsCM reaction, we first determined the necessary simulation parameters: we found that it was sufficient to evaluate ΔE_{pert} every 15 MD steps and to sample for at least 10 ps to obtain

results that are converged well enough. The subsequent DH-FEP QM/MM calculations employed the semiempirical OM3 and SCC-DFTB QM methods for sampling and the *ab initio* MP2/SVP approach for evaluating ΔE_{pert}. In the basic DH-FEP treatment, we constrained only the RC (defined as the difference between the distances of the forming C–O and the breaking C–C bond). The quality of the DH-FEP results was found to depend on the similarity between the low-level and high-level QM/MM structures along the RC: neither OM3 nor SCC-DFTB provided a good match to the MP2-based geometries over the entire RC, while being reasonably accurate in complementary regions of the reaction path. More realistic DH-FEP results could be obtained by a hybrid approach, in which the reaction path was divided into two regions, each described with the most suitable semiempirical method: the computed entropic contribution to the activation free energy was close to the experimental value.

Closer analysis of these DH-FEP QM/MM results for BsCM revealed that the crucial indicator of success is not the RMSD between the low-level and high-level QM/MM structures along the RC, but rather the match of the C–O and C–C distances used to define the RC (see above). Therefore, we applied the more general collective coordinate approach, with separate constraints on these two distances, to ensure an improved sampling of the relevant configurational space. The corresponding results were very close to the those from the hybrid approach, regardless of whether OM3 or SCC-DFTB was used for sampling. We thus recommend to use such a collective RC whenever the analysis of the low-level and high-level QM/MM structures along the RC reveals substantial discrepancies. A suitable collective RC can be defined by proceeding as follows. First, high-level QM/MM calculations are performed to locate the relevant transition state and the reaction path that connects it with the reactants and products. A natural choice for determining the reaction path is to follow the intrinsic reaction coordinate (IRC) starting from the optimized transition state, which can efficiently be done at the QM/MM level by an approximate microiterative scheme.[54] The IRC can then be used to identify the (small) set of internal coordinates, e.g., of individual interatomic distances, that undergo the most drastic changes along the reaction path and that should thus enter the collective RC for the subsequent DH-FEP calculations.

Going beyond this type of RC-based DH-FEP approach, one may attempt to devise procedures that directly control the space being sampled, for example by using MC techniques with update criteria based on the overlap between the two configurational spaces as suggested previously in a different context.[22] Alternatively, one may implement a DH-FEP scheme, in which the geometries and energy differences are stored during MD sampling, with the energy differences being weighted according to phase space overlap criteria at the end. Generally speaking, it is advisable to examine whether there is sufficient similarity of the geometries and sufficient overlap of the configurational phase spaces obtained with the low-level and high-level QM/MM methods used in the DH-FEP approach. If this is the case, DH-FEP offers an efficient opportunity to calculate accurate free energy differences in large QM/MM systems.

This approach can become even more valuable with the increase of computer power that will allow for future large-scale sampling at more expensive first-principles QM/MM levels, which may then enable even more accurate free energy evaluations, e.g., with larger basis sets or coupled cluster QM methods.

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulation. From Algorithms to Applications*, 2nd ed. (Academic Press, 2002).
[2] G. M. Torrie and J. P. Valleau, Chem. Phys. Lett. **28**, 578 (1974).
[3] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[4] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[5] H. M. Senn and W. Thiel, Angew. Chem., Int. Ed. **48**, 1198 (2009).
[6] R. P. Muller and A. Warshel, J. Phys. Chem. **99**, 17516 (1995).
[7] J. Bentzien, R. P. Muller, J. Florin, and A. Warshel, J. Phys. Chem. B **102**, 2293 (1998).
[8] M. Štrajbl, G. Hong, and A. Warshel, J. Phys. Chem. B **106**, 13333 (2002).
[9] M. H. M. Olsson, G. Hong, and A. Warshel, J. Am. Chem. Soc. **125**, 5025 (2003).
[10] E. Rosta, M. Klähn, and A. Warshel, J. Phys. Chem. B **110**, 2934 (2006).
[11] N. V. Plotnikov, S. C. L. Kamerlin, and A. Warshel, J. Phys. Chem. B **115**, 7950 (2011).
[12] N. V. Plotnikov and A. Warshel, J. Phys. Chem. B **116**, 10342 (2012).
[13] T. H. Rod and U. Ryde, Phys. Rev. Lett. **94**, 138302 (2005).
[14] T. H. Rod and U. Ryde, J. Chem. Theory Comput. **1**, 1240 (2005).
[15] J. Chandrasekhar, S. F. Smith, and W. L. Jorgensen, J. Am. Chem. Soc. **106**, 3049 (1984).
[16] J. Chandrasekhar, S. F. Smith, and W. L. Jorgensen, J. Am. Chem. Soc. **107**, 154 (1985).
[17] W. L. Jorgensen, Acc. Chem. Res. **22**, 184 (1989).
[18] R. V. Stanton, M. Perkyl, D. Bakowies, and P. A. Kollman, J. Am. Chem. Soc. **120**, 3448 (1998).
[19] P. A. Kollman, B. Kuhn, O. Donini, M. Perakyla, R. Stanton, and D. Bakowies, Acc. Chem. Res. **34**, 72 (2001).
[20] B. Kuhn and P. A. Kollman, J. Am. Chem. Soc. **122**, 2586 (2000).
[21] O. Donini, T. Darden, and P. A. Kollman, J. Am. Chem. Soc. **122**, 12270 (2000).
[22] R. Iftimie, D. Salahub, D. Wei, and J. Schofield, J. Chem. Phys. **113**, 4852 (2000).
[23] R. Iftimie, D. Salahub, and J. Schofield, J. Chem. Phys. **119**, 11285 (2003).
[24] P. Bandyopadhyay, J. Chem. Phys. **122**, 091102 (2005).
[25] C. J. Woods, F. R. Manby, and A. J. Mulholland, J. Chem. Phys. **128**, 014109 (2008).
[26] Y. Zhang, H. Liu, and W. Yang, J. Chem. Phys. **112**, 3483 (2000).
[27] J. Heimdal and U. Ryde, Phys. Chem. Chem. Phys. **14**, 12592 (2012).
[28] E. Rosta, M. Nowotny, W. Yang, and G. Hummer, J. Am. Chem. Soc. **133**, 8934 (2011).
[29] H. M. Senn, S. Thiel, and W. Thiel, J. Chem. Theory Comput. **1**, 494 (2005).
[30] J. Kästner and W. Thiel, J. Chem. Phys. **123**, 144104 (2005).
[31] J. Ruiz-Pernia, E. Silla, I. Tunon, S. Marti, and V. Moliner, J. Phys. Chem. B **108**, 8427 (2004).
[32] J. J. Ruiz-Pernia, E. Silla, I. Tunon, and S. Marti, J. Phys. Chem. B **110**, 17663 (2006).
[33] J. Kästner, H. M. Senn, S. Thiel, N. Otte, and W. Thiel, J. Chem. Theory Comput. **2**, 452 (2006).
[34] P. Sherwood, A. de Vries, M. Guest, G. Schreckenbach, C. Catlow, S. French, A. Sokol, S. Bromley, W. Thiel, A. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjovoll, A. Fahmi, A. Schäfer, and C. Lennartz, J. Mol. Struct.: THEOCHEM **632**, 1 (2003).
[35] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[36] W. Thiel, *MNDO2005*, version 7.0; Max-Planck-Institut für Kohlenforschung: Mülheim, 2005.
[37] M. Scholten, Ph.D. thesis, Universität Düsseldorf, 2003.
[38] N. Otte, M. Scholten, and W. Thiel, J. Phys. Chem. A **111**, 5751 (2007).
[39] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Phys. Rev. B **58**, 7260 (1998).
[40] R. Ahlrichs, M. Bär, M. Häser, H. Horn, and C. Kölmel, Chem. Phys. Lett. **162**, 165 (1989).

[41]F. Weigend and M. Häser, Theor. Chem. Acc. **97**, 331 (1997).
[42]F. Weigend, M. Häser, H. Patzelt, and R. Ahlrichs, Chem. Phys. Lett. **294**, 143 (1998).
[43]W. Smith and T. Forester, J. Mol. Graphics **14**, 136 (1996).
[44]A. MacKerell, D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[45]N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[46]F. Claeyssens, K. E. Ranaghan, N. Lawan, S. J. Macrae, F. R. Manby, J. N. Harvey, and A. J. Mulholland, Org. Biomol. Chem. **9**, 1578 (2011).
[47]P. Kast, M. Asif-Ullah, and D. Hilvert, Tetrahedron Lett. **37**, 2691 (1996).
[48]H. M. Senn, J. Kästner, J. Breidung, and W. Thiel, Can. J. Chem. **87**, 1322 (2009).
[49]B. Brooks, R. Bruccoleri, D. Olafson, D. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[50]B. R. Brooks, C. L. Brooks III, A. D. MacKerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, J. Comput. Chem. **30**, 1545 (2009).
[51]S. Nose, J. Chem. Phys. **81**, 511 (1984).
[52]W. Hoover, Phys. Rev. A **31**, 1695 (1985).
[53]See supplementary material at http://dx.doi.org/10.1063/1.4817402 for additional evaluation of the effect of sampling time on free energy convergence and for OM3/CHARMM and SCC-DFTB/CHARMM DH-FEP profiles.
[54]I. Polyak, E. Boulanger, K. Sen, and W. Thiel, "A microiterative intrinsic reaction coordinate method for large QM/MM systems," Phys. Chem. Chem. Phys. (published online).

Photochemical Steps in the Prebiotic Synthesis of Purine Precursors from HCN.

Eliot Boulanger, Anakuthil Anoop, Dana Nachtigallova, Walter Thiel, and Mario Barbatti

Angew. Chem. Int. Ed. **2013**, 52, 8000-8003.

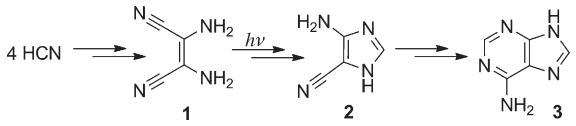
Prebiotic Photochemistry

DOI: 10.1002/anie.201303246

Photochemical Steps in the Prebiotic Synthesis of Purine Precursors from HCN**

*Eliot Boulanger, Anakuthil Anoop, Dana Nachtigallova, Walter Thiel, and Mario Barbatti**

Hydrogen cyanide (HCN) chemistry is believed to be an important part of the abiotic synthesis of organic materials, including nucleobases, amino acids, and oligopeptides.[1] One of the most probable routes for the synthesis of purine nucleobases and nucleotides[2] in the prebiotic world (Scheme 1) involves HCN oligomerization into the tetramer

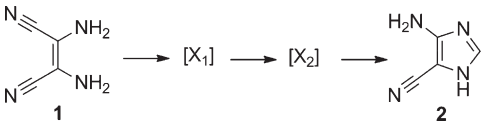


Scheme 1. Conversion of the HCN tetramer *cis*-DAMN (**1**) into AICN (**2**), a key intermediate in the synthesis of purine nucleobases and nucleotides. Adenine (**3**) is shown as one possible product.

cis-2,3-diaminomaleonitrile (*cis*-DAMN, **1**), which may be converted photochemically into an imidazole intermediate (4-amino-1*H*-imidazole-5-carbonitrile, AICN, **2**).[3] Although this reaction has been investigated in detail since its discovery by Ferris and Orgel in 1966,[1,4] the mechanism of the photochemical steps remains unresolved. Herein, we address this issue from a theoretical perspective: by the use of computational chemistry and chemical kinetics we show that among a number of possibilities, including all those previously proposed, there is only one sequence of steps that is thermodynamically and kinetically compatible with the experimental conditions.

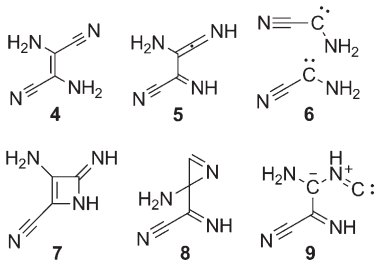
One of the most appealing features of the DAMN→AICN reaction is its robustness.[5] It was observed in a large variety of solvents (polar and nonpolar), with several enaminonitrile derivatives, and at diverse concentrations and temperatures.[6] The imidazole derivative **2** is photostable (5% reduction in absorbance after irradiation at 254 nm for 3 h;[4d] see also Ref. [7] on imidazole photostability) and resistant to hydrolysis (lifetime: 2000 years at pH 8[8]). These characteristics imply that different prebiotic environments, either terrestrial or extraterrestrial, could have been the source of AICN (**2**) in the prebiotic world.[5] The accumulation of AICN, however, requires relatively large HCN concentrations ($>10^{-2}$ M).[9] This requirement sets a first relevant environmental constraint: such high HCN concentrations are only possible in low-temperature environments, such as ice and eutectic water–HCN phases. Therefore, any realistic mechanism cannot count on high thermal energy in addition to the photon energy.

The number of photons and intermediates involved in the photochemical steps is unknown (Scheme 2). The process



Scheme 2. Photochemical steps in the DAMN→AICN reaction.

starts with photoexcitation of *cis*-DAMN (**1**), the first clearly stable HCN oligomer.[10] On the basis of the spectral shifts observed after the irradiation of *cis*-DAMN, Yamada et al.[4e] showed that the first intermediate (X_1) is the *trans*-DAMN isomer **4** (Scheme 3). Irradiation leads to a photostationary



Scheme 3. Previously proposed intermediates.

state with a large predominance of *trans*- (**4**) over *cis*-DAMN (**1**).[4c] Becker et al.[4d] raised the possibility that carbenes, **6**, are the first intermediate, instead of *trans*-DAMN (**4**). However, the absence of cross-products in the experiments of Ferris et al. ruled this hypothesis out.[11]

Little is known about X_2 (Scheme 2), which may represent more than one intermediate. Infrared spectra in a liquid film and a KBr matrix indicated that X_2 may possess a ketenimine group (2000–2020 cm^{-1})[12] and thus indicated a possible

[*] E. Boulanger, Prof. Dr. W. Thiel, Dr. M. Barbatti
Max-Planck-Institut für Kohlenforschung
Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr (Germany)
E-mail: barbatti@kofo.mpg.de
Homepage: http://www.sgk.mpg.de/private/barbatti
Dr. A. Anoop
Department of Chemistry, Indian Institute of Technology
Kharagpur 721302 (India)
Dr. D. Nachtigallova
Institute of Organic Chemistry and Biochemistry, AS CR
Flemingovo nam. 2, 166 10 Praha 6 (Czech Republic)

[**] We acknowledge fruitful discussions with Dr. M. Patil. The research at IOCB was part of the project RVO: 61388963.

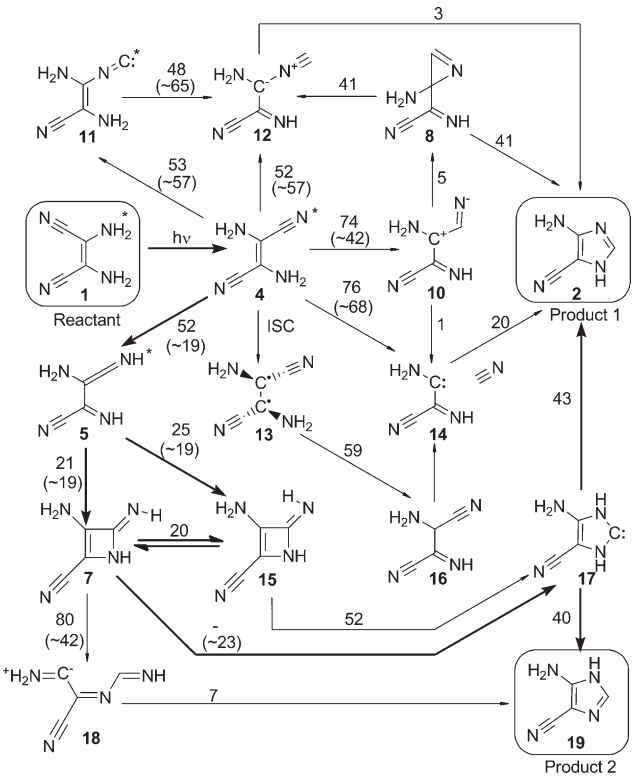
Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201303246.

hydrogen-atom transfer from one of the amino groups of **4** to form 2-amino-3-iminoacrylimidoyl cyanide (AIAC, **5**). The subsequent formation of AICN (**2**) requires CN cleavage, for which pathways via the azetene **7**, azirene **8**, and a formal zwitterion **9** have been proposed, without any consensus on which pathway would be predominant.[4a,6]

In this study, we investigated a large number of possible reaction pathways. We examined both thermodynamic and kinetic aspects of the pathways. One first relevant fact revealed by our simulations is that although energy of 4 eV is added to the system by the photoexcitation, most of this energy is quickly dissipated to the environment after relaxation to the ground state. In water, this dissipation happens within about 0.2 ps after internal conversion to the ground state (see the Computational Section). This ultrafast energy dissipation sets up a second important constraint for the reaction: any hot-ground-state reaction should take place in a very short time span. Naturally, the reaction does not need to occur immediately after the first excitation. Indeed, Koch and Rodehorst[4c] showed that the formation of AICN (**2**) from the photostationary state has a quantum yield of only 0.0034, which means that DAMN (**1** or **4**) is excited about 300 times (on average) before cyclization takes place. These two pieces of information together indicate that any statistical reaction in a hot ground state should occur in less than $300 \times 0.2\ \text{ps} = 60\ \text{ps}$, which corresponds to a maximum free-energy barrier of roughly $30\ \text{kcal mol}^{-1}$ (see the Computational Section).

Scheme 4 summarizes our findings. Starting from *cis*-DAMN, photoisomerization to *trans*-DAMN (**4**) occurs without an energy barrier through internal conversion at a twisted conical intersection. From the *trans* isomer, ground-state reactions leading to all relevant intermediates involve barriers of at least $52\ \text{kcal mol}^{-1}$, which is significantly above our kinetic threshold of $30\ \text{kcal mol}^{-1}$. This finding implies that photoexcitation of the *trans* isomer is required for the reaction to proceed. In the excited state of **4**, CN rearrangement (to **11** or **12**), hydrogen transfer from an amino group to the carbon atom of a cyano group (leading to **10**), and HCN dissociation (to **14**) are again not feasible owing to the high energy barriers (see Scheme 4 and the Supporting Information). Intersystem crossing to the triplet ground state, **13**, can be disregarded on the basis of the triplet-sensitizing experiments reported in Ref. [6], which indicated that the photocyclization takes place in the singlet manifold. The only remaining possibility is an excited-state hydrogen-atom transfer in **4** to form AIAC (**5**) with a computed energy barrier of $19\ \text{kcal mol}^{-1}$.

From AIAC (**5**), an azetene intermediate (**7** or **15**) can be readily formed either in the ground or in the excited state. In the ground state, however, there are large barriers to the subsequent rearrangement of the azetene to **17** or **18**, and moreover, azetenes do not absorb in the wavelength region of interest. Therefore, the only option is an excited-state reaction of AIAC (**5**) via an azetene. From the excited-state minimum of AIAC, **18** is not accessible, whereas the N-heterocyclic carbene (NHC) **17** can be formed after a relatively low barrier of $23\ \text{kcal mol}^{-1}$ has been overcome. Finally, the NHC can tautomerize to AICN (product **2** or **19**).



Scheme 4. Possible mechanisms for the reaction *cis*-DAMN (**1**)→AICN (**2** or **19**). The values near the arrows are the computed free energies of activation (ΔG) in kcal mol^{-1} for the ground-state reaction. When an excited-state reaction is relevant, the corresponding energy barrier is given in parenthesis. Species that can be photoexcited near 300 nm are indicated by an asterisk. The final pathway is indicated by bold arrows. See the Computational Section for a description of the computational methods.

The full proposed reaction, which requires the excitation of *cis*-DAMN (**1**), *trans*-DAMN (**4**), and AIAC (**5**), is shown in more detail in Figure 1. The need for these three excitation steps is not a statistical impediment, as we know that the molecule is excited hundreds of times during the process.[4c] After the excitation of *trans*-DAMN, either the *cis* isomer can be repopulated or the molecule can relax to the S_1 minimum. In fact, the existence of this minimum explains the predominance of the *trans* isomer **4** in the photostationary state[4c] (see the Supporting Information). From the S_1 minimum of *trans*-DAMN, AIAC (**5**) can be formed through internal conversion to the ground state at the CI-2 conical intersection. After the photoexcitation of AIAC, it may relax to its S_1 minimum, from which azetene **7** can be formed by excited-state ring closure. Since the S_1 state of the azetene has near-zero oscillator strength, it has time to reach the CI-3 conical intersection. This intersection is characterized by a C–C ring opening, which helps to guide the rearrangement towards the formation of the five-membered ring of NHC **17**.

At the CI-3 conical intersection and even afterwards in the hot ground state (TS-2), some branching is expected. Part of the population will flow back towards the azetene and may return to AIAC (**5**), which may be excited again. Another part will undergo an internal conversion with C–C bond cleavage

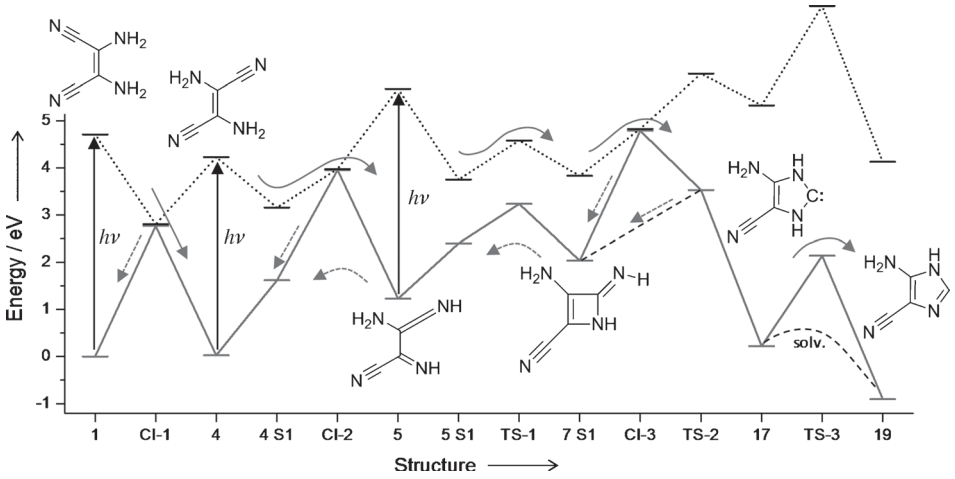


Figure 1. Reaction mechanism for the photoreaction *cis*-DAMN (**1**)→AICN (**19**), including ground (solid lines) and excited states (dotted lines). Dashed arrows indicate back reactions.

in the azetene ring and then directly rearrange to NHC **17**. Such filtering of a reaction at a conical intersection has been observed before for pyrrole, also with the involvement of a ring-opening conical intersection.[13]

From NHC **17**, the final product AICN (**2** or **19**) is obtained by tautomerization. The NHC belongs to the well-studied imidazol-2-ylidene family. Owing to the absence of substituents at the ring nitrogen atoms, it is not stable.[14] In the gas phase, its tautomerization to AICN involves rather high energy barriers, but it should proceed much more readily in polar solvents. Our computations show that the rearrangement to AICN is indeed very facile in solution (see the Supporting Information).

In conclusion, by the use of computational methods, we have identified a multistep mechanism for the DAMN→AICN reaction that is thermodynamically and kinetically compatible with the available experimental data. This mechanism rationalizes the observed ketenimine absorption (at 2020 cm^{-1}) and its disappearance upon heating[12] as well as the preference for the *trans* isomer in the photostationary state. It is consistent with a cold environment, which is required to support a high HCN concentration, and it is also consistent with the lack of luminescence during the reaction.[6] Finally, from a more general perspective, the ultrafast energy dissipation revealed by our simulations provides insight into the time scales that are relevant in photochemical prebiotic reactions.

Computational Section

Energy dissipation: After internal conversion has taken place, the excess electronic energy is transferred to vibrational modes, thus generating a hot ground state. This local hot spot can in principle be the source of energy necessary to overcome a reaction barrier that can normally not be surmounted under standard-temperature conditions. The feasibility of such a process depends on the time during which the molecule is hot enough to undergo the chemical reaction. We employed quantum-mechanical/molecular-mechanical (QM/MM) dynamics simulations to estimate the energy-dissipation time and thus to check whether reactions in the hot ground state may or may not take place in the DAMN→AICN conversion.

We simulated the energy dissipation of a hot *trans*-DAMN ground state into water. Details of the setup for this simulation are specified in the Supporting Information. A QM region composed of DAMN and 9 water molecules was surrounded by a sphere of MM water molecules. The OM2 semiempirical method[15] was used for the QM calculations. MM water was described by the TIP3P model.[16] After equilibration first at MM and then at QM/MM levels, NVE-ensemble simulations were carried out with and without consideration of the hot spot.

To create the hot spot, we kept the direction of the velocities from the initial molecular-dynamics run and only modified their norms to correspond to an excess of 4 eV of photoenergy plus the ground-state zero-point energy, 2.25 eV. Four such trajectories were run, and all of them showed the same tendency. The ratio between the kinetic energy of *trans*-DAMN and the average kinetic energy from the reference simulation (without added energy) is shown in Figure 2a. Evidently,

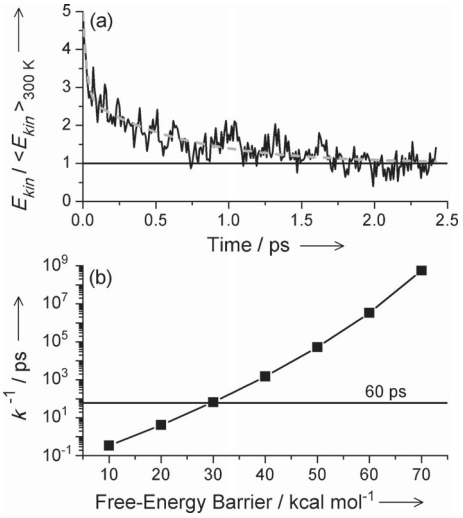


Figure 2. a) Ratio between the kinetic energy of *trans*-DAMN and the average kinetic energy at 300 K as a function of time after creation of the hot spot. The dashed line is a biexponential-decay fitting of the data. b) Inverse of the unimolecular reaction rate as a function of the free-energy barrier.

energy dissipation is extremely fast in water, so that DAMN is already thermalized after about 2 ps. The energy dissipation shows a double-exponential-decay profile, with time constants 0.02 and 0.67 ps. The first dissipation step consists of very fast transfer of about one third of the excess energy to the neighboring water molecules, on the time scale of a couple of N–H stretching oscillations. It is followed by a somewhat slower step, which dissipates the remaining excess energy.

The ultrafast energy-dissipation profile imposes a very short time window for the occurrence of hot-ground-state reactions. For considerations of the reaction rate (see below), we take this window to be

0.2 ps, which corresponds to the time during which the excess energy is reduced to one third of its initial value.

Reaction rate: To estimate the energy barrier that can be overcome within 60 ps given 4 eV of internal energy, we computed unimolecular rates $k(E)$ by using the Rice–Ramsperger–Kassel–Marcus (RRKM) approach.[17] The density and the number of states were estimated with the Beyer–Swinehart direct-count method[18] on the basis of the computed harmonic frequencies for the reactant, *trans*-DAMN (**4**), and the transition state for CN rearrangement, **11**. By solving $k(E)$ for several free-energy-barrier values, we could estimate that the maximum barrier is about 30 kcal mol^{-1} (Figure 2b).

Computational details: Gas-phase minima and transition states in the ground and excited states were determined by density functional theory (DFT) and time-dependent (TD) DFT. The CAM-B3LYP functional[19] with the aug-cc-pVTZ basis set[20] were employed in these calculations. The use of gas-phase model calculations was motivated by the fact that the reaction takes place equally well in a large variety of solvents.[6] Relevant features of the reaction pathways were verified by optimization with the second-order approximate coupled-cluster method (CC2) and the complete-active-space self-consistent-field method (CASSCF) followed by single-point-energy evaluations by second-order perturbation theory (CASPT2). Details of these calculations are described in the Supporting Information. A collection of spectroscopic data obtained at different levels and Cartesian coordinates of all relevant structures are also provided in the Supporting Information. The (TD)DFT and CASPT2 calculations were carried out with the software Gaussian 09[21] and Molcas,[22] respectively.

Received: April 17, 2013
Published online: June 19, 2013

Keywords: computational chemistry · hydrogen cyanide chemistry · photochemistry · prebiotic synthesis · simulations

[1] a) H. L. Barks, R. Buckley, G. A. Grieves, E. Di Mauro, N. V. Hud, T. M. Orlando, *ChemBioChem* **2010**, *11*, 1240 – 1243; b) A. Al-Azmi, A.-Z. A. Elassar, B. L. Booth, *Tetrahedron* **2003**, *59*, 2749 – 2763.

[2] M. W. Powner, J. D. Sutherland, J. W. Szostak, *J. Am. Chem. Soc.* **2010**, *132*, 16677 – 16688.

[3] J. P. Ferris, L. E. Orgel, *J. Am. Chem. Soc.* **1966**, *88*, 1074.

[4] a) B. Bigot, D. Roux, *J. Org. Chem.* **1981**, *46*, 2872 – 2879; b) H. Mizutani, H. Mikuni, M. Takahasi, H. Noda, *Origins Life Evol. Biosphere* **1975**, *6*, 513 – 525; c) T. H. Koch, R. M. Rodehorst, *J. Am. Chem. Soc.* **1974**, *96*, 6707 – 6710; d) R. S. Becker, J. Kolc, W. Rotham, *J. Am. Chem. Soc.* **1973**, *95*, 1269 – 1273; e) Y. Yamada, N. Nagashima, Y. Iwashita, A. Nakamura, I. Kumashiro, *Tetrahedron Lett.* **1968**, *9*, 4529 – 4532.

[5] C. Matthews in *Bioastronomy: The Search for Extraterrestrial Life – The Exploration Broadens*, *Vol. 390* (Eds.: J. Heidmann, M. Klein), Springer, Berlin, **1991**, pp. 85 – 87.

[6] J. P. Ferris, J. E. Kuder, *J. Am. Chem. Soc.* **1970**, *92*, 2527 – 2533.

[7] R. Crespo-Otero, M. Barbatti, H. Yu, N. L. Evans, S. Ullrich, *ChemPhysChem* **2011**, *12*, 3365 – 3375.

[8] R. A. Sanchez, J. P. Ferris, L. E. Orgel, *J. Mol. Biol.* **1968**, *38*, 121 – 128.

[9] a) S. Miyakawa, H. J. Cleaves, S. Miller, *Origins Life Evol. Biosphere* **2002**, *32*, 195 – 208; b) R. A. Sanchez, J. P. Ferris, L. E. Orgel, *J. Mol. Biol.* **1967**, *30*, 223 – 253.

[10] a) J. P. Ferris, L. E. Orgel, *J. Am. Chem. Soc.* **1965**, *87*, 4976 – 4977; b) B. R. Penfold, W. Lipscomb, *Acta Crystallogr.* **1961**, *14*, 589 – 597.

[11] J. P. Ferris, R. S. Narang, T. A. Newton, V. R. Rao, *J. Org. Chem.* **1979**, *44*, 1273 – 1278.

[12] J. P. Ferris, R. W. Trimmer, *J. Org. Chem.* **1976**, *41*, 19 – 24.

[13] B. Sellner, M. Barbatti, H. Lischka, *J. Chem. Phys.* **2009**, *131*, 024312.

[14] R. S. Massey, C. J. Collett, A. G. Lindsay, A. D. Smith, A. C. O'Donoghue, *J. Am. Chem. Soc.* **2012**, *134*, 20421 – 20432.

[15] W. Weber, W. Thiel, *Theor. Chem. Acc.* **2000**, *103*, 495 – 506.

[16] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926 – 935.

[17] T. Baercor, P. M. Mayer, *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 103 – 115.

[18] T. Beyer, D. F. Swinehart, *Commun. ACM* **1973**, *16*, 379.

[19] T. Yanai, D. P. Tew, N. C. Handy, *Chem. Phys. Lett.* **2004**, *393*, 51 – 57.

[20] T. H. Dunning, *J. Chem. Phys.* **1989**, *90*, 1007 – 1023.

[21] Gaussian 09, Revision A.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, N. J. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian, Inc., Wallingford CT, **2009**.

[22] G. Karlström, R. Lindh, P. A. Malmqvist, B. O. Roos, U. Ryde, V. Veryazov, P. O. Widmark, M. Cossi, B. Schimmelpfennig, P. Neogrady, L. Seijo, *Comput. Mater. Sci.* **2003**, *28*, 222 – 239.

A microiterative intrinsic reaction coordinate method for large QM/MM systems.

Iakov Polyak, Eliot Boulanger, Kakali Sen, and Walter Thiel

Phys. Chem. Chem. Phys. **2013**, 15, 14188-14195.

PCCP

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2013, **15**, 14188

A microiterative intrinsic reaction coordinate method for large QM/MM systems

Iakov Polyak, Eliot Boulanger, Kakali Sen and Walter Thiel*

Received 18th April 2013,
Accepted 30th May 2013

DOI: 10.1039/c3cp51669e

www.rsc.org/pccp

Intrinsic reaction coordinate (IRC) computations are a valuable tool in theoretical studies of chemical reactions, but they can usually not be applied in their current form to handle large systems commonly described by quantum mechanics/molecular mechanics (QM/MM) methods. We report on a development that tackles this problem by using a strategy analogous to microiterative transition state optimization. In this approach, the IRC equations only govern the motion of a core region that contains at least the atoms directly involved in the reaction, while the remaining degrees of freedom are relaxed after each IRC step. This strategy can be used together with any existing IRC procedure. The present implementation covers the stabilized Euler, local quadratic approximation, and Hessian predictor–corrector algorithms for IRC calculations. As proof of principle, we perform tests at the QM level on small gas-phase systems and validate the results by comparisons with standard IRC procedures. The broad applicability of the method is demonstrated by IRC computations for two enzymatic reactions using standard QM/MM setups.

I. Introduction

Theoretical and computational studies of chemical reactions often make use of the concept of an intrinsic reaction coordinate (IRC). According to the original definition reported by Fukui,[1,2] it is the steepest-descent pathway in mass-weighted coordinates starting from a transition state (TS) and ending in a local minimum on a potential energy surface (PES). IRC calculations proceed in steps, each of which satisfies:

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s} = -\frac{\boldsymbol{g}(\boldsymbol{x})}{|\boldsymbol{g}(\boldsymbol{x})|}, \qquad (1)$$

where $\boldsymbol{x}$ denotes the mass-weighted Cartesian coordinates of the nuclei, s is the arc length along the IRC, and $\boldsymbol{g}$ is the mass-weighted gradient at $\boldsymbol{x}$.

In quantum-chemical studies of small and medium-sized molecules, IRC path following has become a routine task to establish the connection between optimized stationary points on the PES. There are a number of well-established methods to integrate the basic IRC equation. The Ishida–Morokuma–Komornicki stabilization[3] of the Euler method is the simplest approach since it only requires gradients. The local quadratic approximation (LQA) method[4,5] also utilizes information from the Hessian and is therefore more accurate than Euler methods. The Gonzalez–Schlegel method[6–8] performs a constrained optimization after each Euler step. Finally, the Hessian- and Euler-based predictor–corrector (HPC and EulerPC) methods[9–11] use either an LQA- or an Euler-type predictor step, and a modified Bulirsch–Stoer integrator on a fitted distance-weighted interpolant surface as a corrector step. These various approaches differ in the required order of energy derivatives and in the number of energy and gradient evaluations per IRC step.

In QM/MM studies of large systems with many degrees of freedom, IRC calculations are normally avoided because straightforward application of standard IRC procedures would be quite costly and mostly impractical. Instead, as a pragmatic alternative, one often performs careful energy minimizations that start from two structures generated by perturbing the TS coordinates along the transition mode in both directions and that are supposed to lead to the two nearest local minima. To our knowledge, the IRC technique is implemented at the QM/MM level only in the Gaussian program[12] through a combination of the EulerPC method with the multi-scale ONIOM approach.[13] In this implementation,[14] the IRC is computed for the whole system using first and second derivative information for both the QM and MM part, and special attention is paid to keep the treatment of the Hessian terms tractable by using Hessian updates throughout.

In this paper we present a microiterative method for QM/MM IRC calculations, in which only a subset of QM atoms (the core region) follow the steepest descent path, while all the remaining active atoms are subject to minimization after every IRC step. This is of course an approximation, which will however become increasingly accurate with the growth of the

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-45470 Mülheim an der Ruhr, Germany. E-mail: thiel@mpi-muelheim.mpg.de

core region. In the original paper on microiterative transition state optimization,[15] the possibility of using the same strategy for IRC computations was already mentioned, but without giving any further details (see ref. 16 for an application of the corresponding implementation in the GRACE program to chorismate mutase).

The paper is structured as follows. In Section II we describe the method and implementation details. In Section III we present proof-of-concept QM applications for two small gas-phase systems as well as QM/MM IRC calculations for two enzymatic reactions. We discuss the benefits, pitfalls, and the potential range of applications of the proposed method. Section IV offers a summary and an outlook.

II. Method and implementation

The microiterative IRC method follows the philosophy of the microiterative TS search[15] as implemented in the HDLCopt program.[17] In this kind of TS search, the system is partitioned into the reaction core that follows the P-RFO (partitioned rational function optimizer) algorithm[18,19] uphill towards the transition state, and into the remainder that is minimized using the L-BFGS (low-memory Broyden–Fletcher–Goldfarb–Shanno) algorithm.[20,21] This partitioning is motivated by the need to avoid the calculation and diagonalization of the Hessian for the whole system. The optimization is performed by sequential micro- and macro-iterations such that every single step for the core region is followed by a total relaxation of the environment. This TS search has been demonstrated to be accurate and highly efficient for large systems.[17]

In the microiterative IRC procedure, we perform a full relaxation of the environment after each IRC step for the core atoms (Fig. 1). Thereafter, the resulting gradients and Hessian (if needed) of the core region are used to make the next IRC step. As in the microiterative TS search, this decoupling of the inner and outer region introduces errors, which will be evaluated in the next section. The overall scheme is designed to provide an efficient method for performing approximate IRC calculations on large systems that can be utilized at the QM/MM as well as the pure QM level.

We have implemented the microiterative IRC procedure into the existing HDLCopt module[17] in the ChemShell package.[22] Starting from the transition state, the first step in the core region is taken along the imaginary frequency mode eigenvector,[4] regardless of the chosen IRC integration method. After each IRC step, the outer region is minimized using the L-BFGS optimizer employing user-specified convergence criteria which may play an important role in some cases (see Section III). If needed, the Hessian can be either recalculated numerically or modified by applying one of the two available Hessian updates (Powell[23] or Bofill[24]) at every IRC step. The use of Hessian updates has previously been shown to be accurate enough for IRC calculations,[10] it is fast and the preferred option for routine applications. The IRC steps in the core region are always performed in Cartesian coordinates, while the outer region can be optimized in internal coordinates. For the integration of the IRC equation, we have currently implemented the IMK-stabilized Euler, LQA, and HPC methods.

The IMK-stabilized Euler method[3] starts from a simple Euler IRC step with input step size Δs:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \Delta s \frac{\boldsymbol{g}(\boldsymbol{x}_k)}{|\boldsymbol{g}(\boldsymbol{x}_k)|}. \quad (2)$$

Then a linear search for the energy minimum is performed along the bisector of the gradients to correct the Euler step. This requires additional energy and gradient evaluations (from three to seven energy and two gradient calculations per step). This approach is therefore the least efficient one among those considered here. Nevertheless, it is the simplest way of integrating eqn (1), and with small steps it is expected to work for any system.

The LQA method employs second-order energy derivative information and is thus more accurate. It can be used with larger steps than the Euler methods. An LQA step has the following form:[4]

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + A(t)\boldsymbol{g}(\boldsymbol{x}_k), \quad (3)$$

with

$$A(t) = U_k \alpha(t) U_k^{\dagger} \quad (4)$$

where U_k is the matrix of column eigenvectors of the Hessian (H_k), and $\alpha(t)$ is a diagonal matrix with the following diagonal elements:

$$\alpha_{ii}(t) = (e^{-\lambda_{ii}t} - 1)/\lambda_{ii}. \quad (5)$$

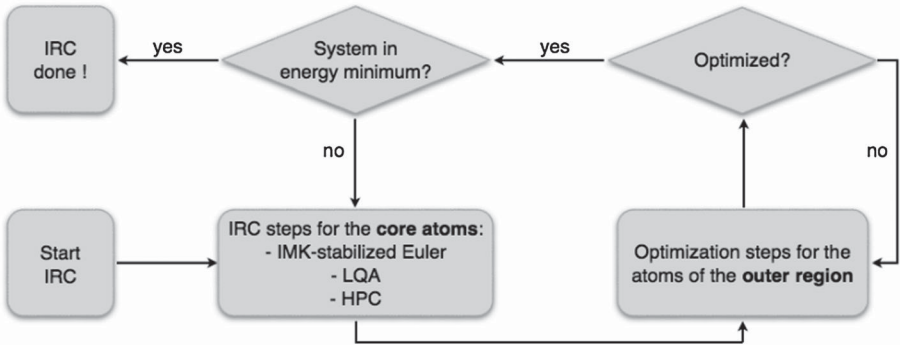


Fig. 1 Scheme of the microiterative IRC procedure as implemented in the HDLCopt program.

The parameter t can be obtained by numerical integration of the following expression:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \sqrt{\sum_i \boldsymbol{g'}_i(\boldsymbol{x}_k)^2 \mathrm{e}^{-2\lambda_{ii}t}}, \quad (6)$$

where $\boldsymbol{g}'(\boldsymbol{x}_k) = U_k^{\dagger}\boldsymbol{g}(\boldsymbol{x}_k)$. The LQA method is both accurate (due to the use of curvature information) and efficient (requiring only one energy and one gradient calculation per step), and is therefore a good choice for the microiterative IRC procedure. The use of Hessian updates improves the computational efficiency.

In the HPC method,[9] an LQA predictor step is performed first. Then the energy and gradients are evaluated at the new coordinates, and the Hessian is updated. By interpolating energy and gradients from the previous and current points along the IRC, the Euler method is used to integrate the IRC equation starting from the previous point N times, with the step size equal to $\frac{\Delta s}{N}$. This integration is performed several times with N growing up to an arbitrarily chosen number. A polynomial extrapolation to a step size of 0 (which corresponds to infinite N) then yields the final, corrected coordinates for this IRC step. This HPC scheme is generally beneficial, since it corrects the LQA step using the available energy and gradient information (from one evaluation per IRC step). It may be expected to be especially efficient for the microiterative IRC approach, since the correction is performed *after* the outer region is optimized, which should decrease the adverse effects of decoupling the inner and outer regions.

III. Examples

Several test systems of varying size and complexity were used to assess the merits and limitations of our approach. These tests include QM studies on the Diels–Alder cycloaddition between 2,4-hexadiene and ethene and on the internal rotation in 1,2-diphenylethane, as well as QM/MM calculations on the enzymatic reactions catalyzed by chorismate mutase and *p*-hydroxybenzoate hydroxylase.

A. Diels–Alder reaction

The Diels–Alder cycloaddition between 2,4-hexadiene and ethene (Fig. 2) was used to validate our implementation against an external standard, to compare the microiterative and full-system IRC treatments, to test the different IRC integration schemes, and to check the influence of the chosen IRC step size. In all calculations, the starting point was a published TS structure[25] that was reoptimized at the B3LYP/SVP[26–32] level using ChemShell in combination with the Gaussian 09 program.

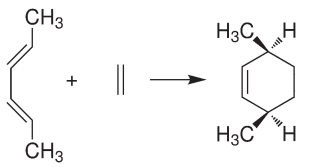


Fig. 2 Diels–Alder cycloaddition reaction of 2,4-hexadiene and ethene.

For the purpose of validation, we compared the full-system LQA IRC paths calculated with HDLCopt and with Gaussian 09 using a step size of $0.15\sqrt{\text{amu}}$ bohr. Except for the region close to the dissociation limit, the two energy profiles overlapped almost perfectly, with a root-mean-square (RMS) deviation of 0.03 kcal mol^{-1}, and the RMS deviations between the geometries along the two pathways were generally in the range of 10^{-3} to 10^{-4} Å occasionally rising up to 0.01 Å. During the last few steps towards the dissociation limit, the 2,4-hexadiene moiety undergoes a slight distortion only in the case of the Gaussian 09 calculation, which gives rise to energy differences up to 0.18 kcal mol^{-1} and RMS deviations up to 0.14 Å in the geometries. When using a smaller step size in the HDLCopt calculation, we find the same slight distortion as in the case of Gaussian 09, indicating that any minor numerical differences can be resolved by tightening the computational options to ensure convergence.

Having validated our present IRC implementation in the HDLCopt module, we performed full-system IRC calculations using the stabilized Euler, LQA, and HPC methods with IRC step sizes of $0.15, 0.10$ and $0.05\sqrt{\text{amu}}$ bohr. The LQA and HPC methods behaved very similarly and gave essentially the same IRC curves regardless of the step size, thus confirming that the largest chosen step size of $0.15\sqrt{\text{amu}}$ bohr is accurate enough for these methods in the case of the Diels–Alder reaction (except close to the dissociation limit, see above). The IMK-stabilised Euler method, in contrast, failed to provide a smooth descending curve for the two larger step sizes. It gave a smooth curve for the step size of $0.05\sqrt{\text{amu}}$ bohr, but the energies were still well above the corresponding LQA or HPC values, which were closely reproduced only after decreasing the step size further to $0.01\sqrt{\text{amu}}$ bohr. These results confirm that the LQA and HPC methods outperform the stabilized Euler method.

Next we carried out microiterative IRC calculations. The Diels–Alder cycloaddition involves a concerted formation of two C–C σ bonds, and hence we adopted an inner core region comprised of the four atoms directly involved in C–C bond formation, which is the smallest chemically meaningful choice. The remaining 18 atoms constituted the outer region and were allowed to move freely during the optimizations. The three IRC integration methods gave essentially the same microiterative IRC energy profiles, and the step size of $0.15\sqrt{\text{amu}}$ bohr was accurate enough for all of them, including the IMK-stabilized Euler method, suggesting that the microiterative scheme tolerates larger steps than the conventional IRC scheme.

Given this situation, we only present comparisons between the microiterative and the full-system IRC results for the LQA approach (see Fig. 3). The red curve in Fig. 3a is a reference IRC energy profile from the calculations on the full system. The green curve is the microiterative IRC energy profile, with the arc length on the abscissa computed from the coordinates of the four core atoms, which must lead to a narrower profile than in the reference curve where the arc length includes the variations in the positions of all 22 atoms. The blue curve is obtained from the microiterative IRC path by plotting the

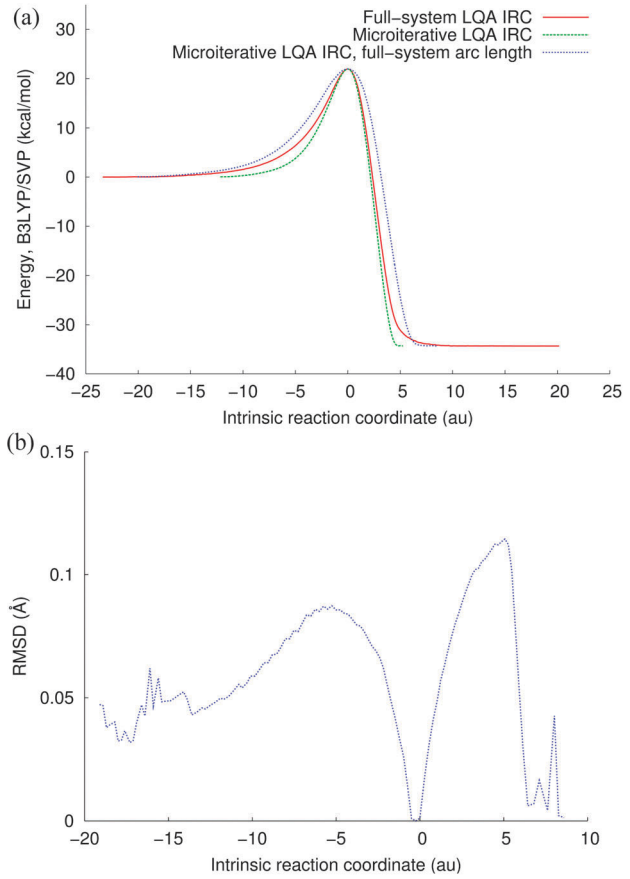


Fig. 3 IRC results for the Diels–Alder reaction obtained using the LQA method and a step size of $0.15\sqrt{\text{amu}}$ bohr. (a) Comparison between full-system and microiterative IRC energy profiles, see the text. (b) RMS deviations between the geometries along the full-system and microiterative IRC pathways, see the text.

energies as a function of the corresponding full-system arc length (22 atoms); it is somewhat broader than the reference curve. For a more direct comparison, we calculated the RMS deviations between the geometries along the microiterative and full-system IRC pathways (see Fig. 3b). Due to differences in the number of steps required to complete the IRC calculation, we compare points on the microiterative and full-system IRC pathways that are closest in energy. The RMS deviation is zero by definition at the transition state (arc length of zero) and then increases up to values of about 0.1 Å since the two methyl groups of 2,4-hexadiene rotate faster in the microiterative approach, whereas the changes in the positions of the other atoms are very similar on both pathways. The RMS deviations decrease again at larger arc lengths as both IRC pathways approach the same reactant and product states.

Overall, the microiterative and full-system IRC pathways for the Diels–Alder reaction between 2,4-cyclohexadiene and ethene are in reasonable agreement, especially when considering the choice of an extremely small core region (4 out of 22 atoms) in the microiterative calculations. In view of the good performance of the LQA integration method in the case of the Diels–Alder reaction, we adopted it in all further IRC calculations.

B. Diphenylethane

We have studied the internal rotation in 1,2-diphenylethane at the B3LYP/SVP level in an attempt to explore the limitations of our microiterative IRC approach: considering the rigidity of the phenyl rings and their steric interaction during internal rotation around the central C–C bond, it should be difficult to define a suitable small core region that is sufficiently decoupled from the remainder of the molecule.

We considered two core regions for the microiterative IRC procedure (see Fig. 4): the first one (1) included only the four central carbon atoms with the adjacent hydrogen atoms, while the second one (2) incorporated four more carbon atoms (the neighbouring two from each phenyl ring). Region 1 is the minimum choice to represent a rotation around the central C–C bond, but is too small to account for the coupled rotation of the phenyl rings.

We performed standard IRC calculations for the full system and microiterative IRC calculations with core regions 1 and 2, using the LQA method and several step sizes including the default value of $0.15\sqrt{\text{amu}}$ bohr. In the case of 1 and 2, the smoothness and shape of the IRC energy profile were found to be sensitive to the convergence criteria for the outer-region optimization steps: the default HDLCopt threshold for the maximum gradient component of 1.5×10^{-4} hartree per bohr was not sufficient and had to be tightened by factors of 3 or even 9 (depending on the IRC step size) to ensure convergence for the overall IRC energy profile.

Already with the default LQA step size, the microiterative IRC procedure for 1 and 2 resulted in reasonable paths that lead to the same product as the standard IRC treatment for the full system 3. Calculations with smaller LQA step sizes showed that the IRC results for the full system 3 are essentially converged for the default step size (see above), while there are still some changes for 1 and 2, with convergence being reached at a step size of $0.05\sqrt{\text{amu}}$ bohr. The use of smaller LQA step sizes in the microiterative IRC procedure for 1 and 2 allows a better and more gradual adaptation of the position of the two phenyl rings during the internal rotation. The resulting IRC

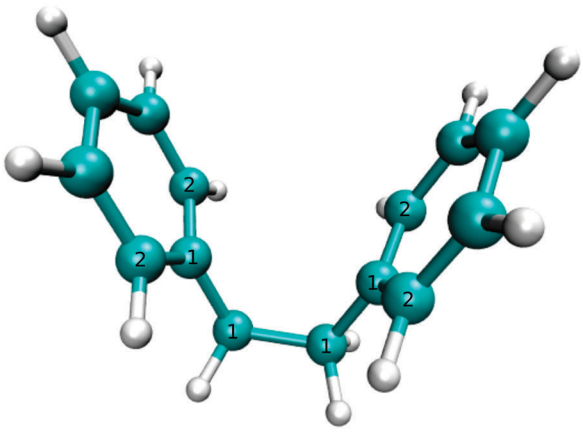


Fig. 4 Transition state for internal rotation in 1,2-diphenylethane, with assignment of carbon atoms to core regions (see text).

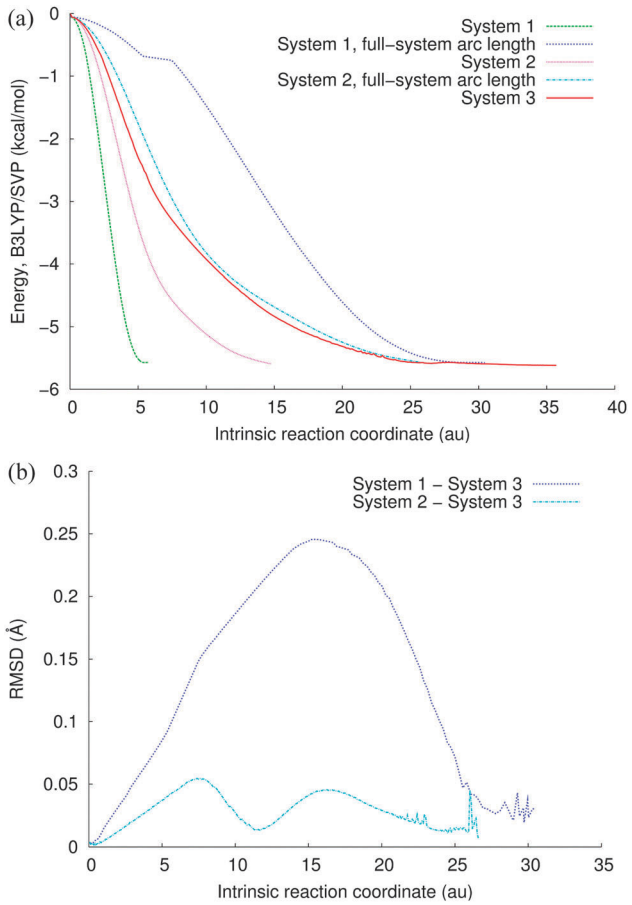


Fig. 5 IRC results for the internal rotation in 1,2-diphenylethane from the microiterative procedure for systems 1 and 2 (see text) and from the standard treatment for the full system 3. LQA step sizes: $0.15\sqrt{\text{amu}}$ bohr for 3 and $0.05\sqrt{\text{amu}}$ bohr for 1 and 2. (a) IRC energy profiles for 1–3 with different arc length definitions. (b) RMS deviations between the geometries along different IRC pathways.

energy profiles for 1–3 are presented in Fig. 5a, in the case of 1 and 2 again for two different definitions of the arc lengths (core region and full system). The RMS deviations between the geometries along the microiterative and full-system IRC pathways are shown in Fig. 5b. It is obvious from these plots that core region 1 is too small to give realistic results in the microiterative IRC treatment: the IRC energy profiles are quite different from the reference curve obtained from the standard full-system treatment, and the RMS deviations reach values of about 0.25 Å at the intermediate stage; visual inspection shows that the rotation of the phenyl rings happens too early in the case of 1. By contrast, for the larger core region 2, the IRC energy profile traces the reference curve closely (when using the full-system arc length definition), the RMS deviations remain small (generally below 0.05 Å), and visual inspection confirms that the microiterative IRC path for 2 closely follows the standard IRC path for 3, with only slight deviations. The microiterative IRC procedure can thus be successfully applied even to complicated coupled systems like 1,2-diphenylethane provided that the core region is chosen appropriately.

C. Chorismate mutase

To assess the performance of the microiterative IRC method in QM/MM calculations of enzymatic reactions, we studied the conversion of chorismate to prephenate (see Fig. 6) catalyzed by chorismate mutase (BsCM) from *Bacillus subtilis*. This reaction is a key step on the shikimate pathway for the synthesis of aromatic amino acids in plants, fungi and bacteria. It has been intensely investigated theoretically.[33]

Using the QM/MM approach we treated the substrate (24 atoms) using the semiempirical OM3 method[34,35] and the rest of the system (including the protein and the solvent shell, 13 421 atoms in total) using the CHARMM22 force field.[36] The initial preparation of the system has been described elsewhere.[37] A snapshot from the previous classical molecular dynamics (MD) simulations[37] was selected and subjected to another MD run in the NVT ensemble using the CHARMM33b1 program.[38,39] One snapshot from this MD run was randomly chosen, and the corresponding transition state for the chorismate–prephenate conversion was optimized. During geometry optimizations and reaction path calculations on BsCM, only the atoms within 16 Å from the ligand were allowed to move (active region), while the remainder of the system was kept frozen, thus enforcing a fixed outer solvent layer and preventing solvent water molecules from escaping into the vacuum.

In BsCM, the ligand is not covalently bound to the protein matrix, and the rearrangement occurs solely within the ligand substrate. This enzymatic reaction is thus ideally suited for applying the microiterative IRC procedure at the QM/MM level: the substrate (24 atoms) serves as a QM region and at the same time as a reference core region during IRC computation. We again compare the corresponding reference IRC results with those for a much smaller core region (4 atoms) composed of the oxygen atom and the three carbon atoms that are directly involved in the bond breaking and bond making processes.

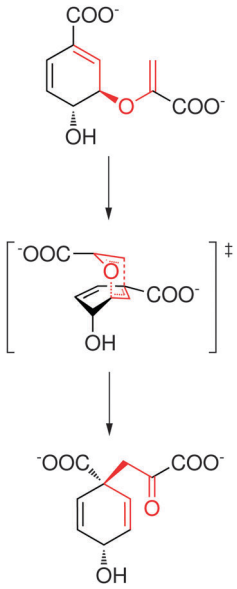


Fig. 6 Claisen rearrangement of chorismate to prephenate in BsCM.

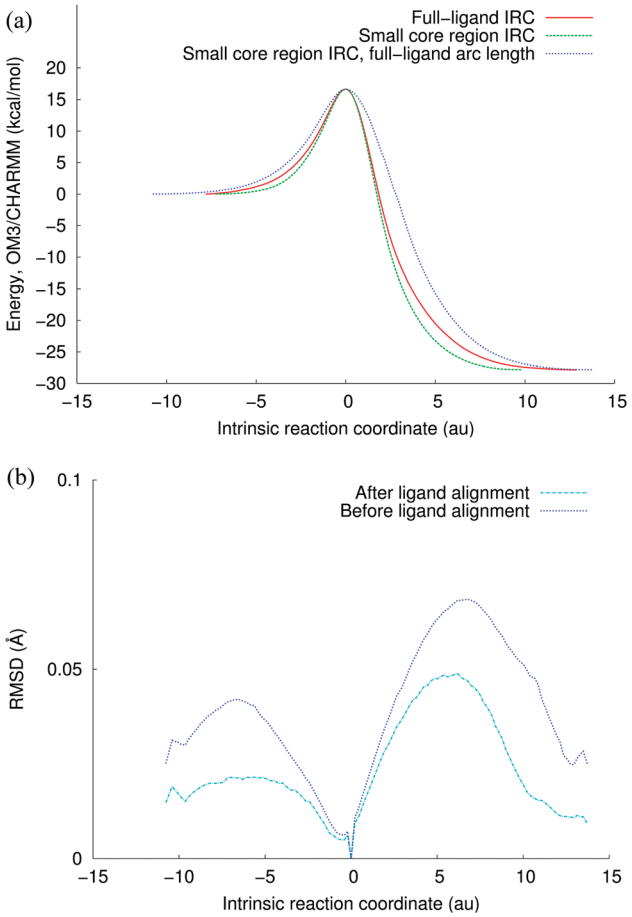


Fig. 7 Microiterative IRC results for the chorismate–prephenate conversion catalyzed by BsCM. LQA step size: $0.15\sqrt{\text{amu}}$ bohr. (a) Energy profiles from full-ligand and small-core IRC calculations (plotted in the latter case with different arc length definitions). (b) RMS deviations between the geometries along the full-ligand and small-core IRC pathways, with and without ligand alignment.

As can be seen from Fig. 7a, the two IRC energy profiles do not deviate much from one another, in spite of testing a very small core region with only four atoms (which clearly undergo the largest displacements during the reaction). Likewise, the RMS deviations between the geometries along the two microiterative IRC pathways are quite small and remain well below 0.1 Å (see Fig. 7b), and visual inspection confirms that the two pathways match very well. When comparing these geometries, it seems appropriate not to align the ligand structures, which are in both cases embedded into a protein matrix with a fixed outer part that provides a structural scaffold. For the sake of completeness, we have also plotted the RMS deviations after ligand alignment, which causes a minor overall rotation/translation of the substrate and leads to somewhat lower curves of similar shape (see Fig. 7b).

For further validation, we performed microiterative IRC calculations for a larger core region containing the substrate (treated at the QM level) and the five residues (treated at the MM level) that form hydrogen bonds with the substrate during the reaction: three arginines, one glutamate, and one tyrosine (see Fig. 8). The resulting IRC energy profile was essentially

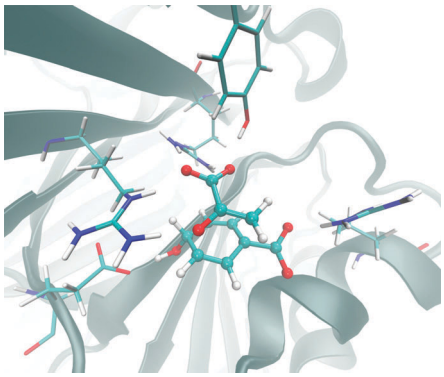


Fig. 8 Active center of BsCM. Shown is the substrate in its transition state and the five hydrogen-bonded active-site residues.

indistinguishable from the reference curve obtained from the full-ligand IRC treatment. This confirms our expectation that surrounding active-site residues need not be included in the core region of the microiterative IRC treatment in the case of BsCM.

D. *p*-Hydroxybenzoate hydroxylase

As a second QM/MM test system, we have chosen another well-studied enzymatic reaction, namely the hydroxylation step in the catalytic cycle of *p*-hydroxybenzoate hydroxylase (PHBH). The theoretical work on PHBH has been reviewed recently.[40] In the course of reaction, the OH group is being transferred from the flavin–adenin hydroperoxide cofactor (FADHOOH) to the *p*-hydroxybenzoate substrate (see Fig. 9).

The initial preparation of the system is described elsewhere.[41] The substrate and the isoalloxazine ring of FADHOOH with the attached hydroperoxide group (48 atoms) were included in the QM region and treated using the semi-empirical AM1 method,[42] while the rest of the system was described by the CHARMM22 force field. A randomly chosen snapshot from those used in ref. 41 was subjected to a restrained potential energy scan along the reaction coordinate defined in ref. 41. The structure with the highest energy on this scan served as a starting point for TS optimization. The subsequent microiterative IRC calculations employed three core regions of different size. The small core region contained only four atoms: the hydroperoxide group (OOH) and the substrate carbon atom, to which the OH group is transferred. The medium core region also included the remaining atoms of the substrate. The large core region comprised nearly all the QM atoms: the isoalloxazine ring, the hydroperoxide group, and the substrate (omitting only the methyl group representing the ribityl side chain in the QM calculations). Technically, the default values for the LQA step size and the outer-region convergence criteria turned out to be accurate enough for each of the three core regions; using smaller values did not lead to any significant changes.

The computed IRC QM/MM energy profiles are depicted in Fig. 10a. They practically coincide when plotted against the arc lengths of the three individual IRC calculations that include only the corresponding core region. This may be taken as an

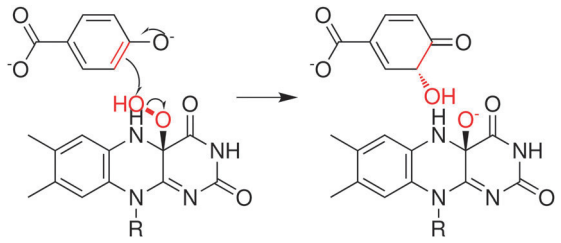


Fig. 9 Hydroxylation reaction catalyzed by *p*-hydroxybenzoate hydroxylase. R denotes the ribityl side chain of the FADHOOH cofactor.

indication that the largest displacements occur just for the few atoms directly involved in the reaction. On the other hand, the curves for the two smaller core regions become broader when plotted against the arc length evaluated for the full QM region. A better assessment is provided by direct comparisons between the geometries of the full QM region along the IRC pathways for the three chosen core regions. The corresponding RMS deviations (Fig. 10b) are quite small and remain below 0.06 Å. Visual inspection confirms that the motions within the QM region along the IRC path are very similar for all three core regions (with 4, 17, and 45 atoms).

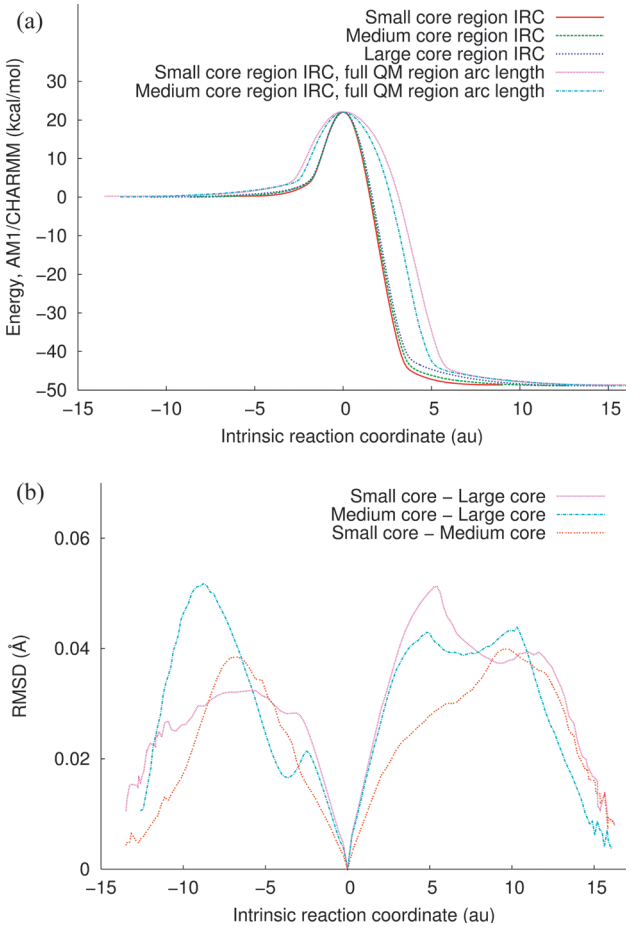


Fig. 10 Microiterative IRC results for the hydroxylation reaction catalyzed by PHBH. LQA step size: $0.15\sqrt{\text{amu}}$ bohr. (a) Energy profiles from the IRC calculations with three different core regions (see text); those for the two smaller core regions are plotted with different arc length definitions. (b) RMS deviations between the geometries along the IRC pathways obtained for different core regions.

IV. Conclusion

We have presented a microiterative procedure to perform IRC calculations on large molecular systems. The method is based on separating the system into a core region and an outer region. The core region moves along the IRC path, while the outer region is minimized after every IRC step following the IRC path adiabatically. This procedure allows large-scale IRC calculations at the QM/MM level. A prototypical example is the determination of IRC paths in enzymatic reactions, with the core region corresponding to the QM region.

Other applications are also possible, of course. The microiterative IRC procedure can be employed at the pure QM level by defining a core region in a medium-sized molecule that encompasses only the atoms directly involved in the reaction. Likewise, in QM/MM studies on large systems, the core region can be chosen to include only the reactive part of the QM region. These options have been examined for two gas-phase test systems and for two enzymatic reactions at the QM and QM/MM level, respectively. These tests confirm that rather small core regions can be used successfully provided that they account for the characteristic bond making and bond breaking processes during the reaction. If this is the case, small-core IRC paths tend to be quite similar to large-core or full-system IRC paths in terms of energies and geometries, and they can thus safely be used to check the connectivity between an optimized transition state and the associated reactant and product states.

Among the three IRC integration methods currently implemented, the LQA approach is recommended as the standard choice, with a default step size of $0.15\sqrt{\text{amu}}$ bohr. In semiempirical QM/MM work, it is generally affordable and recommended to choose the QM region as a core region for microiterative IRC calculations. When using first-principles QM methods, it will often be more practical to use smaller core regions, which is supported by the results of the current test calculations. Apart from characterizing TS connectivity, the resulting IRC paths may also serve as collective coordinates in free energy calculations that are becoming increasingly important in large-scale QM/MM studies. The present implementation of a microiterative IRC treatment should thus be widely applicable.

References

1 K. Fukui, *J. Phys. Chem.*, 1970, **74**, 4161.
2 K. Fukui, *Acc. Chem. Res.*, 1981, **14**, 363.
3 K. Ishida, K. Morokuma and A. Komornicki, *J. Chem. Phys.*, 1977, **66**, 2153.
4 M. Page and J. W. McIver, *J. Chem. Phys.*, 1988, **88**, 922.
5 M. Page, C. Doubleday and J. W. McIver, *J. Chem. Phys.*, 1990, **93**, 5634.
6 C. Gonzalez and H. B. Schlegel, *J. Chem. Phys.*, 1989, **90**, 2154.

7 C. Gonzalez and H. B. Schlegel, *J. Phys. Chem.*, 1990, **94**, 5523.
8 C. Gonzalez and H. B. Schlegel, *J. Chem. Phys.*, 1991, **95**, 5853.
9 H. P. Hratchian and H. B. Schlegel, *J. Chem. Phys.*, 2004, **120**, 9918.
10 H. P. Hratchian and H. B. Schlegel, *J. Chem. Theory Comput.*, 2005, **1**, 61.
11 H. P. Hratchian, M. J. Frisch and H. B. Schlegel, *J. Chem. Phys.*, 2010, **133**, 224101.
12 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09 Revision A.1*, Gaussian, Inc., Wallingford, CT2009.
13 S. Dapprich, I. Komaromi, K. Byun, K. Morokuma and M. Frisch, *THEOCHEM*, 1999, **461**, 1.
14 H. P. Hratchian and M. J. Frisch, *J. Chem. Phys.*, 2011, **134**, 204103.
15 A. J. Turner, V. Moliner and I. H. Williams, *Phys. Chem. Chem. Phys.*, 1999, **1**, 1323.
16 S. Marti, V. Moliner, I. Tunon and I. H. Williams, *Org. Biomol. Chem.*, 2003, **1**, 483.
17 S. R. Billeter, A. J. Turner and W. Thiel, *Phys. Chem. Chem. Phys.*, 2000, **2**, 2177.
18 A. Banerjee, N. Adams, J. Simons and R. Shepard, *J. Phys. Chem.*, 1985, **89**, 52.
19 J. Baker, *J. Comput. Chem.*, 1986, **7**, 385.
20 J. Nocedal, *Math. Comput.*, 1980, **35**, 773.
21 D. C. Liu and J. Nocedal, *Math. Prog.*, 1989, **45**, 503.
22 P. Sherwood, A. H. de Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjovoll, A. Fahmi, A. Schäfer and C. Lennartz, *THEOCHEM*, 2003, **632**, 1.
23 M. J. D. Powell, *Math. Prog.*, 1971, **26**, 1.
24 J. M. Bofill, *J. Comput. Chem.*, 1994, **15**, 1.
25 S. M. Bachrach and P. B. White, *THEOCHEM*, 2007, **819**, 72.
26 J. C. Slater, *Phys. Rev.*, 1953, **91**, 528.
27 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200.
28 A. D. Becke, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1988, **38**, 3098.
29 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648.
30 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623.
31 C. T. Lee, W. T. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785.
32 A. Schäfer, H. Horn and R. Ahlrichs, *J. Chem. Phys.*, 1992, **97**, 2571.
33 F. Claeyssens, K. E. Ranaghan, N. Lawan, S. J. Macrae, F. R. Manby, J. N. Harvey and A. J. Mulholland, *Org. Biomol. Chem.*, 2011, **9**, 1578.
34 M. Scholten, PhD thesis, Universität Düsseldorf, 2003.
35 N. Otte, M. Scholten and W. Thiel, *J. Phys. Chem. A*, 2007, **111**, 5751.
36 A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, **102**, 3586.
37 H. M. Senn, J. Kästner, J. Breidung and W. Thiel, *Can. J. Chem.*, 2009, **87**, 1322.
38 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187.
39 B. R. Brooks, C. L. Brooks III, A. D. MacKerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *J. Comput. Chem.*, 2009, **30**, 1545.
40 H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.*, 2009, **48**, 1198.
41 T. Benighaus and W. Thiel, *J. Chem. Theory Comput.*, 2011, **7**, 238.
42 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902.

Towards QM/MM Simulation of Enzymatic Reactions with the Drude Oscillator Polarizable Force Field.

Eliot Boulanger and Walter Thiel

J. Chem. Theory Comput. **2014**, 10, 1795-1809.

JCTC Journal of Chemical Theory and Computation

Article

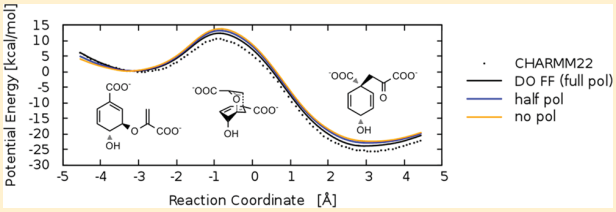
pubs.acs.org/JCTC

Toward QM/MM Simulation of Enzymatic Reactions with the Drude Oscillator Polarizable Force Field

Eliot Boulanger and Walter Thiel*

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

ABSTRACT: The polarization of the environment can influence the results from hybrid quantum mechanical/molecular mechanical (QM/MM) simulations of enzymatic reactions. In this article, we address several technical aspects in the development of polarizable QM/MM embedding using the Drude Oscillator (DO) force field. We propose a stable and converging update of the DO polarization state for geometry optimizations and a suitable treatment of the QM/MM-DO boundary when the QM and MM regions are separated by cutting through a covalent bond. We assess the performance of our approach by computing binding energies and geometries of three selected complexes relevant to biomolecular modeling, namely the water trimer, the N-methylacetamide dimer, and the cationic bis(benzene)sodium sandwich complex. Using a recently published MM-DO force field for proteins, we evaluate the effect of MM polarization on the QM/MM energy profiles of the enzymatic reactions catalyzed by chorismate mutase and by p-hydroxybenzoate hydroxylase. We find that inclusion of MM polarization affects the computed barriers by about 10%.



1. INTRODUCTION

Hybrid quantum mechanical/molecular mechanical (QM/MM) approaches have been established as a reliable tool for computing molecular properties and reaction mechanisms in the condensed phase.[1−4] A prime application example is provided by enzymatic reactions, for which it is difficult to represent the heterogeneous biological environment in an implicit manner.[5−7] QM/MM studies provide qualitative insight into such processes, as well as quantitative predictions that can be confronted with experimentation. In some cases, however, the standard QM/MM approach can fail to give the right answer, which may be due to several reasons. One of them is the neglect of polarization in the standard MM force fields, especially when the investigated reaction involves charged or very polar species.[2]

There are several ways to include polarization at the MM level.[8−10] The most prominent ones make use of induced dipoles,[11−15] fluctuating charges,[16−20] and Drude oscillators (DO).[21−27] The latter approach is adopted in this study and is also called the charge-on-spring[28−32] or shell model.[33−35] In the DO model,[21] a mobile charge, called a Drude particle (DP), is linked to a polarizable atom by a spring. A fixed charge of the same magnitude but opposite sign is added at the position of the atom, thus forming a dipole with the DP. Polarization arises from the electrostatic interactions of the DO with the rest of the system.

We included this model into the QM/MM framework some time ago using the GROMOS charge-on-spring force field.[36] The CHARMM-DO model was interfaced later in a separate development.[37] More recently, we proposed the extension of QM/MM to a fully polarizable three-layer treatment to better account for long-range electrostatics and to increase the computational efficiency.[38] Standard MM-DO polarizable force field parameters for proteins have become available only very recently,[39] and thus QM/MM-DO studies have mostly been carried out up to now for small molecules or ions solvated in water or other solvents.[36,40,41]

There are some special polarizable force fields that have been used in QM/MM studies of enzymatic reactions. For instance, in their pioneering work, Warshel and Levitt proposed a point-dipole approach to include polarization at the MM level.[42] Illingworth et al. employed an induced-charge model to include MM polarization in QM/MM computations on hydrogen-bonded model systems and found effects of approximately 10%.[43] In a later study on chorismate mutase, their induced-charge model for MM polarization gave a significant stabilization of all stationary points in the chorismate-to-prephenate conversion (15−17% of the total QM/MM interaction energies from electrostatics and polarization), but the computed barrier was not affected because of equal MM polarization contributions.[44] In the absence of well-parametrized and generally accepted standard polarizable force fields for proteins, there have not been, to our knowledge, any systematic studies on enzymatic reactions with MM polarization. In this article, we report pilot QM/CHARMM-DO applications to biologically relevant macromolecular systems, namely the enzymatic reactions catalyzed by chorismate mutase and by p-hydroxybenzoate hydroxylase.

QM/MM calculations on enzymatic reactions normally employ first-principles QM methods with QM regions encompassing typically 50 to 150 atoms. They often use geometry optimization techniques to explore the potential surface (PES),[1] even though free energy calculations can also

Received: December 19, 2013
Published: March 13, 2014

dx.doi.org/10.1021/ct401095k | J. Chem. Theory Comput. 2014, 10, 1795−1809

be performed in an approximate manner using various sampling techniques along the reaction path.[45] Here, we focus on geometry optimizations at the QM/MM-DO level. In this case, the polarization of every DO has to be fully updated at every step.[36] In principle, this can be done by solving the corresponding system of equations, but in practice this is not efficient, and an iterative approach is usually preferred. Unfortunately, for DO-type polarizable force fields, which employ the Thole model[46] and include 1−2 and 1−3 bonded interactions in the polarization computation, we find that the iterative approach oscillates and does not converge for systems with a large number of bonded polarizable atoms. This will be the first issue covered in this manuscript.

The QM/MM combination of QM and MM subsystems has been thoroughly discussed in terms of the required embedding and boundary treatments.[1] The former define the QM/MM interactions at various levels of approximation (mechanical, electronic, and polarized embedding), while special protocols for the latter are needed, especially when a covalent bond is cut between the two subsystems. In this study, we focus on improvements for boundary and embedding treatments in the QM/MM-DO case. After briefly reviewing the methodological aspects of including DOs in a QM/MM scheme, we propose a method to update DPs for large systems such as enzymes. We then develop a special boundary treatment using butanol as a standard test case. Thereafter, we assess the QM/MM-DO interactions for three small but typical test cases: the water trimer, the N-methylacetamide dimer, and the cationic bis(benzene)sodium sandwich complex. Finally, we check the sensitivity of the QM/MM results with regard to the polarization of the enzymatic environment treated at the MM-DO level.

2. THEORY

2.1. Polarizable Force Fields. Atomic dipole polarizabilities can be introduced into force fields in several ways.[9,10,47] In most cases, the electrostatic part of the potential function is extended by including induced dipole/static multipole and induced dipole/induced dipole terms, as well as self-energy terms that account for the energy needed to create the dipoles. The static multipoles are usually monopoles, but expansions up to quadrupoles have been considered.[13,14] For force fields with localized polarizable centers, the induced dipoles (μ_i) are obtained using classical electrostatics:

$$\mu_i = \alpha_i E(x_i) \tag{1}$$

where E is the electric field at the position (x_i) of the polarizable atom i and α_i is its polarizability, which is a parameter of the force field. The electric field is comprised of two parts, the static field due to the other permanent multipoles in the system (E^0) and the field due to all other induced dipoles.

$$E(x_i) = E^0(x_i) - \sum_{j \neq i} T_{ij}\mu_j \tag{2}$$

where T_{ij} is the interaction tensor element between μ_i and μ_j that takes into account the interdependence of the induced dipole moments.

For additive force fields, the interactions between bonded atoms (1−2) and between next-nearest neighbor atoms (1−3) are commonly neglected in the computation of the electrostatic part of the potential. On the contrary, for polarizable force fields, it has been shown that such short-range interactions need to be included to obtain the proper polarization state.[46,48] However, their direct inclusion would lead to overpolarization due to the close distance between the polarizable centers, and it is thus necessary to use a damping function. The most popular choice is the Thole function containing additional parameters that are adjusted during the parametrization.[46,48] In this formalism, the product $T_{ij}\mu_j$ in eq 2 is multiplied by a prefactor γ_{ij}, i.e., the damping function for the interactions between nearby atoms.

The induced dipole moments of N polarizable atoms in a given configuration of the system can be obtained by solving the following linear system of equations.

$$(\boldsymbol{\alpha}^{-1}\boldsymbol{\gamma}\mathbf{T})\boldsymbol{\mu} = \mathbf{E}^0 \tag{3}$$

where $\mathbf{T}$ is a $3N \times 3N$ tensor containing the elements T_{ij}, $\boldsymbol{\alpha}^{-1}$ is a diagonal matrix containing the inverse of the atomic polarizability tensors, and $\boldsymbol{\gamma}$ represents the interatomic Thole damping functions; $\boldsymbol{\mu}$ and $\mathbf{E}^0$ are $3 \times N$ matrices containing the Cartesian components of the induced dipoles and of the static electric field at each atom, respectively. The exact solution of eq 3 by matrix algebra is often not practical for large systems with thousands of atoms (N), and an iterative self-consistent (SC) approach is therefore normally preferred.[49]

When using Thole-type models, we find that the straightforward SC approach often oscillates and does not converge for large systems. A common alternative is to use the Successive Over-Relaxation (SOR) method, in which the induced dipole moment μ_i^B of any polarizable center i is taken at each step as

$$\mu_i^B = m_A\mu_i^A + m_B\mu_i^{B'} \tag{4}$$

where $m_A + m_B = 1$, μ_i^A is the dipole moment obtained at the previous step of the iterative cycle, and $\mu_i^{B'}$ is the predicted dipole moment for the current step using the standard SC procedure.[49] This update procedure helps to achieve convergence of the SC method. The required number of cycles strongly depends on the choice of the m coefficients. Their optimum values can be different for different force fields and different systems. Occasionally this approach still fails to converge, and then eq 3 has to be solved by matrix algebra.

Xie et al. proposed a coupled method for converging the induced dipoles of a polarizable force field.[50] In their iterative scheme, intermolecular interactions are taken into account by the standard SC procedure, while intramolecular interactions are handled by matrix inversion. This scheme is directly applicable to solvents, but there is also a variant for polymers such as proteins, in which each monomer is treated separately but with its closest neighbors included in the matrix inversion procedure. We revisit these aspects in section 4.1 when developing a method suitable for geometry optimization at the QM/MM-DO level.

2.2. Drude Oscillators in a QM/MM Framework. The Drude oscillators provide a polarizable force field model, in which the induced dipoles are represented by two point charges of the same magnitude but opposite sign close in space and linked by a spring.[21,51] One of them is maintained at the position of the polarizable atom while the other, the Drude particle (DP), is free to move in the external electric field. In geometry optimizations, the DPs are allowed to adjust, and the ideal polarization state is computed at every step. Since a point-charge approximation is used to represent the induced dipole,

only monopole interactions appear in the electrostatic part of the potential function, which takes the following form:

$$E_{\mathrm{DO}}^{\mathrm{elec}} = \sum_i \sum_{j>i} \frac{q_i q_j}{r_{ij}} + \sum_i \sum_{j'} \left(\frac{q_i q_{j'}}{r_{ij'}} - \frac{q_i q_{j'}}{r_{ij}} \right) + \sum_{i'} \sum_{j'>i'} \left(\frac{q_{i'} q_{j'}}{r_{i'j'}} - \frac{q_{i'} q_{j'}}{r_{ij}} \right) - \sum_{i'} \sum_{j'} \left(\frac{q_{i'} q_{j'}}{r_{ij'}} \right) + \frac{1}{2} \sum_{i'} k_{d,i'} d_{i'}^2 \tag{5}$$

where q_i is the permanent charge at atom i, $q_{i'}$ is the charge associated with the Drude oscillator, r is the distance between the two considered charges, d is the DO charge separation, and $k_{d,i'}$ is the force constant of the DO spring, which is a parameter of the force field. The iterative update of the various positions proceeds as follows: (1) compute the electric field at every DO position; (2) based on the resulting induced dipole moment, update the DP position; (3) check convergence with respect to criteria based on energy, electric field, and/or position, and if not converged, restart at the first step.

DO force fields employ the Thole model and include the 1−2 and 1−3 interactions for induced dipole/induced dipole interactions.[27,52] The screening function is based on an exponential charge distribution and a damping function, which for monopole interactions is given by

$$\gamma_{ij} = 1 - \left(\frac{u_{ij}}{2} + 1 \right) e^{-u_{ij}} \tag{6}$$

where $\mu_{ij} = r_{ij} t_{ij}$, r_{ij} is the inter-DP-distance, and t_{ij} the Thole parameter:

$$t_{ij} = \frac{t_i + t_j}{\sqrt[6]{\alpha_i \alpha_j}} \tag{7}$$

with t_i and t_j being force field parameters of the respective polarizable centers.

The electrostatic potential function of the CHARMM-DO force field is extended not only by a point-charge representation of the atomic dipole polarizability but also by additional point charges that represent lone pairs.[27] These latter charges are rigidly linked to the heteroatoms of the system and allow a better description of the fixed charge distribution. Their positions are determined at every step of a geometry optimization from the associated set of internal coordinates, and their gradient components are distributed among the neighbor atoms such that the total force and the total torque are conserved. Another technical advantage is that the lone pair positions can be used to define local internal coordinate systems centered at heteroatom positions, which allows the use of anisotropic polarization for DOs.[53]

In a QM/MM framework, the mutual polarizations of the QM and MM regions have to be taken into account.[36] The point-charge representation of the DO model allows for a straightforward combination with most of the QM methods, since it can be treated in the same way as the point charges of the additive force fields in the electronic embedding scheme.[49] Several approaches have been proposed to find the ideal polarization state of both the QM and MM parts at every step.[37] We use the dual-SC approach, i.e., we run a full QM calculation and update the DP position as in the standard iterative procedure, but include the QM field contribution in the static electric field at each DP position. This procedure is iterated until the QM energy and DP position both converge. This scheme is expensive but leads to accurate results. It can be easily interfaced with any QM code and can be used for geometry optimization.

Another issue to consider in QM/MM simulations is the boundary treatment in cases when the QM/MM partitioning cuts through a covalent bond. This aspect has not yet been addressed for QM/MM-DO approaches. We will discuss this point in section 4.2 and propose a treatment for this situation.

3. COMPUTATIONAL DETAILS

All computations were run using the ChemShell package interfaced to several programs.[54−56] The MNDO program[57] was used for semiempirical QM calculations, while all other QM results were obtained with TURBOMOLE6.3.[58] The additive parts of the MM potential were computed using DL-POLY.[59] All DO-related computations were implemented separately in the hybrid module of ChemShell.[54] If not stated otherwise, the resolution-of-identity (RI) approximation[60] was applied in the MP2 calculations,[61] as is often done in QM/MM simulations at this level.[62]

4. RESULTS

4.1. Converging the Drude Oscillators. During geometry optimization, the SC procedure may fail to converge the DO positions as soon as a system with bonded polarizable atoms becomes large enough to justify the use of a force field. The SOR approach[49] is a good alternative but requires more iterations than the standard SC procedure, and in some cases it may also fail to converge to the proper polarization state. As the speed of convergence is a major practical issue in QM/MM computations, we investigate other options in the following.

We first consider the noniterative approach of solving eq 3 analytically by matrix algebra. Although computationally demanding, this can be useful in difficult cases, and it should also lead to fewer cycles in QM/MM dual-SCF procedures. Equation 3 can formally be solved by matrix inversion, which becomes very expensive for large systems. Since the considered matrix is symmetric and positive definite, it is much better to solve the corresponding system of linear equations, using Cholesky factorization followed by forward and backward substitutions. For typical system sizes in QM/MM studies of enzymes, this leads to 30−40 fold improvements in efficiency when both procedures are properly implemented. In the following, any reference to an exact or analytic solution by matrix algebra implies the use of the latter Cholesky-based approach.

The analytic procedure does not give exactly the same results as the SC procedure, in which the electric field is computed at the position of the DP for the update of DO; this position changes in each iteration. Note that even if one computes the electric field at the atomic position, as in the GROMOS charge-on-spring force field,[9] the problem remains due to the point-charge approximation so that eq 3 is not solved exactly. As an alternative, we propose an iterative approach in which at each step the electric field is evaluated at the new DP position obtained from the previous step. To assess the method, we used the 30 Å water sphere from our previous work on the solvated glycine test system (903 water molecules without the glycine solute).[38,63] The water molecules were described by the SWM4-NDP model.[64] This system has the advantage of

being large without having any 1−2/1−3 interactions, and thus the iterative approach converges without problems. We also considered computing the permanent electric field (not the one due to other dipoles) at the atomic positions. In terms of polarization energy, the results deviate by less than 0.01% for both techniques. The mean absolute deviation of the gradient components is 2.64×10^{-5} au with a maximum value of 1.25×10^{-4} au If the permanent electric field is also evaluated at the atomic positions, values of 5.68×10^{-5} and 2.78×10^{-4} au are obtained, thus roughly doubling the deviations. In both cases, these values are clearly below the commonly adopted criteria for geometry optimization (3.0×10^{-4} and 4.5×10^{-4} au). This confirms that the point-charge approximation to the induced dipole is accurate enough to simulate point dipoles, which in turn suggests that we can use the iterative Cholesky factorization (ICF) approach as a reference for large systems that do not converge with SC techniques. It can also be used together with the SOR method in the case of nonconvergence. For both types of electric field computation, the ICF procedure took 6 steps for a total of 3.5 h on one 2.9 MHz Xeon processor. We also tried to evaluate the electric field at positions different from the DP, but any deviation from this position decreased the accuracy of the method.

Xie et al. proposed a hybrid approach to tackle systems with bonded polarizable centers.[50] They used the SC technique for intermolecular interactions and solved eq 3 by matrix algebra for intramolecular interactions. In the case of polymers, they suggested to determine intramolecular interactions by matrix algebra including the two neighbors of each monomer in the matrix to be inverted. For a system containing a solute surrounded by small solvent molecules, they discussed the gain in efficiency obtained, thanks to the good convergence of their method, but they did not evaluate its accuracy. Here, we take a bottom-up approach using a more general related model and compare its results with the ICF scheme.

We consider a system of polarizable DO centers which may be bonded to each other. We define blocks as any subsets of these centers. The partitioning into blocks need not be based on chemical intuition, and it is not necessary that they correspond to molecules or residues. To obtain the polarization of each block, the computation is run for a superblock containing also the neighboring blocks (based on connectivity or distance). The polarizable atoms not included in a superblock form the outer shell. The electric field (E) at any DP position (x_i) in a given block can be split in an additive fashion:

$$E(x_i) = E^{PC}(x_i) + E^{SB}(x_i) + E^{OS}(x_i) \quad (8)$$

where E^{PC} is the electric field due to permanent charges (atoms and lone pairs), $E^{SB}(x_i)$ is the electric field due to all other DOs in the superblock, and $E^{OS}(x_i)$ is generated by the DOs in the outer shell. To obtain the ideal electric field at each DP position and to get the polarization state of the system, we use an iterative hybrid approach. We first compute the electric field due to every element of the system at every DP position using monopole interactions such as in the SC approach ($E_{MP}(x_i)$). From the precomputed electric field, we remove all DO electric field contributions from within the superblock using

$$E_{MP}(x_i) - E^{SB}_{MP}(x_i) = E^{PC}_{MP}(x_i) + E^{OS}_{MP}(x_i) = E^{0}_{CF}(x_i) \quad (9)$$

and thus obtain the permanent electric field for the superblock for each DP, which now includes outer-shell DO contributions. We solve the linear system of equations for the superblock with Cholesky factorization using this $E^0_{CF}(x_i)$ as permanent electric field in eq 3. In this way, we get the induced dipole moment for each block separately. If necessary, anisotropy can be included *a posteriori* with an update of the DP positions. The calculation is iterated until we achieve convergence of the DP positions. Note that for QM/MM computations, the field due to QM region should be included into $E^0_{CF}(x_i)$ and kept constant in a given iteration.

In this approach, we need to define the partitioning into blocks and the rules for generating the superblocks. Our only constraint is that all Thole-type interactions (1−2, 1−3) involving a given block should be taken into account within its superblock. To specify the proper combination law, we have chosen to define each polarizable center as a block and to generate the superblock either through connectivity or distance from this atom. As a test system, we have chosen the chorismate mutase enzyme as described below in section 4.4. We have removed the substrate and tested our approach at the full MM level using a preliminary version of the CHARMM-DO protein force field, noting that the conclusions are directly portable to the QM/MM level. To check the dependence of the results on the combination law (connectivity vs distance), we have computed the polarization state of the enzyme in the gas phase and in water. Figure 1 shows the maximum absolute deviation of the gradient components for these two setups (blue/green, with/without water) between the two combination laws (connectivity: plain lines; distance, dashed lines). The units on the x axis are Ångstroms in the case of distance and refer to the connectivity order up to which atoms were included into the superblock (1−5 means that 1−2, 1−3, 1−4, and 1−5 are included). Considering that our standard convergence criterion in geometry optimizations is 4.5×10^{-4} au for the gradient components, it is clear that the connectivity-based approach does not converge properly with increasing superblock size. The distance-based selection seems much better in this regard, as it converges in both cases for distances of more than 6 or 7 Å from the polarizable centers. This is confirmed by comparing the results from the computations in the gas phase and in water: while the distance-based computations converge to the same values, there is a significant gap in the connectivity-based results (implying that distant water molecules play an important role for the polarization state). In terms of efficiency, the ICF scheme took 12 h for the full solvated system and 2 h for the gas-phase system. Solving the linear system of equation using matrix algebra took 385 h for the full system and gave precisely the same result. Convergence of the full system with the distance-based selection of superblocks is reached after 12, 25, and 75 min for cutoffs of 7, 8, and 9 Å, respectively. For comparison, we note that the SOR approach does not converge with a damping factor of 0.1 and takes 15 (20) min with a damping factor of 0.2 (0.4). This similarity in efficiency comes from the fact that the SOR method requires a larger number of steps, each of which is less costly.

To compare with the work of Xie et al.,[50] we considered the case in which the blocks are chosen as amino acid residues or water molecules. For the definition of superblocks, we used their method of taking the bonded blocks as well as the more accurate distance-based approach suggested here, selecting any residue having at least one polarizable center within a predetermined cutoff distance from any polarizable atom of the block. For the protein in the gas phase, the bonded approach gave a mean absolute deviation of 2.6×10^{-5} au with

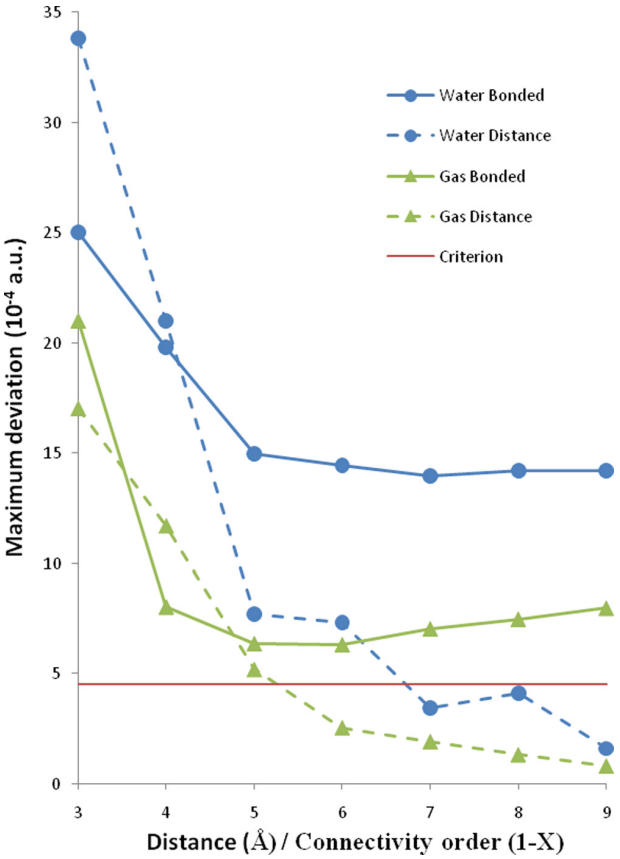


Figure 1. Maximum absolute deviation of the gradient components when comparing the hybrid CF-SC to the ICF approach (see text) for chorismate mutase, without substrate, using a preliminary version of the CHARMM-DO polarizable force field (provided to us by A. D. MacKerell in 2012). Gas phase (green triangles) and solvated enzyme (blue circles) were considered with different selection criteria for the hybrid model. Plain (dashed) lines refer to a connectivity-based (distance-based) superblock selection, with corresponding units of the x axis (maximum connectivity order $1 - X$/distance in Å; see text). We target maximum absolute deviations below the red line, which represents the standard convergence criteria for geometry optimization.

a maximum value of 6.84×10^{-4} au, while the distance-based approach with a cutoff of only 5 Å gave values of 1.07×10^{-5} and 2.4×10^{-4} au, respectively, using the same computation time (86 and 87 s). As in the case of the atom-based block definition, the bonded approach failed at producing sufficiently low gradients while the distance-based approach gave acceptable results (lower than the commonly used convergence criteria). The results were less satisfactory for the solvated enzyme, with a maximum deviation in the computed gradient components of more than 10×10^{-5} au for the distance-based approach (cutoff: 7 Å) and more than 15×10^{-5} au for the bonded approach.

According to our results, the atom-based block definition with a carefully chosen distance criterion should give the best results for this kind of hybrid approach. In studies of enzymes, it can be used together with the SOR technique if there are convergence problems. It could also be attractive for larger systems in general, since it can reduce the computation time significantly. Finally, it may also be useful for other polarizable force fields that do not employ the DO point-charge approximation.

4.2. QM/MM Boundary Treatment for the Drude Oscillator Model. In the development of the QM/MM method, special care has been taken to develop boundary treatments that allow cutting a covalent bond at the frontier between QM and MM regions.[65,66] Several approaches have been proposed, which normally work reasonably well if the frontier is chosen properly (e.g., cutting at an unpolar single bond that is as far away as possible from the electronically relevant part of the QM region).[1]

At a given atom, the DO model may involve the following electrostatic entities: the atomic point charge, the lone pairs, and the DP as well as its counter charge.[51] To define a boundary treatment, we have chosen not to take the lone pairs into account, assuming that no bond to a heteroatom will be cut. For the atomic point charges, we apply the commonly adopted charge shift scheme without any further modification.[67] In this scheme, the charge on the MM atom in the frontier bond (M1) is distributed to the other MM atom(s) that is (are) bonded to it (M2); a point dipole is added at these atoms (M2) to compensate for the charge shift, and the valence of the frontier QM atom is satisfied by adding a QM hydrogen atom (link atom). As this charge shift scheme performs well with additive force fields, we did not see any need to modify it. Therefore, we only have to develop a model for treating the DOs at the boundary.

We have investigated five different models. The first one (model 0) neglects the DO on the M1 atom, without any other modification. In model 1, the polarizability of M1 is transferred to the polarizable centers in the M2 position (without using this polarizability when computing the Thole screening function). As these M2 atoms become extremely polarizable in model 1, it may be more appropriate to compute the Thole function with the full M2 polarizability (including the contributed shifted from M1), which leads to model 2. The full transfer including the t_i parameters gives rise to model 3. Finally, as the M2 atoms are rather close to the virtually bonded QM region, we considered another model, model 4, in which the polarizable M2 atoms interact with the QM region according to the Thole model (without any transfer of parameters).

In the development of boundary treatments, the proton affinity (PA) and deprotonation enthalpy (DE) of n-propanol or n-butanol are commonly used as test systems.[65,68] In these molecules, different C−C bonds can be cut, 2 in the case of n-propanol and 3 for n-butanol. Cutting the C−C bond closest to oxygen (cut1) is considered as an extreme case, while the other options (cut2 and cut3) are more representative of typical QM/MM applications. In previous work, boundary treatments that were successful for these test systems have also performed well in other QM/MM applications.[65,66] Therefore, we only consider n-butanol in the following. In analogy to a previous study on propanol,[65] we used the semiempirical AM1 method[69] to evaluate DE and PA for frozen geometries of butanol, butanolate anion, and butanolium cation. Again following the literature,[65] we also included a sodium cation (Na^+) to simulate an "extreme" environment. In our tests, this cation was put either in the QM or the MM region. Instead of choosing a set of a few predetermined positions for the cation, we generated for each test 100 positions that were randomly picked at distances of at least 3 Å away from any atom of the molecule and within 9 Å of the geometric center of the molecule (the

statistical results remain essentially unchanged when running tests with 150 positions). Note that when the sodium ion is included in the QM region, the computations crashed occasionally when the MM part of n-butanol was situated between the QM part of n-butanol and the QM cation; in these cases (10−15%), extra computations were run to obtain 100 sampling points.

Table 1 compiles the deviations from the full QM results of DE (upper part) and PA (lower part) as well as their standard deviation (in parentheses) for the different boundary treatments (models 0−4), cation treatments (QM vs MM), and positions of the QM/MM boundary (cut1−cut3). The results are consistent with previous experience from additive force fields. The deviations are largest for cut1, and they decrease as the distance of the QM/MM boundary from the OH group increases (smallest for cut3). Regardless of other options, the best boundary treatment is always provided by model 0. Transferring polarizability to the M2 atom in model 1 systematically increases the deviation from the full QM reference results. Applying Thole damping at M2 slightly improves the results in model 2 (but not much), while transferring the complete Thole parameter set in model 3 makes things even worse. Applying Thole damping to the M2/QM interaction in model 4 is also detrimental, especially when the cation is part of the QM region.

Table 1. Average Deviation from the QM Results and Associated Standard Deviation (in Parentheses) for Deprotonation Enthalpy (Upper Part) and Proton Affinity (Lower Part) of *n*-Butanol in the Presence of a Sodium Cation at Different Positions (See Text) Computed with Different QM/MM Boundary Treatments (Models 0−4, See Text)[a]

Na^+	model	cut1		cut2		cut 3	
QM	0	6.91	(0.53)	3.72	(0.20)	2.20	(0.12)
	1	7.24	(0.57)	4.00	(0.28)	2.21	(0.16)
	2	7.21	(0.45)	4.02	(0.24)	2.20	(0.10)
	3	7.48	(1.25)	4.01	(0.28)	2.22	(0.21)
	4	8.45	(1.29)	6.20	(1.04)	1.38	(2.17)
MM	0	9.74	(1.57)	4.59	(0.37)	2.75	(0.31)
	1	10.47	(1.30)	4.49	(0.37)	2.78	(0.33)
	2	10.33	(1.37)	4.51	(0.35)	2.77	(0.32)
	3	11.25	(1.21)	4.50	(0.35)	2.79	(0.34)
	4	9.79	(1.51)	4.60	(0.37)	2.76	(0.31)
QM	0	2.98	(0.48)	3.10	(0.20)	1.49	(0.09)
	1	8.79	(0.46)	4.99	(0.19)	1.51	(0.11)
	2	7.62	(0.28)	4.98	(0.20)	1.50	(0.09)
	3	14.33	(4.75)	4.99	(0.19)	1.51	(0.11)
	4	16.39	(0.78)	11.00	(0.64)	1.55	(0.15)
MM	0	3.04	(0.46)	3.20	(0.26)	1.65	(0.26)
	1	7.79	(1.36)	4.73	(0.50)	1.67	(0.27)
	2	6.85	(0.98)	4.76	(0.47)	1.66	(0.27)
	3	12.92	(4.70)	4.73	(0.48)	1.68	(0.28)
	4	3.08	(0.47)	3.20	(0.26)	1.65	(0.26)

[a]Values are given in kcal/mol, and the full QM computation is taken as a reference. AM1 was used for the QM part and the Drude Oscillator force field for the MM region. The cuts are defined by the number of bonds from the hydroxyl group (cut1 being the nearest, and cut3 the farthest with only one methyl group in the MM region). The sodium cation (Na^+) can be part of the MM or QM region.

The cut1 is obviously not advisable, consistent with the rules of thumb known for additive force fields. As expected, the cut3 gives the most realistic results: the deviations from the QM reference results are smallest for both PA and DE, and the choice of boundary treatment has only little influence (disregarding model 4 with QM Na^+). Overall, the present results suggest that the inclusion of Drude oscillators in the force field has little effect on the boundary treatment, since the quality of the results is similar to what is found for additive force fields with standard boundary treatments. The simplest being the best, we recommend the use of model 0, i.e., simply to remove any DO at the M1 position in QM/MM-DO calculations.

4.3. Biologically Relevant Test Systems. To evaluate QM/MM-DO compatibility, we have chosen a set of simple, biologically relevant systems. The purpose of these tests is not to obtain any insight into the properties of the studied molecules but to systematically investigate the ability of CHARMM-DO to give more accurate results than the CHARMM22 additive force field in a QM/MM framework.[70] Usually, MM molecules are not involved or located in the direct vicinity of the chemical reaction being studied. We therefore assume that reproducing the binding energies and geometries of complexes will give valid insight into QM/MM-DO compatibility.

4.3.1. Water Trimer. Apart from its biological relevance, water is one of the cornerstones of force field parametrization.[22,28,64] Indeed, it is generally the first molecule to be parametrized and is then included in the parametrization of all others. Checking water is thus the logical first step when it comes to evaluating QM/MM-DO compatibility. The SWM4-NDP[64] water model (called SWM4 in this study and used in the CHARMM-DO force field) has been the subject of a few QM/MM studies. The water dimer was investigated in the tests by Lu and Zhang using BLYP/6-31G(d,p) as a QM component.[37] QM/MM studies on the solvation of QM ions have employed SWM4 as a polarizable solvent.[40,41]

Our test system is the cyclic water trimer.[71] This simple complex has the advantage that each of the three water molecules is equivalent, donating one hydrogen bond and accepting another one (see Figure 2). It has been investigated in several theoretical studies.[72−74] It has been included in the preliminary parametrization of SWM4 using a positive DP,[22] but not in the final one with a negative DP.[64] Yu et al. have shown that the SWM4 model properly reproduces both the binding energy (−14.35 vs −14.92 kcal/mol) and the geometry of the complex (RMSD 0.07 Å) using MP2/CBS as a benchmark.[75]

Two reasonable QM/MM separations are possible, with either one QM and two MM water molecules or *vice versa*. We have investigated both cases using the SWM4[64] and TIP3P[76] water models to compare with the standard additive force field (TIP3P). To differentiate between the possible combinations, we adopt a three-letter notation to define which model has been used for each individual water molecule; Q for QM, S for SWM4, and T for TIP3P (e.g., QSS means one QM and two SWM4 water molecules). For the QM part, we used standard DFT methods that are commonly employed for QM/MM simulations of enzymes (BP86,[77,78] BLYP,[79] B3LYP,[80] PBE,[81] and PBE0[82]). As the water trimer is a noncovalent complex, we included the Grimme D2 dispersion correction for BP86, BLYP, B3LYP, and PBE.[83] We also considered *ab initio* methods, Hartree−Fock (HF) and MP2;[60,61] the latter is

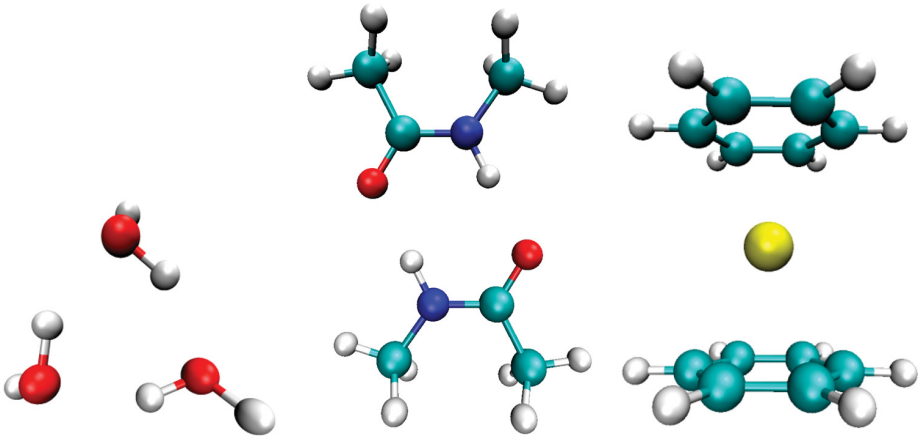


Figure 2. Complexes used in this study to evaluate QM/MM-DO compatibility. From left to right, the water trimer in its most stable cyclic form, the *cis*-NMA dimer, and the cationic bis(benzene)sodium sandwich complex.

known to give very good results for this complex. The SVP,[84] TZVP, and TZVPP[85] basis sets were used in all of these calculations. Further tests were performed using the semiempirical QM methods MNDO,[86,87] AM1,[69] PM3,[88] OM1,[89] OM2,[90] and OM3.[91,92] Since both the TIP3P and SWM4 models have been parametrized with rigid geometries, we constrained the internal geometry of the water molecules during the optimization but also considered the case when they are flexible (for consistency with the QM approach). We discuss the results of the geometry optimizations in terms of the binding energy of the complex and its structure. The binding energy was calculated from the energies of the optimized complex and the optimized water molecule. The geometry of the complex was assessed by two criteria, namely the O−O distances and the angles between the O−O−O plane and the hydrogen atoms not involved in hydrogen bonding. The latter angles are reported in absolute value, without distinguishing between the up or down orientation of the corresponding O−H bonds (note that the up−up−up and down−down−down orientations were never encountered). We do not focus here on the ability of the methods to reproduce experimental or high-level theoretical data but rather on the compatibility of the QM and MM potentials in a QM/MM framework to reproduce the QQQ results obtained with the same QM method.

Binding energies are listed in Table 2. Two key tendencies are observed. The first one is that the QM/MM prediction of the binding energy is improved upon basis set extension. This improvement is systematic for SWM4 when used in combination with DFT, DFT-D2, or MP2. For TIP3P, the binding energy is underestimated with MP2 and DFT, except when applying the D2 correction term. The second general trend concerns the changes in the QM/MM binding energies when replacing TIP3P by SWM4 in the MM part. For the first-principles QM methods, switching from QQT to QQS increases the binding energy by typically 2.05−2.46 kcal/mol (and even by 3.27 kcal/mol for B3LYP-D2/TZVPP). When going from QTT to QSS, this increase is even larger (3.31−3.44 kcal/mol). These changes in the binding energy are not due to the inclusion of polarization in SWM4 (always stabilizing in the complex and zero for an isolated water molecule) but rather to the parametrization of the MM model and the SWM4 treatment of the oxygen lone pair. For the semiempirical QM methods, the changes are in the same direction but less pronounced (see Table 2).

We now briefly address the performance of specific QM/MM combinations with regard to the reproduction of the pure QM reference energies. For DFT QM components, the functionals with Becke exchange tend to perform better than the parameter-free PBE approaches when combined with SWM4, while there is no such clear trend for TIP3P. When applying the D2 dispersion correction, the results for SWM4 deteriorate, and overall the DFT-D2/TIP3P combinations seem to perform better than DFT-D2/SWM4 (without clear distinction between different types of functionals). Concerning *ab initio* methods, HF/MM calculations give rather large deviations from the HF reference energies (regardless of the chosen MM model). The MP2/MM results are satisfactory for QQS but not for QSS, QQT, and QTT. There are no obvious specific patterns when using semiempirical QM methods; here, PM3 and OM3 seem to perform best.

Upon removing the geometry constraints on the MM water molecules during the optimizations (see the results given in parentheses in Table 2), the binding energies are generally increased slightly (as expected). The changes are typically on the order of 0.5 (0.3) kcal/mol for QQS (QQT), between 0 and 1.0 kcal/mol for QSS, and around 0.6 kcal/mol for QTT. These changes do not generally lead to a better reproduction of the QQQ reference results, and for the sake of consistency, it seems preferable to retain the constraints on the MM water geometries in QM/MM geometry optimizations (i.e., to use frozen MM water geometries). We note that this convention is usually not adopted during standard QM/MM minimizations with additive force fields.

With regard to the reproduction of the QQQ reference binding energies, the QM/SWM4 combination is clearly superior to QM/TIP3P for the QQX system when using the BP86, BLYP, and B3LYP functionals with the TZVPP basis set or *ab initio* methods with TZVP or TZVPP. This also holds for the QXX test systems in the case of DFT but not for *ab initio* methods. Overall, the best performance among all tested QM/MM variants is found when combining the SWM4 water model with the following QM components: DFT/TZVPP with DFT = BP86, BLYP, or B3LYP; MP2/TZVP; and MP2/TZVPP. The semiempirical PM3 and OM3 methods also give acceptable results. Generally, QM/SWM4 performs slightly better than QM/TIP3P (compared with the QQQ reference

Table 2. Binding Energies (kcal/mol) of the Water Trimer with Different QM Methods (See Text)[a]

Hamiltonian	basis	QQQ	QQS		QQT		QSS		QTT	
BP86	SVP	−28.45	−19.02	(−19.57)	−21.47	(−21.79)	−14.33	(−15.21)	−17.7	(−18.22)
	TZVP	−18.64	−17.3	(−17.84)	−19.49	(−19.82)	−15.78	(−16.75)	−19.2	(−19.77)
	TZVPP	−16.63	−16.16	(−16.7)	−18.28	(−18.59)	−15.24	(−16.36)	−18.61	(−19.16)
BLYP	SVP	−29.24	−19.45	(−19.98)	−21.83	(−22.15)	−14.09	(−14.21)	−17.44	(−17.95)
	TZVP	−18.04	−17.24	(−17.78)	−19.39	(−19.71)	−15.67	(−16.62)	−19.1	(−19.65)
	TZVPP	−16.14	−16.14	(−16.66)	−18.23	(−18.54)	−15.16	(−16.11)	−18.53	(−19.07)
B3LYP	SVP	−27.41	−19.04	(−19.57)	−21.47	(−21.79)	−14.4	(−15.26)	−17.74	(−18.27)
	TZVP	−18.5	−17.36	(−17.87)	−19.58	(−19.91)	−15.75	(−15.93)	−19.18	(−19.73)
	TZVPP	−16.43	−16.19	(−16.7)	−18.35	(−18.65)	−15.21	(−15.29)	−18.6	(−19.14)
PBE	SVP	−31.37	−20.23	(−20.77)	−22.63	(−22.95)	−14.28	(−15.13)	−17.63	(−18.15)
	TZVP	−20.92	−18.21	(−18.75)	−20.4	(−20.73)	−15.72	(−16.68)	−19.14	(−19.7)
	TZVPP	−18.99	−17.11	(−17.65)	−19.23	(−19.54)	−15.19	(−15.15)	−18.56	(−19.1)
PBE0	SVP	−27.52	−19.15	(−19.71)	−21.61	(−21.94)	−14.57	(−15.39)	−17.94	(−18.48)
	TZVP	−19.72	−17.73	(−18.27)	−19.99	(−20.32)	−15.77	(−15.88)	−19.2	(−19.76)
	TZVPP	−17.63	−16.56	(−17.05)	−18.75	(−19.06)	−15.23	(−16.17)	−18.62	(−19.16)
BPE86-D2	SVP	−31.97	−20.19	(−20.76)	−22.64	(−22.97)	−14.33	(−15.22)	−17.67	(−18.22)
	TZVP	−21.84	−18.37	(−18.94)	−20.57	(−20.91)	−15.77	(−16.75)	−19.18	(−19.77)
	TZVPP	−19.89	−17.25	(−17.8)	−19.38	(−19.69)	−15.24	(−16.2)	−18.59	(−19.16)
BLYP-D2	SVP	−33.31	−20.82	(−21.35)	−23.18	(−23.49)	−14.1	(−14.95)	−17.41	(−17.95)
	TZVP	−21.57	−18.44	(−19)	−20.6	(−20.92)	−15.67	(−15.8)	−19.07	(−19.65)
	TZVPP	−19.71	−17.35	(−17.89)	−19.45	(−19.76)	−15.16	(−15.27)	−18.5	(−19.07)
B3LYP-D2	SVP	−30.87	−20.18	(−20.72)	−22.61	(−22.94)	−14.4	(−15.22)	−17.72	(−18.27)
	TZVP	−21.55	−18.4	(−18.95)	−20.63	(−20.95)	−15.76	(−16.71)	−19.15	(−19.73)
	TZVPP	−19.52	−16.14	(−16.65)	−19.41	(−19.72)	−15.21	(−15.31)	−18.57	(−19.14)
PBE-D2	SVP	−33.89	−21.07	(−21.62)	−23.47	(−23.8)	−14.28	(−15.15)	−17.61	(−18.16)
	TZVP	−23.19	−18.98	(−19.55)	−21.18	(−21.5)	−15.73	(−15.91)	−19.12	(−19.7)
	TZVPP	−21.31	−17.88	(−18.41)	−20.02	(−20.32)	−15.2	(−15.15)	−18.54	(−19.11)
HF	SVP	−17.83	−16.65	(−17.19)	−18.9	(−19.24)	−15.02	(−15.9)	−18.39	(−18.95)
	TZVP	−14.04	−16.02	(−16.55)	−18.15	(−18.48)	−15.84	(−16)	−19.25	(−19.82)
	TZVPP	−12.07	−14.89	(−15.38)	−16.98	(−17.29)	−15.34	(−16.27)	−18.73	(−19.28)
MP2	SVP	−24.29	−18.38	(−18.92)	−20.8	(−21.13)	−14.76	(−15.55)	−18.15	(−18.7)
	TZVP	−18.2	−17.25	(−17.78)	−19.43	(−19.76)	−15.83	(−16.7)	−19.27	(−19.84)
	TZVPP	−17.41	−16.54	(−17.05)	−18.68	(−18.99)	−15.33	(−15.45)	−18.73	(−19.28)
MNDO	/	−1.42	−7.75	(−7.85)	−9.24	(−9.35)	−11.18	(−11.71)	−13.47	(−13.82)
AM1	/	−15.37	−12.07	(−12.22)	−13.55	(−13.68)	−11.44	(−11.9)	−13.91	(−14.28)
PM3	/	−10.07	−9.22	(−9.3)	−10.83	(−10.89)	−10.68	(−11.09)	−13.13	(−13.45)
OM1	/	−10.64	−12.63	(−12.96)	−14.47	(−14.71)	−13.37	(−13.91)	−16.18	(−16.65)
OM2	/	−14.19	−12.7	(−12.99)	−14.58	(−14.81)	−12.95	(−13.62)	−15.77	(−16.22)
OM3	/	−14.66	−13.38	(−13.73)	−15.59	(−15.88)	−13.78	(−14.46)	−16.75	(−17.25)

[a]They were computed at the full QM level (QQQ) or using a hybrid QM/MM approach with the TIP3P water force field (T) or the SWM4-NDP polarizable force field (S). Two kinds of QM regions were considered, one with two water molecules (QQT, QQS) and the other with one water molecule (QTT, QSS). Binding energies were obtained from geometry optimizations, in which the internal structure of each MM water molecule was constrained (values in parentheses from calculations without such constraints).

data). Finally, we note that, thanks to the careful parametrization, the QM/SWM4 binding energies are closer to the experimental values than the full QM or the QM/TIP3P results, regardless of the chosen QM method.

The optimized O−O distances are generally between 2.65 and 2.85 Å, except for the *ab initio* HF and the semiempirical methods. A significant difference is observed when it comes to symmetry. For pure QM calculations (QQQ), all three O−O distances are identical (within the precision of the optimization). This also holds for pure MM calculations, which yield equivalent O−O distances (TIP3P 2.75 Å, SWM4 around 2.80 Å). However, this is no longer true at the QM/MM level, where we generally find some spread between the different O−O distances. In the case of one MM water molecule (QQS and QQT), both MM models show a similar performance. When there are two MM water molecules (QTT and QSS), the spread is much smaller, and the computed O−O distances are close to their values from the full MM calculations. Interestingly, the values obtained do not depend much on the chosen QM method, and the QM/SWM4 results for QSS are thus generally close to experimental results (2.8 Å).[71] Contrary to the binding energies, basis set extension does not affect the O−O distances much.

The difference between the TIP3P and SWM4 water models becomes more pronounced when considering the angles between the O−O−O plane and the hydrogen atoms not involved in hydrogen bonding. In full MM optimizations, SWM4 gives realistic geometries, while TIP3P produces a planar trimer. As TIP4P is also known to give proper geometries,[71] this difference is probably due to the explicit treatment of the lone pairs in SWM4 (rather than polarization effects). Similar trends are observed at the QM/MM level. The

Table 3. Binding Energies (kcal/mol) and Hydrogen Bond Distances (Å) between the Two Monomers in the *cis*-NMA Dimer[a]

		full QM			QM/MM-DO			QM/MM		
Hamiltonian	basis	energy	dist1	dist2	energy	dist1	dist2	energy	dist1	dist2
BP86	SVP	−18.19	1.76	1.76	−17.64	1.72	1.30	−15.35	1.70	1.61
	TZVP	−14.41	1.82	1.82	−19.64	1.62	1.19	−16.58	1.72	1.60
	TZVPP	−14.11	1.80	1.80	−19.37	1.63	1.27	−16.40	1.73	1.60
BLYP	SVP	−17.75	1.81	1.81	−17.22	1.76	1.53	−15.06	1.71	1.61
	TZVP	−13.40	1.87	1.87	−19.38	1.63	1.20	−16.39	1.72	1.60
	TZVPP	−13.00	1.86	1.86	−19.30	1.64	1.25	−16.24	1.73	1.60
B3LYP	SVP	−17.69	1.84	1.84	−17.60	1.75	1.56	−15.43	1.70	1.62
	TZVP	−14.06	1.86	1.86	−19.67	1.60	1.19	−16.62	1.72	1.61
	TZVPP	−13.60	1.86	1.86	−19.30	1.67	1.40	−16.42	1.72	1.61
PBE	SVP	−19.84	1.76	1.76	−17.69	1.76	1.39	−15.30	1.71	1.61
	TZVP	−15.75	1.83	1.83	−19.54	1.62	1.19	−16.51	1.73	1.60
	TZVPP	−15.46	1.80	1.80	−19.29	1.65	1.32	−16.32	1.73	1.60
PBE0	SVP	−18.66	1.79	1.79	−17.83	1.72	1.39	−11.83	1.85	1.86
	TZVP	−15.36	1.83	1.83	−19.62	1.68	1.42	−16.68	1.71	1.61
	TZVPP	−14.96	1.82	1.82	−19.43	1.66	1.35	−16.50	1.72	1.61
BP86-D2	SVP	−22.02	1.72	1.72	−17.94	1.72	1.41	−15.41	1.71	1.61
	TZVP	−18.09	1.77	1.77	−19.78	1.63	1.18	−16.61	1.72	1.60
	TZVPP	−17.88	1.76	1.76	−19.66	1.61	1.20	−16.45	1.73	1.59
BLYP-D2	SVP	−21.83	1.77	1.77	−17.95	1.73	1.36	−15.09	1.71	1.62
	TZVP	−17.36	1.82	1.82	−19.61	1.68	1.40	−16.48	1.73	1.60
	TZVPP	−17.02	1.82	1.82	−20.50	1.69	1.38	−16.28	1.73	1.60
B3LYP-D2	SVP	−21.32	1.79	1.79	−17.71	1.75	1.54	−15.47	1.70	1.62
	TZVP	−17.53	1.82	1.82	−20.27	1.68	1.37	−16.66	1.72	1.61
	TZVPP	−17.22	1.82	1.82	−19.48	1.61	1.19	−16.47	1.72	1.61
PBE-D2	SVP	−22.58	1.73	1.73	−18.08	1.71	1.38	−15.32	1.71	1.61
	TZVP	−18.39	1.78	1.78	−20.36	1.71	1.38	−16.55	1.73	1.60
	TZVPP	−18.20	1.77	1.77	−19.41	1.62	1.22	−16.36	1.73	1.59
HF	SVP	−13.24	1.99	1.99	−18.45	1.73	1.60	−16.32	1.68	1.64
	TZVP	−11.30	2.01	2.01	−19.93	1.65	1.30	−16.94	1.69	1.64
	TZVPP	−7.08	1.86	1.86	−19.63	1.67	1.41	−16.72	1.70	1.63
MP2	SVP	−18.17	1.84	1.84	−17.06	1.86	1.86	−15.46	1.71	1.62
	TZVP	−15.46	1.85	1.85	−17.08	1.64	1.37	−16.30	1.72	1.61
	TZVPP	−16.38	1.82	1.82	−19.28	1.62	1.18	−16.26	1.72	1.61
MNDO		−1.17	3.42	3.42	−10.73	1.84	1.90	−10.43	1.78	1.82
AM1		−7.99	2.07	2.07	−11.82	1.85	1.87	−11.40	1.77	1.76
PM3		−6.58	1.80	1.80	−10.35	1.87	1.87	−10.79	1.84	1.78
OM1		−6.87	2.08	2.08	−13.65	1.79	1.80	−13.15	1.71	1.71
OM2		−13.75	1.64	1.64	−14.31	1.79	1.79	−13.73	1.70	1.70
OM3		−12.93	1.50	1.50	−14.61	1.78	1.74	−13.95	1.69	1.66

[a]Different QM methods are used to describe the QM monomer. Results are given for full QM and for QM/MM computations using either the polarizable Drude Oscillator force field or the CHARMM additive force field to represent the MM monomer. Distances are taken between the hydrogen atom and the oxygen acceptor atom.

QM/TIP3P optimizations always give planar trimer structures, while the QM/SWM4 calculations (in combination with any first-principle QM method) yield realistic out-of-plane angles that are typically 1−3° too large (compared with the QQQ reference data).

In the hybrid QM/MM computations, the SWM4 water model performs better than TIP3 overall, because it can properly reproduce the geometries, thanks to the explicit lone-pair treatment. Being partly parametrized with respect to high-level *ab initio* data, it tends to give accurate binding energies in a QM/MM framework (compared with the full QM data). The best QM/SWM4 results are obtained when employing the QM method and basis set used for its parametrization, namely MP2 with a large basis set. The SWM4 model is also compatible with DFT methods, especially when a large basis set is used. For fast QM/SMW4 computations, the semiempirical PM3 and OMX methods appear to be efficient alternatives for this system.

4.3.2. NMA Dimer. N-methylacetamide (NMA) often serves as a prototype test system when parametrizing a new force field for proteins, as it provides the smallest possible representation of the peptide bond.[93] It has also been used in benchmark studies that target biologically relevant data.[94,95] Here, we investigate the NMA dimer, in which both monomers are in their *cis* conformation. As shown in Figure 2, two equivalent hydrogen bonds are formed between the two monomers. The NMA dimer is particularly relevant for our purposes as this type of hydrogen bond is often encountered in QM/MM studies of enzymes.

In QM/MM work, one normally avoids QM/MM boundaries that cut through a hydrogen bond directly involving the substrate or other reactive species.[1] Here, we deliberately

perform a demanding test on the NMA dimer where this convention is violated, treating one NMA as a QM molecule and the other one as an MM molecule (thus cutting through both hydrogen bonds). Since the dimer is symmetric, the two possible assignments are equivalent. We again compare QM/MM results obtained with a polarizable DO force field and the additive CHARMM force field (QM/MM-DO vs QM/CHARMM). Parameters for the additive force field were taken from the distributed CGenFF set,[96] and DO parameters were a refined version of those developed by Harder et al.[97] We focus on the binding energy and the hydrogen bond lengths. The results are listed in Table 3. Full MM computations lead to a binding energy of −14.88 kcal/mol and a hydrogen bond distance of 1.75 Å for the DO model, compared with −11.90 kcal/mol and 1.74 Å for the CHARMM22 force field.

Concerning the binding energies, the full QM reference results are reproduced very well at the QM/MM-DO level when using DFT/SVP as the QM method (contrary to what has been found for the water trimer). Upon basis set extension from SVP to TZVP, the dimer is destabilized for any QM method (both in pure QM and hybrid QM/MM calculations, regardless of the chosen force field). Including dispersion corrections generally improves the QM/MM results, but not the QM/MM-DO results. The polarizable force field performs better than the additive one whenever SVP is used as the basis. Among the semiempirical QM methods, both force fields give good results in combination with OM2, which is known to perform well for these kinds of systems.

Table 3 also lists the hydrogen bonding distances in the NMA dimer, which are identical by symmetry in the pure QM and MM calculations. When using a hybrid QM/MM model, the symmetry is broken, and the difference (splitting) between the two computed hydrogen bond distances is an excellent criterion to assess the compatibility of the QM and MM descriptions. Compared to the full QM reference results, the QM/CHARMM calculations produce acceptable geometries. The splitting is 0.09 Å on average (maximum: 0.14 Å for BP86-D2/TZVPP and PBE-D2/TZVPP), and the deviation from the QM reference distances amounts to 0.16 Å on average (maximum: 0.37 Å for OM1, disregarding the pure QM results from MNDO which fails to give hydrogen bonds). The QM/MM-DO calculations generally perform less well: the splitting is 0.28 Å on average (maximum: 0.45 Å for BP86/TZVP), and the average deviation from the QM reference distances is 0.29 Å (maximum: 0.53 Å for HF/TZVP). The best QM/MM-DO distances are generally obtained with the SVP basis, which may at least partially explain the good results for the binding energy obtained with this basis set. Comparing the performance of different QM methods in the QM/MM-DO calculations, all DFT functionals fail to reproduce the pure QM(DFT) geometries (best match: splitting of 0.19 Å and deviation of 0.185 Å for B3LYP/SVP), and the inclusion of empirical dispersion corrections does not improve the results at all. Among the *ab initio* QM methods, MP2/SVP gives the best QM/MM-DO results with zero splitting and a deviation of only 0.02 Å, while the others do not perform well. On the other hand, semiempirical QM methods lead to a very small splitting at both the QM/MM-DO and QM/CHARMM levels, but they are less accurate at reproducing the corresponding QM reference distances. In the QM/MM-DO framework, PM3 and OM2 give no splitting and are closer to the reference distances than the corresponding full QM results.

To summarize, QM/MM-DO calculations with MP2/SVP as a QM component best reproduce both the energetic and geometric QM reference results, and they also yield excellent agreement with the experimental values. It should be noted in this context that MP2 with a basis of SVP-type quality has been used as the QM method for parametrizing the geometry of such compounds in the MM-DO force fields.[98] For other first-principles QM methods, QM/MM-DO tends to perform less well than QM/CHARMM, especially when it comes to the geometry of the NMA dimer. Among the semiempirical QM methods, OM2 seems to be the best choice.

4.3.3. π-Cation Interactions. The preceding comparisons indicate that the QM/MM-DO results tend toward those obtained by the DO model alone when the size of the MM region is expanded. We verify those aspects with another type of interaction relevant for enzyme catalysis: the π-cation interactions. We have chosen the cationic bis(benzene)sodium sandwich complex as a representative test system (Figure 2). This complex has no heteroatoms with lone pairs so that we can directly assess the effect of MM polarization.

At the MM-DO level, several benzene−sodium complexes were investigated by Orabi and Lamoureux.[99] Their results show a clear improvement over the CHAMM22/27 results in comparisons with experimental and high-level *ab initio* data.[100] This was achieved by a specific parametrization for the DO model, which is not used here as we would like to assess the difference between the reference data and the DO results. We use the initially proposed DO parameters, designed for condensed phase simulations.[101]

As in the case of the water trimer, we investigated several definitions of the QM and MM regions. We again use a three-letter notation: the first and third letters refer to the benzene moieties and the middle one to the sodium cation. In this scheme, Q stands for QM and D for DO. We tested every possible combination except QDQ, which showed some instability.

For such systems, MP2 is known to give fairly accurate results at reasonable computation cost.[100] It was employed to obtain the geometries of the benzene rings in the DO parametrization.[101] Therefore, we used it here in combination with the SVP, TZVP, TZVPP, and QZVP basis sets. The counterpoise correction for the basis set superposition error was computed for the QQQ and QQD systems. In the QQD case, this correction was evaluated for the QQ system without taking into account the benzene ring represented at the MM-DO level.

As previously, we first consider the binding energy of the complex. Figure 3 shows the MP2-based results for all basis sets and QM/MM partitionings, along with the experimental and the MM-DO value. The pure MP2 results (QQQ) are in good agreement with experimental results, especially for the larger basis sets. The MM-DO approach (DDD) overestimates the binding energy by around 15 kcal/mol when using the original DO parameters that had been calibrated for proper interaction with water molecules (solvation energy).[101] The QM/MM-DO binding energies (QQD, QDD, DQD) apparently interpolate between the QQQ and DDD values, approaching the MM-DO result when including two fragments in the MM region.

When treating only the sodium cation at the QM level (DQD), the computed binding energy is essentially identical to the MM-DO value, regardless of the chosen basis set, indicating that the MM-DO parameters for Na^+ are consistent with its QM description. Improvements in the parametrization should

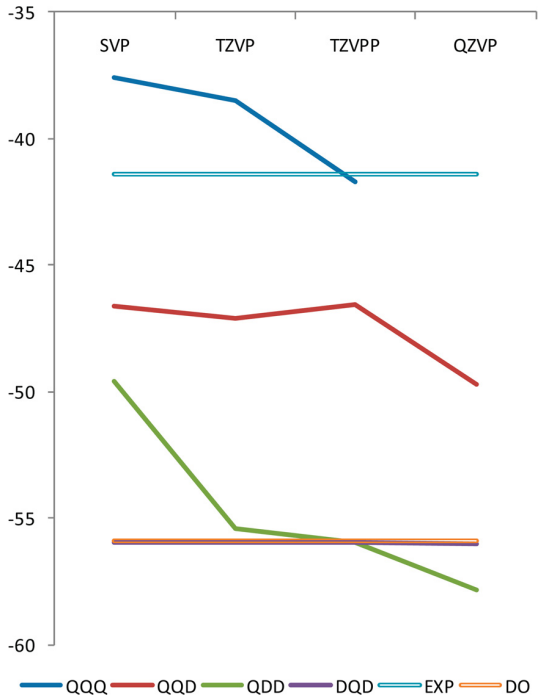


Figure 3. RI-MP2-based results for the energy of the cationic bis(benzene)sodium sandwich complex relative to the isolated fragments (in kcal/mol) obtained with four different basis sets. Shown are the full QM results (QQQ, blue), the QM/MM-DO results for different partitionings (QQD, red; QDD, green; DQD, purple; see text), the experimental value (EXP), and the pure MM-DO value (DO). Energies for complexes with more than one QM fragment were corrected for basis set superposition error (BSSE). For technical reasons, the BSSE could not be determined for the QQQ/QZVP combination.

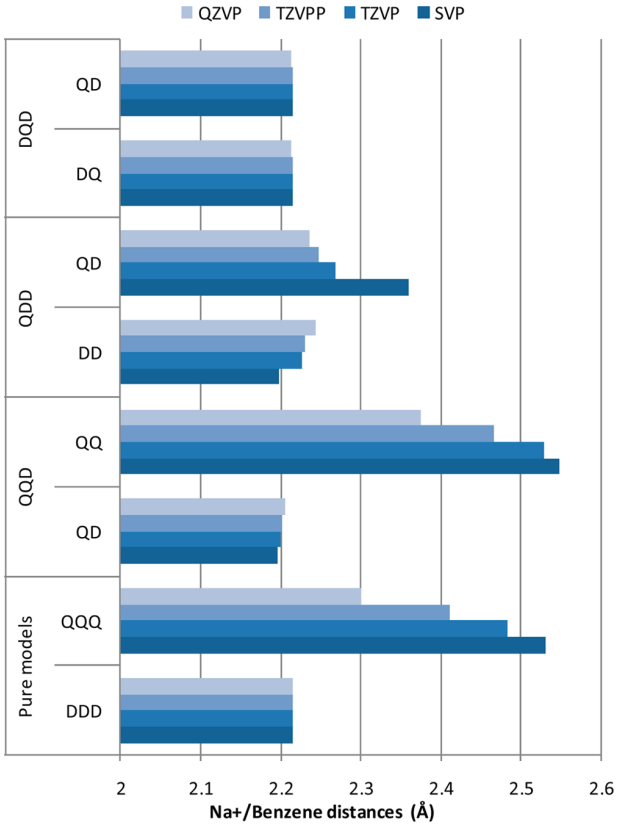


Figure 4. RI-MP2-based distances between the sodium cation and the center of the benzene rings in the cationic bis(benzene)sodium complex for four different basis sets (SVP, TZVP, TZVPP, and QZVP; see color code). Shown are results from the pure QM treatment (QQQ), from the pure MM-DO treatment (DDD), and from QM/MM-DO calculations (QQD, QDD, and DQD; the labels QQ, QD, and DQ specify which distance is plotted). See text for further details.

thus focus on the benzene part.[99] We note in this context that the QM/MM-DO nonbonded interactions are calculated from the same Lennard-Jones potential that is used in the DO model. As known from other QM/MM-DO studies,[99] the readjustment of these parameters for the QM atoms may improve the results through a more realistic description of the Na^+/benzene interaction. With proper reparameterization of the force field, QM/MM-DO calculations should give binding energies of an accuracy similar to the QQQ approach.

Figure 4 shows the optimized distances between the sodium cation and the center of the benzene rings for all currently investigated approaches. In the case of the pure models (QQQ and DDD), the complex is symmetric, and hence only one distance value is given. The symmetry is lost for the hybrid QM/MM models (QQD, QDD, DQD) for which both distances are shown with the corresponding assignments. In the pure QM model (QQQ), basis set extension from SVP to QZVP shortens the distance from 2.5 to 2.3 Å. The QQ distance in QQD shows similar behavior. In the pure MM model (DDD), the optimized distance is 2.2 Å. Similar values are obtained in the QM/MM models for distances involving a benzene ring described at the MM level, regardless of the chosen basis set (QQD, QDD, DQD; see Figure 4). The angle between the three moieties (benzene center−Na^+−benzene center) is always found to be around 180°.

4.4. Influence of Force Field Polarization on Enzymatic Reactions. In this section, we investigate the effect of DO polarization on enzymatic reactions. Our goal is to assess the influence of DO polarization on single-point QM/MM energies along previously determined enzymatic reaction pathways. In this initial study, we refrain from geometry optimizations and from free energy calculations (sampling), and hence also from comparison with experimental data, because we consider it most important to first gauge the basic effect of DO polarization on the QM/MM energetics.

As the main test enzyme, we have chosen chorismate mutase from *Bacillus subtilis*, which catalyzes the pericyclic Claisen rearrangement of chorismate into prephenate through a six-membered transition state (see Figure 5). This enzyme is well characterized experimentally[102] and has been extensively studied by QM/MM computations.[1,103] The rather small substrate (24 atoms) is a natural choice for the QM region; it is bound through noncovalent interactions so that there is no need for cutting bonds at the QM/MM boundary.[104]

For the MM region, we used the recently published CHARMM-DO parameters for proteins[39] and the SWM4 parameters for water.[64] No further DO parameters are required, since there are no other species in the MM region. For the QM-MM nonbonded interactions, we applied the CHARMM22 Lennard-Jones parameters of the substrate as in previous studies. These may not be the optimum choice, but we expect them to be realistic enough for a qualitative

dx.doi.org/10.1021/ct401095k | J. Chem. Theory Comput. 2014, 10, 1795−1809

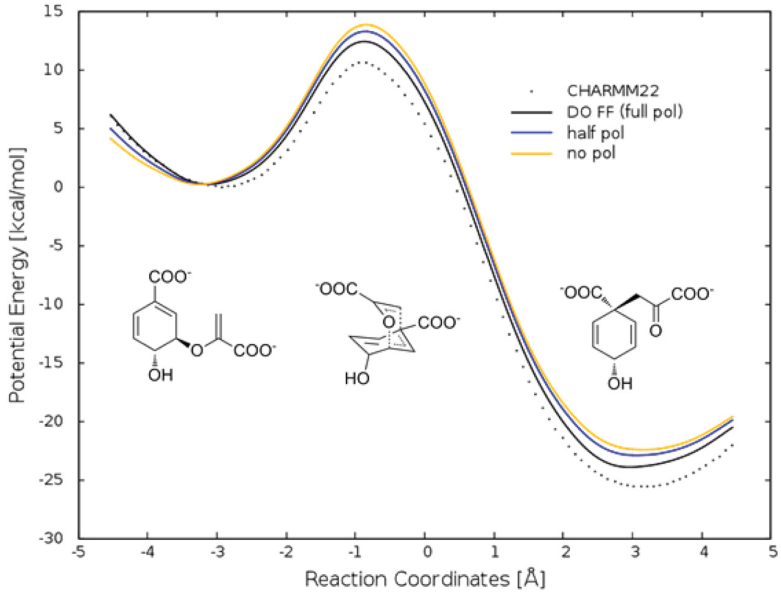


Figure 5. Potential energy profile of the Claisen rearrangement in chorismate mutase from QM(RI-MP2/SVP)/MM optimizations using the fixed-charge CHARM22 and the polarizable CHARMM-DO force fields. Also shown are QM/MM results obtained with the CHARMM-DO force field with the DO contributions switched off (no pol) or scaled down by a factor of 0.5 (half pol). The reaction coordinate is the difference between the distances for the breaking and forming bonds. The insets show the QM regions for the reactant (left), transition state (middle), and product (right). See text for further details.

assessment of the effect of MM polarization on the reaction profile.

In previous work of our group on chorismate mutase,[62] we had taken five snapshots from a classical molecular dynamics run of the reactant system and had used them as starting points for QM(MP2/SVP)/CHARMM22 geometry optimizations to determine reaction paths for the chorismate-prephenate rearrangement. We have now performed single-point calculations at the these previously optimized geometries using the QM(MP2/SVP)/CHARMM-DO approach. These calculations were done for all five pathways at all available points (intervals of 0.05 Å). The choice of QM method is supported by the fact that the MP2/SVP level of theory was the one that gave the best QM/MM-DO results for the NMA dimer (see section 4.2).

We adopted the following procedure for evaluating the effects of MM polarization. At any given geometry, we first replaced the fixed-charged CHARMM22 force field for the MM region by the nonpolarizable part of the CHARMM-DO force field (without the DO contributions but including the lone pair terms at the heteroatoms) and evaluated the corresponding QM/MM single-point energies (labeled "no pol"). We then included the DO contributions and reevaluated the single-point energies with full MM polarization using the SOR approach with a damping factor of 0.2 to obtain the DP positions (results labeled "full pol"). To check for consistency, we also considered the case in which the DO contributions are scaled down by applying a factor of 0.5, with appropriate scaling of the DO charges to preserve the force constant of the DO spring (results labeled "half pol"). We also tried to double the DO contributions, but this invariably led to nonconvergence of the iterative SOR procedure ("polarization catastrophe"). Figure 5 gives an example of the results obtained for one of the snapshots. Since the computed single-point energies show some minor irregular fluctuations ("noise" mostly arising from the replacement of the nonpolarizable terms), we used spline interpolation to produce smooth curves in Figure 5; this does not affect the following qualitative assessment of the effects of MM polarization.

Table 4 lists the changes in the computed barrier height and reaction energy for the five snapshots considered when switching from the "no pol" to the "full pol" model (in parentheses: from "no pol" to "half pol"). Evidently, the inclusion of full MM polarization consistently diminishes the barrier height, and it also tends to make the reaction more exothermic. The "full pol" results are close to those obtained with the additive CHARMM22 force field, which are known to be in good agreement with experimental results. Inclusion of half of the MM polarization normally leads to changes in the same direction, which are however less pronounced than might have been expected (typically ca. 40% except for $\Delta\Delta E$ of snapshot 4, see Table 4), indicating the nonlinear nature of polarization effects. Quantitatively, the effects of MM polarization on the computed barriers and reaction energies in chorismate mutase are rather small: for example, the calculated

Table 4. Effect of MM Polarization on the Barrier ($\Delta\Delta E^{\ddagger}$) and the Reaction Energy ($\Delta\Delta E$) of the Claisen Rearrangement in Chorismate Mutase in Five Independent Snapshots (in kcal/mol)[a]

snapshot	$\Delta\Delta E^{\ddagger}$		$\Delta\Delta E$	
1	−1.3	(−0.29)	−1.21	(−0.41)
2	−1.25	(−0.51)	0.23	(0.09)
3	−1.42	(−0.62)	−2.01	(−0.73)
4	−0.52	(−0.13)	−0.46	(0.02)
5	−1.15	(−0.51)	−2.99	(−1.16)

[a]The values correspond to the differences between "full pol" and "no pol" results (in parentheses: between "half pol" and "no pol"). See text for details.

barriers are lowered by 0.5 to 1.4 kcal/mol. The computed barrier heights for the five snapshots are around 10 kcal/mol; the contributions from MM polarization thus amount to about 5−15%, which is consistent with previous studies.[43]

The Claisen rearrangement catalyzed by chorismate mutase involves relatively little charge transfer,[105] and one may thus suspect that the impact of MM polarization could be more pronounced in enzymatic reactions that exhibit more pronounced charge redistribution. To check for this possibility, we have investigated the enzyme p-hydroxybenzoate hydrolase (PHBH) in a completely analogous manner. PHBH catalyzes the transformation of p-hydroxybenzoate into 3,4-dihydroxybenzoate, by formally moving an "OH^+" moiety from the cofactor to the substrate, with a concomitant charge transfer of one electron in the opposite direction. We selected the four snapshots from our previous QM/MM work on PHBH[105] and performed single-point QM/MM computations with and without MM polarization using the same QM method (B3LYP/6-31G*), the same QM region, and the same geometries as before.[105] Like in the case of chorismate mutase, the effect of MM polarization was assessed by switching it on and off. In these single-point QM/MM energy evaluations, the PHBH protein was represented by the CHARMM-DO force field, whereas the ribityl side chain was described by the standard CHARMM force field (due to the lack of CHARMM-DO parameters). This inconsistency is not expected to be severe, because the main impact of MM polarization should come from the polarizable PHBH residues surrounding the reactive center of the system (and not from the rather distant ribityl side chain at the opposite side of the cofactor). The single-point results for PHBH are collected in Table 5.

Table 5. Effect of MM Polarization on the Barrier ($\Delta\Delta E^{\ddagger}$) and the Reaction Energy ($\Delta\Delta E$) of the Electrophilic Substitution Reaction in p-Hydroxybenzoate Hydroxylase in Four Independent Snapshots (in kcal/mol)[a]

snapshot	$\Delta\Delta E^{\ddagger}$		$\Delta\Delta E$	
1	2.1	(1.7)	1.7	(1.2)
3	0.8	(0.6)	2.8	(1.9)
4	0.8	(0.6)	0.5	(0.3)
5	2.2	(1.3)	2.7	(1.7)

[a]The values correspond to the differences between "full pol" and "no pol" results (in parentheses: between "half pol" and "no pol"). See text for details. Snapshot 2 was discarded already in the original work for technical reasons.[105]

In analogy to chorismate mutase (see Table 4), the effects of MM polarization in PHBH are rather small on an absolute scale both for the barrier (0.8 to 2.2 kcal/mol) and for the reaction energy (0.5 to 2.7 kcal/mol). The QM/CHARMM-DO single-point values range from 7.6 to 11.0 kcal/mol for the barrier and from −26.1 to −31.0 kcal/mol for the reaction energy so that the contributions from MM polarization typically amount to 10−20% and 5−10%, respectively. It seems noteworthy that inclusion of MM polarization leads to a slight decrease of the barrier in chorismate mutase, but to a slight increase for PHBH. Incorporating MM polarization may thus shift barriers in different enzymes into different directions. More importantly, however, our two initial tests indicate that these shifts tend to be rather small regardless of whether the reaction involves little charge redistribution (chorismate mutase) or strong charge transfer (PHBH).

5. CONCLUSIONS

In this article, we have addressed several issues connected with the use of the polarizable DO force fields in QM/MM simulations of enzymatic reactions.

First, we investigated the convergence of the DO scheme in QM/MM geometry optimizations. We suggested and assessed two approaches that can be used in addition or instead of the one previously proposed.[49] The first one consists of iteratively solving the system of equations of polarization using Cholesky factorization instead of matrix inversion (for the sake of efficiency). The second one is a hybrid approach in which short-distance interactions are treated by Cholesky factorization and the remaining ones through an iterative self-consistent approach.

We further studied possible QM/MM boundary treatments involving MM atoms that carry Drude oscillators. By a series of tests for n-butanol in the presence of a sodium cation at different positions, we showed that the simplest possible model, namely the removal of the DO located at the MM frontier position (M1), gave the most satisfactory results, with an accuracy similar to what is normally achieved in standard QM/MM boundary treatments for additive force fields.

The systematic tests on the water trimer and the NMA dimer indicate that the DO model performs best in a QM/MM framework when employing the QM method and basis set used in the underlying MM parametrization. The match with experimental results is not perfect but benefits from the calibration of the MM parameters against experimental data. We thus propose to preferentially use this QM method in QM/MM calculations. This is consistent with the approach of Rowley and Roux who computed the solvation structure of sodium and potassium ions in water using MP2/def2-TZVP QM for the ions and neighboring water molecules in combination with the SWM4 model as a reference system.[40] A similar observation was made by Illingworth et al. when considering the water dimer with different polarizable force fields.[43] They concluded that some force fields were more compatible with certain basis sets than with others. We generalize this further by suggesting that best results will be obtained with the method used during MM parametrization. These conclusions are based on a few representative test systems, and further validation will be needed to support them.

In the selected test systems, the QM/MM-DO results for the binding energies and geometries tend to converge to the pure MM-DO values when describing an increasing part of the system at the MM-DO level. This will often lead to rather accurate results since the MM-DO force field generally gives results close to experimental ones (more so than the CHARMM22/27 additive force field).

In the chorismate mutase case, we find that switching on MM polarization in QM/MM single-point calculations affects the computed barrier height for the enzymatic Claisen rearrangement of chorismate to prephenate, but only to a rather minor extent: the barrier is lowered by 0.5−1.4 kcal/mol (i.e., by 5−15%) in the five snapshots studied. In PHBH, MM polarization leads to a slight increase in the computed barriers by 0.8−2.2 kcal/mol. We expect that inclusion of MM polarization may generally cause small changes on the order of 5−20% in the computed energies, but this will need to be confirmed by further studies that should also include geometry optimizations and free energy calculations.

AUTHOR INFORMATION

Corresponding Author

*E-mail: thiel@mpi-muelheim.mpg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Alex MacKerell and Pedro Lopes for helpful discussions and for providing the CHARMM-DO parameters for proteins prior to publication.

REFERENCES

(1) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198.
(2) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185.
(3) Shurki, A.; Warshel, A. In *Advances in Protein Chemistry*; Valerie, D., Ed.; Academic Press: San Diego, CA, 2003; Vol. *66*, p 249.
(4) Senn, H.; Thiel, W. In *Atomistic Approaches in Modern Biology*; Reiher, M., Ed.; Springer: Berlin, 2007; Vol. *268*, p 173.
(5) Mennucci, B. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6583.
(6) Senn, H. M.; Thiel, W. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182.
(7) Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425.
(8) Lopes, P. E. M.; Roux, B.; MacKerell, A. D. *Theor. Chem. Acc.* **2009**, *124*, 11.
(9) Yu, H.; van Gunsteren, W. F. *Comput. Phys. Com.* **2005**, *172*, 69.
(10) Warshel, A.; Kato, M.; Pisliakov, A. V. *J. Chem. Theory Comput.* **2007**, *3*, 2034.
(11) Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787.
(12) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.
(13) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.
(14) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549.
(15) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. *J. Chem. Theory Comput.* **2013**, *9*, 4046.
(16) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1504.
(17) Rappe, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.
(18) Rick, S. W.; Stuart, S. J.; Bader, J. S.; Berne, B. *J. Mol. Liq.* **1995**, *65*, 31.
(19) Stuart, S. J.; Berne, B. *J. Phys. Chem.* **1996**, *100*, 11934.
(20) Patel, S.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1.
(21) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.
(22) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.
(23) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 1120.
(24) Jiang, W.; Hardy, D. J.; Phillips, J. C.; MacKerell, A. D., Jr.; Schulten, K.; Roux, B. *J. Phys. Chem. Lett.* **2011**, *2*, 87.
(25) Yu, H.; Whitfield, T. W.; Harder, E.; Lamoureux, G.; Vorobyov, I.; Anisimov, V. M.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 774.
(26) Lopes, P. E. M.; Zhu, X.; Lau, A.; Roux, B.; MacKerell, A. D. *Biophys. J.* **2011**, *100*, 612.
(27) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2005**, *1*, 153.
(28) Yu, H. B.; Hansson, T.; van Gunsteren, W. *J. Chem. Phys.* **2003**, 118.
(29) Yu, H.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69.
(30) Straatsma, T. P.; McCammon, J. A. *Mol. Simul.* **1990**, *5*, 181.
(31) Geerke, D. P.; Thiel, S.; Thiel, W.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2007**, *3*, 1499.
(32) Geerke, D. P.; van Gunsteren, W. F. *J. Phys. Chem. B* **2007**, *111*, 6425.
(33) Nüsslein, V.; Schröder, U. *Phys. Status Solidi B* **1967**, *21*, 309.
(34) Schröder, U. *Solid State Commun.* **1993**, *88*, 1049.
(35) de Leeuw, N. H.; Parker, S. C. *Phys. Rev. B* **1998**, *58*, 13901.
(36) Geerke, D. P.; Thiel, S.; Thiel, W.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2007**, *3*, 1499.
(37) Lu, Z.; Zhang, Y. *J. Chem. Theory Comput.* **2008**, *4*, 1237.
(38) Boulanger, E.; Thiel, W. *J. Chem. Theory Comput.* **2012**, *8*, 4527.
(39) Lopes, P. E.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2013**, *9*, 5430.
(40) Rowley, C. N.; Roux, B. t. *J. Chem. Theory Comput.* **2012**, *8*, 3526.
(41) Riahi, S.; Roux, B.; Rowley, C. N. *Can. J. Chem.* **2013**, *91*, 552.
(42) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
(43) Illingworth, C. J. R.; Gooding, S. R.; Winn, P. J.; Jones, G. A.; Ferenczy, G. G.; Reynolds, C. A. *J. Phys. Chem. A* **2006**, *110*, 6487.
(44) Illingworth, C. J. R.; Parkes, K. E.; Snell, C. R.; Marti, S.; Moliner, V.; Reynolds, C. A. *Mol. Phys.* **2008**, *106*, 1511.
(45) Kästner, J.; Senn, H. M.; Thiel, S.; Otte, N.; Thiel, W. *J. Chem. Theory Comput.* **2006**, *2*, 452.
(46) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341.
(47) Antila, H.; Salonen, E. In *Biomolecular Simulations*; Monticelli, L., Salonen, E., Eds.; Humana Press: New York, 2013; Vol. *924*, p 215.
(48) van Duijnen, P. T.; Swart, M. *J. Phys. Chem. A* **1998**, *102*, 2399.
(49) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580.
(50) Xie, W.; Pu, J.; Gao, J. *J. Phys. Chem. A* **2009**, *113*, 2109.
(51) Zhu, X.; Lopes, P. E. M.; MacKerell, A. D. *WIREs Comput. Mol. Sci.* **2012**, *2*, 167.
(52) Baker, C. M.; MacKerell, A. D. *J. Mol. Model.* **2010**, *16*, 567.
(53) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1587.
(54) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjøvoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *J. Mol. Struct.: THEOCHEM* **2003**, *632*, 1.
(55) ChemShell. www.chemshell.org (accessed Dec 17, 2013).
(56) Metz, S.; Kästner, J.; Sokol, A. A.; Keal, T. W.; Sherwood, P. *WIREs Comput. Mol. Sci.* **2014**, *4*, 101−110.
(57) Thiel, W. *MNDO program*; Max-Planck-Institut für Kohlenforschung: Mülheim an der Ruhr, Germany, 2004.
(58) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.
(59) Forester, T. R.; Smith, W. *DL-POLY program*; Daresbury Laboratory: Daresbury, Warrington, England, 1996.
(60) Weigend, F.; Häser, M. *Theor. Chem. Acc.* **1997**, *97*, 331.
(61) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
(62) Polyak, I.; Benighaus, T.; Boulanger, E.; Thiel, W. *J. Chem. Phys.* **2013**, *139*, 064105.
(63) Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2009**, *5*, 3114.
(64) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. *Chem. Phys. Lett.* **2006**, *418*, 245.
(65) Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. *J. Phys. Chem. A* **2000**, *104*, 1720.
(66) König, P. H.; Hoffmann, M.; Frauenheim, T.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 9082.
(67) de Vries, A. H.; Sherwood, P.; Collins, S. J.; Rigby, A. M.; Rigutto, M.; Kramer, G. J. *J. Phys. Chem. B* **1999**, *103*, 6133.
(68) Lin, H.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 3991.
(69) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(70) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe,

M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.
(71) Keutsch, F. N.; Cruzan, J. D.; Saykally, R. J. *Chem. Rev.* **2003**, *103*, 2533.
(72) Xantheas, S. S.; Dunning, J. T. H. *J. Chem. Phys.* **1993**, *98*, 8037.
(73) Fowler, J. E.; Schaefer, H. F., III. *J. Am. Chem. Soc.* **1995**, *117*, 446.
(74) Wales, D. J. *J. Am. Chem. Soc.* **1993**, *115*, 11180.
(75) Yu, W.; Lopes, P. E. M.; Roux, B.; MacKerell, J. A. D. *J. Chem. Phys.* **2013**, *138*, 034508.
(76) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(77) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
(78) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
(79) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
(80) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(81) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
(82) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.
(83) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
(84) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.
(85) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.
(86) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
(87) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4907.
(88) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
(89) Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, *14*, 775.
(90) Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495.
(91) Scholten, M. Ph.D. Thesis, Universität Düsseldorf: Düsseldorf, Germany, 2003.
(92) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751.
(93) Ponder, J. W.; Case, D. A. In *Advances in Protein Chemistry*; Valerie, D., Ed.; Academic Press: San Diego, CA, 2003; Vol. *66*, p 27.
(94) Möhle, K.; Hofmann, H.-J.; Thiel, W. *J. Comput. Chem.* **2001**, *22*, 509.
(95) Gao, J.; Freindorf, M. *J. Phys. Chem. A* **1997**, *101*, 3182.
(96) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. *J. Comput. Chem.* **2010**, *31*, 671.
(97) Harder, E.; Anisimov, V. M.; Whitfield, T.; MacKerell, A. D., Jr.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3509.
(98) Harder, E.; Anisimov, V. M.; Whitfield, T.; MacKerell, A. D.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3509.
(99) Orabi, E. A.; Lamoureux, G. *J. Chem. Theory Comput.* **2011**, *8*, 182.
(100) Lamoureux, G.; Orabi, E. A. *Mol. Simul.* **2012**, *38*, 704.
(101) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2007**, *111*, 2873.
(102) Kast, P.; Asif-Ullah, M.; Hilvert, D. *Tetrahedron Lett.* **1996**, *37*, 2691.
(103) Claeyssens, F.; Ranaghan, K. E.; Lawan, N.; Macrae, S. J.; Manby, F. R.; Harvey, J. N.; Mulholland, A. J. *Org. Biomol. Chem.* **2011**, *9*, 1578.
(104) Senn, H. M.; Kästner, J.; Breidung, J.; Thiel, W. *Can. J. Chem.* **2009**, *87*, 1322.
(105) Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2011**, *7*, 238−249.