QM/MM Investigations of Enzymatic Reactions

Inaugural-Dissertation

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

vorgelegt von

Iakov Polyak
aus Moskau

Mülheim an der Ruhr/Düsseldorf 2013

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
Korreferent:	Prof. Dr. Christel Marian

Tag der mündlichen Prüfung:

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

(Iakov Polyak)

Acknowledgements

Foremost, I would like to thank my PhD supervisor Prof. Thiel for support of my research, fruitful suggestions, and immense knowledge. I am grateful for the opportunities to visit scientific meetings, workshops and schools and for the continuous financial support.

I am grateful to Dr. Waller (Westfälische Wilhelms-Universitä Münster, Germany) for his guidance and introduction into the QM/MM calculations at the beginning of my PhD.

I would like to acknowledge the fruitful collaborations with Prof. Reetz (Philipps-Universität Marburg, Germany), Hansel Gómez and Prof. Laura Masgrau (Universitat Autònoma de Barcelona, Spain), Dr. Elsa Sanchez Garcia (Max-Planck-Institut für Kohlenforschung, Germany) and Prof. Wolfram Sander (Ruhr-Universität Bochum, Germany).

I am thankful to Dr. Kenno Vanommeslaeghe (University of Maryland, USA) for his help in the parametrization of small molecules for the CHARMM (CGenFF) force field.

I also thank all my former and current colleagues from the Max-Planck-Institut für Kohlenforschung for fruitful discussions and scientific collaboration, especially Dr. Andrey Yachmenev, Dr. Tobias Benighaus, Eliot Boulanger, Dr. Kakali Sen, Dr. Gopakumar Gopinadhanpillai, Dr. Axel Koslowski, Dr. Jürgen Breidung, Claudia Loerbroks, Dr. Rachel Crespo-Otero, and Dr. Mario Barbatti.

I am deeply grateful to my wife for her everyday support.

List of papers included in the thesis

(1) "Quantum Mechanical/Molecular Mechanical Study on the Mechanism of the Enzymatic Baeyer-Villiger Reaction", I. Polyak, M. T. Reetz, W. Thiel, *J. Am. Chem. Soc.*, **134**, 2732 (2012).
Carried out and analysed all the calculations in the paper.

(2) "A New Type of Stereoselectivity in Baeyer-Villiger Reactions: Access to *E*-and *Z*-Olefins", Z.-G. Zhang, G.-D. Roiban, J. P. Acevedo, I. Polyak, M. T. Reetz, *Adv. Synth. Catal.*, **355**, 99 (2013).
Performed modelling and QM/MM optimizations of the substituted Criegee intermediate, in an attempt to explain the experimentally observed stereoselectivity towards a specific product.

(3) "Quantum Mechanical/Molecular Mechanical Study on the Enantioselectivity of the Enzymatic Baeyer-Villiger Reaction of 4-Hydroxycyclohexanone", I. Polyak, M. T. Reetz, W. Thiel, *J. Phys. Chem. B*, **117**, 4993 (2013).
Carried out and analysed all the calculations in the paper.

(4) "Retaining Glycosyltransferase Mechanism Studied by QM/MM Methods: Lipopolysaccharyl-α-1,4-galactosyltransferase C Transfers α-Galactose via an Oxocarbenium Ion-like Transition State", H. Gomez, I. Polyak, W. Thiel, J. M. Lluch, L. Masgrau, *J. Am. Chem. Soc.*, **134**, 4743 (2012).
Contributed to the setup, the computation of the potential energy surface, and the free energy QM/MM calculations.

(5) "A Microiterative Intrinsic Reaction Coordinate Method for Large QM/MM Systems", I. Polyak, E. Boulanger, K. Sen, W. Thiel, *Phys. Chem. Chem.*

Phys., (2013), DOI: 10.1039/c3cp51669e.
Implemented the main algorithm and performed the test calculations.

(6) “QM/MM Dual Hamiltonian Free Energy Perturbation”, I. Polyak, T. Benighaus, E. Boulanger, W. Thiel, *J. Chem. Phys.*, revised version.
Implemented the algorithm and performed the test calculations.

Own tasks are specified in italics.

Other publications not submitted as part of the thesis.

(1) “The Phenoxyl Radical-Water Complex - A Matrix Isolation and Computational Study”, W. Sander, S. Roy, I. Polyak, J. M. Ramirez-Anguita, E. Sanchez-Garcia, *J. Am. Chem. Soc.*, **134**, 8222 (2012).

(2) “The Cinchona Primary Amine-Catalyzed Asymmetric Epoxidation and Hydroperoxidation of α,β-Unsaturated Carbonyl Compounds with Hydrogen Peroxide”, O. Lifchits, M. Mahlau, C. M. Reisinger, A. Lee, C. Fares, I. Polyak, G. Gopakumar, W. Thiel, B. List, *J. Am. Chem. Soc.*, **135**, 6677 (2013).

Abstract

The main goal of this work was to apply combined quantum mechanics/molecular mechanics (QM/MM) methods to investigate two different enzymatic reactions and to develop new QM/MM techniques in the field of reaction path and free energy calculations.

Cyclohexanone Monooxygenase. The main target of the current work was the reaction mechanism of Cyclohexanone Monooxygenase (CHMO) - a representative Baeyer-Villiger monooxygenase enzyme, responsible for oxygenation of cycohexanone into ϵ-caprolactone in nature, but also famous for accepting and being enantioselective towards a wide range of substrates. We have for the first time modeled the cyclohexanone oxygenation reaction in the active site of CHMO, and explained the experimentally observed regio- and stereoselectivity of this enzyme towards several substituted substrates.

Lipopolysaccharyl-α-1,4-galactosyltransferase C. A QM/MM study of the reaction mechanism of lipopolysaccharyl-α-1,4-galactosyltransferase C (LgtC), a member of the family of retaining galactosyltransferase (GT) enzymes, was performed. We have assessed the different proposed mechanisms and showed that the S_Ni mechanism, via a front-side attack of the acceptor oxygen to the donor anomeric carbon atom, most probably takes place in the wild type enzyme. We have also analysed key interactions that help this attack by stabilizing the corresponding oxocarbenium ion-like transtion state.

Microiterative IRC. We have implemented and tested a microiterative method to perform intrinsic reaction coordinate (IRC) calculations for QM/MM systems of arbitrary size and complexity. In this approach, the reactive system

is divided into a core and an outer region, with each step of the core region along the IRC being followed by the minimization of the outer region. We have tested the microiterative IRC method for a number of systems with varying size and complexity and showed that it is an efficient and relatively accurate procedure to perform reaction path calculations for large systems.

DH-FEP. We have developed a new QM/MM free energy technique called dual Hamiltonian free energy perturbation (DH-FEP). The method is superior to the conventional QM/MM-FE scheme due to the explicit sampling of the QM region. In order to allow for sufficient sampling of the reactive system, simulations are performed at a semiempirical QM/MM level, while the perturbation energies are evaluated at a higher level of theory (QM = DFT or MP2). The performance of the method strongly depends on the geometrical correspondence between the two levels. We have performed test calculations and suggested suitable strategies to identify the best matching pair of methods. We have also proposed the use of collective reaction coordinates within the DH-FEP method, which makes the problem of configurational space overlap between the two levels less relevant at the price of a slightly reduced sampling of the QM region.

Zusammenfassung

Die vorliegende Arbeit berichtet über kombinierte quantenmechanische und molekülmechanische (QM/MM) Untersuchungen zum Mechanismus von zwei unterschiedlichen enzymatischen Reaktionen und über die Entwicklung neuer QM/MM Methoden zur Berechnung von Reaktionspfaden und freien Energien.

Cyclohexanon-Monooxygenase. Ein Hauptziel der Arbeit war die detaillierte Aufklärung des Reaktionsmechanismus der Cyclohexanon-Monooxygenase (CHMO). Diese prototypische Baeyer-Villiger-Monooxygenase katalysiert nicht nur die Oxidation von Cyclohexanon zu ϵ-Caprolacton in der Natur, sondern auch die enantioselektive Umsetzung einer Vielzahl anderer Substrate. Wir haben erstmals die Oxidation von Cyclohexanon in der aktiven Tasche von CHMO modelliert und können die experimentell beobachtete Regio- und Stereoselektivität des Enzyms bezüglich mehrerer Substrate erklären.

Lipopolysaccharyl-α-1,4-galactosyltransferase C. Der Reaktionsmechanismus von Lipopolysaccharyl-α-1,4-galactosyltransferase C (LgtC), einer unter Retention arbeitenden Galactosyltransferase (GT), wurde mit QM/MM Rechnungen studiert. Wir haben verschiedene vorgeschlagene Mechanismen untersucht und gezeigt, dass der S_Ni Mechanismus, über einen frontalen Angriff des Akzeptorsauerstoffs am anomeren Donorkohlenstoff, im Wildtyp des Enzyms bevorzugt ist. Dabei wurden die wesentlichen Wechselwirkungen analysiert, die diesen Angriff durch Stabilisierung des zugehörigen Oxocarbenium-Übergangszustands unterstützen.

Mikroiterative IRC. Wir haben eine mikroiterative Methode zur Berechnung von intrinsischen Reaktionskoordinaten (intrinsic reaction coordinates, IRC) für QM/MM-Systeme beliebiger Größe und Komplexität implementiert und getestet. Hierbei wird das System in eine Kern- und eine Außenregion aufgeteilt, wobei jeder Schritt in der Kernregion entlang des IRC von einer Minimierung der Außenregion gefolgt wird. Wir haben das mikroiterative IRC Verfahren an verschieden großen und komplexen Systemen getestet und gezeigt, dass es eine effiziente und recht genaue Methode ist, um Reaktionspfadberechnungen für große Systeme durchzuführen.

DH-FEP. Wir haben eine neue QM/MM Freie-Energy-Methode (dual Hamiltonian free energy perturbation, DH-FEP) entwickelt. Diese Methode ist dem konventionellem QM/MM-FE Schema durch die explizite Behandlung der QM-Region überlegen. Um für ein ausreichendes Sampling des reaktiven Bereichs zu sorgen, werden die Simulationen auf einem semiempirischen QM/MM Niveau durchgeführt, während die Störungsenergien auf einem höheren Theorielevel (QM = DFT oder MP2) berechnet werden. Die Genauigkeit der Methode hängt stark von der Übereinstimmung der Strukturen auf den beiden Niveaus ab. Wir haben anhand von Testrechnungen passende Strategien entwickelt, um das beste Methodenpaar zu finden. Empfehlenswert ist die Benutzung von kollektiven Reaktionskoordinaten im Rahmen der DH-FEP Methode, um die Konfigurationsraumüberlappung zu verbessern, auf Kosten eines leicht verringerten Samplings der QM-Region.

Contents

List of Figures

List of Tables

1

Introduction

Hybrid quantum mechanics/molecular mechanics (QM/MM) techniques have in the last decades become an invaluable tool to study biomolecular systems, especially enzymatic reactions. On the one hand, the protein environment plays an important role in the reaction mechanism, and therefore must be included explicitly in calculations. Solvated proteins, containing thousands of atoms, are obviously too large to be treated solely by any of the available QM techniques, calling for the use of classical force fields. On the other hand, MM methods can only describe protein dynamics, but will fail for chemical reactions. Therefore, the natural way to study enzymatic reactions is to perform hybrid calculations, in which the reactive centre is treated with a QM method, as accurately as needed, while the protein and solvent environment are treated classically.

The QM/MM approach was first used in the seminal paper by Warshell and Levitt in 1976 (1), and reentered the world of computational chemistry in 1990 with introduction of hybrid QM/MM calculations in the CHARMM program (2). Since then its popularity has been rising, and a lot of different approaches to perform QM/MM calculations were developed, as well as specialized methods to explore QM/MM potential energy surfaces and to run free energy calculations.

In the QM/MM calculations, one chooses a reactive part of the system that is treated quantum-mechanically, while the rest of the system is described with a force field. The main question is how to handle the coupling terms, arising due to interactions between the QM and MM regions, as well as the boundary between the two subsystems. For a detailed discussion of the existing methods the reader is referred to a review by Senn and Thiel (3) and references therein. In the following, we will cover only the major points.

There are in general no limitations in the choice of the QM and MM methods to be used in the QM/MM calculations. There are only few widely used biomolecular force fields available, which are commonly considered equally suitable for QM/MM studies. On the other hand, one can choose between a large variety of QM methods that may vastly differ in accuracy and computational effort. The selection of the QM method is generally based on the same criteria as in pure QM studies.

Density functional theory (DFT) is most commonly used in QM/MM calculations, due to its favorable cost/performance ratio. Among the exchange-correlation functionals, B3LYP (4, 5, 6, 7, 8, 9) is one of the most popular choices, because it usually gives satisfactory results for many chemical systems. However, other functionals may be superior in specific applications, and one should thus base the decision on available benchmark studies and/or preliminary test calculations to verify the functional that performs best for a given chemical system. Standard DFT does not describe van der Waals dispersive interactions properly, which may play an important role in biological systems (10). Empirical corrections that account for dispersion have therefore been introduced (DFT-D method) (11). In the present work we employ DFT for studying enzymatic reactions. We generally use the B3LYP or BP86 (4, 5, 6, 12, 13) functional for geometry optimizations, and perform single-point calculations at the optimized stationary points using B3LYP-D, and in some cases also M06-2X, a modern functional that describes mid-range dispersion well (14) (Chapters 2 and 3).

QM/MM molecular dynamics (MD) simulations require significantly more energy and gradients evaluations than geometry optimizations. For practical reasons, such MD runs are often done at the semiempirical QM/MM level. In the present work we use the OM3 (15, 16), AM1 (17) and SCC-DFTB (18) methods to perform MD sampling during free energy calculations, and to carry out reaction path calculations for testing purposes (Chapters 4 and 5).

Due to their large computational cost, accurate *ab initio* post-Hartree-Fock methods are still of limited use in QM/MM calculations, since the QM region usually contains a relatively large number of atoms. However, the recent development (19, 20, 21) of linear-scaling local correlation methods (LMP2, LCCSD) expands the applicability of the *ab initio* methods in biological systems (usually at the level of single-point calculations). Second-order Møller-Plesset (MP2) perturbation theory, in combination with the resolution-of-the-identity (RI) approximation, allows us to perform geometry optimizations of QM regions of moderate size nearly as fast as with DFT (B3LYP). In the present work, we use RI-MP2 (22, 23), as implemented in Turbomole program (24), to follow reaction paths and to perform sequences of single-point QM/MM calculations for an enzymatic system that contains 24 atoms in its QM region (Chapter 5).

There are two distinct approaches of how to evaluate the QM-MM interactions: substractive and additive schemes. In the substractive scheme, the energy of the whole system is evaluated at the MM level, while the QM subsystem energy is calculated at both QM and MM levels. The MM energy for the whole system is then added to the QM energy of the subsystem, and the MM energy of the subsystem is substracted out in order to avoid double counting. This method was not used in the current work, and we therefore refrain from discussing it further.

The second and most widely used approach does not require an MM calculation of the QM region and thus avoids the problem of having good force field parameters available for a QM subsystem. In the additive scheme, the force field is only used to evaluate the energy of the MM part of the system, which is added to the QM energy of the QM region. One then has to evaluate the QM-MM interactions that, according to force field notation, consist of bonded, van der Waals and electrostatic contributions. These terms can be evaluated either at the QM or MM level.

The most popular choice for the electrostatic QM-MM interactions is to perform QM calculations in the presence of the MM point charges, which are incorporated as one-electron terms in the QM Hamiltonian. This approach is called electrostatic (electronic) embedding. The QM energy then contains the QM-MM electrostatic interactions, and the QM region gets automatically polarized by the environment. Some care is required when applying electrostatic embedding because the QM region may be overpolarized by closely-lying MM point charges, as will be discussed in Chapter 2.

The van der Waals QM-MM interactions are normally evaluated purely at the MM level. They are usually unproblematic due to their short-range nature. The energy between two atoms is computed from a standard Lennard-Jones potential using parameters that are consistent with those employed for the MM region. Sometimes it may be advantageous to use specially parameterised Lennard-Jones parameters for the QM atoms, e.g. in order to compensate for electrostatic overpolarization at the QM-MM boundary (25). This can be especially helpful for simulating ions in solution, in order to ensure that the hydrogen bonding distances and interaction energies between solvent (water) molecules are compatible in the QM and MM region and match the pure QM results (26).

The bonded interactions are more involved, and there exist many different approaches to treat them. In some fortunate cases (e.g. in QM/MM studies of solvation), there are no bonded interactions between the QM and MM region, but in general, this is unavoidable (e.g. in enzymatic reactions or when a chromophore is covalently bound to the protein backbone). The bonded interaction is then usually evaluated at the MM level, but special care must be taken on how to perform the QM calculations, since the clipped bond at the QM-MM boundary must be capped. There are three most popular treatments, namely the link-atom, boundary-atom, and localized-orbital schemes. We only discuss the link-atom method that was used in the present work.

In the link-atom method, an auxiliary atom (usually hydrogen) is placed in between the boundary QM and MM atoms and serves as a capping atom in the QM calculations. To avoid introducing additional degrees of freedom, it is usually positioned and kept fixed on the line between the boundary atoms. During geometry optimizations or MD simulations, the forces acting on a link atom are distributed over the boundary atoms via the chain rule. Due to their close distance, the boundary QM atom may become overpolarized by the neighboring MM point charge. One of the ways to overcome this problem is to set the frontier MM atom charge to zero and to evenly distribute its original charge over the MM atoms directly bonded to it, with additional correction terms to preserve the corresponding bond dipoles (charge-shift scheme (27, 28)).

QM/MM studies require efficient techniques for exploring both the potential and free energy surfaces of large polyatomic systems. In case of optimization techniques, special attention needs to be paid to the scalability of the method. While the size of QM/MM systems is comparable to the dimensions of usual MM systems, the computational costs for energy and gradient evaluations increase enormously with the use of QM techniques. Therefore, a linear-scaling method, such as the low-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method (29, 30), is desirable. L-BFGS is a quasi-

Newton optimization algorithm, in which only the diagonal of the Hessian matrix is stored, and steps are done with the use of information about gradients and positions at a limited number of previous steps.

Internal coordinates are usually more suitable for faster geometry optimizations than the more strongly coupled Cartesian coordinates (31). However, the conversion from Cartesian to internal coordinates and backwards usually does not scale linearly, which becomes a computational bottleneck with increasing system size. In an attempt to solve this problem, hybrid delocalized internal coordinates (HDLC) have been introduced (31). Here, the system is divided into many fragments, each of which is described with a set of internal coordinates. The relative orientation of these small subsystems is handled by the Cartesian HDLC components. Such partitioning of a whole system into local subsets of internal coordinates decreases the coupling between the fragments and allows for faster coordinate manipulations.

In order to obtain accurate results, QM/MM calculations often require the use of relatively large QM regions (50-150 atoms). When searching for a transition state on a potential energy surface, one often needs to have a sufficiently accurate Hessian at every optimization step, which becomes very expensive with the growth of the QM region. A microiterative transition state optimization (32) is an efficient way to tackle this problem. In this approach, the Hessian is calculated only for a small subset of atoms, called the core region, in which steps are taken towards the transition state following a second-order optimization algorithm, e.g. the partitioned rational function optimizer (P-RFO) (33, 34). After every step by the core atoms, all remaining degrees of freedom are totally relaxed, e.g. with the conventional L-BFGS optimizer.

Specific methods for reaction path determination at the QM/MM level have also been suggested. They are mostly based on constrained optimizations or local reaction path following, performed in parallel on a chain of structures, which are initially obtained by linear interpolation of reactant and product coordinates. Examples are the QM/MM adaptation of the nudged elastic band (35) and the path-optimization algorithm by Ayala and Schlegel (36, 37), or the combination of both (38). There is also a method to perform intrinsic reaction coordinate (IRC) calculation for a whole QM/MM system (39), which requires gradients and Hessian information both for the QM and MM regions.

With increasing system complexity, it becomes harder to describe chemical reactions by only finding a subset of stationary points and performing reaction path calculations to connect them on the potential energy surface. Extensive sampling of the phase space and evaluation of reaction free energies become necessary to "average" the contributions from multiple pathways, which may contribute to the investigated reaction. The resulting activation free energies may be used to obtain reaction constants and to compare the theoretical results to experimental data via transition state theory.

Efficient sampling of phase space by means of either MD or Monte Carlo (MC) calculations is a very demanding task for large QM/MM systems. Such applications are therefore normally done at the semiempirical QM/MM level. Even then, it is computationally very expensive to perform long samplings (of the order of ns) and MD runs are often restricted to tens of picoseconds. Therefore, specialised techniques for free energy QM/MM calculations were also designed to suit QM/MM systems better. We discuss those in Chapter 5.

The current thesis is fully devoted to the QM/MM approach. It is divided into two parts. In the first one (Chapter 2 and Chapter 3) we apply existing QM/MM techniques to unravel the details of the reaction mechanisms of two enzymes, Cyclohexanone Monooxygenase (CHMO) and Lipopolysaccharyl-α-1,4-galactosyltransferase C (LgtC), with the main focus on CHMO (Chapter 2). In the second part, we describe two new techniques that extend the existing arsenal of reaction path and free energy methods for QM/MM studies: the microiterative intrinsic reaction coordinate (Chapter 4) and dual Hamiltonian free energy perturbation methods (Chapter 5).

2

Cyclohexanone Monooxygenase (CHMO)

The Baeyer-Villiger reaction (40) has been known for more than 100 years, but continues to attract the attention of synthetic chemists (41). It is an oxidation reaction that transforms ketones into esters or lactones using stoichiometric amounts of oxidant, e.g. peracids. It proceeds in two steps. In the initial addition step, the carbonyl carbon atom of a ketone is attacked by the peroxide moiety of an oxidant, resulting in a tetrahedral Criegee intermediate. Usually the Criegee intermediate is neutral with the carbonyl oxygen being protonated, since the reaction is normally carried out in an acidic environment. However, base-catalyzed variants also exist, which involve an anionic Criegee intermediate (42). The addition step is followed by a concerted migration of one of the C-C σ bonds adjacent to the carbonyl carbon to the closest oxygen atom from the peroxy group, with cleavage of the O-O bond. Both the type of substrate and the reaction environment determine which of the two steps will be rate-determining. Despite the obvious benefits of the Baeyer-Villiger reaction, such as the retention of configuration at the migrating carbon and the large variety of carbonyl substrates prone to oxidation, its regio- and enantioselectivity are limited (43).

Therefore, Baeyer-Villiger monooxygenases (BVMO) have become an attractive alternative, which perform an O_2-driven enzymatic variant of the Baeyer-Villiger oxygenation (43, 44). These enzymes provide access to many valuable products (e.g. chiral lactones) with high enantioselectivity, using only stoichiometric amounts of oxygen and producing water as an ecologically clean byproduct. Many BVMOs have been

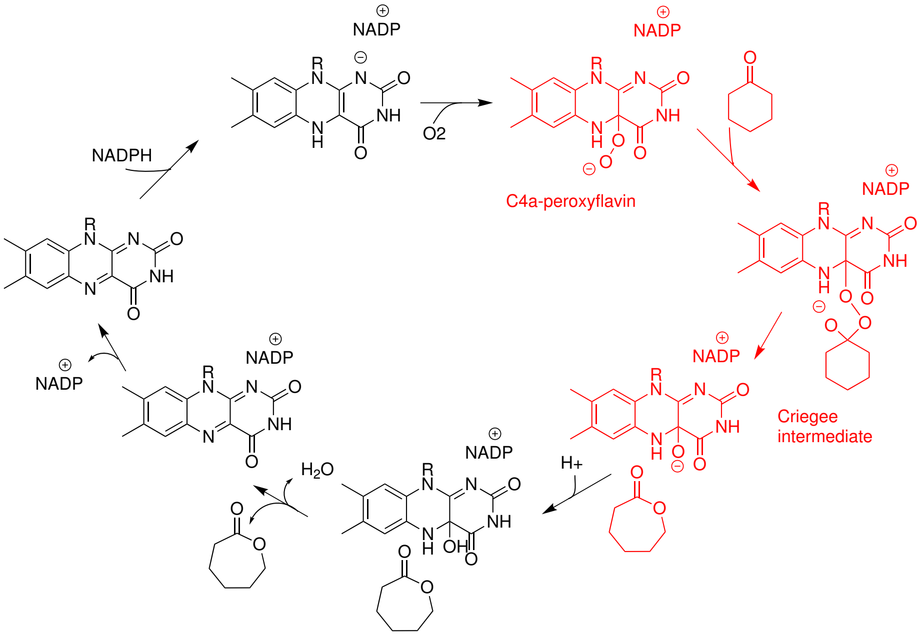


Figure 2.1: Reaction mechanism of CHMO. Shown in red are the steps that have been studied computationally.

characterised and used in synthetic organic chemistry, with cyclohexanone monooxygenase (CHMO) from *Acinetobacter* sp. NCIMB 9871 (45, 46, 47) and phenyl acetone monooxygenase (PAMO) from *Thermobifida fusca* (48) being two prominent examples.

An active BVMO accomodates two cofactors that play important roles in the catalytic cycle: flavin adenine dinucleotide (FAD) that is constantly bound to the enzyme, and a reduced form of the nicotinamide adenine dinucleotide phosphate (NADPH) that binds to the enzyme at the beginning and is released at the end of the reaction cycle. The reaction mechanism of BVMOs has been studied experimentally (49, 50, 51, 52) and its main steps are known (see Fig. 2.1). First, FAD is reduced by NADPH via hydride transfer. Thereafter it reacts with an oxygen molecule, forming a C4a-peroxyflavin intermediate (FADHOO$^-$). It was shown for CHMO (50) that in the absence of substrate, C4a-peroxyflavin slowly transforms into the unreactive C4a-hydroperoxyflavin. When substrate is present in the active site, the Baeyer-Villiger reaction takes place, and the consensus view is that it proceeds via the same two steps as its nonenzymatic analogue. It is not known, however, whether the reaction involves an anionic or neutral Criegee intermediate and whether the addition or migration step is rate-determining.

While the general reaction mechanism is established, the atomistic details of the enzymatic Baeyer-Villiger oxidation are still obscure. The high regio- and enantioselectivity of BVMOs towards a wide range of substrates are of great importance for synthetic chemistry, but their origins remain unclear up to know. Therefore, it is crucial to have structural information on the enzymatic active site during the reaction, which can be obtained by a combination of X-ray and computational techniques.

The first resolved X-ray structure of a BVMO was that of PAMO (53), but it contained only the FAD cofactor bound to the enzyme and lacked the NADP(H) cofactor and a substrate. The first resolved structure of a BVMO containing both the FAD and NADP$^+$ cofactors bound to the protein matrix was reported by Lau, Berghius et al. for CHMO from *Rhodococcus* sp. strain HI-31 (54). It also lacked a ligand or an inhibitor in its active site, but one of the two resolved structures (called a "closed" structure) was considered to reflect the state of reaction after FAD reduction and prior to the Baeyer-Villiger reaction. In this structure, there is a well-defined binding site in the direct proximity to the isoalloxazine ring of FAD, surrounded by hydrophobic residues, that can in principle accomodate a substrate of the size of cyclohexanone or even larger.

This crystal structure of CHMO is a perfect candidate for computational investigation of the BVMO reaction mechanism. While there is more experimental data on CHMO from *Acinetobacter* sp., the two CHMO enzymes share 55% sequence identity (with all residues surrounding the binding site being conserved) and they behave similarly with respect to different substrates concerning stereoselectivity (54). Therefore it seems justified to compare the computational results on the reaction mechanism of CHMO from *Rhodococcus* sp. with the experimental findings for CHMO from *Acinetobacter* sp.

2.1 QM/MM study on the rection mechanism of CHMO

We performed a full-dimensional QM/MM computational study (55) to elucidate the atomistic details of the CHMO reaction mechanism, from the substrate-enzyme complex up to the product. Cyclohexanone, a natural substrate of CHMO, was chosen as ligand.

The system setup and QM/MM methodology were the same throughout all our computation investigations of the CHMO mechanism, which are documented in three publications (55, 56, 57). In the following, we describe the computational procedure briefly; for further details the reader is referred to the corresponding papers.

Initial coordinates were taken from the crystal structure of the "closed" CHMO conformation described above (PDB code 3GWD). We modeled C4a-peroxyflavin based on the FAD geometry present in the X-ray structure. The whole system was solvated in a water sphere and neutralized by Mg^{2+} and Cl^{-} ions by random substitution of solvent water molecules. The solvated neutralized system was then relaxed and subjected to a 1 ns MD equilibration at the pure MM level using the CHARMM22 force field (58) as implemented in the CHARMM program (59, 60). One cyclohexanone molecule in the chair conformation was docked into the enzyme active site of a random snapshot taken from the classical MD trajectory and the resulting structure was used as starting point for the QM/MM calculations.

The QM region comprised the cyclohexanone, all atoms from the isoalloxazine ring of the $FADHOO^{-}$, the side chain of Arg329, and both the nicotineamide ring and adjacent ribose of $NADP^{+}$. It was relatively large and contained 96 atoms. It was treated at the DFT level using the B3LYP functional (4, 5, 6, 7, 8, 9) in combination with the TZVP basis set (61). The rest of the system was described classically with the CHARMM22 force field. The QM/MM calculations were done with the ChemShell package (62), interfacing the TURBOMOLE program (24) used to obtain the QM part energy and gradients, and the DL_POLY program (63) used to treat the MM part of the system. All atoms within 12 Å of the center of the QM system were free to move during optimizations (active region). The calculations employed the QM/MM additive scheme, electronic embedding for the QM/MM electrostatic interactions, and hydrogen link atoms with the charge shift approach at the QM/MM boundary (see Introduction).

In the following we will use the residue notation of CHMO from *Rhodococcus* sp., so that the residue numbers around the active site will differ by two from those of CHMO from *Acinetobacter* sp. For example, Arg329 in CHMO (*Rhodococcus*) corresponds to Arg327 in CHMO (*Acinetobacter*). We will apply the latter notation only when directly referring to that enzyme, e.g. when discussing the available experimental data.

We performed full reaction path calculations and optimized all the relevant stationary points, with the exception of the transition state for the addition step, which

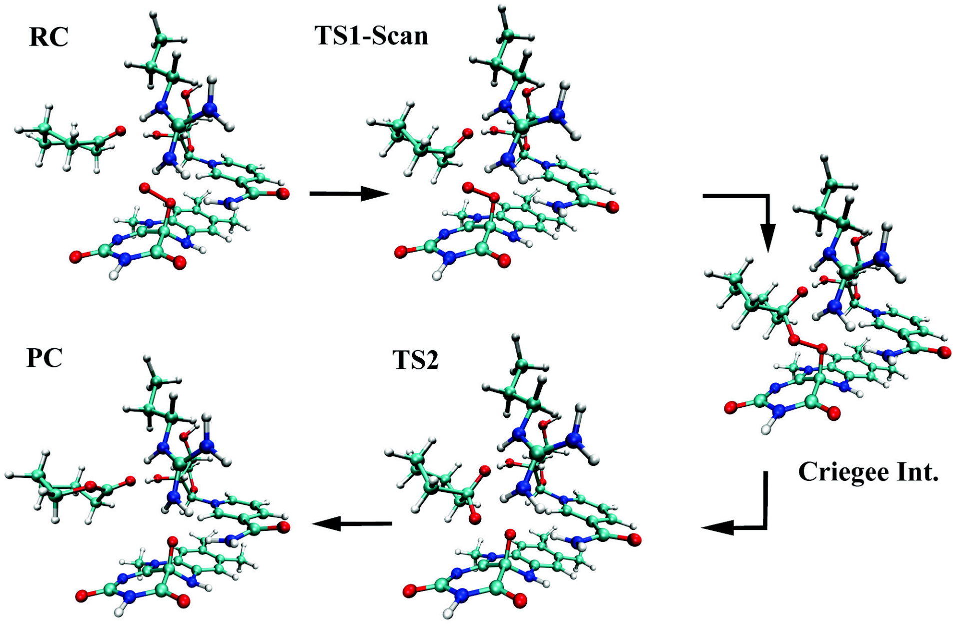


Figure 2.2: Reaction mechanism of CHMO according to the QM/MM calculations. Stationary points along the reaction path (QM region only) optimized at the QM(B3LYP/TZVP)/CHARMM level.

was shown to be kinetically unimportant compared to the transition state for the migration step, as discussed below (see Fig. 2.2). The optimization and reaction path calculations show that cyclohexanone is bound in a distinct orientation with respect to both cofactors and the side chain of Arg329, such that its carbonyl oxygen is oriented towards the hydrogen atoms of Arg329 and of the NADP$^+$ ribose hydroxyl group. This orientation allows the addition step to proceed via a rather small barrier of 8.7 kcal/mol (at the QM(B3LYP/TZVP)/CHARMM level). The resulting Criegee intermediate has a rigid, well-defined structure, occupying a very shallow minimum on the potential surface (7.8 kcal/mol above the reactant state). It should be noted, however, that single-point calculations, performed along the reaction path of the addition step with two QM methods that include mid-range dispersion (B3LYP-D (11) and M06-2X (14)), yield lower relative energies for the transition state and the Criegee intermediate. In both cases the addition step proceeds in a nearly barrierless fashion, and the Criegee intermediate appears to be more stable than the reactant state.

The Criegee intermediate has a pronounced anionic character. Its carbonyl oxygen forms two very short H-bonds with the hydrogen atoms of Arg329 and of the NADP$^+$ ribose group, towards which it is oriented in the reactant configuration. Our calculations thus suggest that the enzymatic Baeyer-Villiger reaction proceeds via an anionic Criegee intermediate and that the two strong H-bonds formed by the carbonyl oxygen make the carbonyl carbon sufficiently electrophilic for the addition reaction to take place. This also confirms the crucial role of Arg329 and NADP$^+$ cofactor in the catalytic activity of BVMOs suggested experimentally (51).

The optimized Criegee intermediate has an asymmetric conformation, in which one of the two C-C bonds adjacent to the carbonyl group is strictly antiperiplanar to the C4a-peroxyflavin O-O peroxy bond (i.e., the C-C-O-O dihedral angle is close to 180°). In accordance with the traditional stereoelectronic requirement of antiperiplanarity (64), our calculations predict that the migration step can proceed only in one direction: the distal oxygen of the peroxy group can be inserted into the cyclohexanone ring only at the C-C bond antiperiplanar to the peroxy bond. The corresponding transition state was optimized and found to be 14.5 kcal/mol above the reactant state at the QM(B3LYP/TZVP)/CHARMM level. The migration step is therefore rate-determining. The formed product, ϵ-caprolactone, lies 69.9 kcal/mol below the reactant state, indicating that the enzymatic Baeyer-Villiger reaction is highly exothermic, just as its non-enzymatic analogue (65). Accounting for dispersion interactions through single-point calculations did not change the qualitative character of the migration step, which remained rate-determining, although the barrier height varied depending on the method used.

2.2 Unveiling the source of enantioselectivity of CHMO

Summarizing the previous section, we have modeled part of the CHMO catalytic cycle, namely the Baeyer-Villiger oxidation of the cyclohexanone substrate. There is one distinct binding mode, in which cyclohexanone may undergo the addition step in its chair conformation, which leads to formation of a Criegee intermediate in a strictly defined conformation. In this intermediate, only one of the two σ-bonds adjacent to the cyclohexanone carbonyl group fulfills the stereoelectronic requirement of antiperiplanarity and can therefore migrate to the oxygen from the C4a-peroxyflavin peroxy

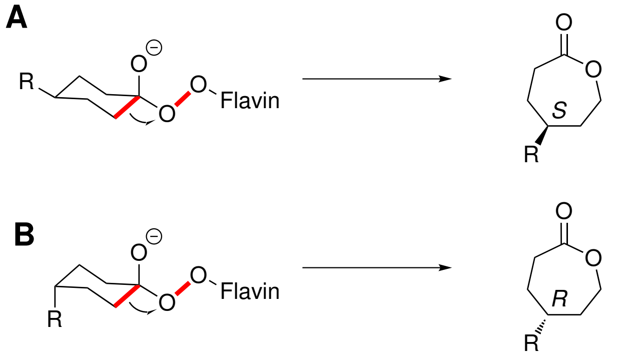


Figure 2.3: CHMO enantioselectivity towards 4-substituted cyclohexanone. According to the stereoelectronic requirements, only the bond antiperiplanar to the O-O bond may migrate in the Criegee intermediate (shown in red). (A) A substituent in equatorial position leads to an (*S*)-lactone. (B) A substituent in axial position leads to an (*R*)-lactone.

group. Armed with this model, we approached the problem of explaining the source of the CHMO enantioselectivity toward diverse substrates.

There have been many speculations about the source of enantioselectivity in BVMO-catalyzed reactions (66, 67, 68, 69, 70, 71). A common basis of all the attempts for explaining the observed enantioselectivity was the stereoelectronic requirement of antiperiplanarity in the Criegee intermediate fragmentation step (see above). Based on this notion and empirical data on the asymmetric transformations of numerous cyclohexanone derivatives catalysed by BVMOs, several models were proposed. The most prominent one is the ”diamond lattice” model (66, 70, 71, 72). It is based on the assumption that the substrate enters the reaction in a chair conformation, in which the substituents occupy only allowed positions (usually equatorial). This model proved to be a useful device in predicting the outcome of several catalyzed reactions (66).

As a representative case of CHMO enantioselectivity, we considered the desymmetrisation of 4-methylcyclohexanone. CHMO from *Acinetobacter* sp. (as well as CHMO from *Rhodococcus* sp.) reacts with the 4-methylcyclohexanone with nearly perfect (*S*)-enantioselectivity, producing (*S*)-4-methyl-ϵ-caprolactone with more than 96% enantiomeric excess (ee). In the gas phase, the 4-methyl group prefers the equatorial position at the cyclohexanone ring, having a chair conformation. If we assume that it retains the same conformation in the active site of CHMO, with the whole ligand adopting an orientation as described above, the resulting structure of the Criegee intermediate immediately suggests formation of the (S)-lactone product due to selective migration of one of the two carbon atoms adjacent to the carbonyl group. A methyl substituent in axial orientation would give rise to the (R)-lactone product, which is not observed in experiment. This principle is illustrated in Fig. 2.3 for a general case.

To verify the prediction made, we reoptimized the Criegee intermediate and transition state for migration, after having manually introduced the methyl group at the cyclohexanone ring (both in equatorial and axial position - four structures altogether). The results confirmed our prediction: the energies of the intermediate and the transition state were lower when the 4-methyl group was in an equatorial rather than axial position at the cyclohexanone ring, by 4.8 and 4.5 kcal/mol, respectively. The actual energy barriers for the migration step were quite similar in the two conformers, but the equatorial pathway is favored on an absolute scale at the rate-determining transition state. Since an isolated 4-methylcyclohexanone molecule with an equatorial 4-methyl group is lower in energy by 1.9 kcal/mol relative to its axial counterpart (calculated at the B3LYP/TZVP level, close to the experimental enthalpy difference of 2.1 kcal/mol between the two conformers (73)), we find an increased preference towards the equatorial conformation when the ligand is in the active site of CHMO.

Another enantioselective reaction in the active site of CHMO was considered in a separate paper (56). Here, wild type (WT) CHMO enantioselectivity towards the E-isomer of the product of 4-ethylidenecyclohexanone desymmetrisation was explained in terms of our computational model. In this case, there are again two poses that the substituent can adopt at the cyclohexanone ring of the Criegee intermediate. In the first conformation that would lead to the Z-product, the 4-ethylidene group points towards a close-lying residue of protein backbone, and its optimization at the QM/MM level leads to significant displacement from the initially guessed geometry (taken from our previous work (55)). The resulting optimized Criegee intermediate was found to be higher in energy than the alternative conformer that would lead to an E-product, in which the substituent is pointing into the relatively empty space in the center of the binding site, by 2.3 kcal/mol. This is in a good agreement with experimental results, which show a strong preference towards the E-isomer with $> 99\%$ ee.

2.3 Effects of mutations on CHMO enantioselectivity.

The effect of point mutations on BVMO enantioselectivity is of great interest to synthetic chemists. Substrate acceptance and enantioselectivity is limited in the WT BVMOs. In case of CHMO, if substituents at the cyclohexanone ring are varied too much, enantioselectivity may decrease, or even get reversed (72). In an attempt to overcome natural limitations of enzymatic catalysis, directed evolution is nowadays frequently applied to achieve the desired enantioselectivity towards selected substrates by means of point mutations in the amino acid chain of an enzyme (74). Detailed knowledge of the reaction mechanism, and the reasons why certain mutations affect the outcome of a reaction, may help to make directed evolution more efficient. In our next paper (57), we therefore tried to explain the effect of two different mutations at the site of a single amino acid residue, which cause notable changes of the observed CHMO enantioselectivity towards 4-hydroxycyclohexanone.

Experimentally (72), the WT CHMO (*Acinetobacter*) exhibits rather low enantioselectivity towards this ligand (9% (*R*)), the Phe432Ser mutant reverses enantiopreference to 79% (*S*), and the Phe432Ile mutant improves the WT enzyme enantioselectivity up to 49% (*R*). According to Kayser and Clouthier (72), 4-hydroxycyclohexanone in the gas phase slightly favors the hydroxyl group in axial rather than equatorial position. They assumed that this is the cause of the observed weak (*R*)-enantioselectivity of the WT CHMO. The drastic change of enantioselectivity towards 4-hydroxycyclohexanone in the Phe432Ser mutant was attributed to a hypothetical H-bond between Ser432 and the substrate hydroxyl group in equatorial position (see Fig. 2.4). The Phe432Ile mutation effect was not discussed.

These qualitative explanations are consistent with our computational model, which translates the preferred substituent position at the cyclohexanone ring into the enantiomeric outcome of the reaction. In WT CHMO (*Rhodococcus*), Phe434 is indeed situated next to the equatorial side of the C4 atom of cyclohexanone, and hence the serine side chain of the Phe434Ser mutant may form an H-bond with the hydroxyl substituent. In case of the Phe434Ile mutant, one may expect steric effects rather than specific interactions. In an attempt to find precisely the reasons for the remarkable changes in the CHMO enantioselectivity, we performed full reaction path calculations and optimized all relevant stationary points for the WT enzyme as well as for both

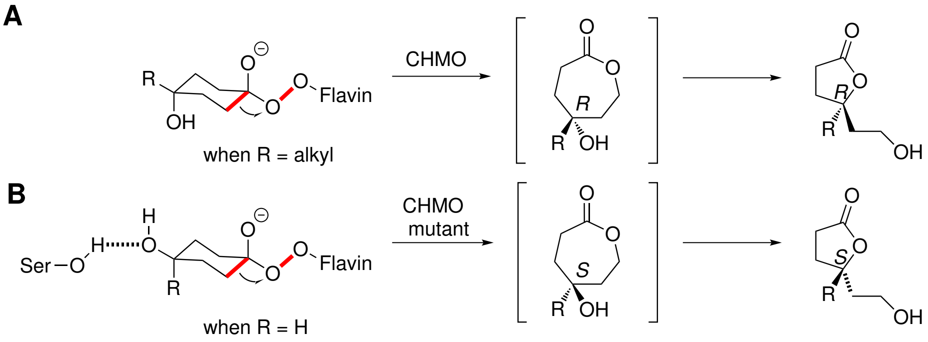
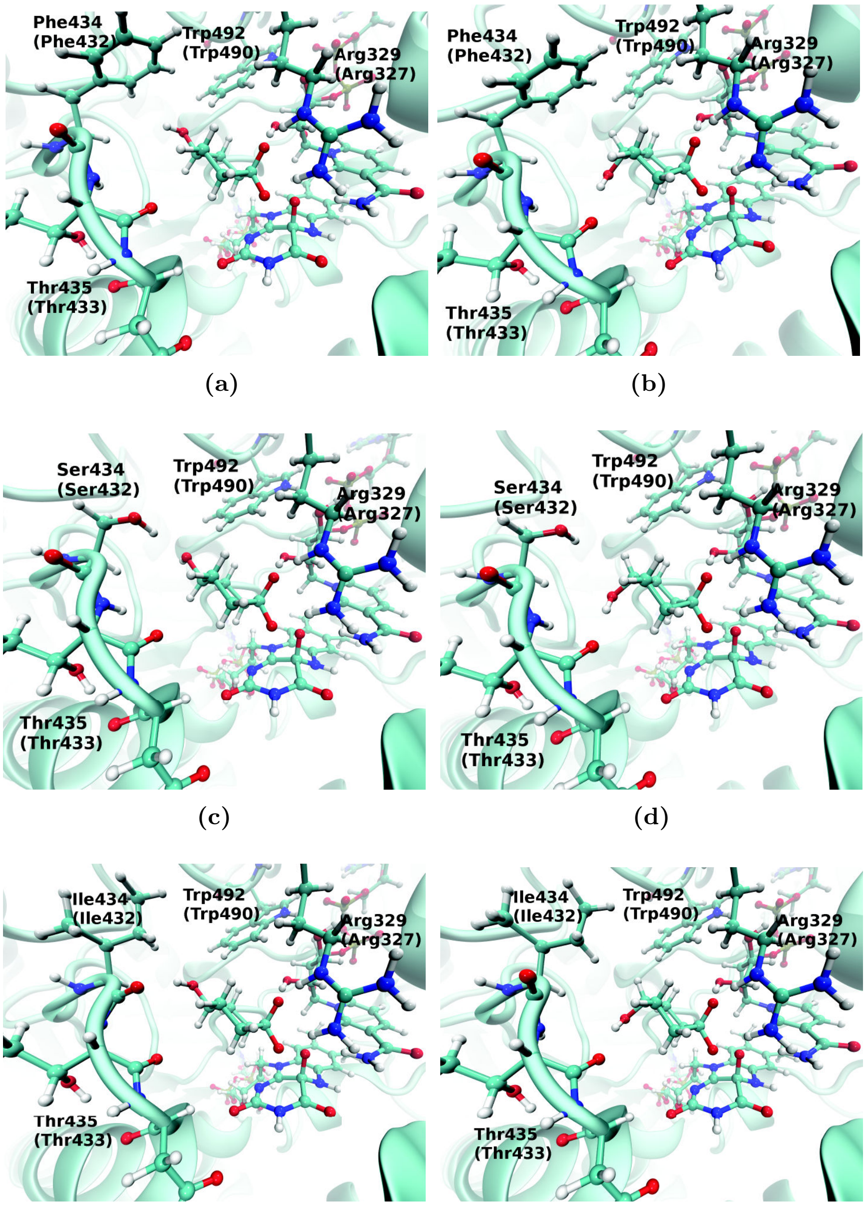


Figure 2.4: CHMO enantioselectivity towards 4-hydroxycyclohexanone. (A) In wild-type CHMO, the hydroxyl group prefers the axial position, which imposes the (*R*) configuration on the lactone product. (B) In the Phe432Ser mutant, an H-bond was proposed (72) between Ser432 and the 4-hydroxy group, provided that the latter occupies an equatorial position, imposing the (*S*) configuration on the lactone product. Subsequent intramolecular translactonization to the thermodynamically more stable 5-membered lactones occurs in both cases.

Phe434Ser and Phe434Ile mutants at the QM(B3LYP/TZVP)/CHARMM level. We then used the difference in the QM/MM energies of the rate-limiting transition state, calculated for both the equatorial and axial orientations of the hydroxyl group (from here on called TS2-eq and TS2-ax, respectively), to estimate the ee value for each mutant.

The results obtained (see Fig. 2.5) indicate a more complex picture of interactions within the active cite of CHMO, than was initially anticipated. In the WT enzyme and in both mutants, the substrate hydroxyl group in axial orientation forms an H-bond with the Thr435 carbonyl oxygen, which is preserved along the whole reaction path. In the WT enzyme TS2-eq, the hydroxyl group in equatorial position is oriented towards the benzene ring of the Phe434 side chain, creating a π-hydroxyl H-bond. This interaction, together with a very weak H-bond between the hydroxyl group oxygen atom and a rather distant Trp492 ring hydrogen atom, nearly counterbalances the strong H-bond that is formed in the case of axial configuration, bringing the energy of TS2-eq only 0.5 kcal/mol above the TS2-ax energy. In the Phe434Ser mutant, as anticipated, the oxygen atom from the substrate equatorial hydroxyl group forms an H-bond with the hydrogen atom of the Ser434 side chain hydroxyl group, and this interaction is preserved along the whole reaction path. The best estimate for the energy difference



(a) (b)

(c) (d)

(e) (f)

Figure 2.5: CHMO reaction with 4-hydroxycyclohexanone: transition state for the migration step. Active sites of the wild type CHMO (*Rhodococcus*) and its two mutants, with the hydroxyl group being either in equatorial or axial position. The residue numbers in parentheses label the analogous residues in CHMO from *Acinetobacter* sp. NCIMB 9871. (a) Wild-type CHMO. Hydroxyl group in equatorial position. (b) Wild-type CHMO. Hydroxyl group in axial position. (c) Phe434Ser mutant of CHMO. Hydroxyl group in equatorial position. (d) Phe434Ser mutant of CHMO. Hydroxyl group in axial position. (e) Phe434Ile mutant of CHMO. Hydroxyl group in equatorial position. (f) Phe434Ser mutant of CHMO. Hydroxyl group in axial position.

between the corresponding TS2-eq and TS2-ax was 1.7 kcal/mol in favor of the equatorial configuration, which corresponds to 89% ee and is in reasonable agreement with the experimental value of 79% ee. In the Phe434Ile mutant, the hydroxyl group of the ligand cannot form any specific interaction with the side chain of Ile434; instead it forms an H-bond with the Ile434 backbone carbonyl oxygen in TS2-eq. This becomes possible because the Ile434 backbone undergoes some rotation, which is not observed in the WT enzyme and the Phe434Ser mutant, but costs some energy. The resulting energy difference between TS2-eq and TS2-ax is 1.7 kcal/mol in favor of the axial conformation, predicting (R)-enantioselectivity with 89% ee, compared to the experimental value of 49% ee. The experimental and computed ee values and the associated energy differences are summarized in Table 2.1.

Table 2.1: Enantiomeric excess of the CHMO-catalysed desymmetrization of 4-methylcyclohexanone as determined from experiment and calculations (QM(B3LYP/TZVP)/CHARMM) for the WT enzyme and two mutants. The corresponding energy differences between the rate-determining transition states are given in parentheses in kcal/mol.

Type of mutant	Type of enantioselectivity	Experiment	Theory
WT	R	9% (0.1)	40% (0.5)
Phe432Ser	S	79% (-1.3)	89% (-1.7)
Phe432Ile	R	49% (0.6)	89% (1.7)

We should note that in order to obtain meaningful results, we had to enlarge the default QM region (see Section 2.1) such that the residues involved in the formation of H-bonds with the substrate were included. Otherwise, especially when there is an H-bond formed between the MM oxygen atom and substrate hydroxyl group, we encountered overpolarization of the QM region by a closely-lying MM charge, which caused an artificial decrease of both the H-bond length and the QM energy. This issue reflects a problem that QM/MM calculations may face when employing electrostatic embedding, as discussed in Chapter 1.

The results of our QM/MM calculations are qualitatively and semi-quantitatively in line with experimental data on CHMO enantiopreference towards 4-hydroxycyclohexanone. We are therefore in a position to provide an explanation for the observed stereochemical outcome of the reaction on the molecular level. The WT CHMO enzyme exhibits a slight (*R*)-enantioselectivity towards 4-hydroxycyclohexanone, because the Thr435-substrate (axial) H-bond stabilizes the TS2 more than the π-hydroxyl interaction between Phe434 and the substrate equatorial hydroxyl group. Enantiopreference of CHMO is reversed in the Phe434Ser mutant, because the equatorial substrate-Ser434 H-bond turns out to be stronger than the axial substrate-Thr435 H-bond. Finally, in the Phe434Ile mutant, the WT enantiopreference is enhanced, because the equatorial substrate-Ile434 interaction is weaker than in the case of the WT enzyme, and can thus not compete effectively with the axial substrate-Thr435 H-bond.

Our results show that CHMO enantioselectivity towards substrates with polar substituents depends on a fine interplay of the weak interactions between the substituent and the protein environment. This makes it difficult to predict the outcome of reactions with an arbitrary substrate, without performing explicit calculations at the QM/MM level. The situation is expected to be simpler for substrates containing only non-polar substituents. In such cases (as with 4-methyl- or 4-ethylidenecyclohexanone), steric effects will dominate over specific electronic effects, and one should be able to explain the origin of enantioselectivity by checking the difference in energy between substrate conformers, as determined in the gas phase, and by considering the generic structure of the migration transition state (or even just the Criegee intermediate) to identify possible steric clashes between non-polar substituents and active-site residues.

3

Lipopolysaccharyl-α-1,4-galactosyltransferase C

Lipopolysaccharyl-α-1,4-galactosyltransferase C (LgtC) from *N. meningitidis* is a member of one of the glycosyltransferase (GT) families (75). GTs are a large group of enzymes, responsible for the biosynthesis of glycosidic bonds, which is accomplished by transferring monosaccharides from donors (usually nucleotide sugars) to different acceptors, such as other saccharides, lipids, proteins or DNA (76). LgtC catalyzes transfer of the α-galactose from uridine 5′-diphospho-α-galactose (UDP-Gal) to a galactose from the terminal lactose (LAT) of the bacterial lipooligosaccharide (LOS), resulting in its elongation. These LOSs cause the great virulence of *N. meningitidis*, since they are able to mimic human glycolipids at the surface of bacteria and thus avoid recognition by the human immune system (77). The reaction proceeds with overall retention of stereochemistry at the anomeric carbon atom, and therefore LgtC is attributed to the retaining GTs, the catalytic mechanism of which is poorly understood up to now (78, 79, 80).

Three different mechanisms were proposed for the retaining GTs (see Fig. 3.1). In the first one, the double-displacement mechanism, the reaction involves formation and subsequent cleavage of a covalent glycosyl-enzyme (CGE) intermediate (79, 80, 81). This scheme requires the presence of an appropriately positioned nucleophilic residue within the active site. The X-ray structure of LgtC (82), however, only contains a glutamine (Gln189) as suitably positioned active-site residue for a nucleophilic attack on the anomeric carbon. The nucleophilic character of glutamine is expected to be rather

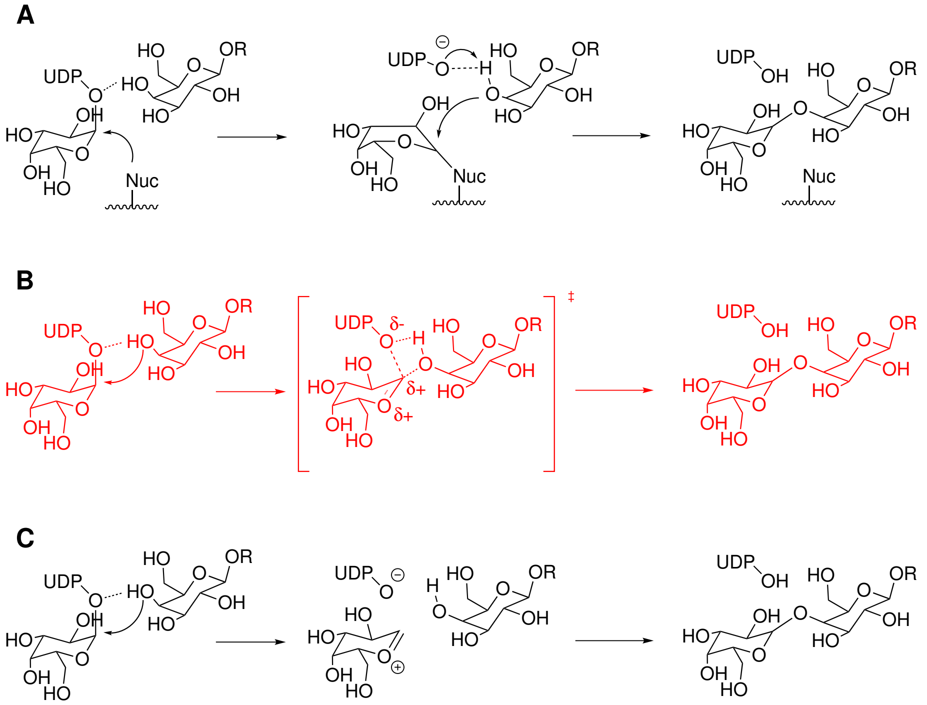


Figure 3.1: Proposed mechanisms for retaining glycosyltransferases. (A) Double-displacement mechanism, via a covalently bound glycosyl-enzyme intermediate. (B) Concerted front-side single displacement (S_Ni) mechanism with an oxocarbenium-like transition state. (C) S_Ni-like mechanism with a short-lived oxocarbenium-phosphate ion pair intermediate.

poor, and taking into account experimental evidence that both Q189A and Q189E LgtC mutants display 3% residual activity (83), the double-displacement mechanism is unlikely to operate in this enzyme. The alternative mechanisms proceed as a front-side attack of the LAT oxygen atom on the anomeric carbon with the formation of an oxocarbenium ion-like transition state (S_Ni mechanism) (84) or an oxocarbenium-phosphate ion pair intermediate (S_Ni-like mechanism) (80).

The LgtC reaction mechanism was previously studied with a reduced cluster model of 136 atoms, including five protein residues (85). This study confirmed the S_Ni mechanism of LgtC. However, many important interactions within the enzyme-substrate complex were neglected in this cluster model, which may prevent full understanding of the role of the enzyme in the reaction.

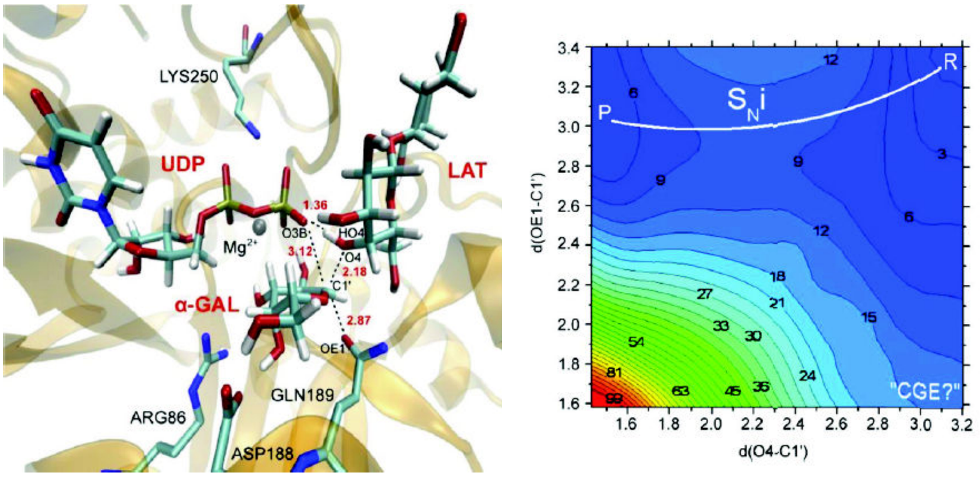
In our paper (86), we presented a full-enzyme QM/MM study of the LgtC catalytic mechanism. The chosen methodology is generally the same as described in Chapter 2, so we omit the details here. The major difference is that we used the B3LYP functional for geometry optimizations in the case of CHMO, while here we employed the BP86 functional for the sake of computational efficiency, and performed single-point energy evaluations at the optimized stationary points with the B3LYP, B3LYP-D, and M05-2X functionals.

We investigated the three different mechanisms proposed for LgtC and showed that the dissociative S_Ni mechanism is most likely in the WT enzyme. The computed barriers of 11.8 kcal/mol at the B3LYP/TZVP level and 14.6 kcal/mol at the M05-2X/TZVP level are in reasonable agreement with experimental kinetic data for the free energy barrier of 16 kcal/mol, as derived from k_{cat} values of 14-34 s^{-1} at 303 K (82, 83). According to our results, the bond between the anomeric carbon (C1') and the UDP oxygen atom (O3B) breaks early in the course of reaction; thereafter the O3B-HO4 bond (involving the hydrogen atom leaving LAT) and the C-O glycosidic bond form simultaneously, immediately after the concerted transition state (see Fig. 3.2a). We thus find that the leaving UDP group acts as a catalyst, as was suggested before (85).

Attempts to locate the covalent glycosyl-enzyme complex in the double displacement mechanism failed. To gain more insight, we performed a two-dimensional potential energy scan, varying the reaction coordinate, represented by the difference of the breaking C1'-O3B bond and the forming glycosidic bond, at fixed values of the distance between the anomeric carbon and Gln189 side chain oxygen atom. The resulting surface (see Fig. 3.2b) indicates a clear S_Ni reaction path as described above, along which the distance between the anomeric carbon and Gln189 changes only slightly, while there is no minimum close to the covalent-bond distance between these two atoms.

On the other hand, in the Q189E mutant, we observed immediate formation of the CGE complex with a covalent bond between the anomeric carbon and Glu189 oxygen atom. However, the following S_N2 attack of LAT at the CGE has a very high activation barrier (more than 30 kcal/mol) that would make this reaction very slow. These findings indicate that mutations in the active site of LgtC, and of retaining GTs in general, may switch the reaction mechanism from S_Ni to a double displacement model, or vice versa.

We have also analysed several other factors that play a role in the reaction, especially the enzyme-substrate and substrate-substrate interactions, and the results obtained



(a) **(b)**

Figure 3.2: Details of the LgtC reaction mechanism revealed by QM/MM calculations. (a) QM(BP86/SVP)/CHARMM optimized transition state of the LgtC reaction following the S_Ni mechanism. (b) Two-dimensional QM(BP86/SVP)/CHARMM potential energy surface. Energies are given in kcal/mol and distances in Å. Contour lines are drawn in intervals of 3 kcal/mol.

indicate that the rate of reaction, and even the mechanistic scenario followed by each enzyme-substrate complex, may be influenced by many factors, including the nature of substrates and their relative orientation. One can therefore not expect that all retaining GTs share the same mechanism, and case-by-case studies are necessary to gain full understanding of these catalytic mechanisms.

4

Microiterative Intrinsic Reaction Coordinate (IRC)

Reaction path determination is an important aspect of theoretical and computational chemistry. The most appropriate way to determine a reaction path is to follow the intrinsic reaction coordinate (IRC), as proposed by Fukui in 1970 (87, 88). According to his definition, the IRC is the steepest-descent pathway in mass-weighted coordinates that proceeds from a transition state in two opposite directions and ends in the two associated local minima. In quantum chemistry, it has become a routine procedure to perform IRC path following for small and medium-size molecules, and a number of different methods have been developed for this purpose (89, 90, 91, 92, 93, 94, 95, 96, 97). However, IRC calculations for large QM/MM systems are usually avoided, as straightforward application of standard IRC techniques will be costly and impractical. Other indirect reaction-path techniques, usually based on the ”chain-of-replicas” approach, have been proposed and adapted specifically for QM/MM systems (35, 37, 38). They are usually based on constrained optimizations or local path following for a set of constrained intermediate coordinates, initially obtained by interpolation of reactant and product structures. Otherwise, if the aim is to ensure that an optimized transition state is connected via a smooth path to the two relevant minima, careful energy minimizations are usually performed, starting from two structures that are obtained by perturbing transition state coordinates along the transition mode in both directions. There exists one IRC implementation at the QM/MM level in the Gaussian program (98), which is used in combiation with the ONIOM approach (99). In this implementation (39), the

IRC is computed for the whole system using gradient and curvature information for both the QM and MM regions.

We have suggested an alternative method called microiterative IRC (100). Here, only a subset of QM atoms (the core region) follows the IRC equation, while all remaining active degrees of freedom are subject to minimization after every IRC step. This approach is derived from the microiterative transition state optimization method (32). In this and few other papers (101, 102), the possibility of performing microiterative IRC computation was mentioned, but no further details were given. We have implemented the microiterative IRC method, investigated its performance in detail, and carefully checked its validity for a number of test systems.

4.1 Microiterative IRC formalism

The IRC equation has the following form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}s} = -\frac{\mathbf{g}(\mathbf{x})}{|\mathbf{g}(\mathbf{x})|}, \tag{4.1}$$

where $\mathbf{x}$ denotes the mass-weighted Cartesian coordinates of the nuclei, s is the arc length along the IRC defined by $\mathrm{d}s^2 = \sum_{i=1}^{3N} \mathrm{d}x_i^2$, and $\mathbf{g}$ is the mass-weighted gradient at $\mathbf{x}$. In order to follow the IRC, Eq. 4.1 is integrated in small steps, from a transition state to the nearest local minima. Below we review the most popular IRC methods, three of which were implemented in our work.

The simplest way to solve Eq. 4.1 is to perform Euler integration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \Delta s \frac{\mathbf{g}(\mathbf{x}_k)}{|\mathbf{g}(\mathbf{x}_k)|}, \tag{4.2}$$

so that only information about the gradient at the previous point along the path is used to make the next step. However, simple Euler integration is too inaccurate and would require very small steps to converge. Therefore, several modifications of the Euler method were suggested, among which we mention only the most prominent approaches. In the Ishida-Morokuma-Komornicki (IMK) stabilized Euler method (89), first an IRC step is done according to Eq. 4.2, followed by a linear search for the energy minimum along the bisector of the gradient at the current and previous point, in order to correct the Euler step. The Gonzalez and Schlegel method (92, 93, 94) also corrects the simple

Euler step; it is based on a constrained optimization on a hypersurface, performed after applying half of the Euler step.

The local quadratic approximation (LQA) method employs second-order energy derivative information and requires reformulation of Eq. 4.1. Following Page and McIver (90), it is convenient to express the IRC as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{g}(\mathbf{x}). \tag{4.3}$$

As initially shown by Pechukas (103), dt is an arbitrary parameter satisfying

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \sqrt{\frac{\mathrm{d}\mathbf{x}^{\dagger}}{\mathrm{d}t}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}}. \tag{4.4}$$

In LQA approximation, the gradient becomes

$$\mathbf{g}(\mathbf{x}) = \mathbf{g}_0 + \mathbf{H}_0(\mathbf{x} - \mathbf{x}_0), \tag{4.5}$$

with $\mathbf{H}_0$ being the Hessian matrix of second energy derivatives, and the steepest descent path can be obtained by integrating

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\mathbf{g}_0 - \mathbf{H}_0(\mathbf{x} - \mathbf{x}_0). \tag{4.6}$$

This can be transformed into

$$\mathbf{x}_{k+1} = \mathbf{x}_k + A(t)\mathbf{g}(\mathbf{x}_k), \tag{4.7}$$

where

$$A(t) = U_k \alpha(t) U_k^{\dagger}, \tag{4.8}$$

with U_k being the matrix of column eigenvectors of the Hessian ($\mathbf{H}_k$), and $\alpha(t)$ being a diagonal matrix with elements

$$\alpha_{ii}(t) = (e^{-\lambda_{ii}t} - 1)/\lambda_{ii}. \tag{4.9}$$

The parameter t can be obtained by numerical integration of an expression that can be derived from Eq. 4.4:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \sqrt{\sum_i \mathbf{g}_i^{'}(\mathbf{x}_k)^2 e^{-2\lambda_{ii}t}}, \tag{4.10}$$

where $\mathbf{g}^{'}(\mathbf{x}_k) = U_k^{\dagger}\mathbf{g}(\mathbf{x}_k)$. Due to the use of curvature information, the LQA method is accurate and efficient (it needs only one energy and one gradient calculation per step), and it can normally be used with larger steps than the Euler methods.

The Hessian and Euler based predictor-corrector (HPC and EulerPC) methods (95, 96, 97) use either an LQA or an Euler method as a predictor step, and a modified Bulirsch-Stoer integrator on a fitted distance-weighted interpolant (DWI) surface as a corrector step. First, a predictor step is performed. Then the energy, gradients and Hessian are evaluated at the resulting coordinates. By interpolating energy and gradients from the previous and current points along the reaction path, the IRC equation is integrated with the Euler method starting from the previous point N times, with the step size equal to $\frac{\Delta s}{N}$. Integration is performed several times with N growing up to an arbitrarily chosen number. Then, a polynomial extrapolation to a step size of 0 (which corresponds to $N = \infty$) is used for the computed sets of coordinates, to get the corrected coordinates for the current IRC step.

The microiterative IRC methodology follows the philosophy of the microiterative transition state search (32) that was previously implemented in the HDLCopt program (31), which is a module within the ChemShell package (62). In both methods, the system is partitioned into the reaction core and the outer regions. While in the transition state search the core region follows the second order P-RFO (partitioned rational function optimizer) optimization algorithm (33, 34) towards a transition state, it undergoes steps along the steepest descent path in the microiterative IRC method, according to the chosen IRC integrator. The outer region is in both cases minimized using the L-BFGS algorithm (29, 30). The calculation is performed via micro- and macroiterations, such that every single step for the core region is followed by a full relaxation of environment.

We have implemented the microiterative IRC method into the HDLCopt module of Chemshell. The first step starts from the optimized transition state and is taken along the imaginary frequency mode eigenvector (90), which is used instead of the gradient, regardless of the IRC method. As will be discussed later, the convergence criterion for the outer-region optimization may play an important role for some systems, and should thus be chosen carefully. In order to make the calculations more efficient, the Hessian matrix, required by some of the IRC methods, can be obtained through Hessian updates, e.g. those proposed by Powell (104) or Bofill (105), which use gradient information at the current point as well as the gradient and Hessian at the previous point. The use of Hessian updates was previously shown to be accurate enough for

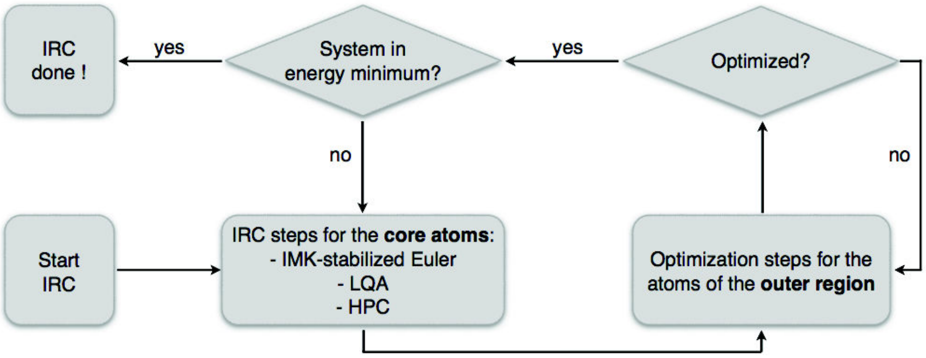


Figure 4.1: Scheme of the microiterative IRC procedure, as implemented in the HDLCopt program.

IRC calculations (106). The IRC steps in the core region atoms are always performed in Cartesian coordinates, while the outer region can be optimized in hybrid delocalized internal coordinates (31).

Among the methods described above, we have implemented the IMK-stabilized Euler, LQA, and HPC approaches into the HDLCopt program. The IMK-stabilized Euler method requires from three to seven energy and two gradient calculations per step, and therefore is the least efficient IRC approach implemented. Nevertheless, it is the simplest way of integrating Eq. 4.1, and with small steps it is supposed to work for any system. LQA is a much more accurate and efficient method, and is a perfect choice for the microiterative scheme, especially when making use of the Hessian updates. The HPC method should in principle be even more beneficial, since it corrects the LQA step using the same number of energy and gradient calculations (one per step). We expected the HPC method to be especially efficient in the microiterative IRC approach, since we perform the correction step *after* the outer region has been optimized, thereby decreasing the decoupling between the two regions. For the chosen test systems, however, we did not observe significant differences in HPC and LQA performance. Therefore, we generally applied the simpler LQA approach in our applications.

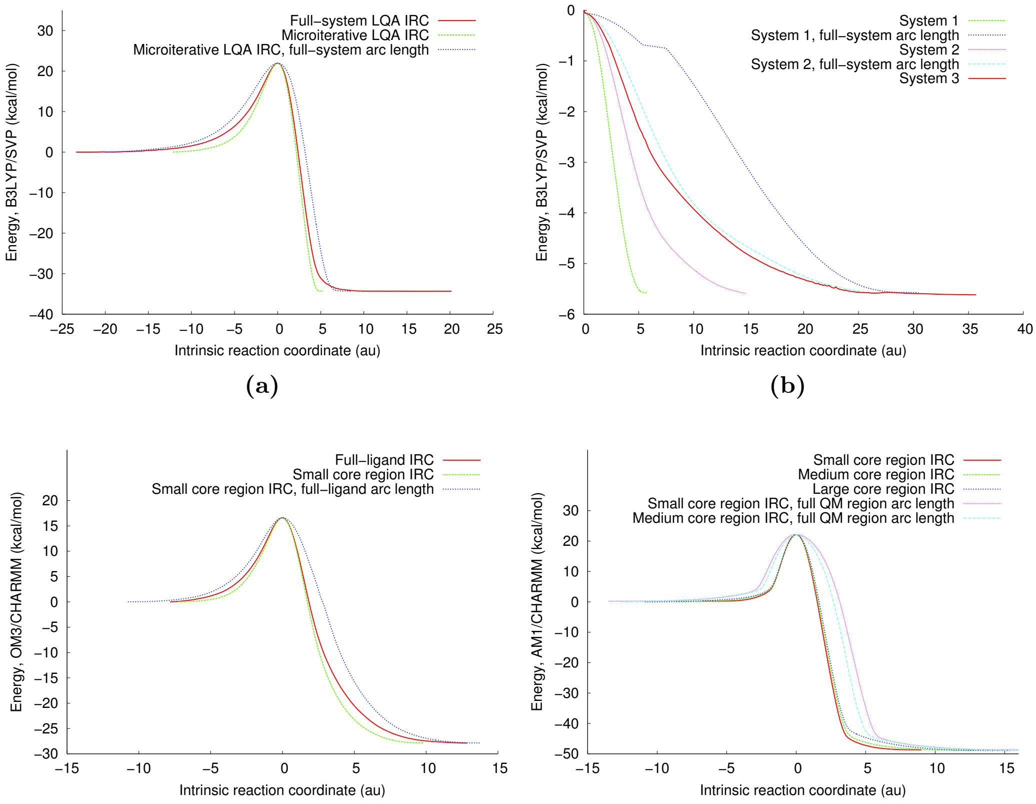
4.2 Microiterative IRC performance

We assessed the benefits and limitations of the implemented algorithms with four different test systems of varying size and complexity. One can not directly compare two microiterative IRC curves with different number of atoms used in the core region, or a microiterative IRC curve with a full-system IRC curve. In our work, the differences between the IRC pathways for the same system were evaluated qualitatively, by comparing the arc length computed for the same number of atoms, and quantitatively, by calculating the root-mean-square (RMS) deviations between geometries along the corresponding IRC paths. Due to common differences in the number of steps required to complete e.g. microiterative and full-system IRC calculations, geometries with the closest energy values along the reaction paths were compared.

We first tested the performance of the three implemented microiterative IRC methods by comparisons with the full-system IRC and with respect to each other. The test case was the Diels-Alder cycloaddition reaction between 2,4-hexadiene and ethene in the gas phase.

In full-system IRC calculations, the LQA and HPC methods behaved equally well, and for IRC step sizes from 0.05 to 0.15 $\sqrt{\text{amu}}$ bohr, the shape of the IRC curve did not change, indicating that the latter step size is accurate enough. The results of the IMK-stabilized Euler method, on the contrary, strongly depended on the step size, and converged only at a very small step size of 0.01 $\sqrt{\text{amu}}$ bohr. For microiterative IRC calculations, the core region consisted of the four carbon atoms involved in the bond breaking and bond making during the reaction. Here, all the three methods gave the nearly identical IRC curves, with converged results already at a step size of 0.15 $\sqrt{\text{amu}}$ bohr.

There are some minor differences between the microiterative and full-system reaction paths, regardless of the IRC method used (see Fig. 4.2a). Visual analysis reveals that the two methyl groups of the 2,4-hexadiene rotate at different rates in the microiterative and full-system setups, giving RMS deviations up to about 0.1 Å, which then decrease again as the two pathways approach the same reactant and product states. On the other hand, within the microiterative IRC framework, all IRC methods yield similar results, with negligible RMS deviations along the reaction pathway of the order of 10^{-3} to 10^{-2} Å. As all the implemented IRC methods performed equally well in



(a) **(b)**

(c) **(d)**

Figure 4.2: Microiterative IRC energy profiles with different arc length definitions. A step size of 0.15 $\sqrt{\text{amu}}$ bohr was used if not stated otherwise. (a) Comparison between full-system and microiterative IRC energy profiles for the Diels-Alder reaction. (b) Comparison between full-system (system **3**) and microiterative IRC energy profiles for a small (system **1**) and medium-size (system **2**) core regions for the internal rotation in 1,2-diphenylethane. LQA step sizes: 0.15 $\sqrt{\text{amu}}$ bohr for the full system and 0.05 $\sqrt{\text{amu}}$ bohr for the small and medium-size core regions. (c) Comparison between microiterative IRC energy profiles for the chorismate–prephenate conversion, catalyzed by BsCM, from the full-ligand and small-core IRC calculations. (e) Comparison between microiterative IRC energy profiles for the hydroxylation reaction conversion, catalyzed by PHBH, for the three different core regions

these tests, we used only LQA for the remaining calculations as the most efficient and straightforward IRC method.

To explore the limits of its performance, we have chosen the internal rotation of 1,2-diphenylethane as our second test. Due to the rigidity of phenyl rings and their steric interaction, the whole system participates in the rotation, and it should thus be difficult to define two uncoupled regions.

We selected two core regions: the first one contained only the four central carbon atoms, while the second one also contained the two adjacent carbon atoms from each phenyl ring. Both core regions resulted in smooth reaction pathways (see Fig. 4.2b) and gave the proper product, even though in the first case, the rotation of the benzene rings was only accounted for during the optimization steps. We note that the convergence criteria value for the outer-region optimization became crucial for the first system, and in order for the microiterative IRC calculation to produce a smooth curve, the default HDLCopt threshold for the maximum gradient component of 15×10^{-5} hartree/bohr had to be decreased by factors of 3 or even 9 (depending on the IRC step size).

As expected, the RMS deviations from the full-IRC geometries along the reaction were much smaller for the microiterative IRC geometries obtained with the large core region (compared with those for the small core region). In small-core case, the rotation of the phenyl rings happens too early during the optimizations of the outer region. Obviously, as soon as the two adjacent carbon atoms from the phenyl rings are also included in the large core region, the rotation of the rings starts to be accounted for during the IRC steps. The microiterative IRC method can thus be successfully applied even to complicated systems like 1,2-diphenylethane, provided that the proper core region is chosen.

Finally, we evaluated the performance of the microiterative IRC method in enzymatic QM/MM systems, for which the method is designed. As our test systems, we chose the conversion of chorismate to prephenate catalyzed by the chorismate mutase (BsCM) enzyme from *Bacillus subtilis*, and the hydroxylation reaction in the catalytic cycle of *p*-hydroxybenzoate hydroxylase (PHBH). In both cases, the microiterative IRC calculations performed very well already with the smallest core region (in both cases the four atoms involved in bond breaking and bond making). The corresponding results were similar to those obtained for larger cores incorporating the whole QM region (see

Fig. 4.2c and 4.2d). Both in BsCM and PHBH, the small-core IRC curves were rather close to the full-QM region IRC results, indicating that the atoms directly involved in the chemical reaction undergo the largest displacements along the reaction pathway. The RMS deviations of the geometries along the IRC paths were not very high, remaining well below 0.1 Å along the reaction path. For BsCM, we also performed a microiterative IRC for a bigger core region, which, apart from the QM atoms, included the five amino acid residues from the MM region that form H-bonds with the substrate. The resulting IRC curve was essentially indistinguishable from the one obtained for the core region comprising only the QM region. This corroborates our assumption that only a limited number of degrees of freedom from a large QM/MM system need to be directly included in the IRC calculations.

The proposed microiterative IRC method can be used to determine reaction paths in large-scale QM/MM calculations, e.g. on enzymatic reactions, but also at the pure QM level, with a core region containing only the atoms directly involved in the chemical reaction. With further validation, it may also be used in the context of variational transition state theory or reaction path Hamiltonian calculations. The resulting IRC paths may serve as a basis for collective coordinates in free energy methods, e.g. the DH-FEP method, described in the next Chapter.

4. MICROITERATIVE INTRINSIC REACTION COORDINATE (IRC)

5

Dual Hamiltonian Free Energy Perturbation (DH-FEP)

Free energy calculations require extensive sampling of configurational space. In a classical system, information about all accessible configurations is contained in the partition function, which can however not be obtained exactly with a limited number of configurations. Free energy differences can be expressed in terms of ensemble averages that can be approximately evaluated with the use of sampling techniques such as Monte Carlo (MC) or Molecular Dynamics (MD) (107).

It is generally not trivial to achieve sufficient sampling. This depends on the system size and on the computational time required to obtain the potential energy and the gradients at a given theory level. Classical force field calculations are fast enough to allow efficient sampling even of large biological systems, but this becomes difficult in QM/MM calculations. Usually large solvated biomolecules are treated with QM/MM techniques, and extensive sampling is required to explore their configurational space, which is computationally demanding, since part of a system is treated at the QM level. Due to this limitation, much effort was spent to develop approximate methods for evaluating free energy differences specifically in QM/MM systems (108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122). These methods usually try to avoid direct sampling at a high theory level, while still aiming at providing an accurate estimate of free energy changes in the course of the process studied (usually a chemical reaction).

In the first approach (108, 109, 110, 111, 112, 113, 114, 115) the free energy of reaction is initially estimated by performing sampling of configurational space with a computationally inexpensive reference Hamiltonian, and then correcting the obtained estimate via the free energy perturbation (FEP) from the reference to a QM/MM Hamiltonian, thus employing a thermodynamic cycle. Another approach is based on sampling acceleration, achieved by using auxiliary MC simulations, performed with a reference Hamiltonian (e.g. a classical force field specially parameterized for the system under investigation); each final structure is then subjected to a MC update test, with the success criterion being based on the overlap of the two Hamiltonian phase spaces (123, 124, 125). This procedure was shown to improve the rate of convergence significantly. The two approaches were also combined (126) such that the free energy difference between two points along a reaction path is evaluated with a reference Hamiltonian, but the free energy correction is determined via thermodynamic integration, employing the Metropolis-Hastings MC algorithm.

The free energy of enzymatic reactions is sometimes also evaluated using a direct sampling of the whole phase space of the QM/MM system on a single potential surface using standard free energy techniques, such as umbrella sampling (127) or thermodynamic integration (128). Usually, sampling is performed with an efficient semiempirical QM method (129). DFT calculations of reaction free energies were also reported (130), but are generally computationally unattainable for systems with a large QM region. The reliability of semiempirical QM methods was questioned in a recent paper by Heimdal and Ryde (115), on the grounds that there may only be a weak overlap of phase space between semiempirical and higher theory levels. In this context, another approach should also be mentioned, in which the configurational space is sampled using a semiempirical QM method, and a high-level energy correction, given as a continuous function of a distinguished reaction coordinate, is applied to modify energies and gradients at every step (101, 102).

The QM/MM free energy (FE) method, developed by Yang et al. (131), has become rather popular for calculating enzymatic reaction free energies. It does not require sampling of the QM degrees of freedom and only uses information from the MM configuration ensemble. In this approach, the FEP formalism of Zwanzig (132) is applied in an unusual way. According to Zwanzig, the free energy difference between two systems, described by an unperturbed (H_1) and a perturbed (H_2) Hamiltonian, can be written as

$$\Delta A = A_2 - A_1 = -\frac{1}{\beta}\ln\int P_1(\mathbf{r})\exp\{-\beta[E_2(\mathbf{r}) - E_1(\mathbf{r})]\}\mathrm{d}\mathbf{r}, \tag{5.1}$$

where $E(\mathbf{r})$ is potential energy and $P_1(\mathbf{r})$ is the probability of finding the unperturbed system in the configuration $\mathbf{r}$. In the QM/MM-FE method, the reaction path is divided into multiple windows, based on the value of a predefined reaction coordinate. The geometries of the QM region in these windows are obtained via a set of initial restrained optimizations. Then, for each window, MD sampling is performed for the MM region only, while the QM region is kept fixed. The coordinates of the QM region of the current window are replaced with the ones from the next window at every step (or every n steps) of MD, and the perturbation energy difference is evaluated. Then, the ensemble of energy differences is used to calculate the free energy difference between the two windows:

$$\begin{aligned}\Delta A(R_c) = \Delta E_{QM}(\mathbf{r}_{QM}^{min}) - \frac{1}{\beta}\ln\int P(R_c^A)\exp\{-\beta[E_{QM/MM}(\mathbf{r}_{QM}^{min}(R_c^B))\\ -E_{QM/MM}(\mathbf{r}_{QM}^{min}(R_c^A))]\}\mathrm{d}\mathbf{r}_{MM}.\end{aligned} \tag{5.2}$$

Here, $\Delta E_{QM}(\mathbf{r}_{QM}^{min})$ is the QM energy difference between the two subsequent windows, and $\mathbf{r}_{QM}^{min}(R_c^A)$ and $\mathbf{r}_{QM}^{min}(R_c^B)$ are the coordinates of the optimized QM region, satisfying constraints at windows A and B, respectively. Note that sampling is performed with the QM region coordinates fixed to their value at window A, and integration is done only over MM degrees of freedom. Summing up the free energy differences between the neighboring windows along the reaction path, one obtains activation and reaction free energies of a given reaction. In order to make sampling inexpensive, the QM region is represented during the sampling with the partial charges obtained by an ESP (electrostatic potential) fit.

This QM/MM-FE procedure assumes that QM and MM degrees of freedom can be sampled individually and that one can avoid sampling the QM region, accounting for its entropic contribution only within the harmonic approximation at the stationary points. QM/MM-FE is widely used and considered to be an appropriate method for calculating enzymatic reaction free energies (133). However, one may expect that it can underestimate the entropic contributions due to the lack of sampling in the QM region.

5.1 DH-FEP formalism

We have developed a modified version of the QM/MM-FE method (134), in which we do not separate the QM and MM degrees of freedom. The perturbation is defined by means of a predetermined reaction coordinate ξ (usually one degree of freedom). As in QM/MM-FE, the reaction path is split into discrete windows, each satisfying a specific ξ_i value, so that ξ_i and ξ_{i+1} are two constraints, which define the two neighboring windows along the reaction coordinate:

$$\Delta E_{pert}^{\xi_i \rightarrow \xi_{i+1}} = E(\mathbf{r}', \xi_{i+1}) - E(\mathbf{r}', \xi_i), \tag{5.3}$$

where $\mathbf{r}'$ denotes any system configuration that fulfills the constraint ξ_i. We thus have a constrained Hamiltonian, which is perturbed by the change of the constraint. Following Eq. 5.1, one can write the free energy difference between two subsequent points along the RC as

$$\Delta A^{\xi_i \rightarrow \xi_{i+1}} = -\frac{1}{\beta} \ln \int P_i(\mathbf{r}', \xi_i) \mathrm{exp}\{-\beta[E(\mathbf{r}', \xi_{i+1}) - E(\mathbf{r}', \xi_i)]\} \mathrm{d}\mathbf{r}'. \tag{5.4}$$

For practical computations, the integration is replaced by a discrete sum over MD steps (like for conventional QM/MM-FE). In the limit of extensive sampling over all $\mathbf{r}'$ one obtains:

$$\Delta A^{\xi_i \rightarrow \xi_{i+1}} = -\frac{1}{\beta} \ln [\frac{1}{N} \sum_{i=1}^{N} \mathrm{exp}\{-\beta \Delta E_{pert}^{\xi_i \rightarrow \xi_{i+1}}\}]. \tag{5.5}$$

As such, the approach presented above has no benefits compared with standard free energy methods, like umbrella sampling or thermodynamic integration. In principle, it enables an exact evaluation of the reaction free energy. In practice, it can be applied for large QM/MM systems only at semiempirical level of QM theory. However, the formalism allows us to introduce two important approximations, which can turn it into a powerful device for calculating reaction free energies of QM/MM systems, with the use of higher-level QM methods.

First we notice that the integration step size in MD simulations is usually chosen such as to guarantee a stable and accurate propagation of the system. Therefore, two consecutive points are very close in geometry, and ΔE_{pert} does not vary significantly within one step. Hence, the MD step size is not suited to sample ΔE_{pert} efficiently. We adopt a procedure, in which ΔE_{pert} is computed only periodically after skipping a predefined number of steps. The intermediate MD steps are neglected in the computation of the free energy. Secondly, we propose to perform QM/MM sampling with a semiempirical QM Hamiltonian, while the ΔE_{pert} values are evaluated using a higher-level QM method (DFT or MP2) for the selected steps. Equation 5.4 can then be reformulated as follows:

$$\Delta A^{\xi_i \to \xi_{i+1}} = -\frac{1}{\beta} \ln \int P_i^{Ham1}(\mathbf{r}', \xi_i) \exp\{-\beta[E^{Ham2}(\mathbf{r}', \xi_{i+1}) \\ -E^{Ham2}(\mathbf{r}', \xi_i)]\} \mathrm{d}\mathbf{r}'. \tag{5.6}$$

In practice, the following algorithm is executed in order to evaluate the free energy difference along the reaction coordinate. The reaction path, obtained from the sequence of restrained optimizations, is divided into windows, with discrete values of the reaction coordinate. Then a constrained QM/MM MD simulation is performed at a semiempirical QM/MM level for each window, with the constraint ξ_i being satisfied at every step via the SHAKE algorythm (135) (see details below). At every predefined number of steps, the constraint is perturbed to ξ_{i+1} and ΔE_{pert} is evaluated, using a higher-level QM Hamiltonian. Then the next MD step is performed from the geometry that satisfied constraint ξ_i. From the ensemble of energy differences obtained, the free energy difference along the reaction coordinate is evaluated as in the conventional QM/MM-FE approach.

We have implemented the DH-FEP method in the developmental version of the ChemShell package. The constraint, applied to the reaction coordinate during MD sampling, is fulfilled using the SHAKE procedure. The SHAKE algorithm is normally employed in MD programs to constrain the X-H bonds (with X being any heavy atom) so that hydrogen atoms do not move too far during a single MD step, but it is designed to satisfy any constraint in a chemical system (135). SHAKE constraints for a single distance, for a torsion angle, and for the difference of two bonds lengths with one common atom were already available in ChemShell (129). In order to be able to run DH-FEP for the chosen test sytem, we also implemented the SHAKE algorithm for the difference of two independent bond lengths involving four different atoms.

Following previous work (129), the difference-of-distances constraint involving four atoms A, B, C and D, has the form

$$\sigma(\mathbf{R}_A, \mathbf{R}_B, \mathbf{R}_C) = |\mathbf{R}_{BA}| - |\mathbf{R}_{DC}| - \xi = 0, \tag{5.7}$$

where $|\mathbf{R}_{IJ}| = |\mathbf{R}_J - \mathbf{R}_I|$ and ξ is the constraint value. The SHAKE algorithm iteratively determines the Lagrangean multiplier γ, which is the force constant that guarantees the constraint to be exactly satisfied in the next time step. For the difference-of-distances constraint, the positions of the atoms are propagated as

$$\mathbf{R}_I(+) = \bar{\mathbf{R}}_I(+) - \frac{\Delta}{M_I}\gamma\nabla_I\sigma|_0, \tag{5.8}$$

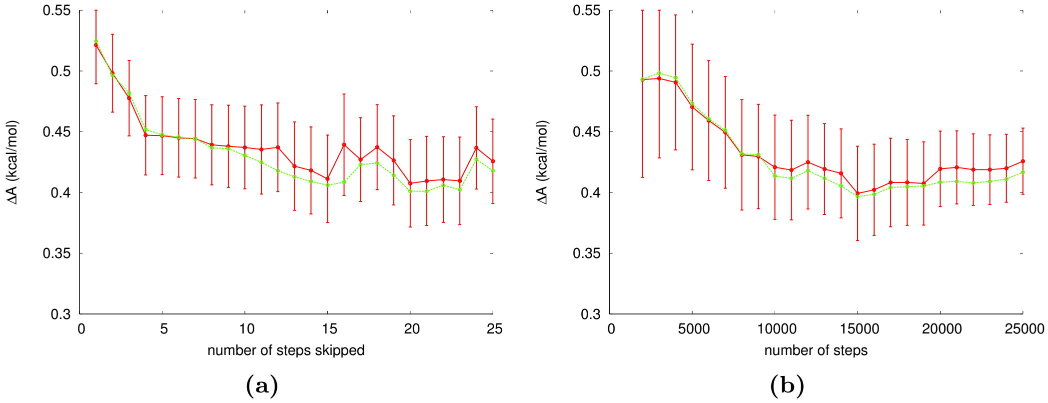
where $\bar{\mathbf{R}}_I(+)$ is the new atom position, propagated without constraints, M_I is the mass of atom I, Δ is the time step, and $\nabla_I\sigma$ is the derivative of Eq. 5.7 with respect to the positions of the atoms involved in the constraint, namely

$$\begin{aligned}
\nabla_A\sigma &= \hat{\mathbf{R}}_{BA} \\
\nabla_B\sigma &= -\hat{\mathbf{R}}_{BA} \\
\nabla_C\sigma &= -\hat{\mathbf{R}}_{DC} \\
\nabla_D\sigma &= \hat{\mathbf{R}}_{DC}.
\end{aligned} \tag{5.9}$$

5.2 DH-FEP performance

With the use of a two-dimensional analytic model potential, sampled by a constrained MC simulation, we have first checked that the free energy perturbation method gives accurate results when used in the reaction coordinate ansatz described above. We studied two similar analytic potentials that differ only slightly in the region of the first-order saddle point, connecting the two minima on the two-dimensional surface. In an assessment of DH-FEP for these model surfaces, MC sampling was performed on one surface, while the perturbations were evaluated on the other one at every step. The accuracy of the computed free energies quickly decreased with increasing difference between the two surfaces. This indicates a pronounced sensitivity of the DH-FEP method to the degree of the configurational space overlap between the two potentials, which necessitates a careful examination of their geometrical correspondence along the reaction coordinate prior to the sampling.

Next, we tested the DH-FEP method performance on a "real-life" QM/MM system, the Claisen rearrangement of chorismate to prephenate in the active site of the *Bacillus subtilis* Chorismate Mutase (BsCM) enzyme. For this system, both the enthalpic and entropic contributions to the activation barrier are known from experiment, with

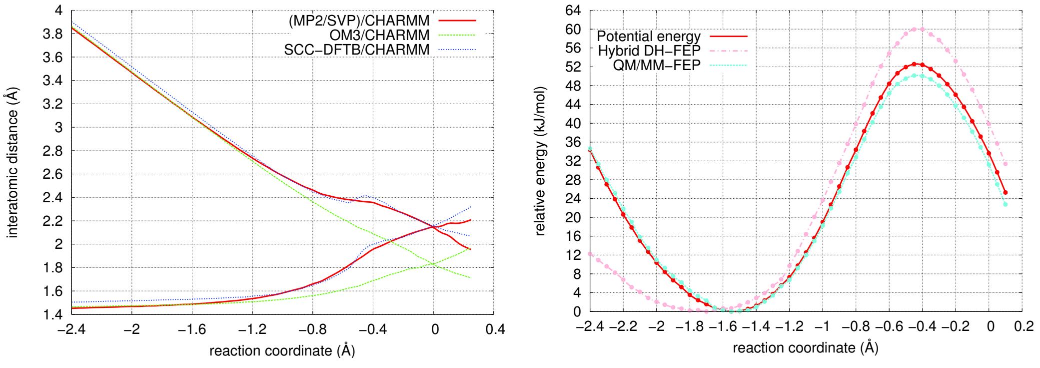


(a) **(b)**

Figure 5.1: Dependence of the free energy differences between two windows on the MD parameters. (a) Free energy difference between the two windows, calculated with different number of steps skipped between two subsequent ΔE_{pert} readings with the overall number of ΔE_{pert} evaluations fixed to 1000. (b) Free energy difference between the two windows, calculated with 14 steps skipped between two subsequent ΔE readings with the overall number of ΔE evaluations being varied. The values in red were obtained after the data was subjected to statistical tests for lack of trend and decorrelation. The values in green were obtained from the direct exponential averaging of all the data points. Data taken starting from the end of the 25 ps MD sampling run of one of the windows along the CM reaction profile at the OM3/CHARMM level.

$T\Delta S = -11.4 \pm 1.5$ kJ/mol at $T = 300$ K (136). The reaction coordinate can be conveniently defined as the difference between the lengths of the breaking C-O and the forming C-C bond.

We first performed test calculations at a semiempirical QM/MM level using OM3 (15, 16), without introducing the dual Hamiltonian approximation. At this level, we analysed the number of steps that should be skipped in between two subsequent perturbation measurements ($N_{skipped}$), and the length of sampling that should be performed in order to obtain converged results with given $N_{skipped}$. Based on the graphs, see Fig. 5.1, we obtain a reasonable value of $N_{skipped} = 15$, for which 10 ps simulation seems to be sufficient to obtain converged results for one window. Ideally, one would take a longer simulation time (20 ps or more), but this will increase the computational time for each window significantly, when perturbations are measured with a higher-level QM method, without improving the results significantly. We then assessed the



(a) **(b)**

Figure 5.2: DH-FEP assessment and performance: Chorismate mutase. (a) Optimized C-C and C-O distances along the reaction coordinate for the three different QM methods. (b) Potential energy, DH-FEP and QM/MM-FE profiles, computed with the QM(MP2/SVP)/CHARMM theory level. DH-FEP sampling was performed with a "hybrid approach": the first part of the reaction path was sampled with OM3/CHARMM, and the second part with SCC-DFTB/CHARMM.

accuracy of the method, again at the semiempirical QM/MM level; as in the case of the analytic model function, we closely reproduced the accurate reference results (obtained here from thermodynamic integration (TI) performed at the same theory level). The computed activation energies were within 0.8 kJ/mol for each of the four snapshots tested, which is within the error bar of the applied TI approach (129).

Moving towards real-life DH-FEP applications for BsCM, we first evaluated the geometrical correspondence between the low-level and high-level methods used for sampling and for perturbation measurements, respectively. This can be done by comparing the individual interatomic distances entering the reaction coordinate or by computing the full QM region root-mean-square (RMS) deviations along the reaction path. We used both criteria and found that the latter criterion is less important than the first one, so we refrain from discussing it here. It should be noted that the SHAKE procedure enforces nearly the same individual interatomic distances in the MD runs as the restraint procedure used during optimizations. Therefore, one can assess the configurational space overlap between the low-level and high-level methods during MD by the information obtained during the restrained optimizations.

We considered the correspondence between the two semiempirical methods used for configurational space sampling, OM3 (15, 16) and SCC-DFTB (18), and the RI-MP2 (22, 23) method used in combination with the SVP basis set (137) for the ΔE_{pert} evaluations. The C-O and C-C distances along the reaction coordinate for all the three methods are given in Fig. 5.2a. Clearly, while the OM3 method gives precisely the same bond lengths as MP2/SVP at the beginning of the reaction path, the distances differ significantly closer to the transition state. On the contrary, SCC-DFTB represents MP2/SVP bond lengths quite accurately in the region around the transition state, while being slightly off in the first part of reaction path.

Accordingly, DH-FEP calculations of the BsCM activation barrier, performed with OM3 or SCC-DFTB based sampling, failed to give an accurate entropic contribution. We therefore performed hybrid DH-FEP calculations, in which the OM3 method was used to sample the configurational space of the system in the first part of reaction path, while SCC-DFTB was used for the second part. This approach gave satisfactory results, with the average entropic contribution ($T\Delta S$) for the six snapshots being equal to -14.5 kJ/mol, after taking into account the difference in the zero-point energy of the reactant and transition state in the harmonic oscillator approximation. The potential energy as well as the DH-FEP and standard QM/MM-FE curves are given in Fig. 5.2b for one of the six snapshots. One can see that while the QM/MM-FE curve practically follows the potential energy curve, the DH-FEP curve differs strongly, reflecting a significant entropic contribution along the reaction path (consistent with experiment).

The DH-FEP method thus seems somewhat tedious to use. It can provide accurate results, but a lot of effort may be required to find the appropriate semiempirical QM/MM method that will produce geometries close to those obtained with the higher-level method used to evaluate perturbation energy differences. One way to tackle this problem may be a special reparameterization of a semiempirical method for each specific case study, but this is quite cumbersome and generally not recommended. Instead, we advocate a modification of the DH-FEP method, in which more degrees of freedom are constrained via the SHAKE procedure, e.g., in case of BsCM, not the difference of two distances, but each of the two distances individually. This introduction of collective reaction coordinates will assure better sampling of configurational space, with only minor effects on the measured entropic contribution. We tested this approach on BsCM and obtained good results when using either OM3 or SCC-DFTB for sampling.

The use of a collective reaction coordinate also simplifies the application of the DH-FEP method in complicated cases. Instead of treating one complex reaction coordinate with a special SHAKE procedure, several simple degrees of freedom, e.g., individual bonds, can be subjected to constraints. We propose the following procedure to identify a collective coordinate and to set up the DH-FEP calculations. First, high-level QM/MM calculations should be performed to find 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 IRC starting from the optimized transition state, which can be efficiently done at the QM/MM level by an approximate microiterative IRC scheme, as discussed in the previous Chapter. The IRC can then be used to identify the limited set of internal coordinates that undergo the most drastic changes along the reaction path and that should thus enter the collective reaction coordinate for the subsequent DH-FEP calculations.

6

Conclusions and outlook

In the current work, we have applied existing QM/MM techniques to uncover the detailed mechanism of two different enzymatic reactions and developed two novel QM/MM techniques for reaction path and free energy calculations.

We have performed the first computational investigation of the reaction mechanism in the Baeyer-Villiger cyclohexanone monooxygenase (CHMO). We have shown that the cyclohexanone oxygenation proceeds via formation of an anionic Criegee intermediate, with the subsequent fragmentation being the rate-limiting step of reaction. We have also rationalized the experimentally observed CHMO enantioselectivity toward several substrates, i.e., 4-methyl-, 4-ethylidene- and 4-hydroxycyclohexanone. The prediction or explanation of experimentally observed CHMO enantioselectivity toward other substituted substrates should be possible in terms of the proposed model. It is the selective binding mode of the ligand in its chair conformation that places it in a chiral environment such that only one of the two enantiotopic C-atoms can migrate. The preferred orientation of substituent, e.g. equatorial or axial, then determines the stereoselective outcome of the reaction. In the case of the non-polar substituents, e.g. 4-methylcyclohexanone, the preferred orientation should be easy to predict, based on the intrinsic preferences among possible substrate conformers and on an assessment of the Criegee intermediate and the subsequent transition state concerning possible steric clashes between the substituent and active-site residues. The situation is more complicated in the case of polar substituents, e.g. 4-hydroxycyclohexanone, where the outcome of reaction depends on a fine interplay of weak interactions between the substrate and protein environment. Explicit calculations are therefore needed to make predictions in such cases.

Further calculations of the reactions between wild-type and mutated CHMO and multiple substrates would be helpful to support and guide directed evolution experiments. Of fundamental interest would be the investigation of the CHMO reaction mechanism with larger substrates, since currently it is not obvious whether the enzyme can accomodate those in the binding site present in the available X-ray structures, or whether the protein structure must first undergo some structural rearrangement. An alternative active site, proposed in another recently resolved X-ray structure of CHMO (138), could be tested by explicit QM/MM calculations as well. A study of the reaction mechanism of another thermostable Baeyer-Villiger monooxygenase, PAMO, is currently underway in our group.

We have also investigated the reaction mechanism of Lipopolysaccharyl-α-1,4-galactosyltransferase C (LgtC). Most likely in the wild-type enzyme is the so-called S_Ni mechanism, which occurs through a front-side attack of the nucleophile (lactose) to the anomeric carbon of the galactose, at the same side as the leaving group (UDP). It proceeds via formation of an oxocarbenium ion-like transition state. We also found evidence that mutations in the active site and/or changes in substrate composition may change the rate and even the nature of the reaction mechanism. Hence, explicit QM/MM calculations are needed in order to explain the reaction mechanisms of other retaining glycosyltransferases (GTs).

We have proposed a method for performing intrinsic reaction coordinate (IRC) calculations for large QM/MM systems. In this approach, only the atoms directly involved in the reaction (core region) follow the steepest descent path, while the remaining degrees of freedom (outer region) are fully relaxed after each IRC step. This microiterative IRC method adheres to the same philosophy as the microiterative transition state search that has been successfully used in previous QM/MM calculations. We tested this method on four different systems, including two enzymatic reactions. We showed that the microiterative IRC approach performs very well already with a very small core region (provided that it is well chosen). There are usually slight differences in reaction paths compared with those obtained with larger core regions, but the RMS deviations of the geometries along the reaction path are quite small, especially in the case of enzymatic reactions. Microiterative IRC calculations can therefore be used to obtain a qualitative picture of the reaction path in large QM/MM systems. We expect that our method may also be used in other types of calculations, e.g. when using reaction path Hamiltonians or variational transition state theory. Further validation of the microiteraive IRC method is desirable for this purpose.

We have developed a modified QM/MM free energy perturbation method, in which the QM region degrees of freedom are explicitly sampled by MD. Normally, only one or two degrees of freedom (reaction coordinate) are constrained and subject to perturbations, as opposed to the original QM/MM-FE formulation, in which only the MM region is sampled and the entropic contribution of the QM region is evaluated only at the stationary points within the harmonic approximation. To make such calculations feasible, we suggest to accelerate sampling by performing it at the semi-empirical QM/MM level, while evaluating the perturbation energies with higher-level QM methods like DFT or MP2. We therefore call our method dual Hamiltonian free energy pertutrbation (DH-FEP). We have assessed DH-FEP performance using an analytic model function with exactly known solutions and one real-life enzymatic QM/MM system, namely chorismate mutase. In the latter case, we have shown that the low-level semiempirical method needs to be chosen carefully such that a configurational space similar to that of the higher-level method is sampled during MD; if so, accurate free energy results can be obtained. We suggest a modified version of DH-FEP, in which a collective reaction coordinate composed of several individual interatomic distances is constrained to the values obtained from restrained higher-level optimizations, which ensures sampling of an appropriate configurational space.

The microiterative IRC method can be used in combination with the DH-FEP method to identify the collective coordinate in cases where the reaction coordinate cannot be easily defined on qualitative grounds in terms of a combination of interatomic distances. After optimizing a transition state, one can run microiterative IRC calculations to get a sequence of structures along the reaction path. These can provide starting coordinates for each of the sampling windows, and the individual degrees of freedom that undergo the most drastic distortions along the reaction path can be constrained and subjected to perturbations. Such a scheme may become an efficient way of performing free energy calculations in arbitrary QM/MM systems.

References

[1] A. Warshel and M. Levitt. *J. Mol. Biol.*, **103**:227–249, 1976. 1

[2] Martin J. Field, Paul A. Bash, and Martin Karplus. *J. Comput. Chem.*, **11**(6):700–733, 1990. 1

[3] Hans Martin Senn and Walter Thiel. *Angew. Chem. Int. Ed.*, **48**(7):1198–1229, 2009. 2

[4] JC Slater. *Phys. Rev.*, **91**(3):528–530, 1953. 2, 10

[5] SH Vosko, L Wilk, and M Nusair. *Can. J. Phys.*, **58**(8):1200–1211, 1980. 2, 10

[6] AD Becke. *Phys. Rev. A*, **38**(6):3098–3100, 1988. 2, 10

[7] AD Becke. *J. Chem. Phys.*, **98**(7):5648–5652, 1993. 2, 10

[8] PJ Stephens, FJ Devlin, CF Chabalowski, and MJ Frisch. *J. Phys. Chem.*, **98**(45):11623–11627, 1994. 2, 10

[9] CT Lee, WT Yang, and RG Parr. *Phys. Rev. B*, **37**(2):785–789, 1988. 2, 10

[10] R. Lonsdale, J. N. Harvey, and A. J. Mulholland. Inclusion of Dispersion Effects Significantly Improves Accuracy of Calculated Reaction Barriers for Cytochrome P450 Catalyzed Reactions. *J. Phys. Chem. Lett.*, **1**:3232–3237, 2010. 2

[11] S Grimme. *J. Comput. Chem.*, **27**(15):1787–1799, 2006. 2, 11

[12] J. P. Perdew. *Phys. Rev. B*, **33**(12, Part 2):8822–8824, 1986. 2

[13] J. P. Perdew. *Phys. Rev. B*, **34**(10):7406, 1986. 2

[14] Y. Zhao and D. G. Truhlar. *Theor. Chem. Acc.*, **120**(1–3):215–241, 2008. 2, 11

[15] M Scholten. PhD thesis, Universität Düsseldorf, 2003. 2, 41, 43

[16] Nikolaj Otte, Mirjam Scholten, and Walter Thiel. *J. Phys. Chem. A*, **111**(26):5751–5755, 2007. 2, 41, 43

[17] Michael J. S. Dewar, Eve G. Zoebisch, Eamonn F. Healy, and James J. P. Stewart. *J. Am. Chem. Soc.*, **107**(13):3902–3909, 1985. 2

[18] M Elstner, D Porezag, G Jungnickel, J Elsner, M Haugk, T Frauenheim, S Suhai, and G Seifert. *Phys. Rev. B*, **58**:7260, 1998. 2, 43

[19] M Schütz, G Hetzer, and HJ Werner. *J. Chem. Phys.*, **111**(13):5691–5705, 1999. 2

[20] G Hetzer, M Schütz, H Stoll, and HJ Werner. *J. Chem. Phys.*, **113**(21):9443–9455, 2000. 2

[21] M Schütz and HJ Werner. *J. Chem. Phys.*, **114**(2):661–681, 2001. 2

[22] F Weigend and M Hser. *Theor. Chem. Acc.*, **97**:331 – 340, 1997. 3, 43

[23] F Weigend, M Hser, H Patzelt, and R Ahlrichs. *Chem. Phys. Letters*, **294**:143 – 152, 1998. 3, 43

[24] R Ahlrichs, M Bär, M Häser, H Horn, and C Kölmel. *Chem. Phys. Lett.*, **162**(3):165–169, 1989. 3, 10

[25] R. B. Murphy, D. M. Philipp, and R. A. Friesner. *J. Comput. Chem.*, **21**(16):1442–1457, 2000. 4

[26] Christopher N. Rowley and Benoit Roux. *J. Chem. Theory Comput.*, **8**(10):3526–3535, 2012. 4

[27] P Sherwood, AH de Vries, SJ Collins, SP Greatbanks, NA Burton, MA Vincent, and IH Hillier. *Farad. Disc.*, **106**:79–92, 1997. 4

[28] AH de Vries, P Sherwood, SJ Collins, AM Rigby, M Rigutto, and GJ Kramer. *J. Phys. Chem. B*, **103**(29):6133–6141, 1999. 4

[29] J Nocedal. *Math. Comp.*, **35**(151):773–782, 1980. 4, 28

[30] DC Liu and J Nocedal. *Math. Programming*, **45**(3):503–528, 1989. 4, 28

[31] SR Billeter, AJ Turner, and W Thiel. *Phys. Chem. Chem. Phys.*, **2**(10):2177–2186, 2000. 5, 28, 29

[32] AJ Turner, V Moliner, and IH Williams. *Phys. Chem. Chem. Phys.*, **1**(6):1323–1331, 1999. 5, 26, 28

[33] A Banerjee, N Adams, J Simons, and R Shepard. *J. Phys. Chem.*, **89**(1):52–57, 1985. 5, 28

[34] J Baker. *J. Comput. Chem.*, **7**(4):385–395, 1986. 5, 28

[35] Li Xie, Haiyan Liu, and Weitao Yang. *J. Chem. Phys.*, **120**(17):8039–8052, 2004. 5, 25

[36] Philippe Y. Ayala and H. Bernhard Schlegel. *J. Chem. Phys.*, **107**(2):375–384, 1997. 5

[37] Haiyan Liu, Zhenyu Lu, G. Andres Cisneros, and Weitao Yang. *J. Chem. Phys.*, **121**(2):697–706, 2004. 5, 25

[38] G. Andres Cisneros, Haiyan Liu, Zhenyu Lu, and Weitao Yang. *J. Chem. Phys.*, **122**(11):114502, 2005. 5, 25

[39] Hrant P. Hratchian and Michael J. Frisch. *J. Chem. Phys.*, **134**(20):204103–204112, 2011. 5, 25

[40] A. Baeyer and V. Villiger. *Ber. Dtsch. Chem. Ges.*, **32**:3625–3633, 1899. 7

[41] M Renz and B Meunier. *Eur. J. Org. Chem.*, (4):737–750, 1999. 7

[42] GJ ten Brink, IWCE Arends, and RA Sheldon. *Chem. Rev.*, **104**(9):4105–4123, 2004. 7

[43] H Leisch, K Morley, and PCK Lau. *Chem. Rev.*, **111**(7):4165–4222, 2011. 7

[44] Daniel E. Torres Pazmino, Hanna M. Dudek, and Marco W. Fraaije. *Curr. Opin. Chem. Biol.*, **14**(2):138–144, 2010. 7

[45] G.E. Turfitt. *Biochem. J.*, **42**(3):376–383, 1948. 8

[46] NA Donoghue, DB Norris, and PW Trudgill. *Eur. J. Biochem.*, **63**(1):175–192, 1976. 8

[47] YCJ Chen, OP Peoples, and CT Walsh. *J. Bacteriol.*, **170**(2):781–789, 1988. 8

[48] MW Fraaije, J Wu, DPHM Heuts, EW van Hellemond, JHL Spelberg, and DB Janssen. *Appl. Microb. Biotechnol.*, **66**(4):393–400, 2005. 8

[49] CC Ryerson, DP Ballou, and C Walsh. *Biochemistry*, **21**(11):2644–2655, 1982. 8

[50] DW Sheng, DP Ballou, and V Massey. *Biochemistry*, **40**(37):11156–11167, 2001. 8

[51] Daniel E. Torres Pazmino, Bert-Jan Baas, Dick B. Janssen, and Marco W. Fraaije. *Biochemistry*, **47**(13):4082–4093, 2008. 8, 12

[52] Roberto Orru, Hanna M. Dudek, Christian Martinoli, Daniel E. Torres Pazmino, Antoine Royant, Martin Weik, Marco W. Fraaije, and Andrea Mattevi. *J. Biol. Chem.*, **286**(33):29284–29291, 2011. 8

[53] E Malito, A Alfieri, MW Fraaije, and A Mattevi. *Proc. Natl. Acad. Sci. USA*, **101**(36):13157–13162, 2004. 9

[54] I. Ahmad Mirza, Brahm J. Yachnin, Shaozhao Wang, Stephan Grosse, Helene Bergeron, Akihiro Imura, Hiroaki Iwaki, Yoshie Hasegawa, Peter C. K. Lau, and Albert M. Berghuis. *J. Am. Chem. Soc.*, **131**(25):8848–8854, 2009. 9

[55] Iakov Polyak, Manfred T. Reetz, and Walter Thiel. *J. Am. Chem. Soc.*, **134**(5):2732–2741, 2012. 9, 10, 14

[56] Zhi-Gang Zhang, Gheorghe-Doru Roiban, Juan Pablo Acevedo, Iakov Polyak, and Manfred T. Reetz. *Adv. Synth. Catal.*, **355**(1):99–106, 2013. 10, 14

[57] Iakov Polyak, Manfred T. Reetz, and Walter Thiel. *J. Phys. Chem. B*, **117**:4993–5001, 2013. 10, 15

[58] AD MacKerell, D Bashford, M Bellott, RL Dunbrack, JD Evanseck, MJ Field, S Fischer, J Gao, H Guo, S Ha, D Joseph-McCarthy, L Kuchnir, K Kuczera, FTK Lau, C Mattos, S Michnick, T Ngo, DT Nguyen, B Prodhom, WE Reiher, B Roux, M Schlenkrich, JC Smith, R Stote, J Straub, M Watanabe, J Wiorkiewicz-Kuczera, D Yin, and M Karplus. *J. Phys. Chem. B*, **102**(18):3586–3616, 1998. 10

[59] B.R. Brooks, R.E. Bruccoleri, D.J. Olafson, D.J. States, S. Swaminathan, and M. Karplus. *J. Comput. Chem.*, **4**:187–217, 1983. 10

[60] 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**(10, Sp. Iss. SI):1545–1614, 2009. 10

[61] A Schäfer, C Huber, and R Ahlrichs. *J. Chem. Phys.*, **100**(8):5829–5835, 1994. 10

[62] P Sherwood, AH de Vries, MF Guest, G Schreckenbach, CRA Catlow, SA French, AA Sokol, ST Bromley, W Thiel, AJ Turner, S Billeter, F Terstegen, S Thiel, J Kendrick, SC Rogers, J Casci, M Watson, F King, E Karlsen, M Sjovoll, A Fahmi, A Schäfer, and C Lennartz. *J. Mol. Struct. (Theochem)*, **632**(Sp. Iss. SI):1–28, 2003. 10, 28

[63] W Smith and TR Forester. DL_POLY_2.0: A general-purpose parallel molecular dynamics simulation package. *J. Mol. Graph.*, **14**(3):136–141, 1996. 10

[64] GR Krow. *Tetrahedron*, **37**(16):2697–2724, 1981. 12

[65] J. Raul Alvarez-Idaboy, Lino Reyes, and Nelaine Mora-Diez. *Org. Biomol. Chem.*, **5**(22):3682–3689, 2007. 12

[66] Margaret M. Kayser. *Tetrahedron*, **65**(5):947–974, 2009. 13

[67] V Alphand and R Furstoss. *J. Org. Chem.*, **57**(4):1306–1309, 1992. 13

[68] G Ottolina, G Carrea, S Colonna, and A Rückemann. *Tetrahedron: Asymmetry*, **7**(4):1123–1136, 1996. 13

[69] D R Kelly. *Tetrahedron: Asymmetry*, **7**(4):1149–1152, 1996. 13

[70] JD Stewart. *Curr. Org. Chem.*, **2**(3):195–216, 1998. 13

[71] JD Stewart, KW Reed, CA Martinez, J Zhu, G Chen, and MM Kayser. *J. Am. Chem. Soc.*, **120**(15):3541–3548, 1998. 13

[72] Margaret M. Kayser and Christopher M. Clouthier. *J. Org. Chem.*, **71**(22):8424–8430, 2006. 13, 15, 16

[73] AR Potts and T Baer. *J. Mol. Struct. (Theochem)*, **419**:11–18, 1997. 14

[74] Manfred T. Reetz. *Angew. Chem. Int. Ed.*, **50**(1):138–174, 2011. 15

[75] JA Campbell, GJ Davies, V Bulone, and B Henrissat. *Biochem. J.*, **329**(Part 3):719, 1998. 21

[76] N. Taniguchi, K. Honke, and M. Fukuda. *Handbook of Glycosyltransferases and Related Genes*. Springer: Tokyo, 2002. 21

[77] YL Tzeng and DS Stephens. *Microb. Infect.*, **2**(6):687–700, 2000. 21

[78] UM Ünligil and JM Rini. *Curr. Opin. Struct. Biol.*, **10**(5):510–517, 2000. 21

[79] C Breton, L Snajdrova, C Jeanneau, J Koca, and A Imberty. *Glycobiology*, **16**(2):29R–37R, 2006. 21

[80] L. L. Lairson, B. Henrissat, G. J. Davies, and S. G. Withers. *Annu. Rev. Biochem.*, **77**:521–555, 2008. 21, 22

[81] GJ Davies. *Nature Struct. Biol.*, **8**(2):98–100, 2001. 21

[82] K Persson, H. D. Ly, M Dieckelmann, W. W. Wakarchuk, S. G. Withers, and N. C. J. Strynadka. *Nature Struct. Biol.*, **8**(2):166–175, 2001. 21, 23

[83] LL Lairson, CPC Chiu, HD Ly, SM He, WW Wakarchuk, NCJ Strynadka, and SG Withers. *J. Biol. Chem.*, **279**(27):28339–28344, 2004. 22, 23

[84] M. L. Sinnott and W. P. Jencks. *J. Am. Chem. Soc.*, **102**(6):2026–2032, 1980. 22

[85] I Tvaroška. *Carbohydrate Res.*, **339**(5):1007–1014, 2004. 22, 23

[86] Hansel Gomez, Iakov Polyak, Walter Thiel, Jose M. Lluch, and Laura Masgrau. *J. Am. Chem. Soc.*, **134**(10):4743–4752, 2012. 23

[87] K Fukui. *J. Phys. Chem.*, **74**(23):4161–4163, 1970. 25

[88] K Fukui. *Acc. Chem. Res.*, **14**(12):363–368, 1981. 25

[89] K Ishida, K Morokuma, and A Komornicki. *J. Chem. Phys.*, **66**(5):2153–2156, 1977. 25, 26

[90] M Page and JW Mciver. *J. Chem. Phys.*, **88**(2):922–935, 1988. 25, 27, 28

[91] M Page, C Doubleday, and JW Mciver. *J. Chem. Phys.*, **93**(8):5634–5642, 1990. 25

[92] C Gonzalez and HB Schlegel. *J. Chem. Phys.*, **90**(4):2154–2161, 1989. 25, 26

[93] Carlos. Gonzalez and H. Bernhard. Schlegel. *J. Phys. Chem.*, **94**(14):5523–5527, 1990. 25, 26

[94] Carlos Gonzalez and H. Bernhard Schlegel. *J. Chem. Phys.*, **95**(8):5853–5860, 1991. 25, 26

[95] HP Hratchian and HB Schlegel. *J. Chem. Phys.*, **120**(21):9918–9924, 2004. 25, 28

[96] HP Hratchian and HB Schlegel. *J. Chem. Theory Comput.*, **1**(1):61–69, 2005. 25, 28

[97] Hrant P. Hratchian, Michael J. Frisch, and H. Bernhard Schlegel. *J. Chem. Phys.*, **133**(22):224101–224108, 2010. 25, 28

[98] 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, Montgomery Jr. J. A., 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, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision A.1. Gaussian Inc. Wallingford CT 2009. 25

[99] S Dapprich, I Komaromi, KS Byun, K Morokuma, and MJ Frisch. *J. Mol. Struct.-THEOCHEM*, **461**:1–21, 1999. 25

[100] I Polyak, E Boulanger, K Sen, and W Thiel. Microiterative intrinsic reaction coordinate method for large QM/MM systems. *online*, DOI:10.1039/c3cp51669e, 2013. 26

[101] JJ Ruiz-Pernia, E Silla, I Tunon, S Marti, and V Moliner. *J. Phys. Chem. B*, **108**(24):8427–8433, 2004. 26, 36

[102] J. Javier Ruiz-Pernia, Estanislao Silla, Inaki Tunon, and Sergio Marti. *J. Phys. Chem. B*, **110**(35):17663–17670, 2006. 26, 36

[103] Philip Pechukas. *J. Chem. Phys.*, **64**(4):1516–1521, 1976. 27

[104] M. J. D. Powell. *Math. Prog.*, **26**:1, 1971. 28

[105] J. M. Bofill. *J. Comput. Chem.*, **15**(1):1–11, 1994. 28

[106] HP Hratchian and HB Schlegel. *J. Chem. Theory Comput.*, **1**(1):61–69, 2005. 29

[107] Daan Frenkel and Berend Smit. *Understanding molecular smulation. From algorithms to Applications.* Academic press, 2nd edition, 2002. 35

[108] Richard P. Muller and Arieh Warshel. Ab Initio Calculations of Free Energy Barriers for Chemical Reactions in Solution. *J. Phys. Chem.*, **99**(49):17516–17524, 1995. 35

[109] J. Bentzien, R. P. Muller, J. Florin, and A. Warshel. *J. Phys. Chem. B*, **102**(12):2293–2301, 1998. 35

[110] Marek Štrajbl, Gongyi Hong, and Arieh Warshel. *J. Phys. Chem. B*, **106**(51):13333–13343, 2002. 35

[111] Nikolay V. Plotnikov, Shina C. L. Kamerlin, and Arieh Warshel. Paradynamics: An Effective and Reliable Model for Ab Initio QM/MM Free-Energy Calculations and Related Tasks. *J. Phys. Chem. B*, **115**(24):7950–7962, 2011. 35

[112] Nikolay V. Plotnikov and Arieh Warshel. Exploring, Refining, and Validating the Paradynamics QM/MM Sampling. *J. Phys. Chem. B*, **116**(34):10342–10356, 2012. 35

[113] Thomas H. Rod and Ulf Ryde. Quantum Mechanical Free Energy Barrier for an Enzymatic Reaction. *Phys. Rev. Lett.*, **94**:138302, Apr 2005. 35

[114] Thomas H. Rod and Ulf Ryde. Accurate QM/MM Free Energy Calculations of Enzyme Reactions: Methylation by Catechol O-Methyltransferase. *J. Chem. Theory Comput.*, **1**(6):1240–1251, 2005. 35

[115] Jimmy Heimdal and Ulf Ryde. *Phys. Chem. Chem. Phys.*, **14**:12592–12604, 2012. 35, 36

[116] Jayaraman Chandrasekhar, Scott F. Smith, and William L. Jorgensen. *J. Am. Chem. Soc.*, **106**(10):3049–3050, 1984. 35

[117] Jayaraman Chandrasekhar, Scott F. Smith, and William L. Jorgensen. *J. Am. Chem. Soc.*, **107**(1):154–163, 1985. 35

[118] William L. Jorgensen. *Acc. Chem. Res.*, **22**(5):184–189, 1989. 35

[119] Robert V. Stanton, Mikael Perkyl, Dirk Bakowies, and Peter A. Kollman. *J. Am. Chem. Soc.*, **120**(14):3448–3457, 1998. 35

[120] Peter A. Kollman, Bernd Kuhn, Oreola Donini, Mikael Perakyla, Rob Stanton, and Dirk Bakowies. *Acc. Chem. Res.*, **34**(1):72–79, 2001. 35

[121] Bernd Kuhn and Peter A. Kollman. *J. Am. Chem. Soc.*, **122**(11):2586–2596, 2000. 35

[122] Oreola Donini, Tom Darden, and Peter A. Kollman. *J. Am. Chem. Soc.*, **122**(49):12270–12280, 2000. 35

[123] Radu Iftimie, Dennis Salahub, Dongqing Wei, and Jeremy Schofield. *J. Chem. Phys.*, **113**(12):4852–4862, 2000. 36

[124] Radu Iftimie, Dennis Salahub, and Jeremy Schofield. *J. Chem. Phys.*, **119**(21):11285–11297, 2003. 36

[125] Pradipta Bandyopadhyay. *J. Chem. Phys.*, **122**(9):091102, 2005. 36

[126] Christopher J. Woods, Frederick R. Manby, and Adrian J. Mulholland. *J. Chem. Phys.*, **128**(1):014109, 2008. 36

[127] Glenn M. Torrie and John P. Valleau. *Chem. Phys. Lett.*, **28**(4):578 – 581, 1974. 36

[128] John G. Kirkwood. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.*, **3**(5):300–313, 1935. 36

[129] Hans Martin Senn, Stephan Thiel, and Walter Thiel. *J. Chem. Theory Comput.*, **1**(3):494–505, 2005. 36, 39, 42

[130] Edina Rosta, Marcin Nowotny, Wei Yang, and Gerhard Hummer. *J. Am. Chem. Soc.*, **133**(23):8934–8941, 2011. 36

[131] Yingkai Zhang, Haiyan Liu, and Weitao Yang. *J. Chem. Phys.*, **112**(8):3483–3492, 2000. 36

[132] Robert W. Zwanzig. *J. Chem. Phys.*, **22**(8):1420–1426, 1954. 36

[133] Johannes Kästner, Hans Martin Senn, Stephan Thiel, Nikolaj Otte, and Walter Thiel. *J. Chem. Theory Comput.*, **2**(2):452–461, 2006. 37

[134] I Polyak, Benighaus T., E. Boulanger, and W Thiel. QM/MM Dual Hamiltonian free energy perturbation. under revision, 2013. 38

[135] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. *J. Comput. Phys.*, **23**(3):327 – 341, 1977. 39

[136] Peter Kast, M. Asif-Ullah, and Donald Hilvert. *Tetrahedron Letters*, **37**(16):2691 – 2694, 1996. 41

[137] A Schäfer, H Horn, and R Ahlrichs. *J. Chem. Phys.*, **97**(4):2571–2577, 1992. 43

[138] Brahm J. Yachnin, Tara Sprules, Michelle B. McEvoy, Peter C. K. Lau, and Albert M. Berghuis. *J. Am. Chem. Soc.*, **134**(18):7788–7795, 2012. 46

Quantum Mechanical/Molecular Mechanical Study on the Mechanism of the Enzymatic Baeyer-Villiger Reaction

I. Polyak, M. T. Reetz, W. Thiel

J. Am. Chem. Soc., **134**, 2732 (2012)

JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

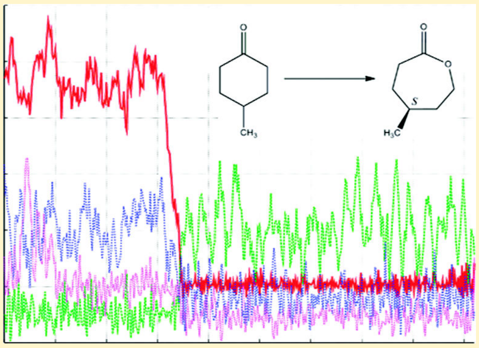
Quantum Mechanical/Molecular Mechanical Study on the Mechanism of the Enzymatic Baeyer–Villiger Reaction

Iakov Polyak,[†] Manfred T. Reetz,[†,‡] and Walter Thiel*[,†]

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

[‡]Fachbereich Chemie, Philipps-Universität Marburg, Hans-Meerwein-Strasse, D-35032 Marburg, Germany

Supporting Information



ABSTRACT: We report a combined quantum mechanical/molecular mechanical (QM/MM) study on the mechanism of the enzymatic Baeyer–Villiger reaction catalyzed by cyclohexanone monooxygenase (CHMO). In QM/MM geometry optimizations and reaction path calculations, density functional theory (B3LYP/TZVP) is used to describe the QM region consisting of the substrate (cyclohexanone), the isoalloxazine ring of C4a-peroxyflavin, the side chain of Arg-329, and the nicotinamide ring and the adjacent ribose of $NADP^+$, while the remainder of the enzyme is represented by the CHARMM force field. QM/MM molecular dynamics simulations and free energy calculations at the semiempirical OM3/CHARMM level employ the same QM/MM partitioning. According to the QM/MM calculations, the enzyme–reactant complex contains an anionic deprotonated C4a-peroxyflavin that is stabilized by strong hydrogen bonds with the Arg-329 residue and the $NADP^+$ cofactor. The CHMO-catalyzed reaction proceeds via a Criegee intermediate having pronounced anionic character. The initial addition reaction has to overcome an energy barrier of about 9 kcal/mol. The formed Criegee intermediate occupies a shallow minimum on the QM/MM potential energy surface and can undergo fragmentation to the lactone product by surmounting a second energy barrier of about 7 kcal/mol. The transition state for the latter migration step is the highest point on the QM/MM energy profile. Gas-phase reoptimizations of the QM region lead to higher barriers and confirm the crucial role of the Arg-329 residue and the $NADP^+$ cofactor for the catalytic efficiency of CHMO. QM/MM calculations for the CHMO-catalyzed oxidation of 4-methylcyclohexanone reproduce and rationalize the experimentally observed (*S*)-enantioselectivity for this substrate, which is governed by the conformational preferences of the corresponding Criegee intermediate and the subsequent transition state for the migration step.

1. INTRODUCTION

The Baeyer–Villiger (B–V) oxidation[1] transforms ketones into esters or lactones using stoichiometric amounts of hydrogen peroxide, peracids, or alkylhydroperoxides. The reaction is catalyzed by acids, bases, or transition metals, asymmetric versions being possible by the use of chiral transition metal catalysts[2] or organocatalysts.[3] Unfortunately, notable activity and high enantioselectivity is restricted to the reaction of strained cyclic ketones such as cyclobutanone derivatives. An attractive alternative is the O_2-driven enzymatic variant that is catalyzed by Baeyer–Villiger monooxygenases (BVMOs).[4] These enzymes have proven to be of great interest for synthetic chemists[4,5] because in many cases their use allows stereoselective access to valuable products such as chiral lactones without the need for hazardous reagents, with only water as an ecologically benign byproduct. In those cases in which a given substrate fails to react stereoselectively, directed evolution[6] can be employed as a tool to control this important catalytic parameter.[7,8] A variety of different BVMOs have been isolated, characterized, and used in synthetic organic chemistry, cyclohexanone monooxygenases (CHMOs) forming an important sub-family with CHMO from *Acinetobacter* sp. NCIMB 9871 (EC 1.14.13.22) as the most prominent member.[9]

The reaction mechanism of BVMOs (see Figure 1) has been explored with the use of kinetic and spectroscopic methods.[10−13] In the initial phase of the reaction, the enzyme-bound FAD cofactor is reduced by NADPH via hydride transfer. The reduced $FADH^-$ then interacts with an oxygen molecule and forms the C4a-peroxyflavin intermediate ($FADHOO^-$). Up to this point, the reaction can proceed without the substrate necessarily occupying the binding site. For CHMO it was shown[11] that the C4a-peroxyflavin is stable for some time, but in the absence of substrate it slowly transforms into the protonated state, C4a-hydroperoxyflavin (pK_a = 8.4), which was found to be unreactive. In the presence of substrate, the B–V reaction takes place: the consensus view is that a tetrahedral Criegee intermediate is formed, which subsequently fragments with concomitant C-migration to give C4a-hydroxyflavin and a product (ester or lactone). The FAD cofactor in its oxidized form is then recovered by the

Received: November 4, 2011
Published: January 5, 2012

dx.doi.org/10.1021/ja2103839 | *J. Am. Chem.Soc.* 2012, 134, 2732−2741

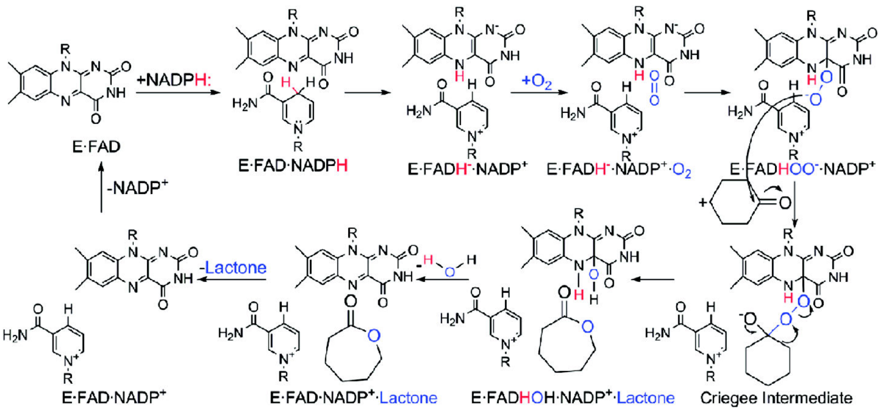


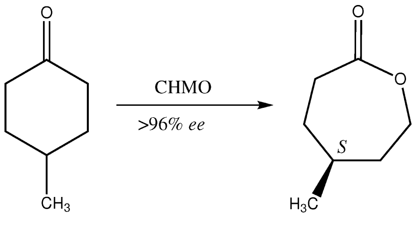
Figure 1. Mechanism of CHMO reaction. Reproduced with permission from ref 14. Copyright 2009 American Chemical Society.

spontaneous elimination of a water molecule. Furthermore, the initially formed NADP$^+$ needs to be released, and a new NADPH cofactor must bind to the enzyme in order to reduce FAD into the catalytically active form before the next reaction cycle can begin. The NADP$^+$ cofactor is considered to play an important role in stabilizing the C4a-peroxyflavin, and the Arg-329 residue is vital for the B−V reaction to proceed,[12] presumably by stabilizing the Criegee intermediate.

Nonenzymatic B−V reactions in synthetic organic chemistry proceed by two steps, both of which also occur in the enzymatic variant (Figure 1).[15,16] In the addition step, the peroxy moiety of an oxidant such as an alkylhydroperoxide attacks the carbonyl carbon of the ketone to generate the so-called Criegee intermediate. In the following migration step, one of the C−C σ bonds adjacent to the carbonyl carbon migrates to the closest oxygen atom of the peroxy group, with the O−O bond being simultaneously cleaved. The so-called anti-periplanar arrangement is part of the stereoelectronic requirement for the reaction to proceed smoothly, leading to the formation of either an ester or a lactone. In solution, the B−V reaction is generally carried out in an acidic environment so that the carbonyl oxygen is protonated during the addition step and loses its proton during or immediately after the migration step. However, base-catalyzed reactions have also been reported,[16] which are believed to proceed via an anionic Criegee intermediate. Whether the addition or migration (fragmentation) step is rate-determining will depend both on the type of the substrate and the reaction environment.[15]

The nonenzymatic B−V reaction has been studied computationally for small molecules in the gas phase and in solution.[17] In the case of the acid-catalyzed B−V oxidation of cyclohexanone,[17d] the migration step was found to be rate-determining, and the overall reaction was highly exothermic (by −67.4 kcal/mol). To our knowledge, the enzyme-catalyzed B−V reaction has not been studied theoretically up to now.

The first X-ray structure of a BVMO was that of phenyl acetone monooxygenase (PAMO) with bound FAD but in the absence of the flavin cofactor and lacking a substrate.[18] The first crystal structure of a BVMO enzyme that contains both FAD and NADP$^+$ cofactors bound to the enzyme was reported for the CHMO from *Rhodococcus* sp. strain HI-31 by Lau, Berghuis, and co-workers.[14] It has no ligand or inhibitor in the binding site. One of the two structures resolved for this CHMO (the "closed" form) was considered to represent the conformation of the enzyme that is best suited for the formation of the Criegee intermediate and the subsequent fragmentation. CHMO from *Rhodococcus* sp. strain HI-31 and CHMO from *Acinetobacter* sp. NCIMB 9871 are homologous BVMOs, sharing an overall 55% sequence identity.[14] The two enzymes display similar profiles with respect to the sense and degree of stereoselectivity, as for example in the desymmetrization of 4-methylcyclohexanone (**1**), both BVMOs leading to >96% ee in favor of the respective (*S*)-lactone (**2**) (Figure 2).



1 (*S*)-**2**

Figure 2. Stereochemical conversion of 4-methylcyclohexanone (**1**) to (*S*)-4-methyl-ε-caprolactone ((*S*)-**2**) in the B−V reaction catalyzed by CHMO.

According to phylogenetic analysis, the sequence of CHMO (*Rhodococcus*) clusters with several other CHMOs as well, showing that it is a good representative of this family of BVMOs.[14] Therefore, the crystal structure of the closed form of CHMO (*Rhodococcus*), hereafter referred to as CHMO, was used as the starting point for our investigation. More recently, further X-ray structures have become available for PAMO (and some PAMO mutants) containing both FAD and NADP$^+$ cofactors, including those for the reduced form of PAMO without and with a bound 2-(*N*-morpholino)ethanesulfonic acid (MES) inhibitor.[13]

In this article, we report quantum mechanical/molecular mechanical (QM/MM) calculations on the B−V oxidation of cyclohexanone in the active site of CHMO. Our aim is to provide a detailed atomistic understanding of the mechanism of this enzymatic reaction starting from the substrate-enzyme complex and up to the formation of the product (ε-caprolactone). Another important goal is to find an explanation of the observed enantioselectivity of CHMO as a catalyst in the desymmetrization of **1** on the basis of the computed structures for the Criegee intermediate and the subsequent transition state.

2. METHODS

Initial coordinates were taken from the X-ray structure of CHMO from *Rhodococcus* sp. strain HI-31 (PDB code 3GWD, resolution 2.3 Å).[14] The protonation states of the titratable residues (His, Glu, Asp) were chosen on the basis of the pK_a values obtained via the H++ Web software[19] and the PROPKA procedure[20] and were verified through visual inspection. Based on the FAD structure in the crystal, the $FADHOO^-$ geometry was built and optimized in the gas phase with all coordinates frozen, except for those of the peroxy group oxygen atoms and the C4a carbon atom of the isoalloxazine ring which is directly bound to the peroxy group. The whole enzyme was solvated in a water ball of 45 Å radius centered at the CHMO center of mass. The total charge of the whole system was $-30e$ at this point; to avoid artifacts, it was neutralized by Mg^{2+} and Cl^- ions via random substitution of solvent water molecules lying at least 5.5 Å away from any protein atom. In molecular dynamics (MD) simulations, a potential was imposed on the water sphere to prevent the outer solvent water molecules from drifting away into the vacuum.

The solvated system was relaxed via energy minimization and subjected to MD simulations at the MM level using the CHARMM22 force field[21] as implemented within the CHARMM program.[22] During the classical energy minimizations and MD simulations, the whole system was moving freely except for the $FADHOO^-$ and $NADP^+$ cofactors, the coordinates of which were kept fixed at the positions of the crystal structure. A random snapshot from the classical MD trajectory was used as starting point for the QM/MM calculations.

Cyclohexanone was docked into the enzyme active site of the chosen snapshot via the AutoDock program.[23] The water molecules overlapping with the docked ligand were deleted as well as the two water molecules close to the distal oxygen of the $FADHOO^-$ peroxy group. The resulting system (see Figure 3) contained 39 913 atoms.

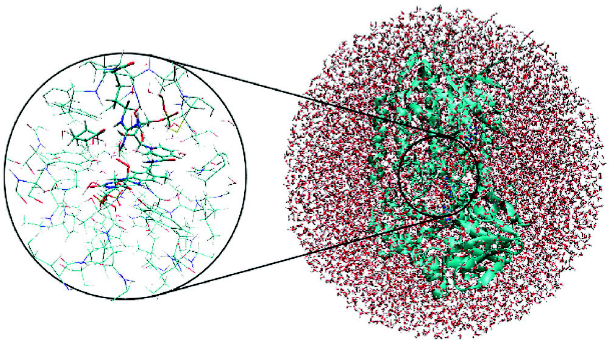


Figure 3. System used for QM/MM calculations. The active region is enlarged (on the left).

The chosen QM/MM methodology[24,25] is analogous to that used in previous studies from our group. Here we mention only the aspects relevant to the present work. In the QM/MM calculations, the QM part was treated by density functional theory (DFT) using the B3LYP functional,[26] while the MM part was described by the CHARMM22 force field. An electronic embedding scheme[27] was adopted in the QM/MM calculations, with the MM point charges being incorporated into the one-electron Hamiltonian during the QM calculation. No cutoffs were introduced for the nonbonding MM and QM/MM interactions. Hydrogen link atoms with the charge shift model[28] were employed to treat the QM/MM boundary. The QM/MM calculations were performed with the ChemShell package[29] using the TURBOMOLE program[30] to obtain the energy and gradients for the QM part and the DL_POLY program[31] to compute the energy and gradients of the MM part represented by the CHARMM22 force field. Within ChemShell, the HDLCopt[32] and DL_FIND[33] optimizers were used as well as the MD module.

The QM region incorporated all atoms from the isoalloxazine ring of C4a-peroxyflavin, cyclohexanone, the side chain of Arg-329, and the nicotinamide ring and the adjacent ribose of $NADP^+$. The latter group was included in the QM region upon inspection of preliminary optimization results with a smaller QM region, which placed the ribose 2′-hydroxyl group into close proximity to the oxygen atom of cyclohexanone. The total charge of the QM region was +1.

The SVP basis set[34] was chosen for initial B3LYP/CHARMM pathway explorations and geometry optimizations. The stationary points thus obtained along the reaction path were reoptimized using the TZVP basis set[35] in B3LYP/CHARMM calculations. Empirical dispersion corrections for DFT[36] (DFT-D2) were applied in single-point calculations at the B3LYP/TZVP level to check for the influence of dispersion. Additional single-point calculations were carried out at the M06-2X/TZVP level for further validation using a modern functional that performs particularly well for main-group thermochemistry and kinetics.[37]

During the QM/MM geometry optimizations, the active region to be optimized (see Figure 3) included all QM atoms as well as all residues and water molecules of the MM region within 12 Å of the C4a atom in the isoalloxazine ring of $FADHOO^-$. This radius was chosen such that all enzyme residues around the binding pocket were part of the active region.

Reaction paths were scanned along suitably defined reaction coordinates by performing sequences of restrained optimizations. The resulting structures served as starting points for subsequent full optimizations of the relevant stationary points. Energy minimizations and transition state (TS) searches were done with the low-memory Broyden−Fletcher−Goldfarb−Shanno (L-BFGS) algorithm[38] and with the microiterative TS optimizer that combines L-BFGS and the partitioned rational function optimizer[39] (P-RFO), both of which are implemented in the HDLCopt module of ChemShell. Frequency calculations were performed for the QM region to confirm that the optimized TS structure is indeed characterized by one imaginary frequency and a suitable transition vector, and subsequent intrinsic reaction coordinate calculations ensured that the TS is indeed connected to the proper minima by a continuous pathway.

QM/MM dynamics was performed with the use of the dynamics module within ChemShell. These MD simulations employed the NVT ensemble with a Nosé-Hoover thermostat.[40] The QM part of the system was treated with the OM3 semiempirical method[41] in these QM/MM MD runs; the active region was the same as in the QM/MM optimizations (see above), with all other atomic coordinates being frozen. The SHAKE procedure[42] was applied at every step for the water O−H bonds. QM/MM free energy calculations were carried out at the OM3/CHARMM level using thermodynamic integration and established procedures described elsewhere.[43]

For the purpose of comparison, pure QM calculations were performed at the B3LYP/TZVP level in the gas phase for the QM region (including link atoms), which was "anchored" in space by freezing the Cartesian coordinates of all atoms that were covalently bound to MM atoms. In this way, all three minima along the reaction path were reoptimized, potential energy scans were carried out along the reaction coordinates for addition and migration, and the TS for the migration step was also reoptimized.

3. RESULTS AND DISCUSSION

In the crystal structure determination of CHMO,[14] two structures were resolved. It was argued that one of these, namely the "closed" structure, is likely to represent the enzyme conformation in the post-flavin reduction state and during the essential chemical transformations.[14] For the current mechanistic investigation, this "closed" structure was chosen as the starting point. Binding of an oxygen molecule to FAD will hardly cause significant conformational changes in the active site, and it is known experimentally[11] that the enzyme remains stable after oxygen binding in the absence of a substrate in the binding pocket. Therefore, the crystal structure was manually modified by adding a peroxy group to the FAD cofactor and subjected to MD equilibration (after adding solvent molecules and counterions, see above). Cyclohexanone was then docked into the binding site of a snapshot taken from the sampling MD trajectory, followed by an energy minimization to adjust the orientation of cyclohexanone and the active-site residues. This procedure is based on the assumption that substrate binding does not cause a significant change in the overall protein structure and, in particular, in the orientation of the FADHOO$^-$ and NADP$^+$ cofactors.

The "closed" crystal structure contains two water molecules in the region where the oxygen molecule is bound, which can form hydrogen bonds with the distal oxygen of the FADHOO$^-$ peroxy group after the latter is inserted manually (see above). Preliminary scans along the reaction coordinate for the addition step indicated that these two water molecules destabilize the forming Criegee intermediate: if they are included in the QM region, the Criegee intermediate dissociates spontaneously after releasing the reaction coordinate restraints in a full optimization. Therefore, these two water molecules were removed from the system, and the reaction mechanism was investigated under the assumption that specific water molecules are not directly involved.

The chosen setup for CHMO is supported by recently published X-ray structures of another BVMO enzyme, namely PAMO containing NADP$^+$ and FAD in its oxidized form (PDB code 2YLR), in its reduced form (PDB code 2YLS), and with a bound MES inhibitor (PDB code 2YLT).[13] After alignment, the three structures match very well, with root-mean-square deviations for the protein backbone atoms (residues 12−542) of only 0.281 (2YLR vs 2YLS), 0.283 (2YLR vs 2YLT), and 0.322 Å (2YLS vs 2YLT). Visual inspection confirms that there are no significant conformational changes upon binding of MES to PAMO. Furthermore, the PAMO crystal structures without bound MES contain water molecules close to the FAD isoalloxazine ring which are absent in the structure with MES, indicating that inhibitor or substrate binding may indeed lead to the removal of water molecules from the active site of BVMO enzymes.

The structure obtained by our setup procedure for CHMO was subjected to QM/MM geometry optimization. The resulting reactant complex is shown in Figure 4, featuring cyclohexanone in the chair conformation in a well-defined position in the binding pocket, with hydrogen bonds between the carbonyl oxygen atom and Arg-329 and the NADP$^+$ ribose 2′-hydroxyl group (NADP:HN2T) (Table 1). The molecule can also be positioned in a different way, namely in a juxtaposition resulting from rotation by 180° around the main axis passing through the CYHN:C1 and CYHN:C4 atoms. The resulting geometry was reoptimized at the QM(B3LYP/SVP)/CHARMM level. During this minimization, cyclohexanone retained its new orientation and did not rotate back, but the total energy of the optimized minimum structure (Supporting Information (SI), Figure S1) was 11 kcal/mol higher than that of the previously optimized complex with cyclohexanone in its original orientation. In addition, we checked the energy required for converting the favored reactant complex (Figure 4) into the alternative binding mode by constrained potential energy scans at the OM3/CHARMM level: the resulting barrier for rotation of the cyclohexanone moiety around the C1−C4 axis is more than 25 kcal/mol and thus prohibitively high in the enzyme environment. As delineated later (section 3.5), these findings are the key to explaining the source of enantioselectivity in the B−V reaction of **1**.

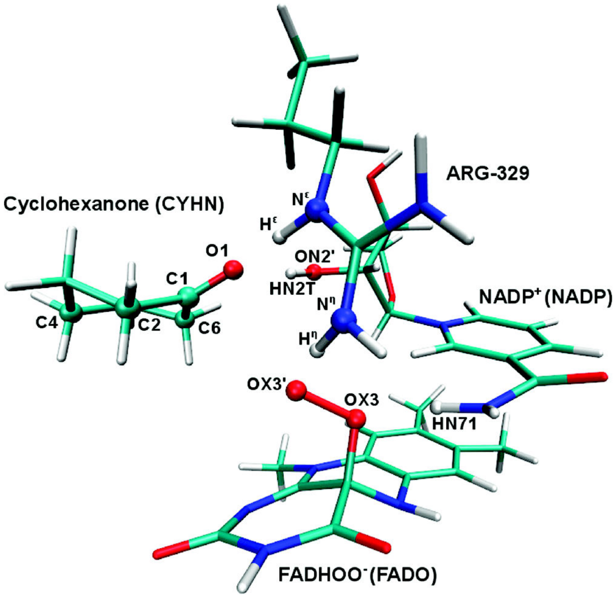


Figure 4. Enzyme−reactant complex (QM region only) optimized at the QM(B3LYP/TZVP)/CHARMM level with definition of atom labels used in the text. Residue names are given in parentheses. For characteristic interatomic distances, see text and Table 1.

Table 1. Characteristic Interatomic Distances in QM(B3LYP/TZVP)/CHARMM Optimized Geometries (in Å)[a]

	reactant	intermediate	TS2	product
CYHN:C1-FADO:OX3′	3.41	1.61	1.38	1.32
CYHN:O1-Arg329:H$^\varepsilon$	1.91	1.60	1.62	1.75
CYHN:O1-NADP:HN2T	1.77	1.53	1.58	1.69
FADO:OX3′-Arg329:H$^\eta$	1.57	1.98	1.97	3.46
FADO:OX3-NADP:HN71	1.88	2.22	2.09	1.80

[a]For the definition of the atomic labels, see Figure 4.

The mechanistic studies at the QM/MM level started from the enzyme−reactant complex. The energy values and interatomic distances given in this section were obtained using the B3LYP functional and, unless noted otherwise, the TZVP basis set as QM treatment in the DFT/CHARMM calculations. The specific atom labels used in the text are defined in Figure 4. Whenever necessary they will be specified in the format Resname:AtomName, for example FADO:OX3′ for atom OX3′ in FADHOO$^-$. Some characteristic interatomic distances in the stationary points along the reaction path are listed in Table 1.

3.1. Addition Step and Criegee Intermediate.

In the optimized enzyme−reactant complex (Figure 4), cyclohexanone and FADHOO$^-$ are still far apart, with a distance of 3.41 Å between the carbonyl carbon of cyclohexanone (CYHN:C1) and the distal oxygen of the FADHOO$^-$ peroxy group (FADO:OX3′). As noted above, the cyclohexanone oxygen atom (CYHN:O1) is oriented toward hydrogen atoms at Arg-329 (H$^\varepsilon$) and the NADP$^+$ ribose 2′-hydroxyl group (NADP:HN2T). The corresponding distances (1.91 and 1.77 Å, respectively) are short and indicate hydrogen bonding. Otherwise cyclohexanone is surrounded by the hydrophobic residues of the binding pocket. One of the NADP$^+$ nicotinamide hydrogen atoms (NADP:HN71) forms a hydrogen bond with the proximal oxygen of the FADHOO$^-$ peroxy group (FADO:OX3) (1.88 Å), consistent with experimental evidence that NADP$^+$ stabilizes the C4a-peroxyflavin, while the distal oxygen of the FADHOO$^-$ peroxy group is close to the Arg-329 H^{η} atom (1.57 Å).

Starting from the optimized enzyme−reactant complex (Figure 4), a potential energy scan was performed for the addition step. The reaction coordinate was defined as the distance between the carbonyl carbon of cyclohexanone and the distal oxygen of the FADHOO$^-$ peroxy group, $d_{C1-OX3'}$, since there is a covalent bond between these atoms in the Criegee intermediate. During the scan, this distance was successively decreased in steps of 0.1 Å until the value of 1.6 Å was reached. The resulting QM/MM potential energy curve (see Figure 5 for QM = B3LYP/TZVP and SI, Figure S2 for QM = B3LYP/SVP) is relatively smooth, with a maximum of about 8.7 kcal/mol at $d_{C1-OX3'}$ close to 1.9 Å. Beyond this point, the energy goes down again because of the formation of the Criegee intermediate. There is some minor unevenness in the computed energy profile arising from the limited precision of the geometry optimizations, particularly with regard to structural features governed by weak non-covalent interactions.

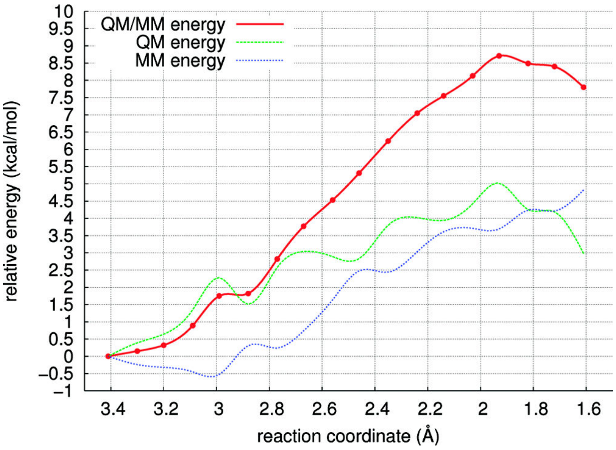


Figure 5. Energy profile for the addition step computed at the QM(B3LYP/TZVP)/CHARMM level using the reaction coordinate $d_{C1-OX3'}$.

Figure 5 also shows the decomposition of the QM/MM energy into QM and MM contributions (with the QM energy including the electrostatic interactions between QM atoms and MM point charges). The MM curve shows an overall rise, with some fluctuations (by typically less than 1 kcal/mol). Since all atoms directly involved in the formation of the Criegee intermediate are part of the QM region, and since there are no significant motions in the binding pocket during the scan (except for the cyclohexanone getting closer to the C4a-peroxyflavin peroxy group and undergoing a slight rotation), the rise of MM energy is probably due to the weakening of the van der Waals interactions between cyclohexanone and the active-site MM residues. The QM energy fluctuates around ∼4 kcal/mol (relative to the reactant complex) at distances between 2.3 and 1.9 Å and then drops slightly. As a consequence of compensating changes in the QM and MM energies, the overall QM/MM energy profile is rather flat in this region, and a TS search starting from its highest point actually did not converge. The final point from the scan at a distance of 1.6 Å is only slightly lower in energy than the highest point, but when subjected to unconstrained minimization, it retains the covalent C1-OX3′ bond and yields the Criegee intermediate. During this minimization, the geometry changes only very slightly.

We also carried out single-point QM(B3LYP-D2/TZVP)/CHARMM calculations with empirical dispersion corrections at geometries taken from the QM(B3LYP/SVP)/CHARMM scan (see SI, Figure S3 for the energy profile) and at the corresponding optimized stationary points. In this case, the QM energy decreased from the reactant complex to the intermediate, with no intervening maximum, because of the inclusion of the attractive dispersion interactions. However, because of the concomitant increase in the MM energy, there is still a barrier such that unconstrained reoptimization of the reactant complex at the QM(B3LYP-D2/TZVP)/CHARMM level essentially retained its structure and did not produce the Criegee intermediate. Given the well-organized active site of CHMO, we did not consider it necessary to perform further reoptimizations at this level, thus assuming that the essential dispersion effects on the energy profile can be captured by single-point QM(B3LYP-D2/TZVP)/CHARMM calculations, in line with previous findings.[44]

In an attempt to follow the dynamics of the addition step, we performed QM/CHARMM MD simulations using the semiempirical OM3 method for the QM part and starting from the reactant complex. After heating the system up to 300 K, the Criegee intermediate was formed spontaneously after about 3.5 ps and remained stable thereafter for the following 5.5 ps of the MD run. Figure 6 shows the variation of several key distances during the MD simulation. The distance $d_{C1-OX3'}$ fluctuates around 3.3 Å during the first 3.2 ps and then quickly shrinks to the covalent bond distance of 1.5 Å which is retained in the remaining time. These dynamics results suggest that there should exist a low-lying TS for the addition step at the OM3/CHARMM level. This TS could indeed be optimized properly on the OM3/CHARMM potential energy surface, as well as the reactant complex and the Criegee intermediate (see SI and Figure S4 for detailed results). The computed OM3/CHARMM barrier for the addition step is 5.7 kcal/mol, which is reasonably close to the value of about 8.7 kcal/mol obtained from the QM(B3LYP/TZVP)/CHARMM energy profile (Figure 5). Free energy calculations using thermodynamic integration lower the OM3/CHARMM barrier for

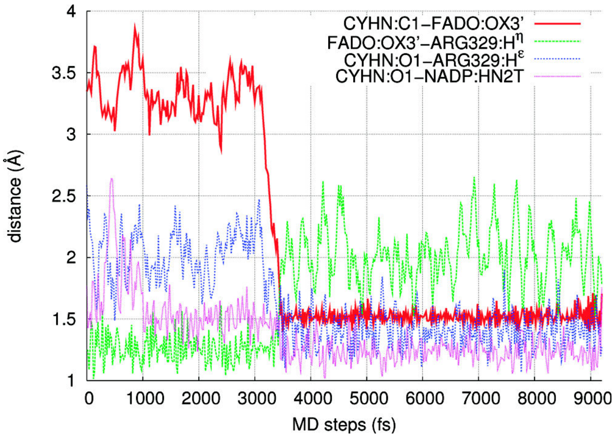


Figure 6. Changes in selected interatomic distances during the OM3/CHARMM MD simulation. For the definition of the atomic labels, see Figure 4.

addition to 4.3 kcal/mol, thus making it feasible to directly observe this reaction in the OM3/CHARMM MD runs.

Contrary to previous theoretical work[17d] on the gas-phase acid-catalyzed B−V oxygenation of cyclohexanone, the carbonyl oxygen of cyclohexanone is not protonated in the Criegee intermediate within the enzyme (see Figure 7). However, it is in close proximity to the hydrogen atoms at Arg-329 (H$^{\varepsilon}$, 1.60 Å) and the NADP$^+$ ribose 2′-hydroxyl group (1.53 Å). These two hydrogen bonds are strong enough to stabilize the intermediate and to make the carbonyl carbon atom more electrophilic. As a further check on the stability of the deprotonated Criegee intermediate, we performed a QM(B3LYP/TZVP)/CHARMM potential energy scan for the proton transfer from Arg-329 to the carbonyl oxygen, using as reaction coordinate the difference of distances in the bonds being formed and broken, $d_{\mathrm{O1-H}\varepsilon}-d_{\mathrm{H}\varepsilon-\mathrm{N}\varepsilon}$. The QM/MM energy was rising monotonously in this scan (up to 4.6 kcal/mol at the last point), indicating that this proton transfer does not occur spontaneously in the enzyme.

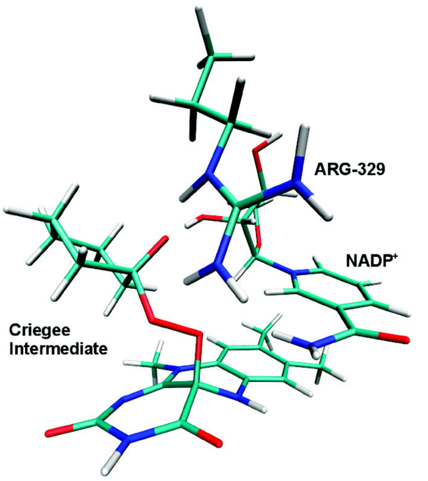


Figure 7. Criegee intermediate (QM region only) optimized at the QM(B3LYP/TZVP)/CHARMM level. For characteristic interatomic distances, see text and Table 1.

For reasons that will become clear when addressing the migration step (section 3.2), we also considered the alternative substrate binding mode (SI, Figure S1) and performed an OM3/CHARMM potential energy scan for the addition step leading to the respective Criegee intermediate. This turned out to be less favorable: in order to undergo nucleophilic C−O bond formation with the peroxy group, cyclohexanone must undergo an energetically costly conformational change such that the carbonyl oxygen atom points in the proper direction (i.e., the same as in the originally discussed approach, Figure 4). Importantly, the Criegee intermediate obtained in this manner contains cyclohexanone in a boat-like conformation and is 4.6 kcal/mol (OM3/CHARMM) higher in energy than the original one (Figure 7). It thus appears that the preferred binding mode of cyclohexanone generally involves an orientation as shown in Figure 4, and the energetically favored pathway will then be the one with cyclohexanone in its most stable chairlike conformation.

Our data support experimental evidence that NADP$^+$ and Arg-329 play a crucial role in stabilizing the C4a-peroxyflavin and the Criegee intermediate. In the initial phase of the addition reaction, there are strong hydrogen bonds between the nicotinamide hydrogen atom and the proximal oxygen of the FADHOO$^-$ peroxy group as well as between the 2′-hydroxyl group of the adjacent ribose group and the carbonyl oxygen of cyclohexanone. The latter stabilizing interaction is retained in the Criegee intermediate, whereas the former is diminished since the nicotinamide ring amino group shifts away (up to 2.2 Å) from the proximal oxygen of the FADHOO$^-$ peroxy group as soon as the intermediate is formed (probably due to the reduced nucleophilicity of the peroxy group).

3.2. Migration Step. In the Criegee intermediate arising from the energetically preferred orientation of cyclohexanone in the binding pocket, the C1−C2 bond of cyclohexanone is antiperiplanar to the peroxy bond of the C4a-peroxyflavin, with the FADO:OX3-OX3′-CYHN:C1−C2 dihedral angle being close to 180°. Because of the stereoelectronic requirement, the migration step of the B−V reaction is expected to involve a bond antiperiplanar to the peroxy bond.[45] Therefore a QM(B3LYP/SVP)/CHARMM potential energy scan was performed using as reaction coordinate the distance between the distal oxygen atom of the peroxy group and the C2 atom of cyclohexanone, $d_{\mathrm{OX3'-C2}}$, which was successively decreased in steps of 0.05 Å until the value of 1.5 Å was reached. During this scan, the QM/MM energy of the system first rises fast and then at some point decreases dramatically (see SI, Figure S5). This sudden drop is associated with the insertion of the distal oxygen from the peroxy group into the cyclohexanone ring and the dissociation of the product from the isoalloxazine ring. The following smooth decrease of the QM/MM energy along the reaction coordinate reflects the gradual approach of the oxygen−carbon distance in the ε-caprolactone ring to its equilibrium value.

A TS search starting from the maximum-energy geometry was successful. In the resulting QM(B3LYP/TZVP)/CHARMM transition structure (see Figure 8), characteristic distances are those between the two former peroxy group oxygen atoms (1.83 Å) and between the distal oxygen of the peroxy group and the C2 atom of cyclohexanone (2.19 Å). The energy barrier to migration is computed to be 6.7 kcal/mol at this level. To verify the type of TS, we distorted its geometry slightly along the transition vector in both directions and

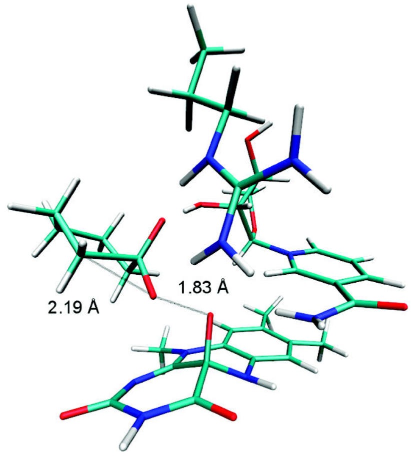


Figure 8. Transition state for the migration step (QM region only) optimized at the QM(B3LYP/TZVP)/CHARMM level. For characteristic interatomic distances, see Table 1.

thereafter performed careful energy minimizations (with smaller maximum step sizes than usual) which led to the Criegee intermediate in one direction and the product (see Figure 9) in the other direction (as expected). The optimized enzyme−product complex is 81.4 kcal/mol lower in energy than the Criegee intermediate, and the overall reaction is highly exothermic with a reaction energy of −66.9 kcal/mol at the QM(B3LYP/TZVP)/CHARMM level. The overall energy profile is depicted in Figure 10, and the energies of all stationary points are given in Table 2 at different levels of theory. We note that the two QM approaches that include midrange dispersion (B3LYP-D2 and M06-2X) yield lower relative energies for the first transition state (TS1) and the Criegee intermediate (compared with B3LYP/TZVP). For the migration step (TS2), the barrier from QM(M06-2X/TZVP)/CHARMM is higher than that from QM(B3LYP/TZVP)/CHARMM, which may correct for the known general tendency of B3LYP to underestimate barriers; however, relative to the reactant complex, both approaches yield a similar energy for TS2 (14.3 vs 12.8 kcal/mol). Finally, all methods predict the overall enzymatic reaction to be very exothermic, with reaction energies between −66.5 and −70.4 kcal/mol.

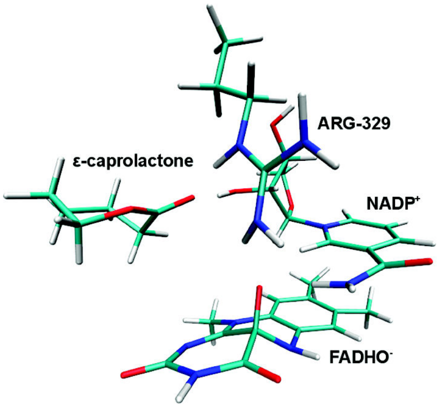


Figure 9. Enzyme−product complex (QM region only) optimized at the QM(B3LYP/TZVP)/CHARMM level. For characteristic interatomic distances, see Table 1.

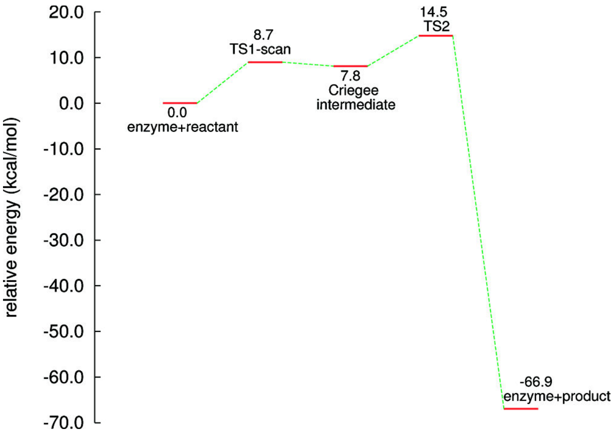


Figure 10. QM(B3LYP/TZVP)/CHARMM energy profile.

Table 2. Calculated QM/MM Energies (in kcal/mol) for Stationary Points along the Reaction Profile Relative to the Reactant Complex, Obtained with Different QM Methods

	reactant	TS1 scan	intermediate	TS2	product
B3LYP/SVP//B3LYP/SVP	0.0	3.6	0.8	9.3	−69.3
B3LYP/TZVP//B3LYP/SVP	0.0	7.9	7.6	14.3	−66.5
B3LYP/TZVP//B3LYP/TZVP	0.0	8.7	7.8	14.5	−66.9
B3LYP-D2/TZVP//B3LYP/SVP	0.0	−0.3	−4.0	3.4	−68.8
M06−2X/TZVP//B3LYP/SVP	0.0	1.7	−6.2	12.8	−70.4

To confirm that the migration step will preferably proceed as described (with the bond antiperiplanar to the peroxy bond being the one to migrate), another potential energy scan was performed, taking as reaction coordinate the distance between the distal oxygen of the peroxy group and the other carbon atom (C6) connected to the carbonyl carbon of cyclohexanone, $d_{\mathrm{OX3'-C6}}$. The resulting QM(B3LYP/SVP)/CHARMM energy profile had a very high maximum of about 26 kcal/mol. Moreover, during a subsequent TS search, the distal oxygen from the peroxy group flipped, assuming an orientation close to the one adopted in the potential energy scan along the $d_{\mathrm{OX3'-C2}}$ reaction coordinate. We thus conclude that the structure of the Criegee intermediate completely determines the preferred mode of migration. This implies that the energetically preferred binding mode of cyclohexanone (Figure 4) sets the stage for all subsequent molecular events. If the alternative binding mode (SI, Figure S1) were to be followed in the subsequent steps, then the other carbon atom (C6) would participate in the migration. Whereas C2 and C6 are equivalent by symmetry in the free cyclohexanone molecule, this is not so when it is complexed in the chiral binding pocket. As shown in section 3.5, this discrimination between C2 and C6 is crucial in determining the sense of stereoselectivity in the oxidative desymmetrization of **1**.

3.3. Gas-Phase Calculations.

Generally speaking, the protein environment in CHMO promotes the B−V reaction by defining the spatial orientation of the two cofactors and by properly placing the catalytic Arg-329 side chain within the active site. To check the influence of the MM region on the reaction mechanism in more detail, we performed pure QM gas-phase reoptimizations of the QM region for the relevant minima, transition states, and reaction paths at the B3LYP/TZVP level, starting from the corresponding QM/MM geometries. In these calculations, we (partially) retained the overall spatial arrangement within the QM region by keeping the Cartesian coordinates of all atoms fixed that are covalently bound to MM atoms in the QM/MM setup.

The most significant difference between the reactant complex in the gas phase and in the enzyme is a proton transfer from Arg-329 to FADHOO$^-$ that happens during the gas-phase reoptimization of the QM region. In the enzyme, Arg-329 is protonated (with a covalent H^{η}—N^{η} bond), and there is a short hydrogen bond between OX3′ and H^{η}, whereas in the gas-phase reoptimization, H^{η} quickly moves from N^{η} to the OX3′ atom in FADHOO$^-$ (with $d_{H^{\eta}-N^{\eta}}$ = 1.57 Å in the equilibrium structure). These gas-phase results are at odds with the experimental evidence for the enzyme where FADHOO$^-$ remains deprotonated in its active state. Apparently the protein environment stabilizes the deprotonated form of FADHOO$^-$ in the reactant complex, presumably via the two strong hydrogen bonds between Arg-329 and Asp-59 which make the nitrogen atoms of the arginine guanidinium group more nucleophilic. Furthermore, the overall geometry of the reactant complex changes substantially upon gas-phase reoptimization because the isoalloxazine and nicotinamide rings rotate (being repelled by each other) and reorient with respect to the cyclohexanone ring, such that the FADO:OX3-OX3′-CYHN:C1−C2 dihedral angle decreases from 170° to 100°; however, cyclohexanone remains hydrogen bonded via the carbonyl oxygen atom with Arg-329 (H^{ε}) and with the NADP$^+$ ribose 2′-hydroxyl group (which are part of the QM region).

The structure of the Criegee intermediate complex changes much less upon gas-phase reoptimization, and the Criegee intermediate itself remains deprotonated and retains its anionic character. The gas-phase reaction from the reactant to the intermediate thus requires significant conformational changes: to enable the formation of the covalent bond between CYHN:C1 and FADO:OX3′ the proton must first be transferred back to Arg-329:N^{η}, and in the Criegee intermediate the FADO:OX3-OX3′-CYHN:C1−C2 dihedral angle must be close to 180° (which is crucial for the reaction to proceed further). This causes the energy barrier for the addition step to be much higher (18 kcal/mol) than in the enzyme. The gas-phase Criegee intermediate lies 12 kcal/mol above the reactant complex, and hence 6 kcal/mol below the TS for its formation. The enzyme thus lowers the energy barrier for the addition step not only by constraining the spatial orientation of the cofactors, but also by allowing the cyclohexanone to be bound in a reactant conformation with the FADO:OX3-OX3′-CYHN:C1−C2 dihedral angle being close to 180° and by stabilizing the deprotonated form of the C4a-peroxyflavin.

The energy barrier for the migration step is also higher in the gas phase than in the enzyme (12 vs 6.7 kcal/mol) even though the corresponding TSs are geometrically not very different. For example, the key distances in the breaking OX3-OX3′ bond and the forming OX3′−C2 bond are 1.93 and 2.17 Å in the gas-phase TS (compared with 1.83 and 2.19 Å in the enzyme, respectively), and it is thus difficult to pinpoint geometrical factors that would account for the lower barrier in the enzyme. The overall reaction energies are rather similar in the gas phase and in the enzyme (−64 vs −66.9 kcal/mol).

3.4. Comparison with Reaction in Solution.

According to our calculations, the mechanism of the CHMO-catalyzed B−V oxidation of cyclohexanone differs from that of the acid-catalyzed reaction in solution that has been described previously.[17d] In the latter case, the presence of a strong acid like trifluoroacetic acid (TFAA) causes protonation of the carbonyl oxygen during the addition step, which increases the electrophilicity of the carbonyl carbon atom and thus facilitates the addition of the peroxyacid (which is at the same time deprotonated by the same TFA that protonates the substrate in a concerted fashion).[17d] Hence this acid-catalyzed solution reaction generates a neutral Criegee intermediate, contrary to the situation in the enzyme where, according to our QM/MM calculations, the carbonyl oxygen is not protonated although doubly H-bonded, and the Criegee intermediate has pronounced anionic character. We note in this context that both the protonated arginine and the hydroxyl group are weak acids with pK_a values higher than 10 and that the hydrogen bonds, which they form in CHMO with the carbonyl oxygen, seem strong enough to make the carbonyl carbon sufficiently electrophilic. As mentioned elsewhere,[16] the anionic nature of the Criegee intermediate is expected to accelerate the migration step, and one important role of the enzyme is thus the effective stabilization of this kind of Criegee adduct, which allows for a faster overall reaction compared with an acidic solution.

3.5. Unveiling the Source of Enantioselectivity of CHMO.

After exploring the mechanism of the CHMO-catalyzed B−V reaction for the parent substrate cyclohexanone, we now turn to a substituted substrate, namely **1**, and study the stereoselectivity of its reaction. It is experimentally known that CHMO (*Rhodococcus*) as well as CHMO (*Acinetobacter*) react with this substrate with nearly perfect (*S*)-enantioselectivity providing (*S*)-**2** with >96% ee (Figure 2). Chemists have long speculated about the source of enantioselectivity in BVMO-catalyzed transformations (including the asymmetric **1** → **2** transformation).[5,46,47] The traditional stereoelectronic requirement regarding anti-periplanarity in the cleavage (fragmentation) of the Criegee intermediate, known to be essential in nonenzymatic reactions, formed a common basis in all attempts to develop a model. Then, upon considering the empirical results of BVMO-catalyzed asymmetric transformations of numerous cyclohexanone derivatives, appropriate models were proposed by Furstoss,[46a] Ottolina and Colonna,[46b] Kelly,[46c] Stewart,[46d,47] and Kayser.[5,7] Especially the “diamond lattice model”, in which the substrate is assumed to have the chair conformation with substituents in the equatorial position, has proven to be a useful mnemonic aid with notable predictive power.[5,7,46d,47] We are now in the position to deliver the underlying explanation for these models on the molecular level. Making the reasonable assumption that the 4-methyl group adopts the equatorial position in the cyclohexanone chair conformation in the preferred binding mode (Figure 4), our optimized structure of the Criegee intermediate (Figure 7) immediately suggests exclusive formation of the (*S*)-lactone product by energetically favored selective migration of C2 rather than C6.

To verify this qualitative prediction, we have built the two 4-methyl-substituted Criegee intermediates by manually adding a methyl group into the optimized geometry of the original Criegee intermediate. The resulting structures were fully reoptimized at the QM(B3LYP/TZVP)/CHARMM level, followed by energy scans and TS searches for the migration step. As expected, the energies of the Criegee intermediate and the TS are lower when an equatorial rather than an axial 4-methyl group is present, by 4.8 and 4.5 kcal/mol, respectively. For an isolated molecule of **1**, gas-phase B3LYP/TZVP calculations favor the equatorial over the axial conformer by 1.9 kcal/mol (energy difference including zero-point vibrational corrections), close to the experimental value of 2.1 kcal/mol for the enthalpy difference between the two conformers.[48] The preference for the equatorial form is thus reinforced both in the Criegee intermediate and the subsequent TS in CHMO. While the energy barriers to migration are quite similar within each conformer, the equatorial pathway is favored overall by an energy difference of 4.5 kcal/mol at the crucial TS, and hence the (*S*)-lactone product will be formed with high enantioselectivity.

Our current mechanistic scenario for the mechanism of the CHMO-catalyzed B−V reaction thus reproduces and explains on a molecular level the observed (*S*)-enantioselectivity for the chiral substrate **1** very well. This pertains to CHMO (*Rhodococcus*) and other homologous CHMOs such as the most used CHMO (*Acinetobacter*). On this basis it should thus be possible to make qualitative predictions concerning the enantioselectivity for other substituted cyclohexanone derivatives and the effect of active-site mutations. One such example is the B−V oxidation of 4-hydroxycyclohexanone by CHMO (*Acinetobacter*) which changes from marginal (*R*)-enantioselectivity (9% ee) in the wild-type enzyme to good (*S*)-enantioselectivity (79% ee) in the Phe432Ser mutant.[7] Our model immediately suggests an explanation for this drastic change in terms of the orientation of the substrate in the binding site and the formation of a hydrogen bond in the mutant between Ser432 and the 4-hydroxy group of the substrate, which will lead to more favorable binding when the latter is in an equatorial rather than an axial position. We are currently studying the B−V oxidation of this substrate by QM/MM calculations to confirm the qualitative conclusions derived from our model and to arrive at quantitative assessments. The results of this work will be reported elsewhere.

4. CONCLUSION

We have reported the first computational investigation of the Baeyer−Villiger reaction proceeding in a native enzyme environment. Reaction pathways for cyclohexanone oxidation by cyclohexanone monooxygenase from *Rhodococcus* (CHMO) were determined as well as the structures of the Criegee intermediate and the transition states. Judging from the QM(B3LYP/TZVP)/CHARMM results, the initial addition reaction has to overcome an energy barrier of about 8.7 kcal/mol, which generates the Criegee intermediate that occupies a shallow minimum on the potential energy surface and can rearrange to the lactone product by surmounting a second energy barrier of 6.7 kcal/mol. The latter transition state for the migration step is the highest point on the energy profile. The CHMO-catalyzed reaction proceeds via a Criegee intermediate of anionic character and is highly exothermic, with a computed overall reaction energy of −66.9 kcal/mol. The QM/MM calculations confirm the crucial role of the Arg-329 residue and the NADP$^+$ cofactor for the catalytic efficiency of CHMO. The experimentally observed enantioselectivity toward **1** with essentially complete formation of the lactone (*S*)-**2** has been rationalized for the first time on a molecular basis by the current QM/MM calculations. It is the selective binding mode of the substrate in the chair conformation with the methyl-substituent in the equatorial position that places it in a chiral environment in a way that only one of the two enantiotopic C-atoms can migrate. This physical model also serves as a guide for interpreting the stereochemical results of other synthetically interesting transformations catalyzed by CHMO and other homologues such as the most used CHMO from *Acinetobacter* sp. NCIMB 9871.[4] Finally, this work provides insight into the possible effects of mutations aimed at tuning the enantio- and regioselectivity of CHMO and other BVMOs by rational design or directed evolution.[6−8]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figure of reactant complex with different substrate orientation, additional energy profiles, detailed results from OM3/CHARMM molecular dynamics, sequence alignment of CHMO (*Rhodococcus*) and CHMO (*Acinetobacter*), and complete refs 21 22b, and 29. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

thiel@mpi-muelheim.mpg.de

■ ACKNOWLEDGMENTS

The authors thank Prof. Albert M. Berghuis for providing the Cartesian coordinates of their X-ray structure[14] before depositing them (PDB code 3GWD).

■ REFERENCES

(1) Baeyer, A.; Villiger, V. *Ber. Dtsch. Chem. Ges.* **1899**, *32*, 3625−3633.

(2) (a) Bolm, C.; Schlingloff, G.; Weickhardt, K. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 1848−1849. (b) Gusso, A.; Baccin, C.; Pinna, F.; Strukul, G. *Organometallics* **1994**, *13*, 3442−3451. (c) Strukul, G. *Angew. Chem., Int. Ed.* **1998**, *37*, 1199−1209. (d) Uchida, T.; Katsuki, T. *Tetrahedron Lett.* **2001**, *42*, 6911−6914. (e) Ito, K.; Ishii, A.; Kuroda, T.; Katsuki, T. *Synlett* **2003**, 643−646. (f) Frison, J.-C.; Palazzi, C.; Bolm, C. *Tetrahedron* **2006**, *62*, 6700−6706. (g) Colladon, M.; Scarso, A.; Strukul, G. *Synlett* **2006**, 3515−3520.

(3) (a) Murahashi, S.; Ono, S.; Imada, Y. *Angew. Chem., Int. Ed.* **2002**, *41*, 2366−2368. (b) Imada, Y.; Iida, H.; Murahashi, S.; Naota, T. *Angew. Chem., Int. Ed.* **2005**, *44*, 1704−1706. (c) Wang, B.; Shen, Y.-M.; Shi, Y. *J. Org. Chem.* **2006**, *71*, 9519−9521. (d) Peris, G.; Miller, S. J. *Org. Lett.* **2008**, *10*, 3049−3052. (e) Malkov, A. V.; Friscourt, F.; Bell, M.; Swarbrick, M. E.; Kocovsky, P. *J. Org. Chem.* **2008**, *73*, 3996−4003. (f) Xu, S.; Wang, Z.; Zhang, X.; Zhang, X.; Ding, K. *Angew. Chem., Int. Ed.* **2008**, *47*, 2840−2843.

(4) For reviews see: (a) Mihovilovic, M.; Muller, B.; Stanetty, P. *Eur. J. Org. Chem.* **2002**, 3711−3730. (b) Wohlgemuth, R. *Eng. Life Sci.* **2006**, *6*, 577−583. (c) Rehdorf, J.; Bornscheuer, M. *Monooxygenases, Baeyer Villiger Applications in Organic Synthesis*; John Wiley & Sons: Hoboken, NJ, 2009. (d) de Gonzalo, G.; Mihovilovic, M. D.; Fraaije, M. W. *ChemBioChem* **2010**, *11*, 2208−2231. (e) Pazmino, D. E. T.; Dudek, H. M.; Fraaije, M. W. *Curr. Opin. Chem. Biol.* **2010**, *14*, 138−144. (f) Leisch, H.; Morley, K.; Lau, P. *Chem. Rev.* **2011**, *111*, 4165−4222.

(5) Kayser, M. M. *Tetrahedron* **2009**, *65*, 947−974.

(6) Review of directed evolution of enantioselective enzymes as catalysts in organic chemistry: Reetz, M. T. *Angew. Chem., Int. Ed.* **2011**, *50*, 138−174.

(7) Kayser, M. M.; Clouthier, C. M. *J. Org. Chem.* **2006**, *71*, 8424−8430.

(8) (a) Reetz, M.; Brunner, B.; Schneider, T.; Schulz, F.; Clouthier, C.; Kayser, M. *Angew. Chem., Int. Ed.* **2004**, *43*, 4075−4078. (b) Bocola, M.; Schulz, F.; Leca, F.; Vogel, A.; Fraaije, M.; Reetz, M. *Adv. Synth. Catal.* **2005**, *347*, 979−986. (c) Mihovilovic, M.; Rudroff, F.; Winninger, A.; Schneider, T.; Schulz, F.; Reetz, M. *Org. Lett.* **2006**, *8*, 1221−1224. (d) Clouthier, C. M.; Kayser, M. M.; Reetz, M. T. *J. Org. Chem.* **2006**, *71*, 8431−8437. (e) Reetz, M. T.; Wu, S. *J. Am. Chem. Soc.* **2009**, *131*, 15424−15432. (f) Wu, S.; Acevedo, J. P.; Reetz, M. T. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 2775−2780. (g) Kirschner, A.; Bornscheuer, U. T. *Angew. Chem., Int. Ed.* **2006**, *45*, 7004−7006. (h) Kirschner, A.; Bornscheuer, U. T. *Appl. Microbiol. Biotechnol.* **2008**, *81*, 465−472. (i) Rehdorf, J.; Mihovilovic, M. D.; Bornscheuer, U. T. *Angew. Chem., Int. Ed.* **2010**, *49*, 4506−4508.

(9) (a) Turfitt, G. *Biochem. J.* **1948**, *42*, 376−383. (b) Donoghue, N.; Norris, D.; Trudgill, P. *Eur. J. Biochem.* **1976**, *63*, 175−192. (c) Chen, Y.; Peoples, O.; Walsh, C. *J. Bacteriol.* **1988**, *170*, 781−789.

(10) (a) Ryerson, C.; Ballou, D.; Walsh, C. *Biochemistry* **1982**, *21*, 2644−2655. (b) Walsh, C.; Chen, Y. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 333−343. (c) Brzostowicz, P.; Walters, D.; Thomas, S.; Nagarajan, V.; Rouviere, P. *Appl. Environ. Microbiol.* **2003**, *69*, 334−342.

(11) Sheng, D.; Ballou, D.; Massey, V. *Biochemistry* **2001**, *40*, 11156−11167.

(12) Pazmino, D. E. T.; Baas, B.-J.; Janssen, D. B.; Fraaije, M. W. *Biochemistry* **2008**, *47*, 4082−4093.

(13) Orru, R.; Dudek, H. M.; Martinoli, C.; Pazmino, D. E. T.; Royant, A.; Weik, M.; Fraaije, M. W.; Mattevi, A. *J. Biol. Chem.* **2011**, *286*, 29284−29291.

(14) Mirza, I. A.; Yachnin, B. J.; Wang, S.; Grosse, S.; Bergeron, H.; Imura, A.; Iwaki, H.; Hasegawa, Y.; Lau, P. C. K.; Berghuis, A. M. *J. Am. Chem. Soc.* **2009**, *131*, 8848−8854.

(15) Renz, M.; Meunier, B. *Eur. J. Org. Chem.* **1999**, 737−750.

(16) ten Brink, G.; Arends, I.; Sheldon, R. *Chem. Rev.* **2004**, *104*, 4105−4123.

(17) (a) Okuno, Y. *Chem.—Eur. J.* **1997**, *3*, 212−218. (b) Grein, F.; Chen, A.; Edwards, D.; Crudden, C. *J. Org. Chem.* **2006**, *71*, 861−872. (c) Alvarez-Idaboy, J.; Reyes, L.; Cruz, J. *Org. Lett.* **2006**, *8*, 1763−1765. (d) Alvarez-Idaboy, J. R.; Reyes, L.; Mora-Diez, N. *Org. Biomol. Chem.* **2007**, *5*, 3682−3689. (e) Alvarez-Idaboy, J. R.; Reyes, L. *J. Org. Chem.* **2007**, *72*, 6580−6583. (f) Mora-Diez, N.; Keller, S.; Alvarez-Idaboy, J. R. *Org. Biomol. Chem.* **2009**, *7*, 3682−3690.

(18) Malito, E.; Alfieri, A.; Fraaije, M.; Mattevi, A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13157−13162.

(19) (a) Gordon, J.; Myers, J.; Folta, T.; Shoja, V.; Heath, L.; Onufriev, A. *Nucleic Acids Res.* **2005**, *33*, W368−W371. (b) Anandakrishnan, R.; Onufriev, A. *J. Comp. Biol.* **2008**, *15*, 165−184.

(20) Li, H.; Robertson, A.; Jensen, J. *Proteins* **2005**, *61*, 704−721.

(21) MacKerell, A.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(22) (a) Brooks, B.; Bruccoleri, R.; Olafson, D.; States, D.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217. (b) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(23) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(24) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227−249.

(25) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198−1229.

(26) (a) Slater, J. *Phys. Rev.* **1953**, *91*, 528−530. (b) Vosko, S.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200−1211. (c) Becke, A. *Phys. Rev. A* **1988**, *38*, 3098−3100. (d) Becke, A. *J. Chem. Phys.* **1993**, *98*, 5648−5652. (e) Stephens, P.; Devlin, F.; Chabalowski, C.; Frisch, M. *J. Phys. Chem.* **1994**, *98*, 11623−11627. (f) Lee, C.; Yang, W.; Parr, R. *Phys. Rev. B* **1988**, *37*, 785−789.

(27) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580−10594.

(28) (a) de Vries, A.; Sherwood, P.; Collins, S.; Rigby, A.; Rigutto, M.; Kramer, G. *J. Phys. Chem. B* **1999**, *103*, 6133−6141. (b) Sherwood, P.; de Vries, A.; Collins, S.; Greatbanks, S.; Burton, N.; Vincent, M.; Hillier, I. *Farad. Disc.* **1997**, *106*, 79−92.

(29) Sherwood, P.; et al. *J. Mol. Struct. (Theochem)* **2003**, *632*, 1−28.

(30) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165−169.

(31) Smith, W.; Forester, T. *J. Mol. Graph.* **1996**, *14*, 136−141.

(32) Billeter, S.; Turner, A.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177−2186.

(33) Kästner, J.; Carr, J. M.; Keal, T. W.; Thiel, W.; Wander, A.; Sherwood, P. *J. Phys. Chem. A* **2009**, *113*, 11856−11865.

(34) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571−2577.

(35) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829−5835.

(36) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787−1799.

(37) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215−241.

(38) (a) Nocedal, J. *Math. Comp.* **1980**, *35*, 773−782. (b) Liu, D.; Nocedal, J. *Math. Programming* **1989**, *45*, 503−528.

(39) (a) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52−57. (b) Baker, J. *J. Comput. Chem.* **1986**, *7*, 385−395.

(40) (a) Nose, S. *J. Chem. Phys.* **1984**, *81*, 511−519. (b) Hoover, W. *Phys. Rev. A* **1985**, *31*, 1695−1697.

(41) (a) Scholten, M. Ph.D. thesis, Universität Düsseldorf, 2003. (b) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751−5755.

(42) Ryckaert, J.; Ciccotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327−341.

(43) Senn, H. M.; Thiel, S.; Thiel, W. *J. Chem. Theory Comput.* **2005**, *1*, 494−505.

(44) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. *J. Phys. Chem. Lett.* **2010**, *1*, 3232−3237.

(45) Krow, G. *Tetrahedron* **1981**, *37*, 2697−2724.

(46) (a) Alphand, V.; Furstoss, R. *J. Org. Chem.* **1992**, *57*, 1306−1309. (b) Ottolina, G.; Carrea, G.; Colonna, S.; Rückemann, A. *Tetrahedron: Asymmetry* **1996**, *7*, 1123−1136. (c) Kelly, D. R. *Tetrahedron: Asymmetry* **1996**, *7*, 1149−1152. (d) Stewart, J. *Curr. Org. Chem.* **1998**, *2*, 195−216.

(47) Stewart, J.; Reed, K.; Martinez, C.; Zhu, J.; Chen, G.; Kayser, M. *J. Am. Chem. Soc.* **1998**, *120*, 3541−3548.

(48) Potts, A.; Baer, T. *J. Mol. Struct. (Theochem)* **1997**, *419*, 11−18.

Supporting information for: QM/MM Study on the Mechanism of the Enzymatic Baeyer-Villiger Reaction

Iakov Polyak,[†] Manfred T. Reetz,[†,‡] and Walter Thiel[*,†]

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1,
D-45470 Mülheim an der Ruhr, Germany, and Philipps-Universität Marburg, Fachbereich
Chemie, Hans-Meerwein-Strasse, D-35032 Marburg, Germany

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

---

[*] To whom correspondence should be addressed

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

[‡] Philipps-Universität Marburg, Fachbereich Chemie, Hans-Meerwein-Strasse, D-35032 Marburg, Germany

Contents

1. Reactant complex with alternative substrate orientation

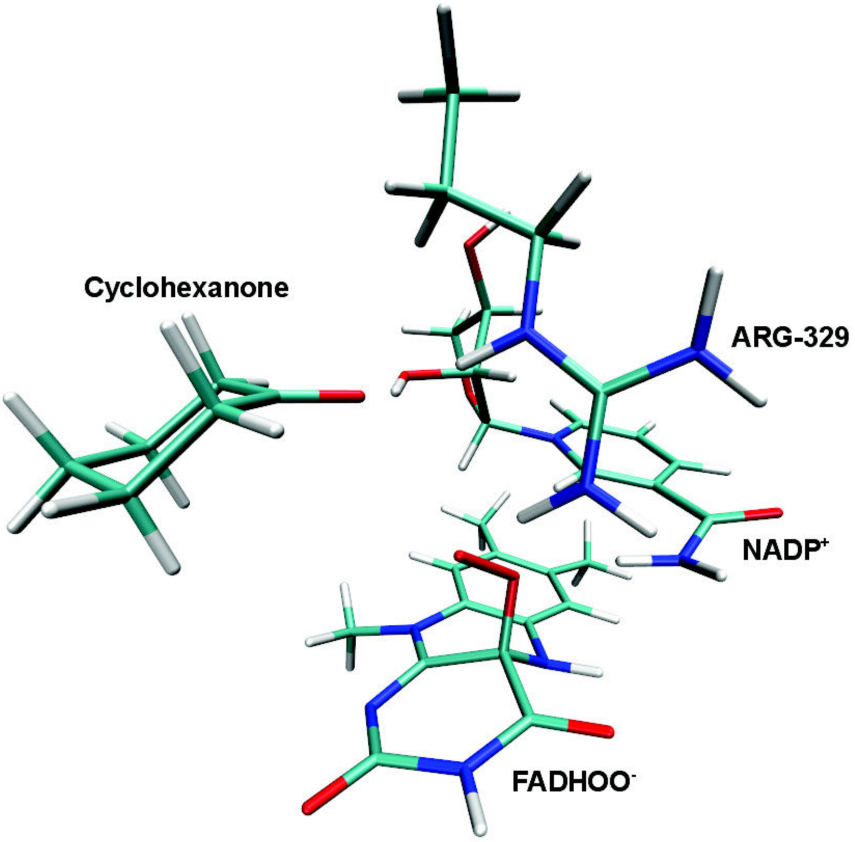


Figure S1: Enzyme-reactant complex (QM region only) optimized at the QM(B3LYP/TZVP)/CHARMM level with the cyclohexanone being flipped around its main axis by 180° (relative to the preferred orientation shown in Figure 4 of the main paper).

2. QM/MM energy profiles for the addition step

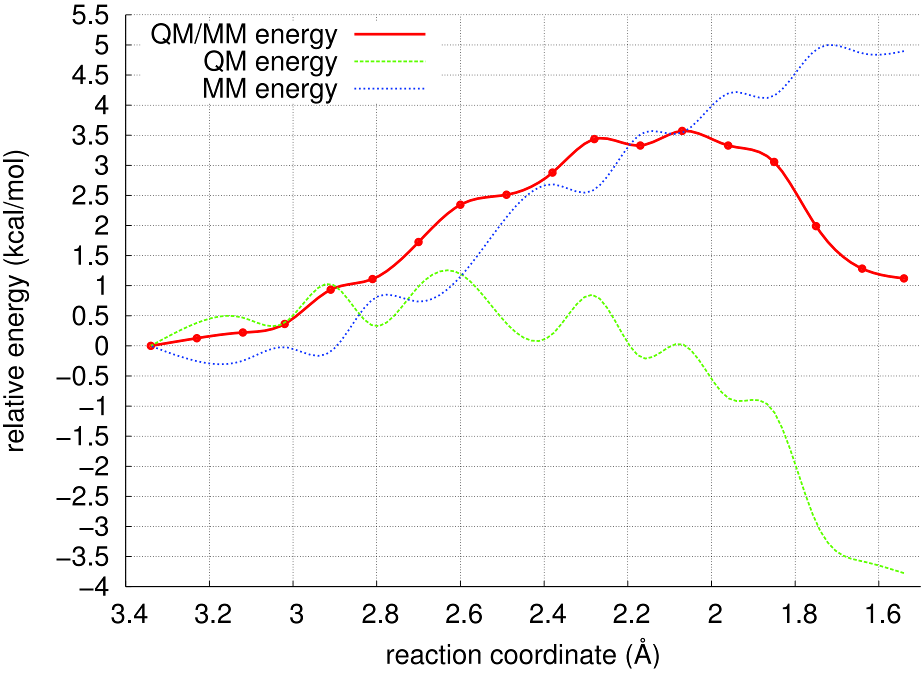


Figure S2: Energy profile for the addition step computed at the QM(B3LYP/SVP)/CHARMM level using the reaction coordinate $d_{C1-OX3'}$.

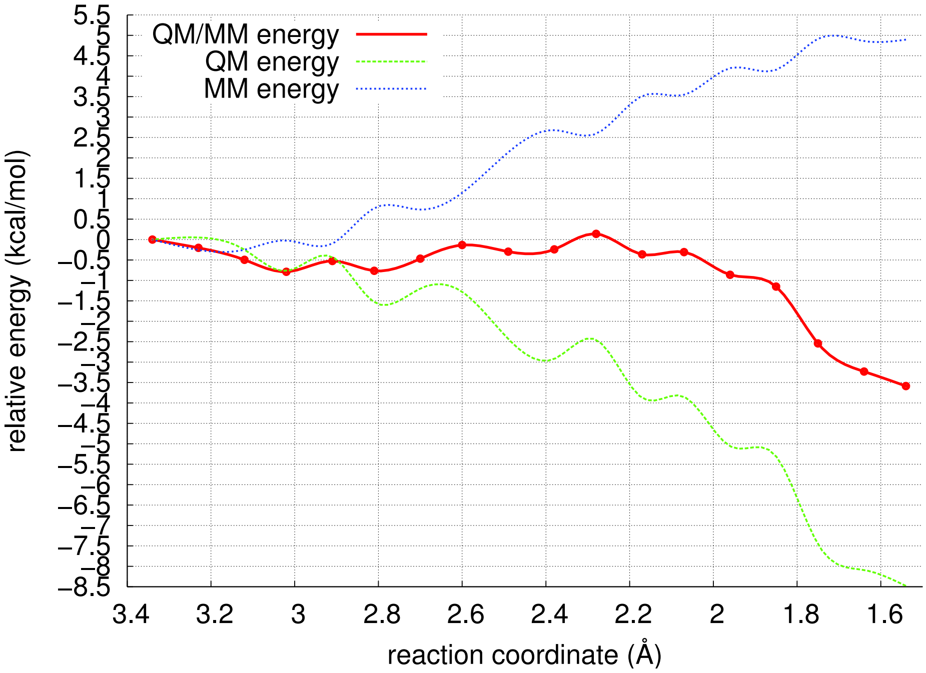


Figure S3: Dispersion-corrected single-point QM(B3LYP-D/TZVP)/CHARMM energies calculated at QM(B3LYP/SVP)/CHARMM optimized geometries for the addition step.

3. OM3/CHARMM calculations: Detailed results

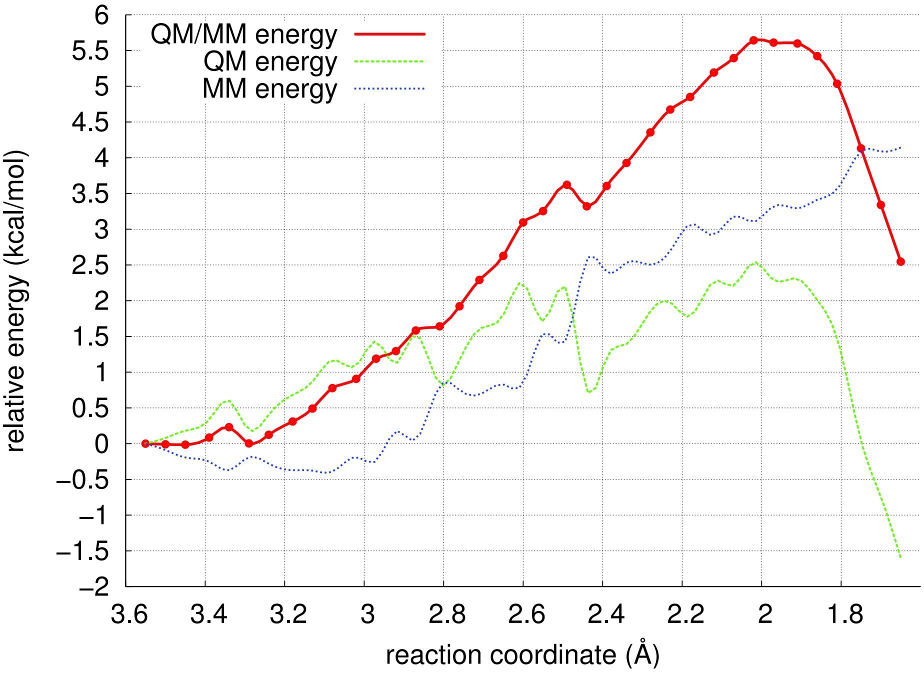


Figure S4: Energy profile for the addition step computed at the OM3/CHARMM level using the reaction coordinate $d_{C1-OX3'}$.

3.1 OM3/CHARMM molecular dynamics

During the QM/MM (OM3/CHARMM) MD simulation (see Figure 6 in the main text), no proton transfer is observed after formation of the Criegee intermediate. However, the ARG-329 H^{ε} atom fluctuates between being at a covalent bond distance $d_{H^{\varepsilon}-N^{\varepsilon}}$ of ∼1.1Å and in a hydrogen bond arrangement with roughly equal distances of ∼1.3Å from ARG-329 N^{ε} and the carbonyl oxygen of cyclohexanone. The same is true for the hydrogen from the NADP^{+} ribose 2′-hydroxyl group, which approaches the carbonyl oxygen even more closely during the dynamics. In the OM3/CHARMM optimized structure of the Criegee intermediate, the corresponding distances are $d_{O1-H^{\varepsilon}}$ = 1.41 Å, $d_{O1-HN2T}$ = 1.27 Å, $d_{H^{\varepsilon}-N^{\varepsilon}}$ = 1.11 Å, and $d_{HN2T-ON2'}$ = 1.12 Å. We also note that prior to the formation of the Criegee intermediate, the OM3/CHARMM MD simulation places the ARG-329 H^{η} atom mostly in between the ARG-329 N^{η} atom and the distal oxygen of the FADHOO$^-$ peroxy group, being on average less than 1.3Å away from either one (see Figure 6). In the OM3/CHARMM optimized structure of the reactant complex, we find $d_{OX3'-H\eta}$ = 1.29 Å and $d_{H\eta-N\eta}$ = 1.16 Å, compared with values of 1.95 Å and 1.03 Å, respectively, in the Criegee intermediate.

3.2 OM3/CHARMM optimization of the transition state for the addition reaction

It was technically quite difficult to locate the transition state for the addition at the OM3/CHARMM level. Using the P-RFO algorithm in the transition state search, we had to calculate the full Hessian of a subsystem including the whole cyclohexanone ring (in total more than 20 atoms) every 25 steps. The initial search gave a structure with two imaginary frequencies, the second of which corresponded to the mode connecting reactant and intermediate. This structure was then manually displaced along the reaction mode until the corresponding imaginary frequency became the leading one, and thereafter another P-RFO search was performed which yielded the correct transition state structure (confirmed by frequency and IRC calculations). With regard to its geometry and energy, the fully optimized OM3/CHARMM transition state resembles the highest point on the OM3/CHARMM energy profile very closely (see Figure S4). It seems reasonable to assume that such similarity also holds at the QM(B3LYP/TZVP)/CHARMM level, where we did not perform the costly manual procedure outlined above to precisely locate the transition state.

4. Energy profile for the migration step

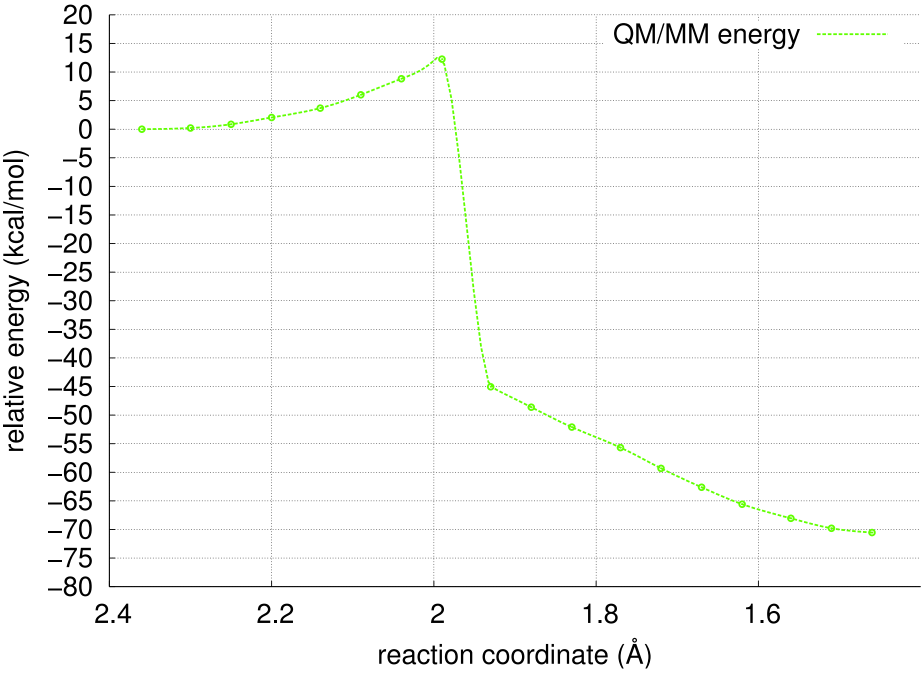


Figure S5: Energy profile for the migration step computed at the QM(B3LYP/SVP)/CHARMM level using the reaction coordinate $d_{OX3'-C2}$.

5. Sequence alignment

```
1      --MSQKMDFDAIVIGGGFGGLYAVKKLRDELELKVQAFDKATDVAGTWYWNRYPGALTDT   58  CHMOac
1      MTAQTTHTVDAVVIGAGFGGIYAVHKLHHELGLTTVGFDKADGPGGTWYWNRYPGALSDT   60  CHMOrh
          . .  .**:***.****:***:**:.** *.. .**** . .************:**

59     ETHLYCYSWDKELLQSLEIKKKYVQGPDVRKYLQQVAEKHDLKKSYQFNTAVQSAHYNEA  118  CHMOac
61     ESHLYRFSFDRDLLQESTWKTTYITQPEILEYLEDVVDRFDLRRHFKFGTEVTSALYLDD  120  CHMOrh
       *:*** :*:*::***.   *..*:  *:: :**::*.::.**:: ::*.* * ** * :

119    DALWEVTTEYGDKYTARFLITALGLLSAPNLPNIKGINQFKGELHHTSRWPDDVSFEGKR  178  CHMOac
121    ENLWEVTTDHGEVYRAKYVVNAVGLLSAINFPNLPGLDTFEGETIHTAAWPEGKSLAGRR  180  CHMOrh
       : ******::*: * *::::.*:***** *:**: *:: *:**  **: **:. *: *:*

179    VGVIGTGSTGVQVITAVAPLAKHLTVFQRSAQYSVPIGNDPLSEEDVKKIKDNYDKSLGW  238  CHMOac
181    VGVIGTGSTGQQVITSLAPEVEHLTVFVRTPQYSVPVGNRPVNPEQIAEIKADYDRIWER  240  CHMOrh
       ********** ****::** .:***** *:.*****:** *:. *:: :** :**:

239    CMNSALAFALNESTVPAMSVSAEERKAVFEKAWQTGGGFRFMFETFGDIATNMEANIEAQ  298  CHMOac
241    AKNSAVAFGFEESTLPAMSVSEEERNRIFQEAWDHGGGFRFMFGTFGDIATDEAANEAAA  300  CHMOrh
       . ***:**.::***:****** ***: :*::**: ******** *******:  **  *

299    NFIKGKIAEIVKDPAIAQKLMPQDLYAKRPLCDSGYYNTFNRDNVRLEDVKANPIVEITE  358  CHMOac
301    SFIRAKVAEIIEDPETARKLMPKGLFAKRPLCDSGYYEVYNRPNVEAVAIKENPIREVTA  360  CHMOrh
       .**:.*:***::**  *:****:.*:***********..:** **.   :* *** *:*

359    NGVKLENGDFVELDMLICATGFDAVDGNYVRMDIQGKNGLAMKDYWKEGPSSYMGVTVNN  418  CHMOac
361    KGVVTEDGVLHELDVLVFATGFDAVDGNYRRIEIRGRDGLHINDHWDGQPTSYLGVSTAN  420  CHMOrh
       :**  *:* : ***:*:* *********** *::*:*::** ::*:*.  *:**:**:. *

419    YPNMFMVLGPNGPFTNLPPSIESQVEWISDTIQYTVENNVESIEATKEAEEQWTQTCANI  478  CHMOac
421    FPNWFMVLGPNGPFTNLPPSIETQVEWISDTIGYAERNGVRAIEPTPEAEAEWTETCTEI  480  CHMOrh
       :** ******************:********* *:  *.*.:**.* *** :**:**::*

479    AEMTLFPKAQSWIFGANIPGKKNTVYFYLGGLKEYRTCASNCKNHAYEGFDIQLQRSDIK  538  CHMOac
481    ANATLFTKGDSWIFGANIPGKKPSVLFYLGGLRNYRAVMAEVAADGYRGFEVKS--AEMV  538  CHMOrh
       *: ***.*.:************ :* ******::**:  ::   ..*.**:::    :::

539    QPANA  543  CHMOac
539    TV---  540  CHMOrh
```

Figure S6: Sequence alignment of the two CHMOs from *Acinetobacter* sp. NCIMB 9871 and Rhodococcus sp. strain HI-31.

The sequence alignment shows that the two enzymes have the same active-site residues in the binding pocket (145-146, 248, 279, 329, 434-435, 437, 492, and 507 according to the CHMOrh numbering scheme). The overall sequence similarity is 55% (see ref. 14 in the main paper).

6. Complete references 21, 22b and 29 of the main paper

[21] MacKerell, Jr. A. D.; Bashford, D.; Bellott, M.; Dunbrack, Jr., 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.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *112*, 3586-3616.

[22b] Brooks, B. R.; Brooks, C. L. III.; MacKerell, A. D. Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545-1614.

[29] 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; Sjovoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *J. Mol. Struc. (Theochem)* **2003**, *632*, 1-28

A New Type of Stereoselectivity in Baeyer-Villiger Reactions: Access to E-and Z-Olefins

Z.-G. Zhang, G.-D. Roiban, J. P. Acevedo, I. Polyak,

M. T. Reetz,

Adv. Synth. Catal., **355**, 99 (2013)

DOI: 10.1002/adsc.201200759

A New Type of Stereoselectivity in Baeyer–Villiger Reactions: Access to *E*- and *Z*-Olefins

Zhi-Gang Zhang,[a,b,d] Gheorghe-Doru Roiban,[a,b,d] Juan Pablo Acevedo,[c,d]
Iakov Polyak,[a] and Manfred T. Reetz[a,b,]*

[a] Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany
E-mail: reetz@mpi-muelheim.mpg.de
[b] Philipps-Universität Marburg, Fachbereich Chemie, Hans-Meerwein-Strasse, 35032 Marburg, Germany
[c] Facultad de Medicina y Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Santiago, Chile
[d] Equal contributions

Received: August 24, 2012; Revised: October 16, 2012; Published online: January 7, 2013

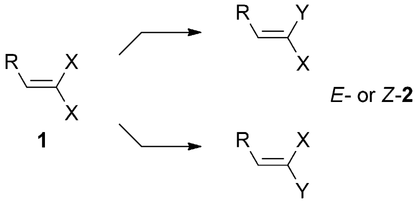
Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/adsc.201200759.

Abstract: A new concept for accessing configurationally defined trisubstituted olefins has been developed. Starting from a common ketone precursor of the type 4-ethylidenecyclohexanone, Baeyer–Villiger monooxygenases are employed as catalysts in diastereoselective Baeyer–Villiger reactions leading to the corresponding *E*- or *Z*-configurated lactones. Wild-type cyclohexanone monooxygenase (CHMO) as catalyst delivers the *E*-isomers and a directed evolution mutant the opposite *Z*-isomers. Subsequent transition metal-catalyzed chemical transformations of a key product containing a vinyl bromide moiety provide a variety of different trisubstituted *E*- or *Z*-olefins. A model based on QM/MM sheds light on the origin of this unusual type of diastereoselectivity. In contrast to this biocatalytic approach, traditional Baeyer–Villiger reagents such as *m*-CPBA fail to show any selectivity, 1:1 mixtures of *E*- and *Z*-olefins being formed.

Keywords: alkenes; Baeyer–Villiger oxidation; diastereoselectivity; directed evolution; palladium

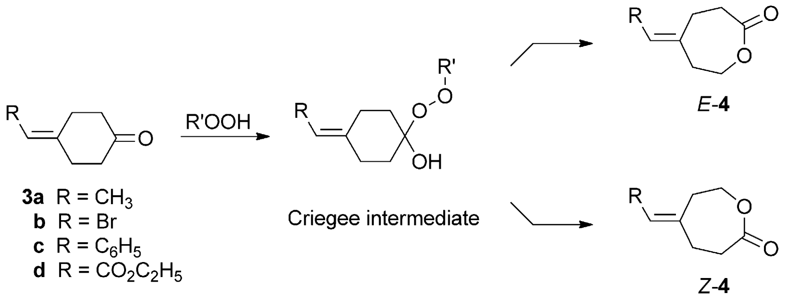
Introduction

A commonly occurring structural feature of many natural and unnatural organic compounds is the presence of configurationally well defined olefinic moieties, the respective double bonds having either the *E*- or *Z*-configuration. Controlling the *E*/*Z*-selectivity of olefin-forming processes with formation of trisubstituted products as in Wittig-type reactions, Julia–Lythgoe olefination, sigmatropic rearrangements, olefin metathesis, or C–C bond-forming reactions of alkynes has attracted a great deal of attention.[1] Variations of these themes as well as new approaches continue to be developed.[2] Making the appropriate choice of methods depends upon the particular starting material and synthetic goal under investigation.



Scheme 1.

A different approach is to start with easily accessible trisubstituted olefins **1** in which the 1,1-substituents are identical (Scheme 1). Such substituents are diastereotopic, which means that they may react with different rates leading to *E*- or *Z*-configurated olefins **2** (assignment of *E*- or *Z*-configuration depends on the priority of the particular substituents). Prominent examples include Pd-catalyzed C–C bond formation in which the *trans* C–X bond of 1,1-dihalides **1** (X = Cl or Br) reacts preferentially with formation of the respective *Z*-configurated trisubstituted olefins, as first developed by Tamao.[3] Subsequently it was generalized using a variety of coupling processes which include Kumada–Tamao–Corriu, Suzuki, Stille or Sonogashira reactions.[2f,4] It is also possible to consider a "reversed" approach in which 1,1-dimetalated reagents are utilized (e.g., X = boron in compounds **1**).[5] Another case pertains to the lipase-catalyzed hydrolysis of diesters of the type **1** (X = CO_2Et) which proceeds likewise in a *Z*-selective manner.[6] In all of



3a R = CH_3
b R = Br
c R = C_6H_5
d R = $CO_2C_2H_5$

Scheme 2.

these systems the sterically less shielded diastereotopic group reacts preferentially, but a reversal of the stereoselectivity has yet to be achieved. Nevertheless, the possibility of subsequent chemoselective transformations makes these processes synthetically attractive.[3–6]

We envisioned a different approach to selectively activate diastereotopic groups in appropriate trisubstituted olefins bearing identical substituents at the 1,1-position. The plan foresaw Baeyer–Villiger (BV) reactions[7] of keto-olefins of the type **3**, in which the two diastereotopic σ-bonds flanking the ketone function could react with different rates, thereby diastereoselectively providing olefins *E*- or *Z*-**4** from a common precursor, ideally on an optional basis (Scheme 2). Of particular interest was the vinyl bromide **3b**, because the respective *E*- or *Z*-lactones **4b** could serve as key compounds for transition metal-catalyzed cascade reactions with formation of a variety of different configurationally defined trisubstituted olefins. This type of stereoselectivity has not been considered previously, perhaps due to the anticipated difficulty in designing selective reagents and/or catalysts for such BV reactions. The reactive diastereotopic groups in compounds of the type **3**, namely the two σ-bonds flanking the ketone function in the respective Criegee intermediate, are spatially far removed from the olefinic moiety and are thus likely to migrate with similar rates leading to an undesired mixture of thermodynamically similar *E*- and *Z*-olefins. Indeed, upon treating keto-olefin **3b** with *m*-CPBA, a 1:1 mixture of lactones *E*- and *Z*-**4b** was formed.

Results and Discussion

Utilizing Biocatalysis as an Option

We therefore turned to biocatalytic BV reactions by considering Baeyer–Villiger monooxygenases (BVMOs)[8] as potentially diastereoselective catalysts. In previous work a number of enantioselective transformations have been reported, including oxidative kinetic resolution of 2-substituted cyclohexanone derivatives and desymmerization of 4-alkylcyclohexanones using wild-type (WT) cyclohexanone monooxygenase from *Acinetobacter* sp. NCIMB 9871 (designated here as CHMO), which is the most commonly employed BVMO in enantioselective BV reactions.[8] Regiodivergent oxidation has also been observed in some cases.[8h–k] In these enzymatic transformations, dioxygen in air reacts with an enzyme-bound (reduced) flavin (FAD) to form an anionic intermediate FAD-OO^- which initiates the BV reaction by nucleophilic addition to the carbonyl function, the respective Criegee intermediate then undergoing the usual fragmentation and σ-bond migration.[8] Subsequently, NADPH transforms the oxidized FAD back to the reduced form, thereby closing the catalytic cycle. When selectivity proves to be poor, the methods of directed evolution[9] can be applied with generation of enantioselective BVMO mutants.[10]

In the present study we chose WT CHMO as the catalyst in a whole cell *E. coli* system, the initial model transformation being the reaction of keto-olefin **3a**. Gratifyingly, essentially complete stereodifferentiation was observed, *E*-selectivity being favored (*E*-**4a**:*Z*-**4a** = 98:2) with excellent conversion and essentially no side-products. The other ketones were then tested under similar conditions (Table 1). Substrate **3b** likewise reacted with complete *E*-selectivity (*E*-**4b**:*Z*-**4b** ≥ 99:1). Unambiguous configurational assignment was made on the basis of the crystal structure of the key compound *E*-**4b** (Figure 1) and its transformation into *E*-**4a** by Suzuki coupling (see the Supporting Information). In sharp contrast, the other two ketones **3c** and **3d** reacted sluggishly, allowing substantial amounts of side-products to be formed which were identified as the alcohols corresponding to the reduction of the ketones. We suspect that in the whole cell process unidentified alcohol dehydrogenases are involved. From a synthetic viewpoint, poor conversion to the desired lactones **4c** and **4d** is not a serious problem, because the vinyl bromide *E*-**4b**, as already delineated, can serve as a key intermediate in the transition metal catalyzed transformation into the respective *E*-configurated products (see below).

Table 1. WT-CHMO as a biocatalyst in the Baeyer–Villiger reaction of ketones **3a–d** with formation of lactones **4a–d** in a whole-cell process.

Ketone	Product	*E*:*Z*	Conversion [%][a]	Time [h]	Other product [%][a]
3a	**4a**	98:2	>99	5	1
3b	**4b**	99:1	>99	5	4
3c	**4c**	>99:1	39	12	84
3d	**4d**	96:4	42	12	70

[a] By GC analysis of the crude products, the side-products being the alcohols corresponding to the ketones.

A possible strategy for reversing the diastereoselectivity in favor of the *Z*-isomers could be the systematic screening of different BVMOs, but a preliminary search beginning with the thermally robust phenyl acetone monooxygenase (PAMO)[11] and known mutants[10c] thereof proved to be unsuccessful. Rather than continuing along this strategy, we turned to directed evolution of WT CHMO using ketone **3b** as the model substrate, although changing such a highly *E*-selective enzyme into a completely *Z*-selective mutant appeared challenging. In earlier studies we had used error-prone PCR in order to enhance and invert the enantioselectivity of CHMO as a catalyst in the oxidative desymmetrization of 4-hydroxycyclohexanone.[10a] Later the much improved directed evolution method based on structure-guided saturation mutagenesis, the *combinatorial active-site saturation test* (CAST),[9i,12] was applied to PAMO with the creation of enantioselective mutants.[10c] Accordingly, sites around the binding pocket of an enzyme as revealed by the respective crystal structure or homology model are subjected to saturation mutagenesis, meaning the focused introduction of all of the other 19 canonical amino acids. If the initial libraries provide only moderately improved mutants, iterative saturation mutagenesis (ISM) can be applied.[9i] In the present study this structure-based approach to laboratory evolution was invoked. Since the X-ray structure of CHMO from *Acinetobacter* sp. NCIMB 9871 has not been reported to date, we used as a guide for choosing appropriate randomization sites the crystallographic data of the homologous CHMO from *Rhodococcus* recently reported by Lau, Berghuis and co-workers.[13] A number of potential randomization sites are possible, but only a limited number of experiments were performed, guided by exploratory docking simulations employing substrate **3a**. These indicated steric clashes between the methyl group and certain residues of the protein environment when considering the *Z*-selective process (see the Supplementary Information). On this basis three sites were selected, namely A (Phe432/Thr433), B (Leu143 and C (Phe505) as illustrated in Figure 2.

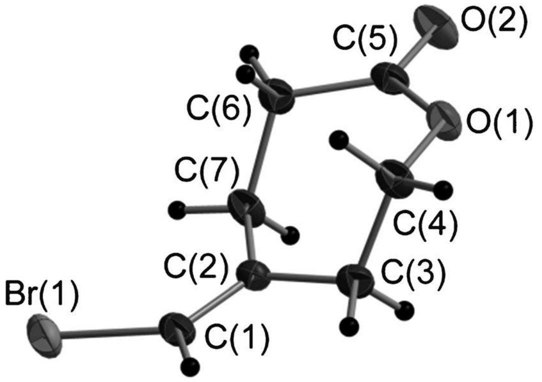


Figure 1. X-Ray structure of lactone *E*-**4b**.

In the case of the single-residue sites B and C, only about 100 transformants need to be screened for 95% library coverage.[9i] In order to minimize the screening effort in the case of saturation mutagenesis at the 2-residue site A, a reduced amino acid alphabet was used comprising 12 building blocks (Phe, Tyr, Cys, Leu, His, Arg., Ile, Asn, Ser, Val, Asp and Gly) as defined by NDT codon degeneracy. This requires the assessment of only ~430 transformants for 95% library coverage (instead of about 3000 when employing the standard NNK codon degeneracy encoding all 20 canonical amino acids).[9i] Six different ISM pathways are possible in a 3-site ISM scheme. In the present case we started with site A. Upon screening 500 transformants, a double mutant Phe432Ile/Thr433Gly (variant I) was identified leading to a shift toward *Z*-selectivity, but not yet to the desired reversal of diastereose-

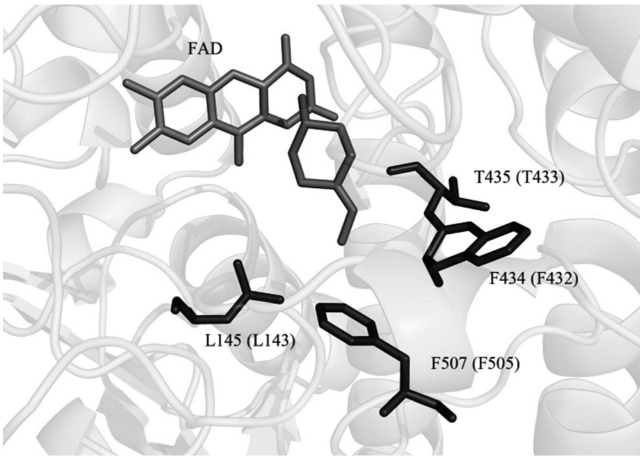


Figure 2. Possible saturation mutagenesis sites A (Phe432/Thr433), B (Leu143) and C (Phe505) in CHMO from *Acinetobacter* sp. NCIMB 9871, chosen on the basis of the crystal structure of CHMO from *Rhodococcus*,[13] PDB code 3GWD. The residue numbers not in parentheses correspond to the analogous residues in CHMO from *Rhodococcus*. The structure of cofactor FAD cofactor is marked in light grey and the substrate **3b** is marked in dark grey.

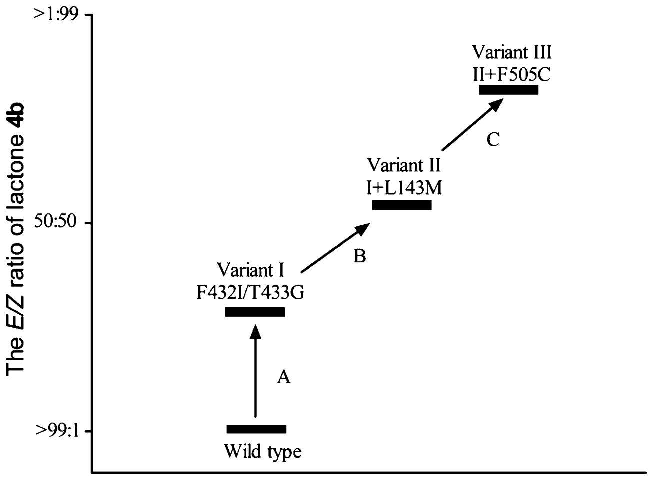


Figure 3. Chosen ISM pathway for evolving a *Z*-selective CHMO mutant as the catalyst in the BV reaction of ketone **3b**.

lectivity (*E*/*Z* = 71:29). The gene of this variant was then used as a template for saturation mutagenesis at single residue site B, this time utilizing NNK codon degeneracy. This provided a triple mutant Phe432Ile/Thr433Gly/Leu143Met (variant II) resulting in a slight reversal of selectivity (*E*-**4b**:*Z*-**4b** = 46:54). Finally, upon continuing the upward climb to site C, reversal of diastereoselectivity was further improved to 82% *Z*-selectivity (*E*-**4b**:*Z*-**4b** = 18:82) and 19% conversion after 5 h, catalyzed by quadruple mutant Phe432Ile/Thr433Gly/Leu143Met/Phe505Cys (variant III). An increased reaction time 12 h afforded conversions up to 30% with similar *Z*-selectivity (*E*-**4b**:*Z*-**4b** = 20:80) (see the Supporting Information, Table S3). Column chromatography and alternatively preparative HPLC provided essentially pure *Z*-**4b**. The arbitrarily chosen ISM pathway A→B→C is pictured in Figure 3. Exploring other pathways may well provide even better diastereoselectivities, but at this point we settled for the present results.

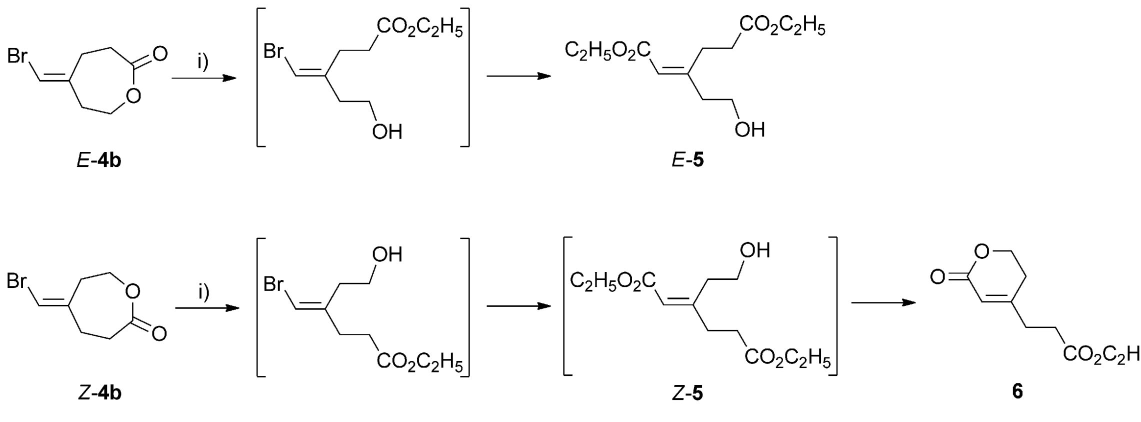
Testing the Best Mutant with Other Substrates

In order to explore the catalytic efficiency of the quadruple mutant (variant III), specifically evolved for compound **3b**, we tested it with the other ketones **3a** and **3c** and **3d**. In the case of substrate **3d**, the biocatalyst performs very well, the original *Z*-selectivity (WT CHMO) being fully reversed (*E*-**4d**:*Z*-**4d** = 4:96) at 98% conversion within 12 h and only 7% side-product (1.2 mmol scale). In the reactions of the other substrates, the results proved to be less encouraging (see the Supporting Information).

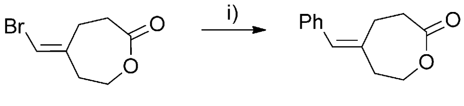
Exploiting the Vinyl Bromide 4b as a Key Compound in Cascade Reactions

As already alluded to, we planned to utilize *E*- and *Z*-**4b** as key compounds in further transformations. For example, upon subjecting them to Pd-catalyzed carbonylation[14] in the presence of ethanol, trisubstituted olefins *E*-**5** and **6** were obtained, respectively (Scheme 3). The presence of ethanol induces rapid transesterifying ring-opening under the reaction conditions, probably before the actual carbonylation occurs. We observed this kind of ring-opening for compounds of the type **5** whenever they were treated with mild nucleophiles such as water or alcohols in the absence of transition metal catalysts (see the Supporting Information). These carbonylation reactions stand in contrast to the poor results when attempting the BV reaction of **3c** (Table 1).

In other reactions, Pd-catalyzed Suzuki-type arylation[15] of *E*-**4b** provided the respective lactone *E*-**4c** again with retention of configuration (Scheme 4), which is not readily accessible starting from keto-olefin **3c** (Table 1). The products in these schemes bear functional groups which can in principle be manipulated with formation of a wide variety of different trisubstituted olefins for potential applications.



Scheme 3. Pd-catalyzed carbonylation with retention of double bond configuration. *Reagents and conditions:* i) EtOH, $Pd(PPh_3)_4$/CsF, CO, 80 °C, 48 h: *E*-**5**, 61%; **6**, 78%.



E-**4b** *E*-**4c**

Scheme 4. Stereospecific Suzuki transformation of lactone *E*-**4b** with retention of configuration. *Reagents and conditions:* i) $PhB(OH)_2$, $Pd(PPh_3)_4$/CsF, THF/70 °C/24 h; 92%.

Developing a Model to Explain the Observed Diastereoselectivity

In order to gain insight into the source of diastereoselectivity of WT CHMO on a molecular level, we first performed docking computations and molecular dynamics (MD) simulations[16] using ketone **3a** as the substrate. In this endeavor, our recent QM/MM study of CHMO as a catalyst in the BV reaction of the parent compound cyclohexanone was considered, in which it was shown that the chair form of this ketone can bind in only one way which ensures a low-energy pathway to the Criegee intermediate.[17] Analysis of this intermediate revealed a crucial characteristic, namely that only one of the two σ-bonds flanking the carbonyl group is capable of undergoing rapid migration because here the traditional stereoelectronic requirement is fulfilled (anti-periplanar arrangement of the reactive C–C–O–O segment),[7,18] in contrast to the geometric arrangement when the other σ-bond migrates. This hypothesis was then supported by QM/MM calculations of the respective transition states. In the case of enantioselective desymmetrization of 4-methylcyclohexanone, the same applies, but two different chair forms are possible, one in which the methyl group is equatorial and the other in which it is axial. The difference in energy between the two chair forms in the protein environment determines the degree of enantioselectivity (≥96% *ee* in favor of *S*-lactone).[17]

This model is not directly applicable to the CHMO-catalyzed BV reaction of ketone **3a** in the present study, because the axial/equatorial structural element is not relevant. Nevertheless, we used the geometric features of the previously calculated Criegee intermediate of cyclohexanone as a template for manually constructing the analogous intermediate involving **3a**. As shown by force field calculations and molecular dynamics (MD) simulations, there are two poses of different energy, in each case only one σ-bond reacting due to the stereoelectronic requirement. The energetically lower pose leads to the experimentally observed formation of *E*-**4a** (see the Supporting Information). In the higher energy pose, the methyl group clashes with residues in a nearby loop (Supporting Information). This conclusion was corroborated by more accurate QM(B3LYP/TZVP)/CHARMM optimizations of the respective Criegee intermediates (Figure 4, a and b). Accordingly, the difference in energy amounts to 2.3 kcal mol^{-1}, which correlates well with the experimentally observed diastereoselectivity (see the Supporting Information).

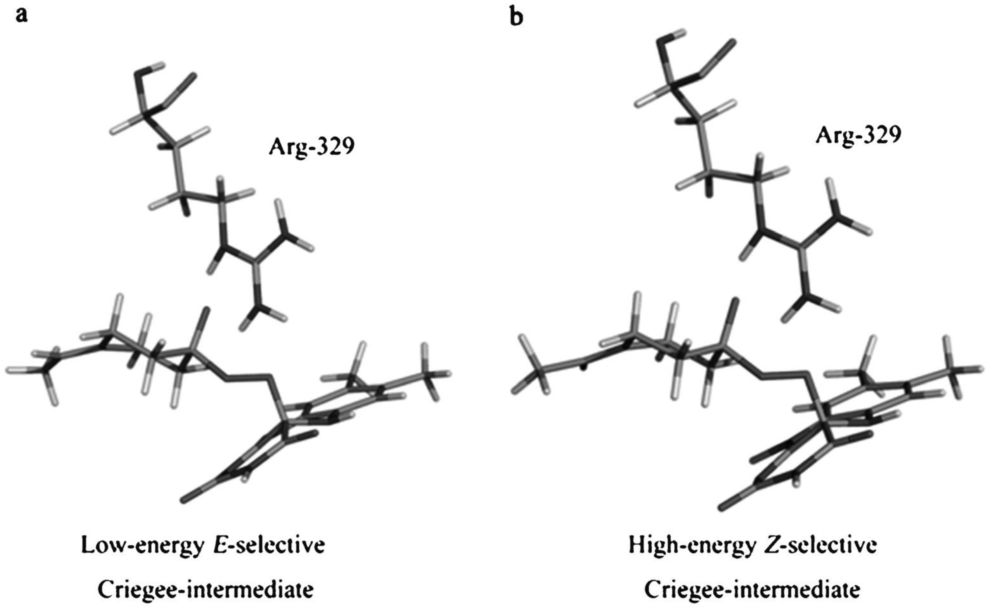


Figure 4. QM-calculated Criegee intermediates involving 4-ethylidenecyclohexanone (**3a**) in WT CHMO. **a**) Low-energy intermediate leading to the experimentally observed lactone *E*-**4a.** **b**) High-energy intermediate which would provide *Z*-**4a** (not observed).

Conclusions

The present study contributes to the continuing search for ways to prepare configurationally defined trisubstiuted olefins. In doing so, we have demonstrated a new type of stereoselective transformation in Baeyer–Villiger reactions, which has not been considered previously using synthetic reagents/catalysts or enzymes. Utilizing a Baeyer–Villiger monooxygenase, specifically CHMO from *Acinetobacter* sp. NCIMB 9871, a biocatalytic approach was developed which enables the control of *E*/*Z*-configuration in the formation of structurally different trisubstituted olefins starting from the respective 4-alkylidenecyclohexanone derivatives. WT CHMO is fully *E*-selective in the reaction of three different substrates. Especially the vinyl bromide **3b** proved to be a synthetically important substrate because the respective diastereoselectively formed lactone *E*-**4b** serves as a key compound in subsequent stereospecific cascade transformations such as Pd-catalyzed carbonylation or Suzuki coupling with formation of configurationally defined trisubstituted olefins. A model was developed which explains the source of diastereoselectivity on a molecular level. We also succeeded in reversing the selectivity in favor of *Z*-configurated olefins up to 82%. Future work will focus on directed evolution of the thermally stable Baeyer–Villiger monooxygenase PAMO[11] as a catalyst in this type of stereoselective transformation. It also remains to be seen whether appropriate derivatives of other even-membered cyclic ketones such as those of cyclobutanone or cyclooctanone can be used in a similar way.

Experimental Section

Biology: Creation of Mutant Libraries

The whole length of the CHMO gene from *Acinetobacter* sp. NCIMB 9871 was cloned into *E.coli* expression vector pET-22b (+) just as described previously,[19] which was also used as template for saturation mutagenesis. Mutant libraries shown in Table S1 (Supporting Information) were created by employing the QuikChange PCR method (Stratagene). Table S2 (Supporting Information) provides the oligonucleotide primers used for the creation of mutant libraries. PCR reaction mixtures (50 μL final volume) consisted of 10×KOD buffer (5 μL), $MgSO_4$ (2 μL, 25 mM), dNTP (5 μL, 2 mM each), primers (5 μL, 2.5 μM each), template plasmid (2 μL, 10 ngμL^{-1}) and KOD Hot Start DNA polymerase (1 U). The PCR reaction started at 95 °C for 3 min, followed by 18 cycles of denaturing step at 95 °C for 1 min, annealing at 52–58 °C (depending on the particular pairs of primers) for 1 min and elongation at 72 °C for 8 min. After cycling a final extension step at 72 °C for 16 min was performed. To ensure the removal of the template plasmid, PCR products were digested during 4 h at 37 °C after adding 1 unit of *Dpn*I (New England Biolabs) at 37 °C for 4 h. The digested PCR product was used to transform electrocompetent *E.coli* BL21-Gold (DE3) cells. The cells were spread on the LB agar plate with 100 μg mL^{-1} carbenicilline.

Library Screening

Single colonies were picked into 2.2-mL 96 deep-well plates containing 800 μL LB media with 100 μg mL^{-1} carbenicilline. Four wells per plate were used to culture wild-type CHMO as a control. After shaking at 37 °C, 800 rpm overnight, 100 μL of this overnight culture were used to inoculate a new 96 deep-well plate containing 700 μL LB media with relevant antibiotics. To conserve copy of libraries at −80 °C, the glycerol stock plates containing culture solution with 30% glycerol were prepared at this stage. The duplicate plates were cultivated for 1 h at 37 °C, 800 rpm. Then, 100 μL of the substrate mixture (80 μL LB with the relevant antibiotics and 20 μL of a stock solution of substrate 0.1 M in acetonitrile) supplemented with 0.2 mM IPTG were added to the plates. The incubation was continued at 30 °C shaking at 800 rpm for additional 15 h. The reaction was stopped and reaction product was extracted by adding 400 μL ethyl acetate to each well of the plates. After centrifugation, the organic phase was transferred to 96-well glass plate (Zinsser, Analytic) for GC analysis. The final positive mutants were collected and confirmed by reproducing the biotransformation in a flask scale. The plasmids of the mutants were extracted from overnight cultures and sequenced.

Biotransformations: General Upscaling Procedure for Biotransformations with the *E. coli* Wild-Type CHMO or Mutant CHMO

The wild-type CHMO or mutant CHMO were stored in the form of glycerol frozen stocks of *E.coli* BL21-Gold (DE3) (Stratagene) harboring the pET-22b (+) expression vector. Fresh LB medium (50 mL) containing carbenicillin (100 μg mL^{-1}) was inoculated with 1% of an overnight preculture of the CHMO-expressing *E.coli* strains in 300-mL Erlenmeyer flasks. The culture was incubated at 37 °C with shaking at 200 rpm until the OD_{600} reached 0.7. Then, isoproxypropyl-thio-β-D-galactoside (IPTG) was added to a final concentration of 0.2 mM, followed by adding ketone substrate in acetonitrile with the final concentration of 20 mM. The final reaction mixture was shaken at 30 °C and 200 rpm. The whole process of the reaction was monitored by GC analysis of the aliquot samples every 2 h until the reaction completion. Then, the reaction was stopped and biotransformation mixtures were extracted with ethyl acetate. The organic phase was afterwards dried with Na_2SO_4, evaporated and the residue subjected to column chromatography for further purification and products separation.

Chemistry: General Remarks

Starting compounds and all other reagents including dry solvents were purchased from Acros, Sigma–Aldrich and Alfa and used without further purification. NMR spectra were recorded on a Bruker Avance 300 or DRX 400 (^{1}H: 300 MHz or 400 MHz, ^{13}C: 75 MHz or 101 MHz) spectrometer using TMS as internal standard ($d=0$) unless otherwise noted. High-resolution EI mass spectra were measured on a Finnigan MAT 95S spectrometer. High-resolution mass spectra recorded in ESI and APCI mode were performed on a ThermoScientific LTQ-FT spectrometer. Conversion and diastereoisomeric composition were determined by achiral gas chromatography as described. Alternatively, product **4b** diastereomeric mixture was determined by HPLC. Diastereoisomeric assignments were determined as described in the Supporting Information. Analytical thin layer chromatography was performed on Merck silica gel 60 F254q while for column chromatography Merck silica gel 60 (230–400 mesh ASTM) was used. Reactions that required inert atmosphere (nitrogen or argon) were carried out using standard Schlenk techniques. GC analyses were performed on a HP 6890 series while HPLC data were measured on a Shimadzu LC-8 A. Crystallographic data for the structure of ***E*-4b** has been deposited with the Cambridge Crystallographic Data Centre (CCDC 880686). These data can be obtained free of charge from The Cambridge Crystallographic Data Centre *via* www.ccdc.cam.ac.uk/data_request/cifor on application to CCDC, 12 Union Road, Cambridge CB2 1EZ, U.K. [fax: +44-(0)1223-336033 or e-mail: deposit@ccdc.cam.ac.uk].

Acknowledgements

*We thank Stephanie Dehn for GC analyses, Heike Hinrichs and Alfred Deege for HPLC analyses, and Dr. Richard Goddard for the X-ray structure determination of lactone E-**4b**. Financial support by the Max-Planck-Society and the Arthur C. Cope Foundation is gratefully acknowledged.*

References

[1] Selected reviews of stereospecific and stereoselective olefin syntheses based on different reaction types: a) J. Reucroft, P. G. Sammes, *Q. Rev. Chem. Soc.* **1971**, *25*, 135–169; b) J. F. Normant, A. Alexakis, *Synthesis* **1981**, 841–870; c) B. E. Maryanoff, A. B. Reitz, *Chem. Rev.* **1989**, *89*, 863–927; d) R. W. Hoffmann, *Angew. Chem.* **1979**, *91*, 625–634; *Angew. Chem. Int. Ed. Engl.* **1979**, *18*, 563–572; e) T. Nakai, K. Mikami, *Chem. Rev.* **1986**, *86*, 885–902; f) E. Negishi, D. Y. Kondakov, *Chem. Soc. Rev.* **1996**, *25*, 417–426; g) J. Schwartz, J. A. Labinger, *Angew. Chem.* **1976**, *88*, 402–409; *Angew. Chem. Int. Ed. Engl.* **1976**, *15*, 333–340; h) P. R. Blackmore, *J. Chem. Soc. Perkin Trans. 1* **2002**, 2563–2585; i) R. H. Grubbs, *Adv. Synth. Catal.* **2007**, *349*, 34–40.

[2] a) Y. Schmidt, B. Breit, *Chem. Eur. J.* **2011**, *17*, 11780–11788; b) J. K. Belardi, G. C. Micalizio, *J. Am. Chem. Soc.* **2008**, *130*, 16870–16872; c) S. J. Meek, R. V. O'Brien, J. Llaveria, R. R. Schrock, A. H. Hoveyda, *Nature* **2011**, ***471***, 461–466; d) B. K. Keitz, K. Endo, P. R. Patel, M. B. Herbert, R. H. Grubbs, *J. Am. Chem. Soc.* **2012**, *134*, 693–699; e) A. Fuerstner, K. Radkowski, H. Peters, *Angew. Chem.* **2005**, *117*, 2837–2841; *Angew. Chem. Int. Ed.* **2005**, *44*, 2777–2781; f) O. Reiser, *Angew. Chem.* **2006**, *118*, 2904–2906; *Angew. Chem. Int. Ed.* **2006**, *45*, 2838–2840; g) J. Prunet, *Angew. Chem.* **2003**, *115*, 2932–2936; *Angew. Chem. Int. Ed.* **2003**, *42*, 2826–2830.

[3] A. Minato, K. Suzuki, K. Tamao, *J. Am. Chem. Soc.* **1987**, *109*, 1257–1258.

[4] a) T. Sugihara, in: *Handbook of Organopalladium Chemistry for Organic Synthesis*, (Eds.: E.-i. Negishi, A. de Meijere), Wiley-VCH, Weinheim, **2002**, pp 649–656; b) Z. Tan, E.-i. Negishi, *Angew Chem.* **2006**, *118*, 776–779; *Angew. Chem. Int. Ed.* **2006**, *45*, 762–765.

[5] I. Marek, *Chem. Rev.* **2000**, *100*, 2887–2900.

[6] T. Schirmeister, H. H. Otto, *J. Org. Chem.* **1993**, *58*, 4819–4822.

[7] Reviews and pertinent studies of Baeyer–Villiger reactions: a) G. R. Krow, *Org. React.* **1993**, *43*, 251–798; b) M. Renz, B. Meunier, *Eur. J. Org. Chem.* **1999**, 737–750; c) G.-J. ten Brink, I. W. C. E. Arends, R. A. Sheldon, *Chem. Rev.* **2004**, *104*, 4105–4123; d) C. Bolm, G. Schlingloff, K. Weickhardt, *Angew. Chem.* **1994**, *106*, 1944–1946; *Angew. Chem. Int. Ed. Engl.* **1994**, *33*, 1848–1849; e) G. Strukul, *Angew. Chem.* **1998**, *110*, 1256–1267; *Angew. Chem. Int. Ed.* **1998**, *37*, 1198–1209; f) S.-I. Murahashi, S. Ono, Y. Imada, *Angew. Chem.* **2002**, *114*, 2472; *Angew. Chem. Int. Ed.* **2002**, *41*, 2366–2368; g) G. Peris, S. Miller, J. *Org. Lett.* **2008**, *10*, 3049–3052; h) S. Xu, Z. Wang, X. Zhang, X. Zhang, K. Ding, *Angew. Chem.* **2008**, *120*, 2882–2885; *Angew. Chem. Int. Ed.* **2008**, *47*, 2840–2843.

[8] Reviews and selected studies of BVMOs: a) H. Leisch, K. Morley, P. C. K. Lau, *Chem. Rev.* **2011**, *111*, 4165–4222; b) G. de Gonzalo, M. D. Mihovilovic, M. W. Fraaije, *ChemBioChem* **2010**, *11*, 2208–2231; c) M. M. Kayser, *Tetrahedron* **2009**, *65*, 947–974; d) M. D. Mihovilovic, in: *Enzyme Catalysis in Organic Synthesis*, Vol. 3, (Eds.: K. Drauz, H. Gröger, O. May), Wiley-VCH, Weinheim, **2012**, pp 1439–1485; e) C. T. Walsh, Y.-C. J. Chen, *Angew. Chem.* **1988**, *100*, 342–352; *Angew. Chem. Int. Ed. Engl.* **1988**, *27*, 333–343; f) R. Wohlgemuth, *Eng. Life Sci.* **2006**, *6*, 577–583; g) M. D. Mihovilovic, F. Rudroff, B. Grötzl, *Curr. Org. Chem.* **2004**, *8*, 1057–1069; see selected studies on regioselectivity of certain BVMOs: h) V. Alphand, A. Archelas, R. Furstoss, *Tetrahedron Lett.* **1989**, *30*, 3663–3664; i) W. Carruthers, J. P. Prail, S. M. Roberts, A. J. Willetts, *J. Chem. Soc. Perkin Trans. 1* **1990**, 2854–2856; j) S. Wang, M. M. Kayser, V. Jurkauskas, *J. Org. Chem.* **2003**, *68*, 6222–6228; k) M. J. Fink, T. C. Fischer, F. Rudroff, H. Dudek, M. W. Fraaije, M. D. Mihovilovic, *J. Mol. Catal. B: Enzym.* **2011**, *73*, 9–16.

[9] Reviews of directed evolution: a) N. J. Turner, *Nat. Chem. Biol.* **2009**, *5*, 568–574; b) T. W. Johannes, H. M. Zhao, *Curr Opin. Microbiol.* **2006**, *9*, 261–267; c) S. Lutz, U. T. Bornscheuer, *Protein Engineering Handbook*, Vols. 1–2, Wiley-VCH, Weinheim, **2009**; d) C. Jäckel, P. Kast, D. Hilvert, *Annu. Rev. Biophys. Biomol. Struct.* **2008**, *37*, 153–173; e) P. A. Romero, F. H. Arnold, *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866–876; f) A. V. Shivange, J. Marienhagen, H. Mundhada, A. Schenk, U. Schwaneberg, *Curr. Opin. Chem. Biol.* **2009**, *13*, 19–25; g) L. G. Otto, F. Hollmann, I. W. C. E. Arends, *Trends Biotechnol.* **2010**, *28*, 46–54; h) M. T. Reetz, in: *Enzyme Catalysis in Organic Synthesis*, Vol. 1, (Eds.: K. Drauz, H. Gröger, O. May), Wiley-VCH, Weinheim, **2012**, pp 119–190; i) review on directed evolution of stereoselective enzymes with emphasis on iter-

ative saturation mutagenesis: M. T. Reetz, *Angew. Chem.* **2011**, *123*, 144–182; *Angew. Chem. Int. Ed.* **2011**, *50*, 138–174.

[10] Examples of directed evolution of BVMOs: a) M. T. Reetz, B. Brunner, T. Schneider, F. Schulz, C. M. Clouthier, M. M. Kayser, *Angew. Chem.* **2004**, *116*, 4167–4170; *Angew. Chem. Int. Ed.* **2004**, *43*, 4075–4078; b) M. D. Mihovilovic, F. Rudroft, A. Winninger, T. Schneider, F. Schulz, M. T. Reetz, *Org. Lett.* **2006**, *8*, 1221–1224; c) M. T. Reetz, S. Wu, *J. Am. Chem. Soc.* **2009**, *131*, 15424–15432; d) A. Kirschner, U. T. Bornscheuer, *Appl. Microbiol. Biotechnol.* **2008**, *81*, 465–472.

[11] a) E. Malito, A. Alfieri, M. W. Fraaije, A. Mattevi, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13157–13162; b) M. W. Fraaije, J. Wu, D. P. H. M. Heuts, E. W. van Hellemond, J. H. L. Spelberg, D. B. Janssen, *Appl. Microbiol. Biotechnol.* **2005**, *66*, 393–400; c) D. E. Torres Pazmino, M. Winkler, A. Glieder, M. W. Fraaije, *J. Biotechnol.* **2010** *146*, 9–24.

[12] M. T. Reetz, M. Bocola, J. D. Carballeira, D. Zha, A. Vogel, *Angew. Chem.* **2005**, *117*, 4264–4268; *Angew. Chem. Int. Ed.* **2005**, *44*, 4192–4196.

[13] I. A. Mirza, B. J. Yachnin, S. Wang, S. Grosse, H. Bergeron, A. Imura, H. Iwaki, Y. Hasegawa, P. C. K. Lau, A. M. Berghuis, *J. Am. Chem. Soc.* **2009**, *131*, 8848–8854.

[14] a) H. M. Colquhoun, D. J. Thompson, M. V. Twigg, *Carbonylation, Direct Synthesis of Carbonyl Compounds*, Plenum, New York, **1991**; b) R. Grigg, S. P. Mutton, *Tetrahedron* **2010**, *66*, 5515–5548.

[15] K. Suzuki, *Angew. Chem.* **2011**, *123*, 6854–6869; *Angew. Chem. Int. Ed.* **2011**, *50*, 6722–6737.

[16] a) *Schrödinger, LLC, Maestro 8.5 user manual*, Schrödinger Press, New York, **2008**; b) F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson, W. C. Still, *J. Comput. Chem.* **1990**, ***11***, 440–467.

[17] I. Polyak, M. T. Reetz, W. Thiel, *J. Am. Chem. Soc.* **2012**, *134*, 2732–2741.

[18] a) Y. Okuno, *Chem. Eur. J.* **1997**, *3*, 212–218; b) N. Mora-Diez, S. Keller, J. R. Alvarez-Idaboy, *Org. Biomol. Chem.* **2009**, *7*, 3682–3690; c) F. Grein, A. C. Chen, D. Edwards, C. M. Crudden, *J. Org. Chem.* **2006**, *71*, 861–872.

[19] G. Chen, M. M. Kayser, M. D. Mihovilovic, M. E. Mrstik, C. A. Martinez, J. D. Stewart, *New J. Chem.* **1999**, *23*, 827–832.

Quantum Mechanical/Molecular Mechanical Study on the Enantioselectivity of the Enzymatic Baeyer-Villiger Reaction of 4-Hydroxycyclohexanone

I. Polyak, M. T. Reetz, W. Thiel

J. Phys. Chem. B, **117**, 4993 (2013)

THE JOURNAL OF
PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

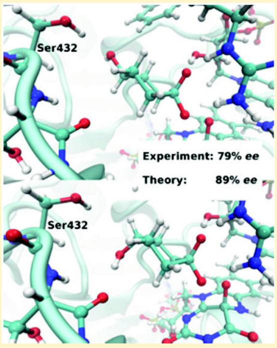
Quantum Mechanical/Molecular Mechanical Study on the Enantioselectivity of the Enzymatic Baeyer–Villiger Reaction of 4-Hydroxycyclohexanone

Iakov Polyak,[†] Manfred T. Reetz,[†,‡] and Walter Thiel*[,†]

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

[‡]Fachbereich Chemie, Philipps-Universität Marburg, Hans-Meerwein-Straße, D-35032 Marburg, Germany

Supporting Information



ABSTRACT: We report a combined quantum mechanical/molecular mechanical (QM/MM) study of the effect of mutations of the Phe434 residue in the active site of cyclohexanone monooxygenase (CHMO) on its enantioselectivity toward 4-hydroxycyclohexanone. In terms of our previously established model of the enzymatic Baeyer–Villiger reaction, enantioselectivity is governed by the preference toward the equatorial ((*S*)-selectivity) or axial ((*R*)-selectivity) conformation of the substituent at the C4 carbon atom of the cyclohexanone ring in the Criegee intermediate and the subsequent rate-limiting transition state for migration (TS2). We assess the enantiopreference by locating all relevant TS2 structures at the QM/MM level. In the wild-type enzyme we find that the axial conformation is energetically slightly more stable, thus leading to a small excess of (*R*)-product. In the Phe434Ser mutant, there is a hydrogen bond between the serine side chain and the equatorial substrate hydroxyl group that is retained during the whole reaction, and hence there is pronounced reverse (*S*)-enantioselectivity. Another mutant, Phe434Ile, is shown to preserve and enhance the (*R*)-selectivity. All these findings are in accordance with experiment. The QM/MM calculations allow us to explain the effect of point mutations on CHMO enantioselectivity for the first time at the molecular level by an analysis of the specific interactions between substrate and active-site environment in the TS2 structures that satisfy the basic stereoelectronic requirement of anti-periplanarity for the migrating σ-bond.

1. INTRODUCTION

Cyclohexanone monooxygenase (CHMO) is a member of the family of Baeyer–Villiger (B–V) monooxygenases (BVMO),[1] responsible for conversion of cyclic or linear ketones into lactones or esters, respectively. The natural ligand for CHMO is cyclohexanone, but this enzyme can accept a wide range of ligands and is highly enantioselective toward different substrates.[2] This feature makes CHMO very useful for enzymatic catalysis,[3] with CHMO from *Acinetobacter* sp. NCIMB 9871 being the most thoroughly investigated member of BVMOs.

However, substrate acceptance and enantioselectivity of CHMO are still limited. When the substituents at the cyclohexanone ring are varied too much, enantioselectivity may become reduced or even reversed.[4] In order to overcome the natural limitations of enzymes in general, directed evolution[5] has been applied to essentially all enzyme types, making it possible to achieve enantioselectivity toward desired substrates by means of mutations in the amino acid chain of the enzyme.[6] Knowledge about the structure of the active site and the reaction mechanism helps to make directed evolution more efficient by designing improved strategies for the required mutations.[5,6] This type of protein engineering has also been applied to CHMO.[7]

The reaction mechanism of BVMOs has been thoroughly studied experimentally with the use of kinetic and spectroscopic methods.[2b,f,8] In brief, the enzyme-bound flavin reacts in its reduced form with oxygen and forms the peroxyflavin, which in the deprotonated form then generates a Criegee intermediate via nucleophilic addition to the carbonyl carbon of the substrate. The Criegee intermediate subsequently fragments by σ-bond migration, yielding the hydroxyflavin and the product (ester or lactone). This is followed by water elimination from the oxidized flavin, which is subsequently reduced by the NADPH cofactor via hydride transfer.

Several X-ray structures have been resolved up to now, for both CHMO and another homologous BVMO, phenylacetone monooxygenase (PAMO).[8d,9] The first two crystal structures[9b] (called "open" and "closed") of CHMO from *Rhodococcus* sp. strain HI-31 contained both FAD and NADP$^+$ cofactors bound to the enzyme but had no ligand or inhibitor in the binding pocket. The "closed" structure was considered to resemble the geometries adopted during the intermediate stages of the enzymatic reaction. Another structure (called "rotated"), also from *Rhodococcus* sp. strain HI-31, was recently resolved,[9c] in

Received: February 20, 2013
Revised: April 2, 2013
Published: April 5, 2013

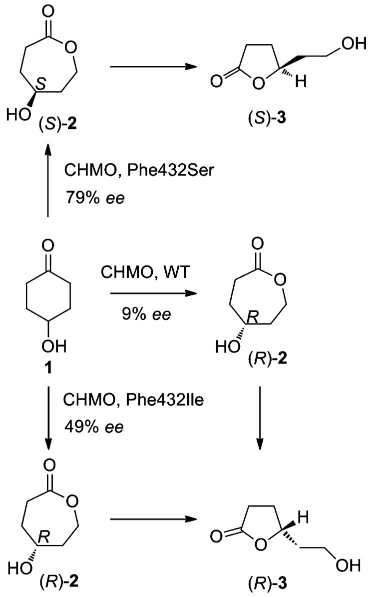
which the NADP$^+$ cofactor shows significant rotation compared with the previously published structures,[9b] and what is believed to be cyclohexanone is located in a position occupied by the NADP$^+$ nicotinamide moiety in the other structures. Because of the proximity of cyclohexanone to the flavin ring, the authors proposed that this "rotated" structure represents the conformation of the enzyme at the stage of the Criegee intermediate and the subsequent fragmentation.[9c] However, in such a configuration, the residues that are known to determine the substrate specificity of CHMO and that line the binding pocket in the "closed" conformation are no longer in the vicinity of the ligand. Therefore, the authors suggested that the substrate first binds to the enzyme in its "open" conformation, followed by rearrangement into the tight "closed" structure; thereafter, the substrate is supposed to be driven into the catalytic position, with the enzyme adopting the "rotated" conformation.[9c]

Several X-ray structures of PAMO were resolved as well,[8d,9a] which contain either only one or both cofactors bound to the enzyme, both with and without inhibitor at the binding site. The PAMO crystal structure with the inhibitor bound in the active site[8d] resembles the "closed" conformation of CHMO[9b] rather than the "rotated" one.[9c] One cannot exclude, of course, that substrate binding and reaction mechanism may be different in these two enzymes.[9c]

Before the "rotated" conformation of the CHMO had been published, we performed a computational quantum mechanical/molecular mechanical (QM/MM) study on the reaction mechanism of CHMO using the "closed" conformation of the enzyme as starting point.[10] We modeled both the addition and fragmentation steps of cyclohexanone oxygenation. We found that the "closed" conformation of CHMO is well-suited for the reaction to proceed with reasonable barriers. According to the QM/MM calculations, the cyclohexanone substrate, being in the chair conformation, binds to the enzyme in a specific manner, which allows for nucleophilic addition on an optimal trajectory, yielding the Criegee intermediate in a strictly defined conformation. In this geometry, only one of the two cyclohexanone σ-bonds adjacent to the carbonyl carbon atom fulfills the stereoelectronic requirement of anti-periplanarity and can migrate. On the basis of this model, we could also explain the experimentally observed enantioselectivity toward 4-methylcyclohexanone.[10] The optimized structures of the Criegee intermediate and the transition state for fragmentation are generally similar to those obtained for unsubstituted cyclohexanone, but two distinct conformations are possible in the case of 4-methylcyclohexanone, with the methyl group either in equatorial or axial orientation. For both conformers, it is the same σ-bond of the ring that can migrate due to the requirement of anti-periplanarity, and it is thus the orientation of the methyl group that ultimately determines the enantioselectivity. In the QM/MM calculations on the Criegee intermediate and the subsequent transition state, the structures with the methyl group of 4-methylcyclohexanone in the equatorial position proved to be more stable than their axial counterparts, thus correctly predicting the experimentally observed[7k] (*S*)-product. This provided for the first time an explanation at the molecular level for the enantioselectivity in the CHMO-catalyzed desymmetrization of a substituted cyclohexanone substrate. Recently, we have also been able to rationalize the enantioselectivity of CHMO toward 4-ethylidenecyclohexanone on the basis of the relative energies of the (*E*)- and (*Z*)-conformers of the corresponding Criegee intermediate.[11]

The current work is a continuation of our previous investigations on the origin of enantioselectivity in CHMO-catalyzed reactions. Here we study the desymmetrization of another substituted cyclohexanone, 4-hydroxycyclohexanone (**1**), catalyzed by the wild-type (WT) enzyme from *Rhodococcus* sp. as well as the Phe434Ser and Phe434Ile mutants. This substrate is of special interest, since the two particular point mutations at this position either increase or reverse the enzyme's enantioselectivity.[4] It is known experimentally[4] that WT CHMO enzyme from *Acinetobacter* sp. NCIMB 9871 exhibits rather low enantioselectivity for **1** (9% (*R*)), while the Phe432Ser mutant reverses the enantiopreference (79% (*S*)) and the Phe432Ile mutant improves the enantioselectivity (49% (*R*)) observed in the WT (Scheme 1). It needs to be pointed out that the 7-membered lactones shown in Scheme 1 are formed in the rate and stereoselectivity determining step, but they are not the final products due to subsequent fast intramolecular translactonization to the thermodynamically more stable 5-membered lactones. Understanding stereocomplementary behavior of enzyme homologues being mutated at a single amino acid site is of theoretical and practical interest and may pave the way to designing improved strategies for directed evolution that target the synthesis of novel chiral products.

Scheme 1. Stereochemical Conversion of 4-Hydroxycyclohexanone (1) to (*R*)-4-Hydroxy-ε-caprolactone ((*R*)-2) and to (*S*)-4-Hydroxy-ε-caprolactone ((*S*)-2) in the B–V Reaction Catalyzed by WT CHMO and Two Mutants (*Acinetobacter*) with Subsequent Intramolecular Translactonization to the More Stable 5-Membered Lactones ((*R*)-3 and (*S*)-3)



As in our previous work,[10] we use the CHMO structure from *Rhodococcus* sp. strain HI-31 as the starting point. It is the only CHMO for which crystal structures are available, and it is homologous to CHMO from *Acinetobacter* sp. NCIMB 9871 (55% sequence identity). Both enzymes display similar profiles with respect to the sense and degree of enantioselectivity.[9b] We continue to employ the "closed" conformation of CHMO in the present calculations. First of all, our previous calculations show that this conformation allows the reaction to occur, and the binding pocket is large enough to accommodate cyclohexanone substrates with suitable (not too large) substituents. One may envision in the case of larger ligands that a less tight "closed" conformation may be formed to enable the reaction. Second, the relevance of the "closed" conformation is supported by the finding that CHMO enantioselectivity is determined by the residues lining its binding pocket. Given this situation, we do not see the need to consider the complicated structural rearrangements proposed by Yachnin et al.[9c] that would drive the ligand from the binding site in the "closed" state to an active site in the "rotated" state, in which the reaction is supposed to happen.[9c] Hence, the current study can also be regarded as a test of the generality of our model[10] by exploring whether QM/MM calculations for the "closed" conformation can elucidate the observed changes in enantioselectivity upon mutation.[4]

2. METHODS

In this paper we adopt the residue notation of CHMO from *Rhodococcus* sp., and hence the numbers of residues around the active site differ by two from those of CHMO from *Acinetobacter* sp.; for example, Phe434 in CHMO (*Rhodococcus*) corresponds to Phe432 in CHMO (*Acinetobacter*). We use the latter notation only when directly referring to that enzyme.

To generate starting structures for the WT complexed with substrate **1**, we manually modified the optimized reactant structure from our previous study[10] by inserting the hydroxyl group as substituent in the 4-position of cyclohexanone, in both the equatorial and axial position, and reoptimized the resulting geometries. In addition, for the Phe434Ser and Phe434Ile mutants, we made the necessary substitutions in the side chain of residue 434.

In the case of the Phe434Ser mutant, the manually prepared structure of the reactant complex was subjected to classical molecular dynamics (MD) simulations for 1 ns, which led to the expected hydrogen-bonding orientation of the Ser434 hydroxyl group toward the equatorial hydroxyl group of **1**, and the final structure from the MD run was taken as starting point for further optimizations. In the reactant complex containing **1** with an axial hydroxyl group, we did not expect specific interactions with the protein side chains, and we thus took the final structure from the MD run for the equatorial conformer, manually shifted the hydroxyl group to the axial position, and optimized the resulting structure.

In the case of the Phe434Ile mutant, there might be some effect of the bulky Ile434 side chain on the local structure of the binding site. Therefore, we subjected the reactant complex for each conformer of **1** to separate MD simulations for 1 ns to generate the starting structures for the subsequent geometry optimizations.

Since we aim at explaining enantioselectivity, which is a matter of small energy differences of the order of 1−3 kcal/mol between the enantiomers, we must be careful to locate the lowest relevant minima and transition states on the potential energy surface. Therefore, we adopted a procedure, in which we follow the reaction path from the initially optimized reactant complex to the rate-limiting transition state for the migration step (TS2)[10] where we try to find the lowest-energy conformer and then follow the reaction path back to the reactant (and forward to the product) to refine the overall path. In the following, we present results for the lowest-energy paths only. The chosen procedure is essential to avoid artifacts arising, e.g., from the rotation of the hydroxyl group of **1**, which could lead to different minima and could thus spoil any attempt to model and understand enantioselectivity.

In the setup phase, we adopted the same MD procedure as in our previous work[10] using the CHARMM22 force field[12] as implemented within the CHARMM program.[13] During the MD simulations the whole system was moving freely except for the cofactors, the coordinates of which were kept fixed at the positions of the crystal structure. A potential was imposed on the surrounding water sphere to prevent the outer solvent water molecules from drifting away into the vacuum. CHARMM parameters for **1** were assigned using the CHARMM general force field (CGenFF) philosophy[14] on the basis of the cyclohexanone parameters that had previously been deposited into CGenFF.

We used exactly the same QM/MM methodology[15] as in our previous work.[10] Briefly, the QM part was treated by density functional theory (DFT) with the B3LYP functional,[16] and the MM part was described by the CHARMM22 force field, as in the classical MD simulations. The QM/MM calculations were performed using the ChemShell package,[17] combining the TURBOMOLE program[18] for the QM part and the DL_POLY program[19] for the MM part. An electronic embedding scheme[20] was adopted, in which the MM point charges are incorporated into the one-electron part of the Hamiltonian during the QM calculation. The QM/MM boundary was treated with the charge shift model and hydrogen link atoms.[21]

In the present study we worked with several different QM regions. By default, we employed for all reaction-path optimizations the same QM region as in our previous study,[10] incorporating all atoms from the isoalloxazine ring of the peroxyflavin, the whole ligand (**1**), the side chain of Arg-329, and the nicotineamide ring and the adjacent ribose of $NADP^+$ cofactor (97 atoms, "small" QM region). Aiming for accurate relative energies of the rate-limiting transition state (TS2), we increased the QM region in the corresponding optimizations by including parts of the interacting active-site residues 434−436, to an extent depending on the particular system. In the case of the WT enzyme and the Phe434Ile mutant, the "large" QM region (109 atoms) also contained the Phe434/Ile434 carbonyl group, the full Thr435 backbone, and the Asn436 amino and α-CH groups; this choice avoids a QM/MM boundary across a polar C−N bond and conserves neutrality of the CHARMM atom groups. In the case of the Phe434Ser mutant, the "large" QM region (108 atoms) included the Ser434 side chain, the Thr435 carbonyl group, and the Asn436 amino and α-CH groups. Finally, in the WT enzyme, we used an even larger QM region (127 atoms) that also incorporates the side chain of Phe434 to check for the effect of the weak π-hydroxyl interaction. This was not deemed necessary in the Phe434Ile mutant where the Ile434 side chain was treated at the MM level because of the lack of specific interactions with the substrate. The chosen QM regions are fully specified in the Supporting Information.

The TZVP basis set[22] was used in all B3LYP/CHARMM geometry optimizations. Single-point calculations were performed for TS2 at the B3LYP-D2/TZVP[23] and M06-2X/TZVP[24] levels.

Optimizations were done with the use of the HDLCopt optimizer[25] as implemented in the ChemShell program. The active region to be optimized included all QM atoms as well as all residues and water molecules of the MM region within 12 Å of the C4a atom in the isoalloxazine ring of the peroxyflavin. Reaction paths were scanned along the same reaction coordinates as in our previous work[10] by performing sequences of restrained optimizations. The resulting starting structures for the stationary points were subjected to full unrestrained optimizations. Energy minimizations were performed with the low-memory Broyden−Fletcher−Goldfarb−Shanno (L-BFGS) algorithm,[26] while transition state (TS) searches were carried out using a microiterative TS optimizer that combines an uphill TS search in the core region (a small subset of atoms directly involved in the reaction) using the partitioned rational function method[27] (P-RFO) with an L-BFGS downhill relaxation of the environment.

Frequency calculations were performed for the core region to confirm that the optimized TS structures are characterized by a single imaginary frequency and a suitable transition vector. Thereafter, as in our previous work,[10] we displaced the optimized TS structure slightly along the direction of the transition vector and subjected it to unrestrained optimizations to ensure that the TS is indeed connected to the proper minima by a continuous pathway.

3. RESULTS AND DISCUSSION

The WT CHMO enzyme shows weak (*R*)-enantiopreference in the reaction with **1**, converting it to the corresponding lactone with only 9% *ee*, in contrast to other 4-substituted cyclohexanone derivatives such as the 4-methyl or 4-chloro compounds that usually afford the corresponding (*S*)-lactones with high enantioselectivity exceeding 96% *ee*. As part of a directed evolution project we have previously found that a single mutation at a particular residue (Phe432) in CHMO (*Acinetobacter*) may drastically change the enantioselectivity in the B−V reaction of 1.[6] For example, the WT enantioselectivity is reversed in the Phe432Ser mutant but enhanced in the Phe432Ile mutant.

Before the active-site structure of CHMO was known experimentally, an empirical model serving as a useful mnemonic device[3a,4,7k,28] was used to explain and predict enantioselectivity, focusing on the specific conformation of the Criegee intermediate. The model assumes that it is the preferred orientation of the substituent that determines enantioselectivity in WT CHMO. For example, a methyl substituent at the C4 atom of cyclohexanone prefers to be in the equatorial orientation, which is reflected in (*S*)-enantioselectivity. According to *ab initio* calculations,[4] compound **1** in the absence of CHMO has a slightly lower energy when the hydroxyl substituent at C4 is in the axial (rather than the equatorial) conformation, which may then be associated with the observed weak (*R*)-enantioselectivity of WT CHMO toward **1**. The reaction mechanism established in our previous study[10] supports this notion. However, the energy difference between the two chair conformers of **1** (hydroxy equatorial versus axial) may be different in the binding pocket of WT CHMO and mutants due to the possibility of H-bonding. In the rigid asymmetric Criegee intermediate, only one of the two σ-bonds adjacent to the carbonyl carbon is allowed to migrate for stereoelectronic reasons. Therefore, the preferred orientation of the substituent determines enantioselectivity. Taking into account the hydrophobic nature of the residues around the WT CHMO binding site, there are no obvious specific interactions between substrate and protein environment, and the relative stabilities of the equatorial and axial conformers of the substrate are thus expected to be qualitatively similar as in the gas phase. In our previous study, the preference for the equatorial conformer of 4-methylcyclohexanone was however found to be more pronounced in the CHMO binding site than in the gas phase.[10]

According to Kayser and Clouthier,[4] Phe432 must be situated in the active site of CHMO (*Acinetobacter*) in close vicinity to the ligand, so that its mutation may alter the WT enantioselectivity of CHMO by introducing a specific H-bond interaction with **1**. In the "closed" crystal structure of CHMO (*Rhodococcus*) and in our optimized geometries (see Figure 1), Phe434 is indeed one of the residues that constitutes the binding site and remains in direct proximity to the substrate during the whole reaction. One can easily see that it is situated next to the equatorial side of the C4 atom of cyclohexanone. In the Phe434Ser mutant, a serine residue in this position should be able to form a hydrogen bond with substrate **1** that has a hydroxyl group at the C4 atom of cyclohexanone. In the case of the Phe434Ile mutant, steric effects are likely rather than specific interactions.

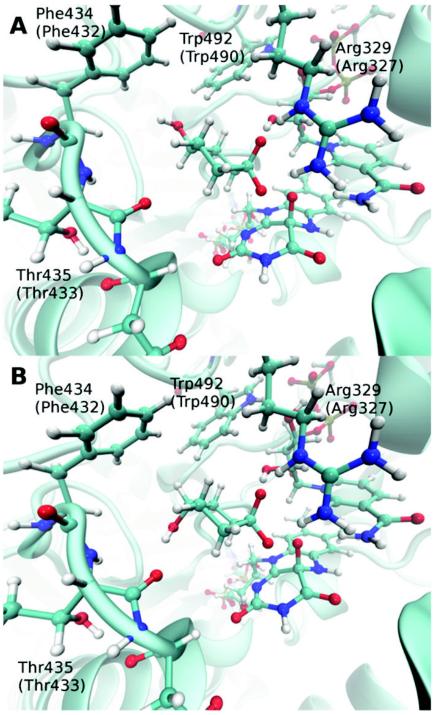


Figure 1. Transition state for the migration step: active site in WT CHMO (*Rhodococcus*) with the hydroxyl group of **1** being in (A) equatorial and (B) axial position. The residue numbers in parentheses label the analogous residues in CHMO from *Acinetobacter* sp. NCIMB 9871.

For our target systems, we performed full reaction path calculations and optimized all relevant stationary points, except for the kinetically unimportant[10] transition state for the initial addition step. We covered the WT enzyme as well as the Phe434Ser and Phe434Ile mutants, and we considered in each case both the equatorial and axial orientations of the hydroxyl group in substrate **1**. The computed overall energy profiles are generally close to each other and to the reference results previously obtained for cyclohexanone oxygenation in the WT CHMO enzyme[10] (see Supporting Information, Figures S1−

S3). The relative energies of the stationary points along the reaction path are generally well separated for the two substrate conformers (equatorial versus axial hydroxyl group in **1**). We use the difference in the corresponding QM/MM energies of the rate-limiting transition state (TS2), calculated with the largest QM region, to estimate the *ee* value (enantiomeric excess). We thus approximate the relevant free energy difference (ΔG) by the corresponding energy difference (ΔE), assuming that the contributions from zero-point vibrational and thermal corrections are negligibly small (see Supporting Information for further details and formulas).

We should again emphasize at this point that our main objective is to understand the origin of enantioselectivity in the enzymatic Baeyer–Villiger reaction of 4-hydroxycyclohexanone. Given the great overall similarity between the relevant transition states for the equatorial and axial substrate isomers (see below), we do not expect conformational sampling to change the qualitative preferences among them, thus obviating the need for demanding free energy calculations. Likewise, for analogous reasons, we believe that computations for a single snapshot are sufficient for our purposes. While it is well-known that individual QM/MM barriers may vary by a few kcal/mol between different snapshots and when including corrections from sampling, these variations are expected to be very similar for the two closely related transition states that govern enantioselectivity in the present case.

3.1. WT CHMO. In the WT enzyme (Figure 1), substrate **1** with the hydroxyl group in axial orientation forms an H-bond with the Thr435 carbonyl oxygen already in the reactant complex, and this interaction is preserved along the whole reaction path. At TS2 the H-bond distance is 1.92 (1.73) Å if the Thr435 carbonyl group is part of the QM (MM) region. In the substrate conformer with an equatorial hydroxyl substituent (see Figure S4), the lowest-energy TS2 structure has the hydroxyl group oriented toward the benzene ring of the Phe434 side chain, creating a weak π-hydroxyl H-bond; in addition, there appears to be a very weak interaction between the hydroxyl oxygen and a rather distant Trp492 ring hydrogen atom (2.72 Å).

We now consider the relative energies of TS2 for the axial and equatorial substrate conformers. The experimentally observed weak (*R*)-enantioselectivity of 9% *ee* translates into a preference of 0.1 kcal/mol for TS2 with an axial substrate hydroxyl group. The QM/MM energies also predict (*R*)-enantioselectivity but overestimate its extent. The rate-limiting transition state TS2 with an axial substrate hydroxyl group is favored by 1.7 kcal/mol when using the small QM region with Phe434 and Thr435 in the MM part and by 0.5 kcal/mol with the two larger QM regions that incorporate the backbones of these two residues. The larger QM regions are expected to be more realistic because inclusion of the Thr435 carbonyl group may help to avoid overpolarization of the substrate hydroxyl group. The energy difference between the axial and equatorial TS2 structures (0.5 kcal/mol) does not change when the Phe434 side chain is incorporated in the largest QM region to improve the description of the corresponding π-hydroxyl interaction.

3.2. Phe434Ser Mutant. As already discussed, the replacement of Phe434 by serine in WT CHMO is anticipated to allow formation of a hydrogen bond between serine and substrate **1** with the hydroxyl group in equatorial position, immediately after ligand binding. Therefore, we decided to perform an MD simulation of the manually prepared reactant complex to enable spatial adjustment of the Ser434 side chain and the substrate. The anticipated H-bond formation could indeed be observed during the 1 ns MD run between the hydrogen atom of the Ser434 hydroxyl group and the oxygen atom of the equatorial substrate hydroxyl group. However, this H-bond was not present during the whole MD run, since the serine–substrate H-bonding distance varied between 2.0 and 5.5 Å, due to the rotation of the Ser434 hydroxyl group (see Figure S5). This H-bond is thus apparently rather weak, as already concluded[4] from the experimental observation that this Phe-Ser substitution does not reverse the enantioselectivity for related substrates, e.g., disubstituted 4-hydroxy-4-methylcyclohexanone.

In the final geometry of the MD simulation, the serine–substrate H-bonding distance was 2.02 Å. This structure was taken as starting point for QM/MM geometry optimization of the reactant complex, which gave an even shorter H-bonding distance of 1.89 Å. Because of the flexibility of the Ser434 side chain, this H-bond is preserved during the full reaction path. At TS2 (Figure 2), the H-bonding distance is computed to be 1.90 Å for the small QM region (Ser434 in the MM part) and 1.97 Å for the large QM region (including Ser434).

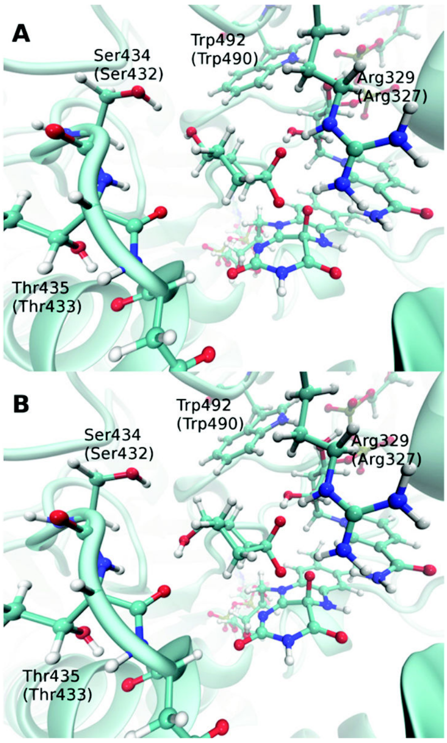


Figure 2. Transition state for the migration step: active site in the Phe434Ser mutant of CHMO (*Rhodococcus*) with the hydroxyl group of **1** in (A) equatorial and (B) axial position. The residue numbers in parentheses label the analogous residues in CHMO from *Acinetobacter* sp. NCIMB 9871.

In the case of axial orientation, a hydrogen bond can be formed between the substrate hydroxyl group and the Thr435 carbonyl oxygen atom (as in the WT enzyme). Qualitatively, the reaction paths for the WT enzyme and the Phe434Ser mutant are found to be completely analogous.

We now address the relative energies of TS2 for the axial and equatorial substrate conformers. The experimentally observed

(*S*)-enantioselectivity of 79% *ee* corresponds to a preference of 1.3 kcal/mol for TS2 with an equatorial substrate hydroxyl group. The rate-limiting transition state TS2 with an equatorial substrate hydroxyl group is *disfavored* by 0.5 kcal/mol when using the small QM region with Ser434 and Thr435 in the MM part and is *favored* by 1.7 kcal/mol for the large QM region that incorporates the Ser434 side chain and the Thr435 backbone. The QM/MM energies thus predict the right (*S*)-enantioselectivity only when employing the large QM region (89% *ee* in reasonable agreement with the experimental value). Qualitatively, the reversal of enantioselectivity upon going from the WT enzyme to the Phe434Ser mutant is due to the serine–substrate hydrogen bond, which is already formed in the reactant complex and is maintained throughout the whole reaction path. For a reliable quantitative assessment of the relative strengths of different H-bond interactions at the QM/MM level, it is thus important to include all H-bonding partners (substrate, Ser434, Thr435) in the QM region. When using the small QM region (Ser434 and Thr435 in the MM part), there is apparently some overpolarization of the QM region by the MM point charges of these residues, which occurs to a different extent for the axial and equatorial substrate conformers. Judging from the computed distances, the H-bond is stronger in both cases when using the small QM region, but this effect is more pronounced for the axial conformer (Thr435-substrate H-bond distances of 1.75 Å versus 1.95 Å with the small and large QM region, respectively) than for the equatorial conformer (Ser434-substrate H-bond distances of 1.90 Å versus 1.97 Å). A realistic assessment of such small energy differences at the QM/MM level thus requires the use of large QM regions that capture all relevant interactions.

3.3. Phe434Ile Mutant.

MD simulations for the reactant complex between substrate **1** and the Phe434Ile mutant of CHMO did not reveal any spatial clashes between **1** and the Ile434 side chain because the side chain is partially pointing out of the binding pocket and thus allows the ligand to adopt its usual orientation (as in the WT enzyme). In the absence of substrate, however, the Ile434 side chain can flip from the outside to the inside of the active site. This may cause problems for the substrate to bind with the hydroxyl group being in equatorial position. A ligand-free 1 ns MD simulation indeed shows such spontaneous flipping of the Ile434 side chain.

As in the case of the WT enzyme, the axial substrate conformer forms a hydrogen bond between the hydrogen atom of its hydroxyl group and the Thr435 carbonyl oxygen atom, which is maintained throughout the whole reaction path (including TS2, see Figure 3B). In the equatorial substrate conformer, the hydroxyl group is oriented toward the Ile434 residue. In the optimized TS2 structure (see Figure 3A), the whole Ile434 backbone has undergone some rotation so that its carbonyl group points roughly into the direction of the substrate and can engage in an H-bonding interaction (2.14 Å) with the equatorial hydroxyl group; the corresponding energy gain apparently outweighs the energy penalty arising from backbone rotation. In this regard, the situation is different for the WT enzyme where the Phe434 backbone carbonyl group points outside the active side (see Figure 1A)—there is no need for a rotation of the Phe434 backbone since the Phe434 residue (unlike Ile434) can offer the phenyl π system as an alternative H-bonding partner to the equatorial hydroxyl group. Test calculations indicate that this latter interaction accounts for about 2 kcal/mol of stabilization because the TS2 energy increases by this amount in the WT case when rotating the equatorial hydroxyl group away from the Phe434 phenyl ring.

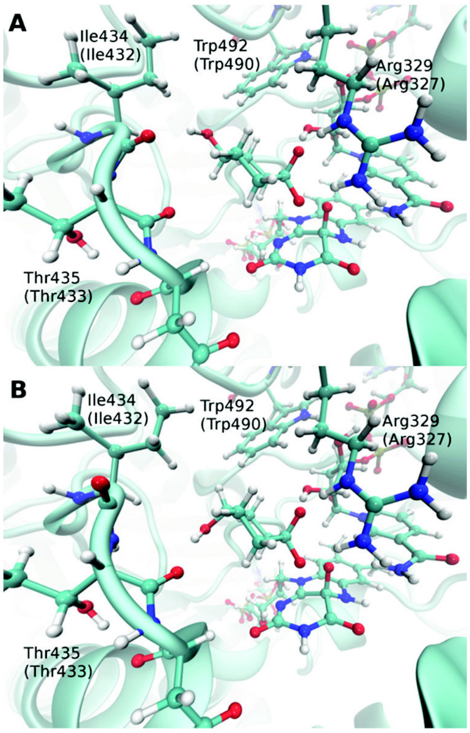


Figure 3. Transition state for the migration step: active site in the Phe434Ile mutant of CHMO (*Rhodococcus*) with the hydroxyl group of **1** in (A) equatorial and (B) axial position. The residue numbers in parentheses label the analogous residues in CHMO from *Acinetobacter* sp. NCIMB 9871.

Given these counteracting effects, it is obviously difficult to predict the enantioselectivity solely by qualitative arguments. The QM/MM calculations with the large QM region (including the Ile434 and Thr435 backbone atoms) favor the TS2 structure with the axial substrate conformer by 1.7 kcal/mol. Hence, they predict (*R*)-enantioselectivity of the Phe434Ile mutant toward **1** with an *ee* value of 89% (compared with the experimental *ee* value of 49%). The calculations thus qualitatively reproduce the observed increase of (*R*)-enantioselectivity when going from the WT enzyme to the Phe434Ile mutant, without being accurate quantitatively.

3.4. Summary and Discussion.

Table 1 collects the QM/MM energy differences between the TS2 structures containing the equatorial and axial substrate conformers for the WT

Table 1. Calculated QM/MM Energy Differences (in kcal/mol) between the TS2 Structures Containing the Equatorial and Axial Substrate Conformers (Equatorial − Axial) for the WT CHMO Enzyme and Two Mutants[a]

	B3LYP/TZVP//B3LYP/TZVP	B3LYP-D2/TZVP//B3LYP/TZVP	M06-2X/TZVP//B3LYP/TZVP
WT	0.5	0.7	1.1
Phe434Ser	−1.7	−1.4	−0.2
Phe434Ile	1.7	0.5	0.8

[a] Results are given for three different QM methods (see text). The residue numbers refer to CHMO (*Rhodococcus*).

CHMO enzyme as well as the Phe434Ser and Phe434Ile mutants, calculated with the largest QM region at the QM(B3LYP/TZVP)/CHARMM optimized geometries. The values for the two DFT approaches that include midrange dispersion (B3LYP-D2 and M06-2X) were obtained from single-point calculations. In many respects, they give a qualitatively similar picture as the B3LYP-based results, but there are also some differences. Since the computed energy differences are generally quite small, their values may be nontrivially affected by reoptimization with dispersion included (B3LYP-D2 and M06-2X), and hence we refrain from further discussion of these single-point results.

The results of our QM/MM calculations are qualitatively in line with the experimental data on CHMO enantioselectivity toward **1**. The basic mechanism of the CHMO-catalyzed oxygenation of cyclohexanone has been elucidated in atomistic detail by our previous QM/MM study,[10] which established the transition state (TS2) for the migration step in the Criegee intermediate to be rate-limiting. The stereochemical outcome of the reaction is thus governed by the lowest-energy TS2 structure. In the case of substrate **1**, the reaction exhibits (*R*)- or (*S*)-enantioselectivity if the lowest-energy TS2 structure contains the substrate with an axial or equatorial hydroxyl group, respectively. This preference is largely determined by the possible interactions between a given substrate conformer and the active-site environment in the TS2 region. The WT CHMO enzyme shows a slight (*R*)-enantiopreference because the axial substrate−Thr435 H-bond is stronger than the equatorial substrate−Phe434 π-hydroxyl H-bond (Figure 1). This enantiopreference is enhanced in the Phe434Ile mutant with an analogous axial substrate−Thr435 H-bond because the competing equatorial substrate−Ile434 interaction is weaker than in the case of the WT enzyme (Figure 3). The Phe434Ser mutant shows the opposite (*S*)-enantiopreference because the equatorial substrate−Ser434 H-bond prevails over the alternative axial substrate−Thr435 H-bond (Figure 2).

To generalize, CHMO enantioselectivity toward substrates with polar substituents is the result of a fine interplay of the weak interactions between the substrate and the protein environment. This makes it difficult to give explanations or to make predictions for any particular substrate/enzyme combination without doing explicit QM/MM calculations, which can capture and quantify these specific interactions. The situation should be easier for substrates with nonpolar substituents (like methyl) where steric effects should dominate over specific electronic effects. In such cases, one may draw conclusions about the origin of enantioselectivity by considering the generic structure of the Criegee intermediate and the rate-limiting transition state for migration,[10] by examining possible steric clashes between nonpolar substituents and active-site residues, and by taking into account the intrinsic preferences among substrate conformers as determined in the gas phase.

Finally, the overall agreement of our calculations with the experimental data[7b] in the present study supports the validity of the underlying model,[10] namely, that the enzymatic Baeyer−Villiger reaction takes place in the "closed" conformation of CHMO. The observed enantioselectivity can be explained in our model without the need to invoke the recently proposed "rotated" conformation,[9c] in which the substrate becomes separated from the residues that govern enantioselectivity according to the present work.

4. CONCLUSION

We have performed a computational investigation on the origin of enantioselectivity in the B−V reaction of 4-hydroxycyclohexanone (**1**) catalyzed by cyclohexanone monooxygenase (CHMO) from *Rhodococcus* and by two mutants (Phe434Ser and Phe434Ile). As shown in our previous QM/MM work,[10] the migration step is rate-limiting and stereoselective because of the requirement of anti-periplanarity for the migrating σ-bond so that a given orientation of a substituent at the cyclohexanone ring gives rise to a particular enantiomeric product. Judging from the QM(B3LYP/TZVP)/CHARMM results, the energy difference between the rate-limiting migration transition states (TS2) for the equatorial and axial chair conformers of substrate **1** is 0.5 kcal/mol in the WT CHMO enzyme, in favor of the axial form and implying slight (*R*)-enantiopreference. In the Phe434Ser mutant, the energetic order of these transition states is reversed, and the equatorial form is favored by 1.7 kcal/mol leading to (*S*)-enantioselectivity. By contrast, the Phe434Ile mutant again shows (*R*)-enantioselectivity that is enhanced compared to the WT enzyme, with a preference of 1.7 kcal/mol for TS2 with an axial substrate conformer. These results agree qualitatively and to some extent even semiquantitatively with the experimental data[7b] on the enantiopreference observed for CHMO from *Acinetobacter* toward **1**. The enantioselectivity in CHMO-catalyzed reactions is determined by a fine interplay of weak interactions between the ligand and surrounding active-site residues and can be rationalized by an analysis of these interactions in the TS2 structures that fulfill the basic stereoelectronic requirement of anti-periplanarity for the migrating σ-bond.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

QM/MM energy profiles for the reactions studied, relative energies for all stationary points, detailed definition of the chosen QM regions, energy profile for rotation of the equatorial substrate hydroxyl group in the Criegee intermediate of the WT CHMO, Ser434-ligand hydrogen bond dynamics, evaluation of *ee* values, and complete refs 12, 13b, 14, and 17. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) (a) Mihovilovic, M. D. Hydrolysis and Formation of Epoxides. In *Enzyme Catalysis in Organic Synthesis*; Drauz, K., Gröger, H., May, O., Eds.; Wiley-VCH: Weinheim, 2012; Vol. 3. (b) Wohlgemuth, R. Tools for Selective Enzyme Reaction Steps in the Synthesis of Laboratory Chemicals. *Eng. Life Sci.* **2006**, *6* (6), 583. (c) Rehdorf, J.; Bornscheuer, M. *Monooxygenases, Baeyer Villiger Applications in Organic Synthesis*; John Wiley & Sons: Hoboken, NJ, 2009. (d) de Gonzalo, G.; Mihovilovic, M. D.; Fraaije, M. W. Recent Developments in the Application of Baeyer−Villiger Monooxygenases as Biocatalysts. *ChemBioChem* **2010**, *11* (16), 2208−2231. (e) Torres Pazmiño, D. E.; Dudek, H. M.; Fraaije, M. W. Baeyer−Villiger Monooxygenases: Recent Advances and Future Challenges. *Curr. Opin. Chem. Biol.* **2010**, *14* (2), 138−144. (f) Leisch, H.; Morley, K.; Lau, P. C. K. Baeyer−Villiger Monooxygenases: More Than Just Green Chemistry. *Chem. Rev.* **2011**, *111* (7), 4165−4222.

(2) (a) Chen, Y. C. J.; Peoples, O. P.; Walsh, C. T. Acinetobacter Cyclohexanone Monooxygenase - Gene Cloning and Sequence Determination. *J. Bacteriol.* **1988**, *170* (2), 781−789. (b) Brzostowicz, P. C.; Walters, D. M.; Thomas, S. M.; Nagarajan, V.; Rouviere, P. E. mRNA Differential Display in a Microbial Enrichment Culture: Simultaneous Identification of Three Cyclohexanone Monooxygenases from Three Species. *Appl. Environ. Microbiol.* **2003**, *69* (1), 334−342. (c) Donoghue, N. A.; Norris, D. B.; Trudgill, P. W. Purification and Properties of Cyclohexanone Oxygenase from Nocardia-Globerula Cl 1 and Acinetobacter NCIB-9871. *Eur. J. Biochem.* **1976**, *63* (1), 175−192. (d) Schwab, J. M.; Li, W. B.; Thomas, L. P. Cyclohexanone Oxygenase - Stereochemistry, Enantioselectivity, and Regioselectivity of an Enzyme-Catalyzed Baeyer-Villiger reaction. *J. Am. Chem. Soc.* **1983**, *105* (14), 4800−4808. (e) Turfitt, G. E. The Microbiological Degradation of Steroids. 4. Fission of the Steroid Molecule. *Biochem. J.* **1948**, *42* (3), 376−383. (f) Walsh, C. T.; Chen, Y. C. J. Enzymic Baeyer-Villiger Oxidations by Flavin-Dependent Monooxygenases. *Angew. Chem., Int. Ed. Engl.* **1988**, *27* (3), 333−343.

(3) (a) Stewart, J. D. Cyclohexanone Monooxygenase: A Useful Reagent for Asymmetric Baeyer-Villiger Reactions. *Curr. Org. Chem.* **1998**, *2* (3), 195−216. (b) Alphand, V.; Carrea, G.; Wohlgemuth, R.; Furstoss, R.; Woodley, J. M. Towards Large-Scale Synthetic Applications of Baeyer-Villiger Monooxygenases. *Trends Biotechnol.* **2003**, *21* (7), 318−323. (c) Mihovilovic, M. D. Enzyme Mediated Baeyer-Villiger Oxidations. *Curr. Org. Chem.* **2006**, *10* (11), 1265−1287. (d) Hollmann, F.; Arends, I. W. C. E.; Buehler, K.; Schallmey, A.; Buehler, B. Enzyme-Mediated Oxidations for the Chemist. *Green Chem.* **2011**, *13* (2), 226−265.

(4) Kayser, M. M.; Clouthier, C. M. New Bioorganic Reagents: Evolved Cyclohexanone Monooxygenases - Why Is It More Selective? *J. Org. Chem.* **2006**, *71* (22), 8424−8430.

(5) Recent reviews of directed evolution: (a) Quin, M. B.; Schmidt-Dannert, C. Engineering of Biocatalysts: from Evolution to Creation. *ACS Catal.* **2011**, *1*, 1017−1021. (b) Brustad, E. M.; Arnold, F. H. Optimizing Non-natural Protein Function with Directed Evolution. *Curr. Opin. Chem. Biol.* **2011**, *15*, 201−210. (c) Turner, N. J. Directed Evolution Drives the Next Generation of Biocatalysts. *Nat. Chem. Biol.* **2009**, *5*, 567−573. (d) Jäckel, C.; Kast, P.; Hilvert, D. Protein Design by Directed Evolution. *Annu. Rev. Biophys. Biomol. Struct.* **2008**, *37*, 153−173. (e) Bommarius, A. S.; Blum, J. K.; Abrahamson, M. J. Status of Protein Engineering for Biocatalysts: How To Design an Industrially Useful Biocatalyst. *Curr. Opin. Chem. Biol.* **2011**, *15*, 194−200. (f) Bershtein, S.; Tawfik, D. S. Advances in Laboratory Evolution of Enzymes. *Curr. Opin. Chem. Biol.* **2008**, *12*, 151−158. (g) Dalby, P. A. Strategy and Success for the Directed Evolution of Enzymes. *Curr. Opin. Struct. Biol.* **2011**, *21*, 473−480. (e) Nair, N. U.; Denard, C. A.; Zhao, H. Engineering of Enzymes for Selective Catalysis. *Curr. Org. Chem.* **2010**, *14*, 1870−1882.

(6) Reetz, M. T. Laboratory Evolution of Stereoselective Enzymes: A Prolific Source of Catalysts for Asymmetric Reactions. *Angew. Chem., Int. Ed.* **2011**, *50* (1), 138−174.

(7) (a) Zhang, Z. G.; Parra, L. P.; Reetz, M. T. Protein Engineering of Stereoselective Baeyer-Villiger Monooxygenases. *Chem.—Eur. J.* **2012**, *18* (33), 10160−10172. (b) Reetz, M. T.; Brunner, B.; Schneider, T.; Schulz, F.; Clouthier, C. M.; Kayser, M. M. Directed Evolution as a Method to Create Enantioselective Cyclohexanone Monooxygenases for Catalysis in Baeyer-Villiger Reactions. *Angew. Chem., Int. Ed.* **2004**, *43* (31), 4075−4078. (c) Bocola, M.; Schulz, F.; Leca, F.; Vogel, A.; Fraaije, M. W.; Reetz, M. T. Converting Phenylacetone Monooxygenase into Phenylcyclohexanone Monooxygenase by Rational Design: Towards Practical Baeyer-Villiger Monooxygenases. *Adv. Synth. Catal.* **2005**, *347* (7−8), 979−986. (d) Clouthier, C. M.; Kayser, M. M.; Reetz, M. T. Designing New Baeyer-Villiger Monooxygenases Using Restricted CASTing. *J. Org. Chem.* **2006**, *71* (22), 8431−8437. (e) Kirschner, A.; Bornscheuer, U. T. Kinetic Resolution of 4-hydroxy-2-ketones Catalyzed by a Baeyer-Villiger Monooxygenase. *Angew. Chem., Int. Ed.* **2006**, *45* (42), 7004−7006. (f) Mihovilovic, M. D.; Rudroff, F.; Winninger, A.; Schneider, T.; Schulz, F.; Reetz, M. T. Microbial Baeyer-Villiger Oxidation: Stereopreference and Substrate Acceptance of Cyclohexanone Monooxygenase Mutants Prepared by Directed Evolution. *Org. Lett.* **2006**, *8* (6), 1221−1224. (g) Kirschner, A.; Bornscheuer, U. T. Directed Evolution of a Baeyer-Villiger Monooxygenase to Enhance Enantioselectivity. *Appl. Microbiol. Biotechnol.* **2008**, *81* (3), 465−472. (h) Reetz, M. T.; Wu, S. Laboratory Evolution of Robust and Enantioselective Baeyer-Villiger Monooxygenases for Asymmetric Catalysis. *J. Am. Chem. Soc.* **2009**, *131* (42), 15424−15432. (i) Rehdorf, J.; Mihovilovic, M. D.; Bornscheuer, U. T. Exploiting the Regioselectivity of Baeyer-Villiger Monooxygenases for the Formation of beta-Amino Acids and beta-Amino Alcohols. *Angew. Chem., Int. Ed.* **2010**, *49* (26), 4506−4508. (j) Wu, S.; Acevedo, J. P.; Reetz, M. T. Induced Allostery in the Directed Evolution of an Enantioselective Baeyer-Villiger Monooxygenase. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (7), 2775−2780. (k) Kayser, M. M. 'Designer Reagents' Recombinant Microorganisms: New and Powerful Tools for Organic Synthesis. *Tetrahedron* **2009**, *65* (5), 947−974.

(8) (a) Ryerson, C. C.; Ballou, D. P.; Walsh, C. Mechanistic Studies on Cyclohexanone Oxygenase. *Biochemistry* **1982**, *21* (11), 2644−2655. (b) Sheng, D. W.; Ballou, D. P.; Massey, V. Mechanistic Studies of Cyclohexanone Monooxygenase: Chemical Properties of Intermediates Involved in Catalysis. *Biochemistry* **2001**, *40* (37), 11156−11167. (c) Pazmino, D. E. T.; Baas, B.-J.; Janssen, D. B.; Fraaije, M. W. Kinetic Mechanism of Phenylacetone Monooxygenase from Thermobifida Fusca. *Biochemistry* **2008**, *47* (13), 4082−4093. (d) Orru, R.; Dudek, H. M.; Martinoli, C.; Pazmino, D. E. T.; Royant, A.; Weik, M.; Fraaije, M. W.; Mattevi, A. Snapshots of Enzymatic Baeyer-Villiger Catalysis Oxygen Activation and Intermediate Stabilization. *J. Biol. Chem.* **2011**, *286* (33), 29284−29291.

(9) (a) Malito, E.; Alfieri, A.; Fraaije, M. W.; Mattevi, A. Crystal Structure of a Baeyer-Villiger Monooxygenase. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (36), 13157−13162. (b) Mirza, I. A.; Yachnin, B. J.; Wang, S.; Grosse, S.; Bergeron, H.; Imura, A.; Iwaki, H.; Hasegawa, Y.; Lau, P. C. K.; Berghuis, A. M. Crystal Structures of Cyclohexanone Monooxygenase Reveal Complex Domain Movements and a Sliding Cofactor. *J. Am. Chem. Soc.* **2009**, *131* (25), 8848−8854. (c) Yachnin, B. J.; Sprules, T.; McEvoy, M. B.; Lau, P. C. K.; Berghuis, A. M. The Substrate-Bound Crystal Structure of a Baeyer-Villiger Monooxygenase Exhibits a Criegee-like Conformation. *J. Am. Chem. Soc.* **2012**, *134* (18), 7788−7795.

(10) Polyak, I.; Reetz, M. T.; Thiel, W. Quantum Mechanical/Molecular Mechanical Study on the Mechanism of the Enzymatic Baeyer-Villiger Reaction. *J. Am. Chem. Soc.* **2012**, *134* (5), 2732−2741.

(11) Zhang, Z.-G.; Roiban, G.-D.; Acevedo, J. P.; Polyak, I.; Reetz, M. T. A New Type of Stereoselectivity in Baeyer−Villiger Reactions: Access to E- and Z-Olefins. *Adv. Synth. Catal.* **2013**, *355* (1), 99−106.

(12) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586−3616.

(13) (a) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187−217. (b) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545−1614.

(14) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31* (4), 671−690.

(15) (a) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions - Dielectric, Electrostatic and Steric Stabilization of Carbonium-ion in Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103* (2), 227−249. (b) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48* (7), 1198−1229.

(16) (a) Slater, J. C. A Generalized Self-Consistent Field Method. *Phys. Rev.* **1953**, *91* (3), 528−530. (b) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin-Density Calculations - A Critical Analysis. *Can. J. Phys.* **1980**, *58* (8), 1200−1211. (c) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098−3100. (d) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Phys. Rev. B* **1988**, *37* (2), 785−789. (e) Becke, A. D. Density-Functional Thermochemistry. 3. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648−5652. (f) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab-initio Calculation of Vibrational Absorption and Circular-Dichroism Spectra Using Density-Functional Force-Fields. *J. Phys. Chem.* **1994**, *98* (45), 11623−11627.

(17) Sherwood, P.; deVries, 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.; et al. QUASI: A General Purpose Implementation of the QM/MM Approach and Its Application to Problems in Catalysis. *THEOCHEM* **2003**, *632*, 1−28.

(18) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic-Structure Calculations on Workstation Computers - the Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162* (3), 165−169.

(19) Smith, W.; Forester, T. R. DL_POLY_2.0: A General-Purpose Parallel Molecular Dynamics Simulation Package. *J. Mol. Graphics* **1996**, *14* (3), 136−141.

(20) Bakowies, D.; Thiel, W. Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches. *J. Phys. Chem.* **1996**, *100* (25), 10580−10594.

(21) (a) Sherwood, P.; deVries, A. H.; Collins, S. J.; Greatbanks, S. P.; Burton, N. A.; Vincent, M. A.; Hillier, I. H. Computer Simulation of Zeolite Structure and Reactivity Using Embedded Cluster Methods. *Faraday Discuss.* **1997**, *106*, 79−92. (b) de Vries, A. H.; Sherwood, P.; Collins, S. J.; Rigby, A. M.; Rigutto, M.; Kramer, G. J. Zeolite Structure and Reactivity by Combined Quantum-Chemical-Classical Calculations. *J. Phys. Chem. B* **1999**, *103* (29), 6133−6141.

(22) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian-Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829−5835.

(23) Grimme, S. Semiempirical GGA-type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787−1799.

(24) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120* (1−3), 215−241.

(25) Billeter, S. R.; Turner, A. J.; Thiel, W. Linear Scaling Geometry Optimisation and Transition State Search in Hybrid Delocalised Internal Coordinates. *Phys. Chem. Chem. Phys.* **2000**, *2* (10), 2177−2186.

(26) (a) Nocedal, J. Updating Quasi-Newton Matrices with Limited Storage. *Math. Comput.* **1980**, *35* (151), 773−782. (b) Liu, D. C.; Nocedal, J. On the Limited Memory BFGS Method for Large-Scale Optimization. *Math. Program.* **1989**, *45* (3), 503−528.

(27) (a) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. Search for Stationary-Points on Surface. *J. Phys. Chem.* **1985**, *89* (1), 52−57. (b) Baker, J. An Algorithm for the Location of Transition-States. *J. Comput. Chem.* **1986**, *7* (4), 385−395.

(28) Stewart, J. D.; Reed, K. W.; Martinez, C. A.; Zhu, J.; Chen, G.; Kayser, M. M. Recombinant Baker's Yeast as a Whole-Cell Catalyst for Asymmetric Baeyer-Villiger Oxidations. *J. Am. Chem. Soc.* **1998**, *120* (15), 3541−3548.

Supporting Information

Quantum Mechanical/Molecular Mechanical Study on the Enantioselectivity of the Enzymatic Baeyer-Villiger Reaction of 4-Hydroxycyclohexanone

Iakov Polyak[†], *Manfred T. Reetz*[†,‡], *and Walter Thiel**[†]

[†]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-45470 Mülheim an der Ruhr, Germany; [‡]Fachbereich Chemie, Philipps-Universität Marburg, Hans-Meerwein-Straße, D-35032 Marburg, Germany

Contents

Energy diagrams

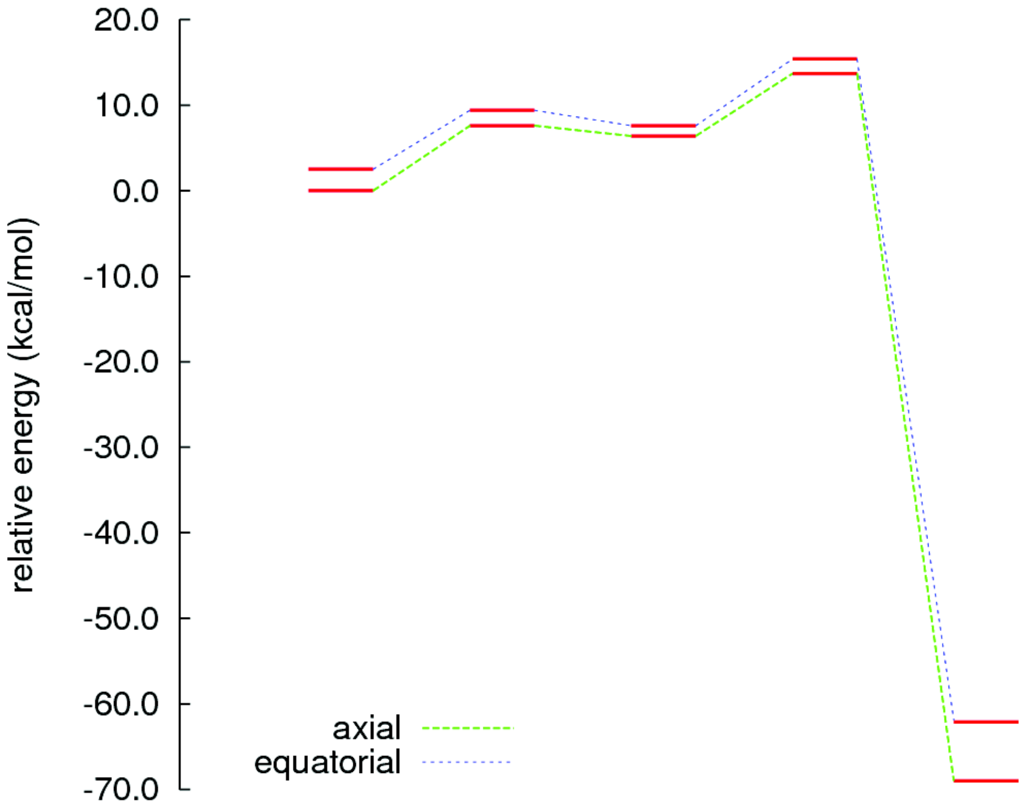


Figure S1. QM(B3LYP/TZVP)/CHARMM energy profiles for the WT CHMO (*Rhodococcus*) with the hydroxyl group of **1** in equatorial and axial position. From left to right, the energy levels correspond to the reactant complex, the transition state for the addition step (TS1), the Criegee intermediate, the transition state for the migration step (TS2), and the product complex. The energy profiles have been computed using the small QM region 1(see below). For energy values, see Table S1.

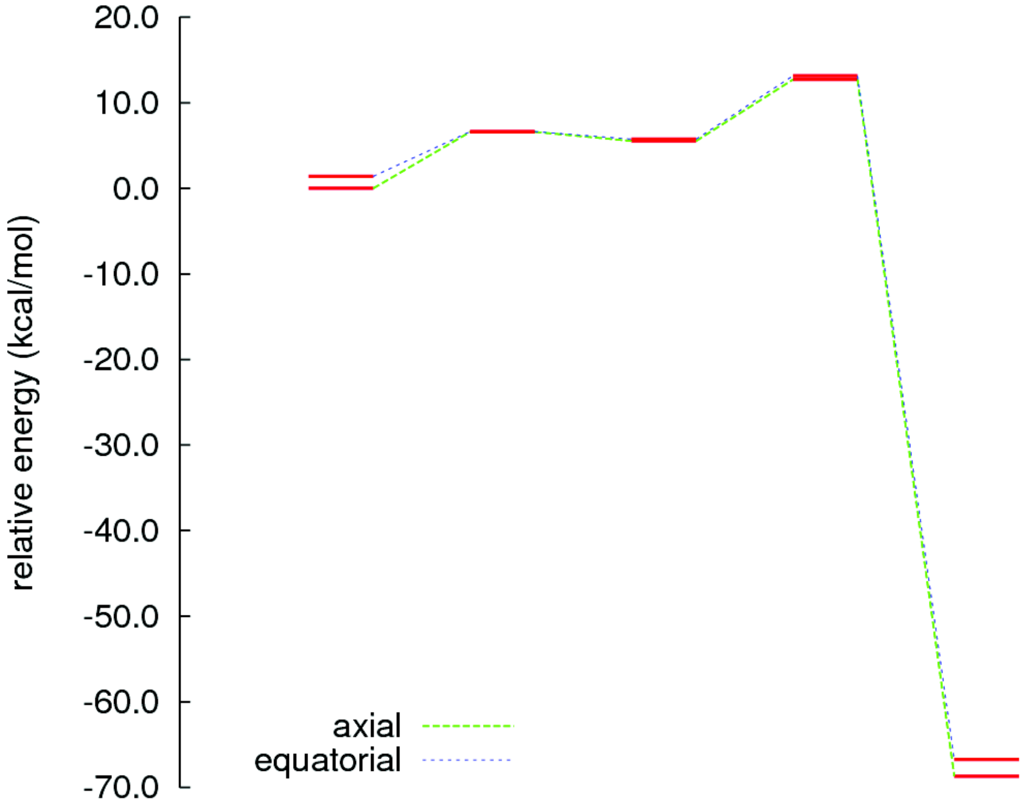


Figure S2. QM(B3LYP/TZVP)/CHARMM energy profiles for the Phe434Ser mutant of CHMO (*Rhodococcus*) with the hydroxyl group of **1** in equatorial and axial position. From left to right, the energy levels correspond to the reactant complex, the transition state for the addition step (TS1), the Criegee intermediate, the transition state for the migration step (TS2), and the product complex. The energy profiles have been computed using the small QM region 1 (see below). For energy values, see Table S2.

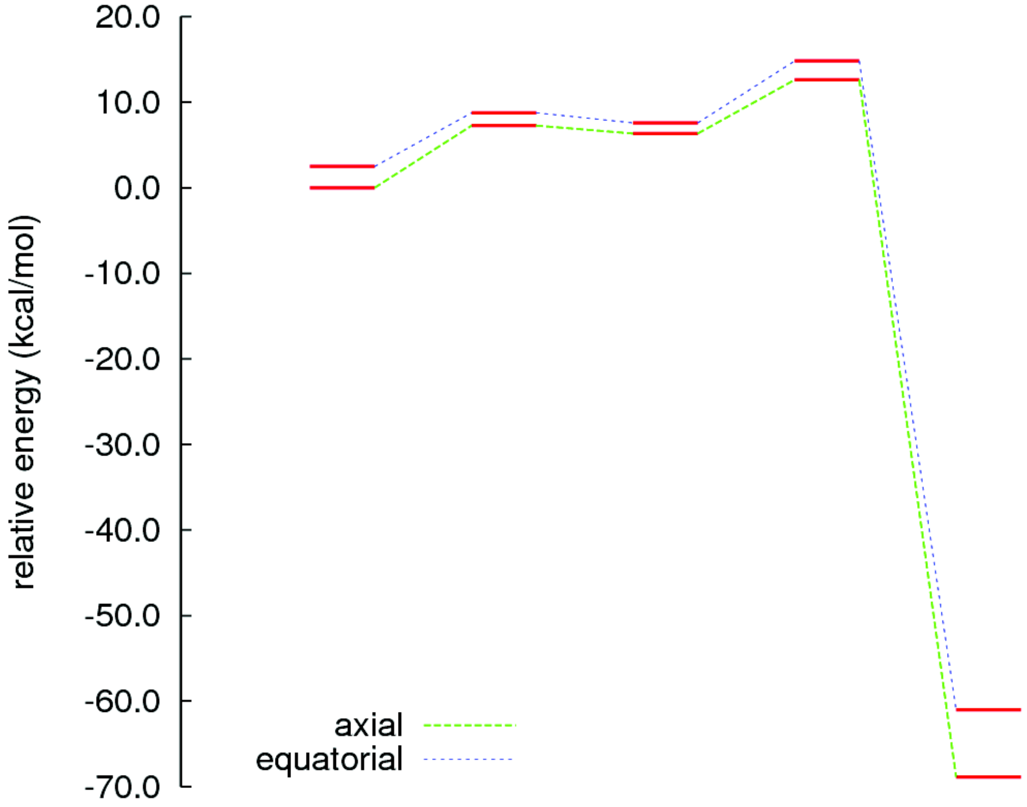


Figure S3. QM(B3LYP/TZVP)/CHARMM energy profiles for the Phe434Ile mutant of CHMO (*Rhodococcus*) with the hydroxyl group of **1** in equatorial and axial position. From left to right, the energy levels correspond to the reactant complex, the transition state for the addition step (TS1), the Criegee intermediate, the transition state for the migration step (TS2), and the product complex. The energy profiles have been computed using the small QM region 1 (see below). For energy values, see Table S3.

	Reactant	TS1 scan	Intermediate	TS2	Product
Equatorial	2.5	9.4	7.6	15.4	-62.1
Axial	0.0	7.6	6.4	13.7	-69.0

Table S1. Calculated QM(B3LYP/TZVP)/CHARMM energies (in kcal/mol) for stationary points along the reaction profile for the WT CHMO (*Rhodococcus*) with the hydroxyl group of **1** in equatorial and axial position. The energy profiles have been computed using the small QM region 1 (see below). Energies are given relative to the reactant complex with axial configuration.

	Reactant	TS1 scan	Intermediate	TS2	Product
Equatorial	1.4	6.6	5.7	13.2	-66.7
Axial	0.0	6.6	5.5	12.7	-68.7

Table S2. Calculated QM(B3LYP/TZVP)/CHARMM energies (in kcal/mol) for stationary points along the reaction profile for the Phe434Ser mutant of CHMO (*Rhodococcus*) with the hydroxyl group of **1** in equatorial and axial position. The energy profiles have been computed using the small QM region 1 (see below). Energies are given relative to the reactant complex with axial configuration.

	Reactant	TS1 scan	Intermediate	TS2	Product
Equatorial	2.5	8.8	7.6	14.8	-61.0
Axial	0.0	7.3	6.3	12.6	-68.9

Table S3. Calculated QM(B3LYP/TZVP)/CHARMM energies (in kcal/mol) for stationary points along the reaction profile for the Phe434Ile mutant of CHMO (*Rhodococcus*) with the hydroxyl group of **1** in equatorial and axial position. The energy profiles have been computed using the small QM region 1 (see below). Energies are given relative to the reactant complex with axial configuration.

Please note:

The transition states TS2 have been reoptimized at the QM(B3LYP/TZVP)/CHARMM level using the larger QM regions 2 and 3 (see below). The resulting energy differences between the TS2 structures containing the equatorial and axial substrate conformers are given in Table 1 of the main paper for the WT CHMO enzyme and the two mutants. They are considered more reliable than those derived from Tables S1-S3 because of the use of larger QM regions.

QM regions

WT CHMO

QM region 1 : Isoalloxazine ring of the peroxyflavin, ligand (**1**), the side chain of Arg329, the nicotineamide ring and the adjacent ribose of NADP$^+$ cofactor (97 atoms).

QM region 2 : Isoalloxazine ring of the peroxyflavin, ligand (**1**), the side chain of Arg329, the nicotineamide ring and the adjacent ribose of NADP$^+$ cofactor; the carbonyl group of Phe434; the full backbone of Thr435; the amino group and the α-CH group of Asn436 (109 atoms).

QM region 3 : Isoalloxazine ring of the peroxyflavin, ligand (**1**), the side chain of Arg329, the nicotineamide ring and the adjacent ribose of NADP$^+$ cofactor; the carbonyl group, the α-CH group, and the side-chain of Phe434; the full backbone of Thr435; the amino group and the α-CH group of Asn436 (127 atoms).

Phe434Ser mutant of CHMO

QM region 1 : Isoalloxazine ring of the peroxyflavin, ligand (**1**), the side chain of Arg329, the nicotineamide ring and the adjacent ribose of NADP$^+$ cofactor (97 atoms).

QM region 2 : Isoalloxazine ring of the peroxyflavin, ligand (**1**), the side chain of Arg329, the nicotineamide ring and the adjacent ribose of NADP$^+$ cofactor; the side chain of Ser434; the carbonyl group of Thr435; the amino group and the α-CH group of Asn436 (108 atoms).

Phe434Ile mutant of CHMO

QM region 1 : Isoalloxazine ring of the peroxyflavin, ligand (**1**), the side chain of Arg329, the nicotineamide ring and the adjacent ribose of NADP$^+$ cofactor (97 atoms).

QM region 2 : Isoalloxazine ring of the peroxyflavin, ligand (**1**), the side chain of Arg329, the nicotineamide ring and the adjacent ribose of NADP$^+$ cofactor; the carbonyl group of Ile434; the full backbone of Thr435; the amino group and the α-CH group of Asn436 (109 atoms).

Ligand hydroxyl group rotation in the Criegee intermediate of the WT CHMO.

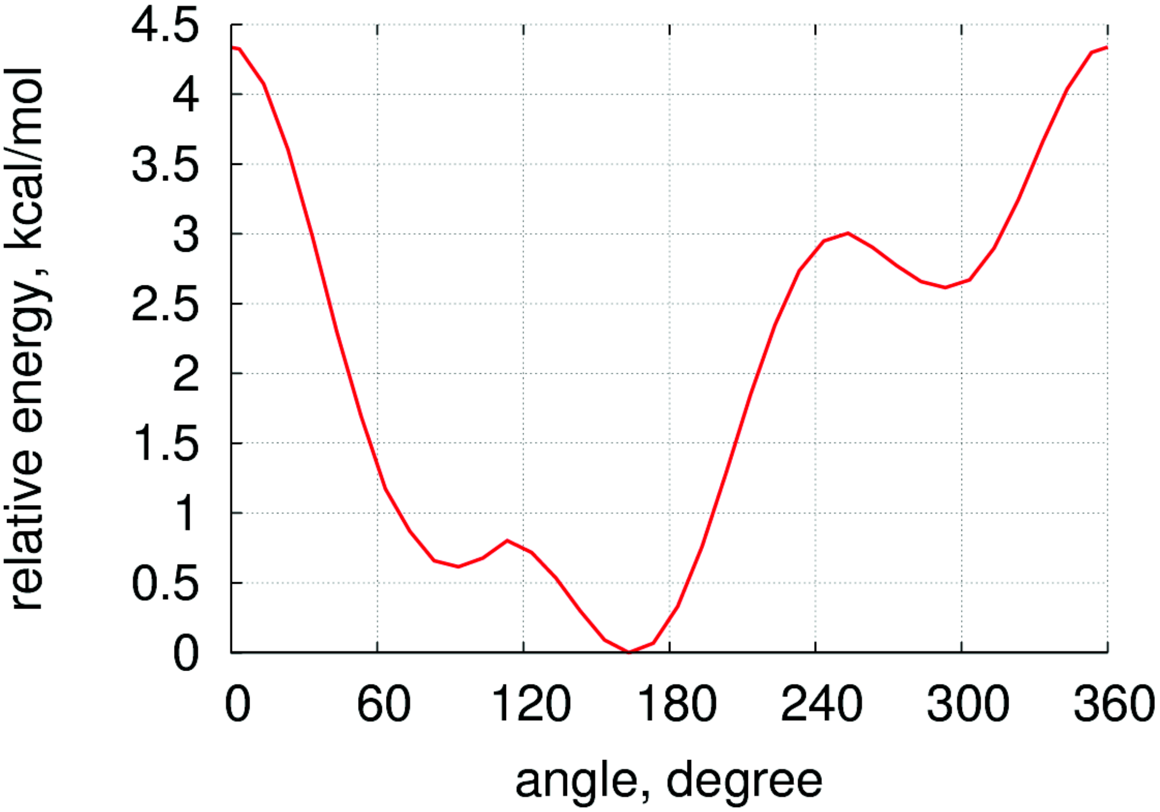


Figure S4. Energy profile for rotation of the equatorial hydroxyl group of the ligand (**1**) in the Criegee intermediate of the WT CHMO (*Rhodococcus*) from QM/MM geometry optimizations using the small QM region 1. At the energy minimum, the hydroxyl group is oriented towards the benzene ring of the Phe434 side chain, creating a weak π-hydroxyl H-bond. This orientation was adopted in the computation of the energy profile of the enzymatic Baeyer-Villiger reaction.

Ser434 – ligand hydrogen bond dynamics

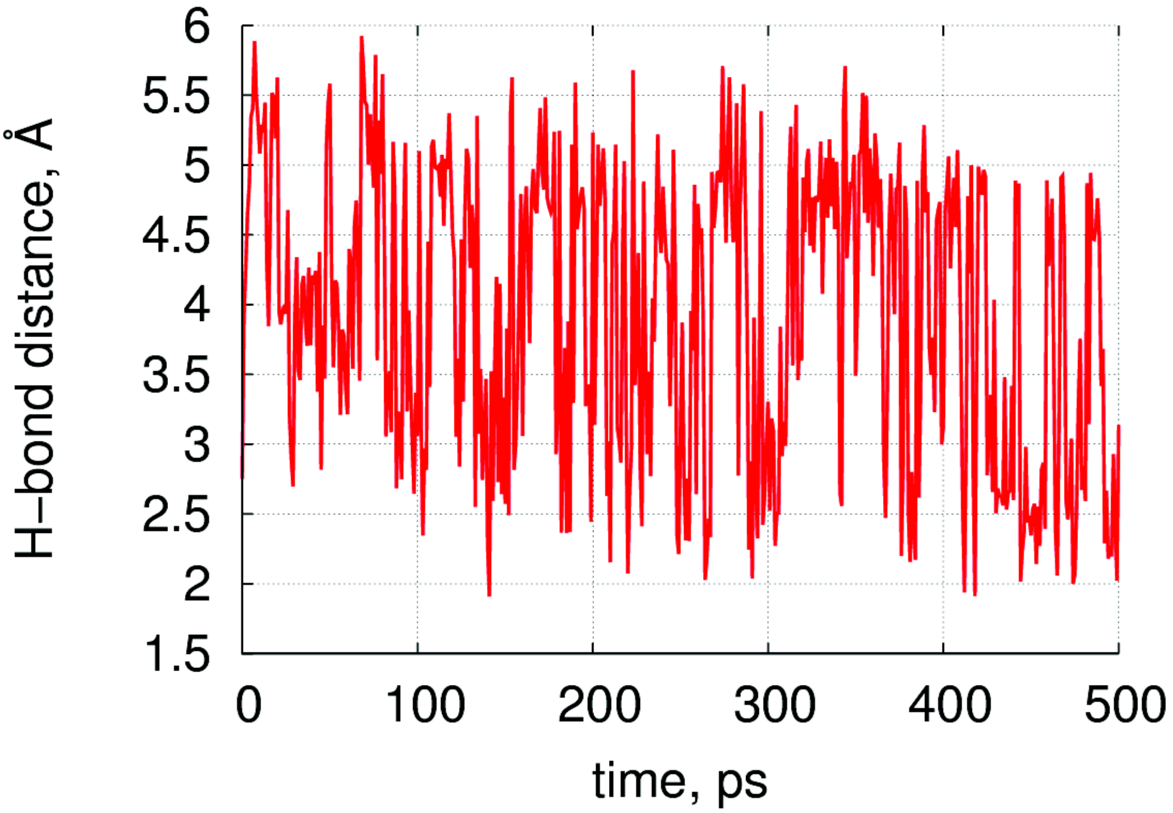


Figure S5. Changes in the H-bonding distance between the hydrogen atom of the Ser434 hydroxyl group and the oxygen atom of the equatorial ligand hydroxyl group during the 500ps classical MD sampling run.

Evaluation of the *ee* value (enantiomeric excess)

We evaluate the *ee* value from the following formula:

$$ee = \frac{\exp(-\frac{\Delta E^{TS}}{RT}) - 1}{\exp\left(-\frac{\Delta E^{TS}}{RT}\right) + 1} * 100\%$$

Here, $\Delta E^{TS} = E^{TS}_{favoured} - E^{TS}_{disfavoured}$ is the difference between the QM/MM energies of the rate-determining transition states TS2 containing the equatorial and axial substrate conformers. We take the difference between the absolute TS2 energies because at equilibrium the initial reference state is the same for both conformers, namely the lowest-energy reactants at infinite separation (enzyme and substrate). In this manner we also account for the energy difference between the two isolated conformers.

Strictly speaking, we would need to evaluate the *ee* value from differences in free energies (ΔG) rather than energies (ΔE). In the above formula, we approximate the free energies of the transition states by their potential energies, assuming that the differences in the corresponding zero-point vibrational energies and thermal corrections are negligibly small.

Conversely, we can determine the energy difference associated with the experimental *ee* value (given as a decimal number in the range 0-1 rather than in percent):

$$\Delta E^{TS} = RT \ln\left(\frac{1 + ee}{1 - ee}\right)$$

Complete references 12, 13b, 14 and 17 of the main paper

12. 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.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998,** *102* (18), 3586-3616.

13. (b) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009,** *30* (10), 1545-1614.

14. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010,** *31* (4), 671-690.

17. 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.; Sjovoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C., QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis. *Theochem* **2003,** *632*, 1-28.

Retaining Glycosyltransferase Mechanism Studied by QM/MM Methods: Lipopolysaccharyl-α-1,4-galactosyltransferase C Transfers α-Galactose via an Oxocarbenium Ion-like Transition State

H. Gomez, I. Polyak, W. Thiel, J. M. Lluch, L. Masgrau

J. Am. Chem. Soc., **134***,* 4743 (2012)

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
pubs.acs.org/JACS

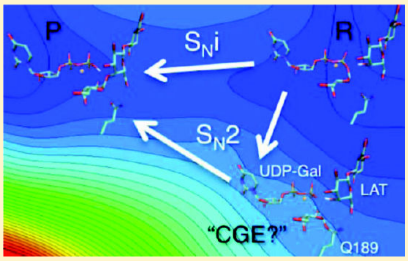
Retaining Glycosyltransferase Mechanism Studied by QM/MM Methods: Lipopolysaccharyl-α-1,4-galactosyltransferase C Transfers α-Galactose via an Oxocarbenium Ion-like Transition State

Hansel Gómez,[‡,†] Iakov Polyak,[§] Walter Thiel,[§] José M. Lluch,[‡,†] and Laura Masgrau*[,‡]

[‡]Institut de Biotecnologia i de Biomedicina and [†]Departament de Química, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

[§]Max-Planck-Institut für Kohlenforschung, D-45470 Mülheim an der Ruhr, Germany

Supporting Information



ABSTRACT: Glycosyltransferases (GTs) catalyze the highly specific biosynthesis of glycosidic bonds and, as such, are important both as drug targets and for biotechnological purposes. Despite their broad interest, fundamental questions about their reaction mechanism remain to be answered, especially for those GTs that transfer the sugar with net retention of the configuration at the anomeric carbon (retaining glycosyltransferases, ret-GTs). In the present work, we focus on the reaction catalyzed by lipopolysaccharyl-α-1,4-galactosyltransferase C (LgtC) from *Neisseria meningitides*. We study and compare the different proposed mechanisms (S_Ni, S_Ni-like, and double displacement mechanism via a covalent glycosyl-enzyme intermediate, CGE) by using density functional theory (DFT) and quantum mechanics/molecular mechanics (QM/MM) calculations on the full enzyme. We characterize a dissociative single-displacement (S_Ni) mechanism consistent with the experimental data, in which the acceptor substrate attacks on the side of the UDP leaving group that acts as a catalytic base. We identify several key interactions that help this front-side attack by stabilizing the transition state. Among them, Gln189, the putative nucleophile in a double displacement mechanism, is shown to favor the charge development at the anomeric center by about 2 kcal/mol, compatible with experimental mutagenesis data. We predict that using 3-deoxylactose as acceptor would result in a reduction of k_{cat} to 0.6−3% of that for the unmodified substrates. The reactions of the Q189A and Q189E mutants have also been investigated. For Q189E, there is a change in mechanism since a CGE can be formed which, however, is not able to evolve to products. The current findings are discussed in the light of the available experimental data and compared with those for other ret-GTs.

INTRODUCTION

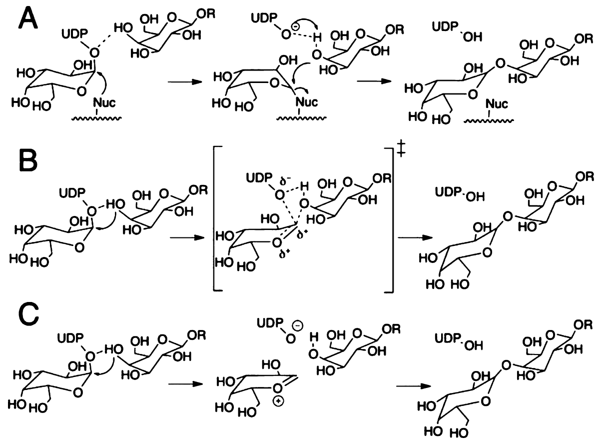
Glycosyltransferases (GTs) are a large family of enzymes that catalyze the biosynthesis of glycosidic linkages by the stereo- and regiospecific transfer of monosaccharides from donors, typically nucleotide sugars, to a variety of acceptors (other saccharides, lipids, proteins, DNA).[1] GTs are responsible for the synthesis of biopolymers, therapeutically important glycosylated natural product antibiotics, and anticancer agents. Moreover, they are part of important biological pathways and are implicated in many diseases. Despite the relevance of their reactions and products in different fields, the understanding of their catalytic mechanism remains one of the fundamental challenges in glycoscience.[2−4] This is specially the case for retaining GTs, which catalyze the sugar transfer with net retention of the configuration at the anomeric carbon.

Neisseria meningitidis is a bacterium that lives within the nasopharyngeal tract of humans and can become a human-specific pathogen causing invasive, life-threatening infections, such as meningitis and septicemia.[5] Its great virulence is related to the presence of lipooligosaccharides (LOSs) on its cell surface that can mimic those of human glycolipids, thus, avoiding the recognition by the human immune system. Lipopolysaccharyl-α-1,4-galactosyltransferase C (EC 2.4.1.44, LgtC) from *N. meningitidis* (a member of family 8 of GTs)[6] plays a key role in the biosynthesis of the oligosaccharide part of LOS structures. Therefore, it represents a very attractive therapeutic target for new antibiotics.[7] This enzyme catalyzes the transfer of an α-galactose from uridine 5′-diphospho-α-galactose (UDP-Gal) to a galactose of the terminal lactose (LAT) on the bacterial LOS to yield an elongated oligosaccharide with an overall retention of stereochemistry at the anomeric carbon atom. Thus, LgtC can be classified as a retaining GT. Moreover, it presents a GT-A fold, one of the three folds reported for GTs, characterized by the presence in the active site of the Asp-X-Asp (DXD) signature implicated in the binding of the donor sugar substrate and/or the coordination of a divalent cation.[4]

Virtually all retaining glycosidases (GHs) utilize a double-displacement reaction mechanism. By analogy, catalysis by retaining GTs was initially believed to proceed in a similar way (Scheme 1A) with the formation and subsequent cleavage of a

Received: November 8, 2011
Published: February 21, 2012

Scheme 1. Proposed Mechanisms for the Retaining GTs[a]



[a](A) Double-displacement mechanism with formation of a covalently bound glycosyl-enzyme intermediate. (B) Concerted front-side single displacement (S_Ni) with an oxocarbenium ion-like transition state. (C) S_Ni-like mechanism in which a short-lived oxocarbenium-phosphate ion pair intermediate is formed.

covalent glycosyl-enzyme (CGE) intermediate, which requires the existence of an appropriately positioned nucleophile within the active site.[3,4,8] However, conclusive evidence for the formation of CGE intermediates has been elusive for many years although several experiments point in that direction.[9−11]

The problem is even more challenging for those retaining GTs in which no good candidate for the catalytic nucleophile seems to exist within the active site. This is the case for LgtC from *N. meningitidis*, whose X-ray crystal structure was the first one determined for a glycosyltransferase in the form of a ternary complex with both donor and acceptor substrate analogues.[12] This 2.0 Å crystal structure contains the inert UDP-2′-deoxy-2′-fluoro-*α*-galactose (UDP-2′FGal) and the nonreactive 4-deoxylactose.[12] Quite surprisingly, the most suitably positioned active-site residue on the *β*-face of the donor sugar for a nucleophilic attack on the anomeric carbon is Gln189 (whose side chain oxygen atom is 3.5 Å away from the anomeric carbon). However, the nucleophilic character of the amide of Gln189 should be rather poor. Experimentally, both the Q189A and Q189E LgtC mutants display 3% residual transferase activity, indicating the limited relevance of Gln189 as catalytic nucleophile.[11,12]

An alternative mechanism involving retention of configuration at the anomeric carbon is the so-called S_Ni ("i" for "internal return") mechanism,[13] which occurs through a front-side attack of the nucleophile (the O4 atom of lactose) on the same side as the leaving group (UDP) forming an oxocarbenium ion-like transition state (Scheme 1B), or even an oxocarbenium-phosphate short-lived ion pair intermediate (Scheme 1C) in an S_Ni-like mechanism.[4] A front-side attack might seem unfavorable in view of the limited free space available in the active site of LgtC and the bulky nature of both the ingoing and the outgoing groups, but such steric problems could be alleviated by having a highly dissociative S_Ni transition state (Scheme 1B) or even a short-lived ion pair intermediate (Scheme 1C). Very recently, experimental work on trehalose-6-phosphate synthase (OtsA), a retaining GT of the GT20 family, has provided evidence for a dissociative mechanism with charge development at the anomeric center.[14,15]

Very little computational work has been devoted to the study of retaining GTs and, to the best of our knowledge, only one paper has been published about LgtC: Tvaroška[16] used density functional theory (DFT) methods to study a cluster model based on the crystallographic structure of LgtC. This reduced model contained 136 atoms, including those from five protein residues (three of them coordinating the Mg^{2+} cation). To mimic the positioning effect of the missing protein environment, the *α*-carbons of the five included amino acids were constrained to their crystallographic positions. Tvaroška predicted a one-step S_Ni mechanism with an energy barrier of 31.3 kcal/mol at the B3LYP/6-311++G**//B3LYP/6-31G* level of theory. Treating the environment as a dielectric continuum with dielectric constants of 2 (78) reduced this barrier by 7 (21) kcal/mol and thus led to rough estimates for the barrier in the range of 10−24 kcal/mol. However, the choice of a reduced cluster model neglects the effects of many important residues and interactions that are actually present in the enzyme−substrate complex, which causes uncertainties and prevents a true understanding of the role of the enzyme in driving the reaction (see Figure 1 for a view of all the enzyme−substrate interactions seen in the crystallographic structure).[12]

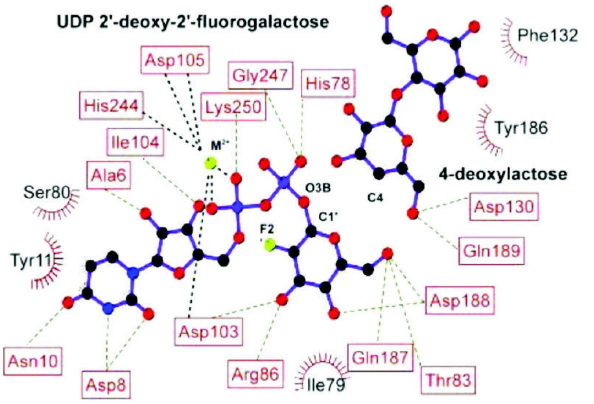


Figure 1. LIGPLOT[17] diagram of UDP-Gal and LAT analogues bound in the LgtC active site (PDB code: 1GA8). The metal coordination and hydrogen bonds between the substrates and the surrounding amino acids are represented by black or green dashed lines, respectively. Nonligand residues in hydrophobic contacts with the substrates are represented by red semicircles with radiating spokes.

In this paper, we present for the first time a full-enzyme hybrid quantum mechanical/molecular mechanical (QM/MM) study of the catalytic mechanism of a retaining glycosyltransferase with the GT-A fold, that is, LgtC from *N. meningitidis*. We explore the different mechanistic alternatives outlined above and analyze the role of the most relevant active-site residues and interactions that support the catalytic action of LgtC despite the absence of a good nucleophilic agent. In addition, we consider the effect of different mutations in the enzyme and of modifications in the substrates. Finally, a comparison with other retaining GTs is made, and possible relationships between specificity and mechanism are discussed.

While this paper was being written, a computational QM/MM study of OtsA was published[18] which concluded that OtsA reacts via an S_Ni-like mechanism with a very short-lived ion-pair intermediate. OtsA differs from LgtC in having the GT-B fold, which lacks the DXD signature and does not require a divalent metal for catalysis. Given these differences, it is unclear whether to expect similar or different mechanistic behavior in LgtC and OtsA.

■ MODELS AND METHODS

Initial coordinates for the wild-type enzyme were taken from the X-ray structure[12] (PDB Code: 1GA8, resolution 2.0 Å). The protonation states of the titratable residues (His, Glu, Asp, Arg, Lys) were chosen based on the pK_a values given by the empirical PROPKA procedure[19] and verified through visual inspection. The substrate analogues present in the crystal structure (UDP 2-deoxy-2′-fluoro-galactose and 4-deoxylactose) were transformed manually into the original substrates, UDP-galactose (UDP-Gal) and lactose (LAT), respectively, while the manganese ion was modeled by Mg^{2+}. This latter choice was also made in the previous model study.[16] In the retaining glycosyltransferases with the GT-A fold, the M^{2+} ion is known to act as a Lewis acid that facilitates the departure of the leaving group,[4] and it is known that both Mn^{2+} and Mg^{2+} can adopt this role in a similar manner.[20] Generally, according to previous theoretical work, both cations behave remarkably similar and exhibit the same coordination preferences.[21] More specifically, both divalent ions have given almost the same results in a thorough validation study[22] of α-glycosidic bond dissociation in sugar phosphates derived from the crystal structure of LgtC; for example, the M05-2X energies (bond lengths) differ between the studied Mg^{2+} and Mn^{2+} complexes by less than 0.2 kcal/mol (0.01 Å). This has been taken as justification for using enzymatic models in which Mn^{2+} is replaced by a computationally more convenient Mg^{2+} cation.[22] Still, test calculations have been done for the most relevant pathways to ensure that this replacement is valid here (see below).

A partial solvation scheme was used to solvate the region of 24 Å around the anomeric center by overlaying a water sphere on the enzyme. The solvated system was first relaxed by performing energy minimizations at the MM level using the CHARMM22 force field[23−25] as implemented in the CHARMM program.[26,27] For the sugar moieties, the topology and parameters from the CHARMM force field for carbohydrates were used, including the recently released ones for glycosidic linkages between hexopyranoses.[28] The solute atoms were initially frozen for 10 000 conjugate gradient optimization steps. In the subsequent minimizations, the restraints on the protein and ligand atoms were gradually released. The prepared system (see Figure 2) contained 6728 atoms, including 755 TIP3P water molecules, and served as starting point for the QM/MM calculations.

The QM/MM calculations were done with the modular program package ChemShell[29] using TURBOMOLE,[30] Gaussian03[31] or MNDO2005[32] to obtain the QM energies and gradients at the DFT (BP86,[33−37] B3LYP[33−35,38−40] and M05-2X[41] functionals) or SCC-DFTB[42,43] levels, respectively. MM energies and gradients were evaluated by DL_POLY,[44] which was accessed through the ChemShell package, using the CHARMM force field. An electronic embedding scheme[45] was adopted in the QM/MM calculations with the MM point charges being incorporated into the one-electron Hamiltonian during the QM calculation. No cutoffs were introduced for the nonbonding MM and QM/MM interactions. Seven hydrogen link atoms were employed to treat the QM/MM boundary with the charge shift model.[46,47]

All residues and water molecules within 12 Å of the anomeric center were included in the optimization process (1225 atoms) as the active region while the remaining atoms were kept fixed. The QM region incorporated 101 atoms: those from the α and β galactose rings (from UDP-Gal and LAT, respectively), Mg^{2+} and its first coordination sphere (phosphate groups from UDP and the side chains of residues Asp103, Asp105 and His244), as well as the side chain of Gln189 (see Scheme 2). The total charge of the QM region was −2.

Reaction paths were scanned by performing constrained optimizations along properly defined reaction coordinates in steps of 0.2 Å. This provided starting structures for subsequent full optimization of all relevant stationary points, employing the low-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)[48,49] algorithm in the case of minimizations and the microiterative optimizer combining both the partitioned rational function optimizer (P-RFO)[50,51] and L-BFGS during the transition state search. All these algorithms are implemented in the HDLCopt[52] module of ChemShell. Frequency calculations on the QM region confirmed that all reported transition states are characterized by a single imaginary frequency and a suitable transition vector that corresponds to the investigated reaction. We ensured by intrinsic reaction coordinate (IRC) calculations and visual inspection of the optimized structures that the computed stationary points are connected by continuous pathways.

Geometry optimizations were generally carried out at the QM(BP86/SVP)/CHARMM level using the BP86 functional and the SVP[53] basis set in combination with the resolution-of-the-identity (RI) approximation.[54,55] For more accurate energy evaluations, we performed single-point energy calculations with other functionals and

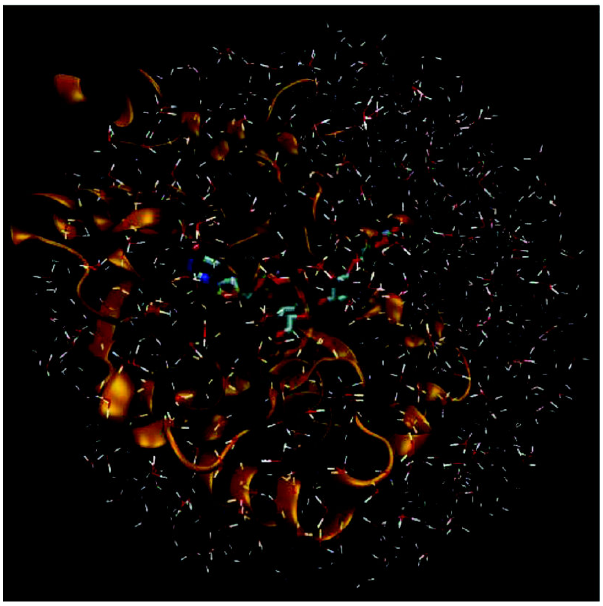
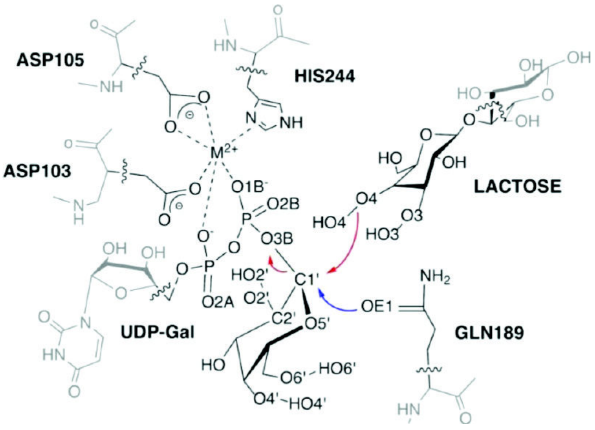


Figure 2. Representation of the system studied: UDP-Gal, LAT, and Mg^{2+} (sticks) bound to LgtC (orange cartoon) and solvated with a 24 Å radius sphere of water molecules.

Scheme 2. Active Site Representation Showing the QM/MM Partition Used in the Present Work[a]



[a]QM (MM) atoms are depicted in black (grey). The boundary between the QM and MM regions is indicated by wavy lines. The first step of the double displacement mechanism and the front-side attack mechanism are illustrated by blue and red arrows, respectively. Atoms mentioned in the text are labeled.

larger basis sets: B3LYP, M05-2X, and DFT-D[56] methods; TZVP[57] and def2-TZVPP(d)[58] basis sets (see Supporting Information (SI) for more details). Natural population analysis (NPA)[59] charges were determined from QM/MM calculations with QM = B3LYP/TZVP and BP86/TZVP.

Umbrella sampling at the SCC-DFTB/CHARMM22 level was performed to compute the potential of mean force (PMF) and the free energy profile for the front-side mechanism using the dynamics module within ChemShell (see SI for more details).

The contribution of different residues to the QM/MM energy in the front-side attack mechanism was examined by setting their point charges to zero in additional energy calculations along the QM(BP86/SVP)/CHARMM reaction path. In the case of Gln189, extra QM/MM calculations were performed in which Gln189 was assigned the charges of a glutamate.

The Molefacture plugin for VMD[60] was used to generate the mutants of the enzyme and the modified substrates. This program also served to generate all the drawings showing molecular structures.

■ RESULTS AND DISCUSSION

Front-Side Attack (S_Ni or S_Ni-like Mechanism). The use of the reaction coordinate (RC = *d*(O3B−C1′) − *d*(O4−C1′)) to model the front-side attack of LAT at UDP-Gal results in a smooth energy profile with a single energy barrier of ~12 kcal/mol at the QM(B3LYP/TZVP//BP86/SVP)/CHARMM level (see Figure 3). The changes in the distances between the reactive atoms along this reaction path (Figure 3) indicate that the computed S_Ni mechanism has a highly dissociative character: the C1′−O3B bond breaks early in the reaction, the attacking O4 atom from LAT slowly approaches C1′, and both the HO4−O3B and O4−C1′ bonds form simultaneously and concomitant with the rapid drop in energy that is observed right after the maximum in Figure 3. It should be noted that, in the reactant complex, the O3B oxygen of the phosphate leaving group is already well oriented to act as the base to deprotonate the acceptor in the S_Ni mechanism (*d*(HO4−O3B) = 1.67 Å). This possible role of the leaving group as the base catalyst has initially been invoked by Sinnott and Jecks[13] for the solvolysis of glycosyl fluoride by trifluoroethanol and has recently also been suggested for OtsA.[14] In the cluster model by Tvaroška, UDP was also found to accept the HO4 proton, but in the reactant state, the O3B−HO4 distance was significantly longer than in the present work (*d*(HO4−O3B) = 2.33 Å).[16] The distances at the energy maximum of Figure 3 are *d*(O3B−C1′) = 3.22 Å, *d*(O4−C1′) = 2.16 Å, *d*(O4−HO4) = 1.07 Å, and *d*(HO4−O3B) = 1.36 Å, clearly indicating a highly dissociative character. This rationalizes the finding that the use of a single distance (*d*(O4−C1′) or *d*(O3B−C1′)) as reaction coordinate for this chemical event did not work satisfactorily (SI, Figure S2).

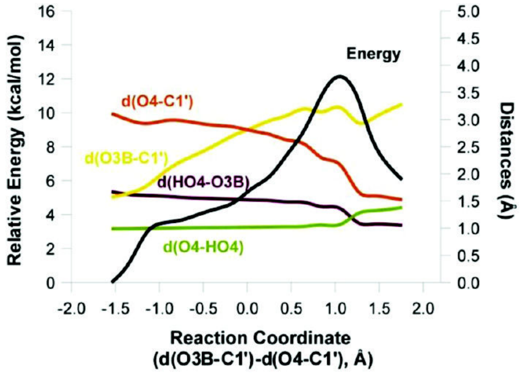


Figure 3. QM(B3LYP/TZVP//BP86/SVP)/CHARMM energy profile for the front-side attack mechanism. The variation of several interatomic distances involved in the reaction is also depicted. See Scheme 2 for a clearer picture of the reaction under study.

The structure of the energy maximum in Figure 3 was used as the starting point for a TS search (see Models and Methods). The computed energy barriers and reaction energies are given in Table 1, while key bond distances and NPA charges of reactants, TS, and products are listed in Table 2. The TS structure is shown in Figure 4. In the following discussion, we focus on single-point QM/CHARMM energies obtained with QM = M05-2X/TZVP and B3LYP/TZVP at geometries optimized with QM = BP86/SVP, which are expected to be our most reliable energy data (SI, Section 1.2). A more detailed comparison of the energies and geometries obtained at different levels is available in the Supporting Information (Section 2 and Table S2).

Table 1. QM/MM Potential Energy Barrier and Reaction Energy (in kcal/mol) for the Proposed S_Ni Mechanism at Different Levels of Theory

QM treatment	$V^{\ddagger}$	ΔV_{R}
BP86/SVP	9.2	3.3
BP86/TZVP//BP86/SVP	8.1	4.6
BP86/def2-TZVPP(d)//BP86/SVP	8.8	4.6
B3LYP/SVP// BP86/SVP	12.9	4.6
B3LYP/SVP	11.1	4.6
B3LYP/TZVP//BP86/SVP	11.8	6.1
B3LYP-D/TZVP//BP86/SVP	9.5	2.8
B3LYP/def2-TZVPP(d)//BP86/SVP	12.5	6.1
M05-2X/SVP//BP86/SVP	16.6	3.6
M05-2X/TZVP//BP86/SVP	14.6	4.6
SCC-DFTB	32.7	9.7
B3LYP/TZVP//SCC-DFTB	15.4	10.3
M05-2X/TZVP//SCC-DFTB	17.2	10.5

Table 2. Selected QM/CHARMM Bond Distances *d* (Å) and Atomic Charges *q* (*e*) in the Optimized Reactants, Transition State, and Products for the Front-Side Single Displacement Mechanism, with QM = BP86/SVP (B3LYP/SVP) for the Distances and QM = B3LYP/TZVP//BP86/SVP for the Charges

	reactant	TS	product
d(O3B−C1′)	1.57(1.53)	3.12(3.19)	3.28(3.30)
d(O4−C1′)	3.11(3.14)	2.18(2.06)	1.52(1.49)
d(O4-HO4)	0.99(0.97)	1.07(1.06)	1.38(1.44)
d(HO4−O3B)	1.67(1.71)	1.36(1.36)	1.06(1.02)
d(OE1−C1′)	3.38(3.40)	2.87(2.89)	3.05(3.09)
d(C1′−O5′)	1.36(1.36)	1.29(1.29)	1.38(1.38)
q(C1′)	0.38	0.49	0.34
q(O5′)	−0.47	−0.41	−0.50

The energy barriers obtained with M05-2X/TZVP and B3LYP/TZVP are 14.6 and 11.8 kcal/mol, respectively. Both values slightly underestimate the phenomenological free energy barrier of ~16 kcal/mol (derived from the experimental k_{cat} values of 14−34 s^{-1} at 303 K),[11,12,61] but they do show that an

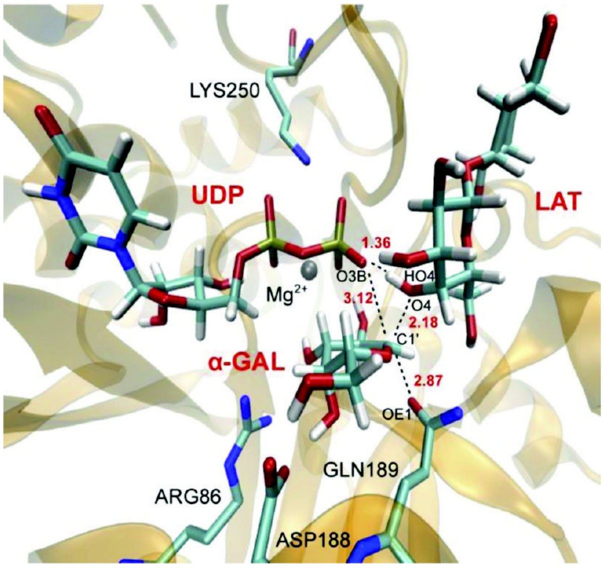


Figure 4. QM(BP86/SVP)/CHARMM optimized transition state. The donor and acceptor substrates, together with some relevant residues, are represented as sticks. Selected distances (in Å) are indicated in red.

S_Ni mechanism is plausible and consistent with the experimental data. Interestingly, the distances obtained for the optimized TS structure [d(O3B−C1′) = 3.12 Å, d(O4−C1′) = 2.18 Å, d(O4−HO4) = 1.07 Å, and d(HO4−O3B) = 1.36 Å] are very similar to those reported for the energy maximum in Figure 3, confirming that the reaction coordinate used to describe the reactive process is adequate. The sugar ring goes from the initial distorted $^4C_1$ chair (φ = 236°, θ = 9°) in the reactant complex to a conformation between an ^4E envelope and a $^4H_5$ half-chair at the TS (φ = 250°, θ = 41°). The oxocarbenium nature of the TS is also reflected in a d(C1′−O5′) distance that is slightly shorter than in the reactant (by 0.07 Å) and in the charge development during the reaction (Table 2). When going from reactant to TS, the charge of the α-Gal moiety increases by $\Delta q(\alpha\text{-Gal}) = 0.30\ e$, with the main contributions coming from the ring atoms C1′ and O5′, $\Delta q(\mathrm{C1'}) = 0.11\ e$ and $\Delta q(\mathrm{O5'}) = 0.06\ e$. This is accompanied by an increase in the negative charge of the UDP moiety ($\Delta q(\mathrm{UDP}) = -0.36\ e$) dominated by the large change at the leaving oxygen atom ($\Delta q(\mathrm{O3B}) = -0.24\ e$), and by a smaller change in the lactose, mainly at the attacking oxygen atom ($\Delta q(\beta\text{-Gal}) = 0.09\ e$, $\Delta q(\mathrm{O4}) = 0.05\ e$).

Our results for LgtC thus indicate that there is a direct one-step path connecting reactant, TS, and products, which corresponds to a highly dissociative S_Ni mechanism (Scheme 1B). This is consistent with recent experiments on another retaining GT. In one of these experiments, Davies and co-workers[14] report the crystal structure of the retaining glycosyltransferase OtsA in a complex with UDP and a bisubstrate analogue, in which the leaving-group phosphate oxygen is sufficiently close to the acceptor nucleophile (analogous to O4 in LgtC) to act as a base to deprotonate the acceptor in a front-side attack mechanism (Scheme 1B or 1C). Kinetic isotope-labeling experiments[15] further support this type of mechanism and indicate an important charge development at the anomeric center at the transition state or intermediate of the reaction. In the computational work of Tvaroška[16] on a cluster model of LgtC, an S_Ni mechanism was also suggested although some structural differences with the present study should be highlighted: the gas-phase cluster model predicts an earlier TS with shorter d(O3B−C1′) = 2.66 Å and longer d(C1′−O4) = 2.34 Å distances, and in the reactant complex, the d(C1′−O4) = 3.50 Å and d(O3B−HO4) = 2.33 Å distances are much longer than those obtained here, probably due to a different donor−acceptor orientation because of the missing enzyme environment. In the very recent computational work on OtsA,[18] which has a GT-B fold and lacks a putative nucleophile residue that could act in a double-displacement mechanism, it has been concluded that the cleavage of the α,α-1,1 linkage is catalyzed in a two-step mechanism via an extremely short-lived ion-pair intermediate (Scheme 1C, S_Ni-like mechanism).

During the front-side attack of LAT on UDP-Gal, the anomeric carbon and the OE1 atom of Gln189 (the putative nucleophile in a double displacement mechanism) get closer by ~0.5 Å. This is mainly caused by the change in the ring puckering on the way from the reactants to the TS (see above). The decrease in the d(C1′−OE1) distance, from 3.38 to 2.87 Å, may help to stabilize the increasing positive charge at the anomeric center. Like in the cluster model,[16] Gln189 (via Nε) forms a hydrogen bond with the O6 atom of β-Gal of LAT, with d(H−O6) distances of 2.0 and 1.9 Å in the reactants and the TS, respectively. This interaction is thus involved both in the binding of LAT and in keeping the proper orientation of the substrate during the reaction. The previously reported[16] interaction between Nε (Gln189) and O5′ is not seen in our QM/MM calculations where Nε (Gln189) is permanently hydrogen bonded to the hydroxyl group of Tyr151. Again, this difference may arise from the limitations of the cluster model that does not include Tyr151.

Umbrella sampling molecular dynamics simulations at the SCC-DFTB/CHARMM22 level were used to estimate the free energy barrier of the front-side mechanism. Although this electronic structure method severely overestimates the energy barrier (Table 1), it is still expected to provide a reasonable estimate of the differences between the potential energy and free energy profiles. These and the corresponding barriers are found to be practically identical, indicating that entropic effects are minor in the reaction under study (SI, Figure S3). This is consistent with the highly dissociative character of the TS.

Double Displacement Mechanism. All attempts to locate the covalent glycosyl-enzyme complex (CGE) for the wild-type enzyme failed. The d(OE1−C1′) distance would seem to be a natural reaction coordinate for driving the reactants toward the CGE intermediate. A corresponding reaction path calculation indeed yields a CGE-type structure, with d(OE1−C1′) = 1.58 Å and d(O3B−C1′) = 3.27 Å. However, the energy profile is monotonously increasing (SI, Figure S4), and an unrestrained minimization of the last point leads back to the reactants again. Along this path, the d(O4−C1′) distance stays close to its initial value of 3.11 Å, showing just some small fluctuations (0.1 Å).

Figure 5 illustrates the different mechanistic proposals for LgtC using a two-dimensional potential energy surface (PES) diagram. At fixed values of the d(OE1−C1′) distance, the QM/MM energy was computed by constrained optimizations along the reaction coordinate RC = d(O3B−C1′) — d(O4−C1′); note that the x-axis in Figure 5 is just d(O4−C1′). Obviously, the only energy minima in Figure 5 are those for the reactants (R, top-right corner) and for the products (P, top-left corner). There is no minimum in the region of the putative intermediate

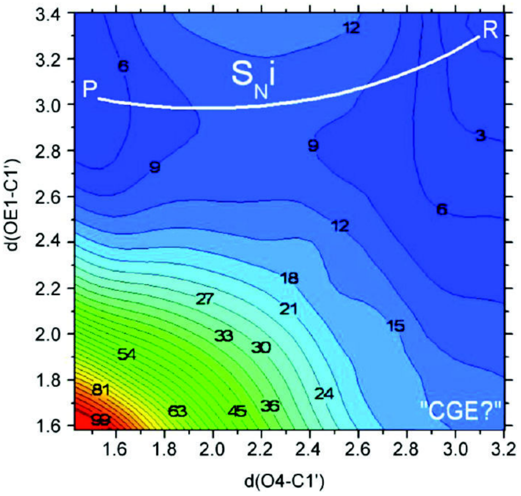


Figure 5. Two-dimensional QM(BP86/SVP)/CHARMM potential energy surface. Energies are given in kcal/mol and distances in Å. Contour lines are drawn in intervals of 3 kcal/mol.

(CGE, bottom-right corner), and hence no evidence for the proposed double-displacement mechanism in the wild-type enzyme that would require CGE formation by nucleophilic attack of Gln189. By contrast, one can easily identify a path connecting reactants and products via an oxocarbenium ion-like transition state (Figure 5). The corresponding transition state is of dissociative character and belongs to a one-step S_Ni reaction.

Test calculations with Mn^{2+} in place of Mg^{2+} have been done to ensure that the replacement used in this work is valid and does not compromise the mechanistic conclusions drawn. The potential energy profiles for both mechanisms remain essentially the same (SI, Figure S5).

Analysis of Factors Contributing to Catalysis. The QM/MM results presented so far show that the front-side attack of LAT at UDP-Gal (S_Ni mechanism) is preferred and can proceed at a reasonable energetic cost in the LgtC active site. The energies and distances computed along the S_Ni reaction path (Figure 3) show that the O3B−C1′ bond starts to break already in the initial stage of the reaction, while the d(O4−C1′) distance still remains above 2.5 Å; the QM/MM energy then reaches its maximum with the O3B−C1′ distance getting even longer and the O4 atom beginning to attack the anomeric center; finally, the energy decreases abruptly after the TS, when the O4−C1′ bond is formed and HO4 is transferred from O4 to O3B. This proton transfer seems crucial for the decrease of the energy in the final stage of this reaction. There are different factors that facilitate the reaction and contribute to the stabilization of the transition state. UDP is known to be a good leaving group, especially if coordinated to a metal cation that can stabilize the negatively charged phosphate. Moreover, it has been suggested[4] that it is because of the use of high-energy donor substrates that retaining GTs can catalyze the reaction via an a priori more expensive S_Ni type mechanism, instead of using a double-displacement mechanism as the retaining glycosidases do. The question then arises whether there are also other factors that may be helpful, for example, intrasubstrate and enzyme−substrate interactions in the TS structure (Figure 4).

Interactions between Substrates. Three hydrogen bonds between the donor and the acceptor substrates can be seen in Figure 4: one involving O4 (as well as HO4) and O3B; another one between O2′ (HO2′) and O1B (i.e., the oxygen atom of the β phosphate coordinated to the metal cation); and a third one between O3 (HO3) and O3B. All of them are present in the reactants (with O−H···O distances of 1.67, 1.68, and 1.94 Å, respectively, at the QM(BP86/SVP)/CHARMM22 level). These hydrogen bonds get shorter at the TS (1.36, 1.64, and 1.65 Å, respectively) and thereby stabilize the increasing negative charge in UDP. Another interaction that may help stabilizing the highly charged phosphate at the TS is the H-bond between O2A of the α phosphate and O3 of the UDP ribose.

To estimate the energy contribution provided by these H-bonds, we produced in silico variants of the substrates (UDP-2′-deoxygalactose and 3-deoxylactose) and calculated the front-side attack pathway for each of them at the QM(BP86/SVP)/CHARMM level using the d(O3B−C1′) — d(O4−C1′) reaction coordinate. This was followed by optimization and characterization of the corresponding transition states. The results are presented in Figure 6 (see also SI, Tables S4 and S6 for other levels of theory). The computed barrier heights confirm that substituting one of these OH-groups by a hydrogen atom impedes the reaction significantly: it increases the barrier by 2.3−3.3 kcal/mol compared with the unmodified substrates. Assuming that these differences are maintained in the free energy barriers, this would reduce the k_{cat} value of the mutants to 0.6−3% of that of the wild type at room temperature. The use of modified substrates leads to somewhat earlier TSs, as indicated by the optimized distances d(O3B−C1′) = 3.09/3.06 Å, d(O4−C1′) = 2.39/2.35 Å, d(O4−HO4) = 1.11/1.08 Å, and d(HO4−O3B) = 1.29/1.35 Å for the substrates UDP-2′-deoxyGal/3-deoxyLAT, respectively.

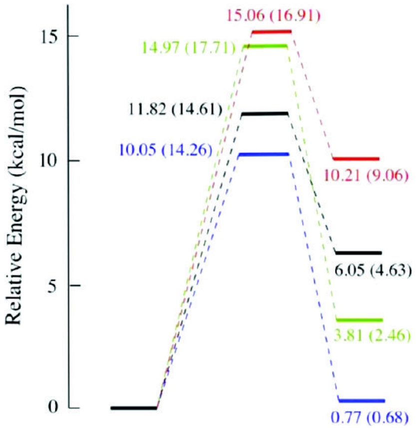


Figure 6. QM/CHARMM potential energy barriers and reaction energies for the front-side attack mechanism (S_Ni), with QM = B3LYP/TZVP//BP86/SVP (M05-2X/TZVP//BP86/SVP). Color code: black, unmodified wild-type enzyme and substrates; red, wild type with modified substrates UDP-2′-deoxygalactose and lactose; green, wild type with modified substrates UDP-Gal and 3-deoxylactose; blue, Q189A mutant with unmodified substrates.

The changes in the computed TS geometries are similar for both modifications, although the replacement of the 3-OH group by H in lactose tends to increase the computed barrier slightly more. By contrast, the reaction energy is affected much more when the 2′−OH group of the α-galactose of UDP-Gal is replaced by H, leading to a rather high endoergicity of ~9 kcal/mol. If we would substitute the 2′−OH group of UDP-Gal by fluorine instead of hydrogen, the oxocarbenium ion-like transition state should be further destabilized inductively, and the result can easily be an inert UDP-2′FGal (which acts as a competitive inhibitor with respect to UDP-Gal), as has been observed experimentally.[12]

Enzyme–substrates Interactions; Key Enzyme Residues. The TS structure in Figure 4 also indicates several interactions between the enzyme and the substrates that may contribute to TS and/or product stabilization and thus facilitate the reaction via this mechanism. These interactions are also present in the reactants and in the crystallographic structure (Figure 1).[12] We have estimated the contribution of individual residues to the (electrostatic) stabilization/destabilization of the system by charge deletion analysis (see Models and Methods). The results for the most relevant residues are shown in Figure 7.

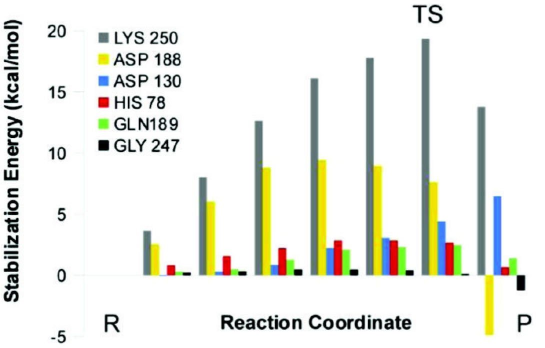


Figure 7. Electrostatic contribution to the stabilization of the QM region by selected residues surrounding the substrates. The reaction coordinate corresponds to the front-side attack mechanism [QM-(BP86/SVP)/CHARMM data].

The most prominent contribution comes from Lys250, with ~19 kcal/mol of TS stabilization compared to the reactants. This strong effect is caused by the interaction of the charged Lys side chain with the α and β phosphates of the leaving UDP: in fact, a new H-bond involving O2B appears at the TS (H···O2B distance: 2.98 Å in the reactants and 2.18 Å in the TS). On the other hand, the H-bond between Lys250 and O2A lengthens from 1.67 Å in the reactants to 1.80 Å in the TS. Lys250 is highly conserved in the GT8 family and belongs to one of the two loops that fold over the donor substrate and are thought to be disordered in its absence;[12] hence, Lys250 is also crucial for substrate binding. The H-bond of the β phosphate (O2B) with Gly247 is very weak, while that with His78 provides a stabilization of 2.5 kcal/mol around the TS (and up to 3 kcal/mol earlier). The electrostatic interactions involving Asp188 (H-bonded to O4′ and O6′ of α-Gal) are most important (up to 9 kcal/mol) at an early stage of the reaction, while those involving Asp130 (H-bonded to O6 of β-Gal of lactose) become more prominent as the reaction proceeds (6 kcal/mol around the TS). Although some residues like Asp188 and Lys250 exhibit strong stabilization effects, their inclusion in the QM region results in negligible differences when comparing the potential energy profiles (SI, Figure S6). These results confirm that the QM/MM partition used in this work satisfactorily describes such enzyme–substrates interactions.

Gln189, the putative nucleophile in a double displacement mechanism, provides very little stabilization in the reactants, which increases up to 2.5 kcal/mol on the route to the TS, where it amounts to ~2 kcal/mol before it drops off again (Figure 7). This stabilization mostly comes from the interaction of the anomeric center C1′ and the OE1 atom of Gln189 and is correlated with the changes in the C1′–OE1 distance (see Table 2). The transient decrease of this distance helps to stabilize the charge development at C1′ during the reaction: this charge shows a quick initial rise and then levels off before decreasing again around the TS (Figure 8).

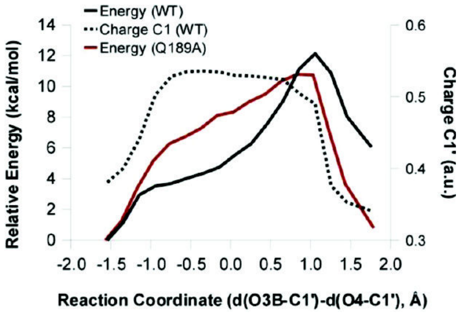


Figure 8. QM(B3LYP/TZVP)/CHARMM energy profile along the d(O3B–C1′) − d(O4–C1′) reaction coordinate at QM(BP86/SVP)/CHARMM optimized geometries, for the wild-type enzyme and the Q189A mutant. The charges at the anomeric center of the wild type enzyme were computed at the same level of theory.

Q189A Mutant. Experimentally, the Q189A mutant has its k_{cat} value reduced to 3% of that of the wild-type enzyme,[12] which translates into an increase of about 2 kcal/mol in the phenomenological free energy barrier. At first sight, this agrees well with the computed electrostatic TS stabilization due to Gln189 (see above) which will be absent in the Q189A mutant. Of course, the Q189A replacement causes other changes too, and therefore, we decided to build the Q189A mutant in silico starting from the coordinates of the wild-type enzyme and making the required substitutions for residue 189. The energy profile for the front-side attack reaction was then recalculated (see Figure 8 and SI, Figure S7), and the corresponding reactants, TS, and products were reoptimized.

Contrary to expectation, the computed QM/CHARMM energy barrier for the Q189A mutant (see Figure 6 and SI, Table S6) is slightly smaller than that for the wild-type enzyme, by 1.8 kcal/mol for QM = B3LYP/TZVP//BP86/SVP and by 0.4 kcal/mol for QM = M05-2X/TZVP//BP86/SVP. The effect of the mutation on the potential energy barrier is thus relatively small, especially for M05-2X, but still in the wrong direction.

Inspection of the computed energy profiles shows that the initial breaking of the C1′–O3B bond and approach of the attacking O4 atom is more difficult for the Q189A mutant, making its energy profile much wider and higher initially (Figure 8). However, the TS occurs somewhat earlier in the mutant compared with the wild type, and the energy thus has to rise more in the latter before the TS is reached (Figure 8). In the optimized QM(BP86/SVP)/CHARMM TS structure, the key distances are d(O3B–C1′) = 3.03 Å, d(O4–C1′) = 2.40 Å, d(O4–HO4) = 1.04 Å, and d(O3B-HO4) = 1.66 Å, that is, the breaking bond is shorter and the forming bond is longer than in the case of the wild-type enzyme. The Q189A mutant thus has an earlier TS with less dissociative character. One may speculate that this could enhance entropic effects (a tighter TS would make $\Delta S^{\ddagger}$ more negative and $\Delta G^{\ddagger}$ more positive, because of the larger $-T\Delta S^{\ddagger}$ term). We have not quantified such effects, however, and therefore refrain from speculating whether this may explain the wrong trend in the computed barriers (see above).

In summary, the Q189A mutant still follows the S_Ni mechanism although with a somewhat less dissociative TS. Whether the enzyme has a single dissociative oxocarbenium ion-like TS (S_Ni) or a short-lived ion-pair intermediate (S_Ni-like) obviously depends on the actual shape of the potential (or free) energy surface along the reaction coordinate (from RC ~ −1.5 to 1.0 Å, see Figure 8), or in other words, on the extent and timing of the making and breaking of bonds.[15] The Q189A mutation with the replacement of a weak nucleophile already significantly affects the energy profile and the location of the TS (see Figure 8 and SI, Figure S7). It is conceivable that other mutations lead to larger changes, up to a point where another local minimum for an ion-pair intermediate appears. Therefore, the tuning by the environment may determine the exact mechanism followed by the enzyme (S_Ni vs S_Ni-like). We note in this connection again that the recent QM/MM work on OtsA suggests an S_Ni-like mechanism with a very shallow ion-pair intermediate.[18]

Q189E Mutant. If Gln189 assists the reaction by stabilizing the charge development at the anomeric center, the Q189E mutation would be expected to provide a much better stabilization and, thus, favor the S_Ni mechanism even more. This is exactly what we find if we transform in silico Gln189 to a pseudo-glutamate (Q189E*) by simply including this residue in the MM part of the system and giving it the charges of a glutamate, without changing the geometries obtained for the WT enzyme (SI, Figure S8). However, when we actually build the Q189E mutant in silico and perform reoptimizations (in analogy to the procedure outlined above for the Q189A mutant), another scenario emerges.

Energy minimization of the Q189E mutant takes the system straight to the CGE complex, indicating that in this case CGE formation is barrierless. At the QM(BP86/SVP)/CHARMM level, the C1′−O(Glu189) distance in the formed CGE is 1.49 Å. The interaction with the Ala154 amide is maintained, but not the hydrogen bond with the O6 atom of LAT (which is oriented to make a hydrogen bond with Asp130, Figure 9A). The S_N2 attack of LAT at the CGE, in what would be the second step of the double displacement mechanism, was studied by scanning the d(O4−C1′) reaction coordinate at different levels of theory (SI, Figure S9). In all DFT-based energy profiles, the computed barrier is higher than 30 kcal/mol. TS optimization has only been successful for SCC-DFTB/CHARMM, yielding a barrier height of 28.8 kcal/mol that increased to 33.9 and 39.5 kcal/mol in single-point energy calculations with QM = B3LYP/TZVP and M05-2X/TZVP, respectively. According to these QM/MM results, the overall reaction for the Q189E mutant would thus be very slow, and moreover, the reaction energy is also computed to be very high (26−30 kcal/mol). Key distances in the SCC-DFTB/CHARMM TS structure are d(C1′−OE1) = 2.31 Å, d(O4−C1′) = 2.19 Å, d(O4−HO4) = 1.18 Å, and d(O3B−HO4) = 1.23 Å (Figure 9B).

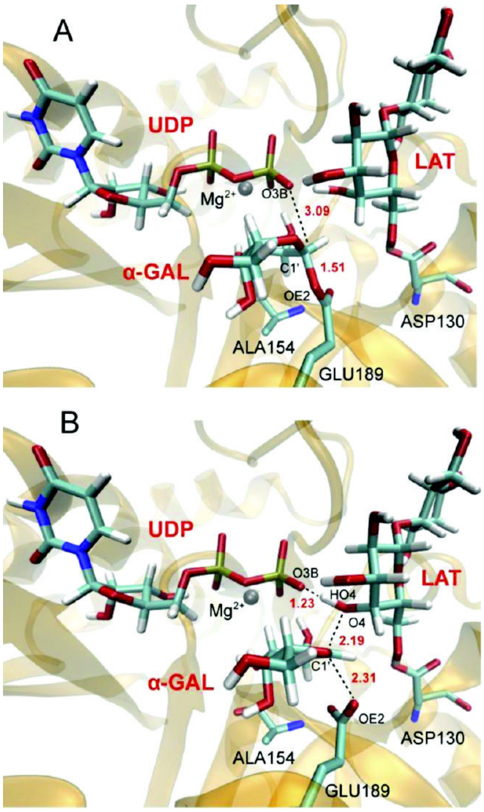


Figure 9. Optimized (A) covalent glycosyl−enzyme complex and (B) transition state for the second step of a double-displacement mechanism in the LgtC Q189E mutant optimized at the SCC-DFTB/CHARMM level. Selected distances (in Å) are indicated in red.

The case of the Q189E mutant would be an example of a change in mechanism introduced by a mutation. Recently, Goedl and Nidetzky[62] have remodeled sucrose phosphorylase to change its kinetics and chemical mechanism from a double-displacement to a direct front-side nucleophilic displacement reaction. In our calculations, we observe CGE formation in the Q189E mutant, but galactosyl transfer via a double displacement mechanism involving Glu189 has a high barrier and is thus too slow to be feasible, at least in the presently studied conformation of the enzyme. We note that O4 access to C1′ is hampered by H1′ and also limited by the need to keep the O3B−HO4 interaction. This enforces a TS structure with a dissociated C1′−OE2 bond, but in the resulting species, Glu189 does not seem to be stabilized enough by the environment. In fact, kinetic experiments on the Q189E mutant of LgtC found a reduction in k_{cat} to 3% of that of the wild-type enzyme and a reduction in its (low) hydrolytic activity by a factor of 10.[11] Formation of a CGE intermediate was indeed detected but involved Asp190, whose side chain carboxylate is located as far as 8.9 Å away from the anomeric carbon C1′ in the Q189E:UDP-2′FGal crystal structure. These experimental results suggest the possibility of a double-displacement mechanism in the Q189E mutant, in which the remote residue Asp190 acts as the catalytic nucleophile. This would obviously require significant conformational changes (relative to the wild-type crystal structure) to correctly position Asp190, which might occur as a consequence of the mutation or upon acceptor binding. A detailed understanding of the reaction in the Q189E mutant would clearly require further experimental and computational studies that are beyond the scope of the present work.

Further Comparisons. There are other retaining GTs which do contain a bona fide nucleophile in the active site. This is the case for the family 6 GT bovine α-1,3-galactosyltransferase (α3GalT),[63] which has a glutamate (Glu317) at the position equivalent to Gln189. It has been proposed that the presence of Glu317 is required for proper acceptor-substrate orientation. α3GalT utilizes the same substrates as LgtC, namely, UDP-Gal and LAT, but catalyzes the formation of an α-1,3 glycosidic linkage. An overlay of the structures of these two enzymes (SI, Figure S10) shows that the substrates indeed adopt different relative orientations and that Glu317 and Gln189 participate in correctly orienting the acceptor. Consequently, some of the interactions identified in this work that favor the S_Ni mechanism in LgtC are not present in α3GalT, for example, the O3(LAT)−O3B(UDP) hydrogen bond. Apparently, as expected, there are connections between the identity of the substrates, the specificity of the glycosidic bond to be formed, the interactions and relative orientation of the bound substrates, and the active-site residues. And even more interestingly, all these factors could conspire to determine the mechanistic strategy followed by the enzyme. Experimental data on α3GalT and two other ret-GTs of the family 6 GTs suggest that these retaining enzymes could actually operate by a double-displacement mechanism. The second step of the double displacement mechanism for retaining α-galactosyltransferases with a glutamate nucleophile has been investigated using a cluster model and QM(DFT) calculations.[64] This study concludes that such a mechanism is plausible and that the hydroxyl group of the acceptor would attack the CGE in the second step, with the UDP acting as the general base that deprotonates the hydroxyl group. This role of UDP thus seems to be a common feature in different ret-GT mechanisms.

CONCLUSIONS

Getting a clear picture of the reaction mechanism used by retaining glycosyltransferases is very difficult experimentally and remains one of the fundamental challenges in glycosciences. Here we have used QM(DFT)/MM calculations on the full enzyme to study the reaction catalyzed by LgtC as well as the corresponding reaction with alternative substrates and with LgtC mutant enzymes. This provides us with a detailed description of the reaction catalyzed by this enzyme, which we expect will be valuable to the experimental groups working in the field. The different mechanisms proposed in the literature (S_Ni, S_Ni-like, and double displacement mechanism via the formation of a CGE intermediate) have been investigated and compared. We find a dissociative S_Ni mechanism for the wild-type enzyme with the most reliable QM/CHARMM barriers ranging between 11.8 kcal/mol (B3LYP/TZVP) and 14.6 kcal/mol (M05-2X/TZVP), in reasonable agreement with the experimental kinetic data. We have identified several factors that help the front-side mechanism, in particular enzyme−substrate and substrate−substrate interactions. Among them, the largest effects come from Lys250, which is also involved in binding. Gln189, the putative nucleophile in a double displacement mechanism, is found to favor the charge development at the anomeric center during the reaction by about 2 kcal/mol. We predict that 3-deoxylactose as acceptor will increase the barrier height by 2−3 kcal/mol (reduction of k_{cat} to 0.6−3% of that for the unmodified substrates). The reactions of the Q189A and Q189E mutants have also been investigated. Comparison of LgtC wild type and mutant systems, as well as other ret-GTs, suggest that the mechanistic strategy followed by each enzyme−substrate complex may be influenced by several factors, including the nature of the substrates, the specificity of the glycosidic linkage to be formed (and thus the interactions and relative orientation of the bound substrates), or even the GT fold. Thus, it should not be expected that all retaining GTs share the same mechanism. In OtsA, no putative nucleophile is present and an S_Ni-like mechanism was proposed; a poor nucleophile is present in wild-type LgtC, and the enzyme is found to follow a dissociative S_Ni mechanism; and when Gln189 is substituted by Ala (Q189A), an S_Ni mechanism is still predicted but with a less pronounced maximum and a wider and flatter barrier top. For the LgtC Q189E mutant, an even more drastic change in mechanism is computed, from a front-side attack to the formation of a CGE with Glu189 that cannot evolve to the products, at least not with the present enzyme conformation. Finally, ret-GTs of family 6 GTs, with a well-defined nucleophile present in the active site, seem to follow a double-displacement mechanism. Therefore, a full understanding of the mechanism used by retaining glycosyltransferases would seem to require case-by-case studies. In an attempt to shed more light on these mechanisms, other ret-GT systems are now being investigated in our laboratory.

ASSOCIATED CONTENT

Supporting Information

Computational details, QM/MM results (geometries and energies) for all methods used in each of the reactions studied, including the tests with Mn^{2+}; complete refs 23, 27, 29, and 31 (PDF); and optimized geometries of the 18 stationary points determined in this study plus the optimized reactants complex with Mn^{2+} (pdb). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

Laura.Masgrau@uab.cat

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge financial support from the Spanish "Ministerio de Ciencia e Innovación" through project CTQ2008-02403/BQU, the "Ramon y Cajal" program (L.M.), and from the"Generalitat de Catalunya", project 2009SGR409.

REFERENCES

(1) Taniguchi, N.; Honke, K.; Fukuda, M. *Handbook of Glycosyltransferases and Related Genes*; Springer: Tokyo, 2002.

(2) Ünligil, U.; Rini, J. M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 510−517.

(3) Breton, C.; Snajdrová, L.; Jeanneau, C.; Koca, J.; Imberty, A. *Glycobiology* **2006**, *16*, 29R−37R.

(4) Lairson, L. L.; Henrissat, B.; Davies, G. J.; Withers, S. G. *Annu. Rev. Biochem.* **2008**, *77*, 521−555.

(5) Tzeng, Y. L.; Stephens, D. S. *Microbes Intect.* **2000**, *2*, 687−700.

(6) Campbell, J.; Davies, G.; Bulone, V.; Henrissat, B. *Biochem. J.* **1998**, *329* (Pt. 3), 719.

(7) Takayama, S.; Chung, S.G-; Igarashi, Y.; Ichikawa, Y.; Sepp, A.; Lechler, R. I.; Wu, J.; Hayashi, T; Siuzdak, G.; Wong, C.-H. *Bioorg. Med. Chem.* **1999**, *7*, 401−409.

(8) Davies, G. J. *Nat. Struct. Biol.* **2001**, *8*, 98−100.

(9) Monegal, A.; Planas, A. *J. Am. Chem. Soc.* **2006**, *128*, 16030−16031.

(10) Soya, N.; Fang, Y.; Palcic, M. M.; Klassen, J. S. *Glycobiology* **2011**, *21*, 547−552.

(11) Lairson, L. L.; Chiu, C. P. C; Ly, H. D.; He, S.; Wakarchuk, W. W.; Strynadka, N. C. J.; Withers, S. G. *J. Biol. Chem.* **2004**, *279*, 28339−28344.

(12) Persson, K.; Ly, H. D.; Dieckelmann, M.; Wakarchuk, W. W.; Withers, S. G.; Strynadka, N. C. *Nat. Struct. Biol.* **2001**, *8*, 166–175.

(13) Sinnott, M. L.; Jencks, W. P. *J. Am. Chem. Soc.* **1980**, *102*, 2026–2032.

(14) Errey, J. C.; Lee, S. S.; Gibson, R. P.; Martinez Fleites, C.; Barry, C. S.; Jung, P. M. J.; O'Sullivan, A. C.; Davis, B. G.; Davies, G. J. *Angew. Chem., Int. Ed.* **2010**, *49*, 1234–1237.

(15) Lee, S. S.; Hong, S. Y.; Errey, J. C.; Izume, A.; Davies, G. J.; Davis, B. G. *Nat. Chem. Biol.* **2011**, *7*, 631–638.

(16) Tvaroška, I. *Carbohydr. Res.* **2004**, *339*, 1007–1014.

(17) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. *Protein Eng.* **1995**, *8*, 127–134.

(18) Ardèvol, A.; Rovira, C. *Angew. Chem., Int. Ed.* **2011**, *50*, 10897–10901.

(19) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.

(20) Fraústo da Silva, J. J. R.; Williams, R. J. P. *The Biological Chemistry of Elements. The Inorganic Chemistry of Life*; Clarendon Press: Oxford, England, 1991.

(21) Bock, C. W.; Katz, A. K.; Markham, G. D.; Glusker, J. P. *J. Am. Chem. Soc.* **1999**, *121*, 7360–7372.

(22) Kóňa, J.; Tvaroška, I. *Chem. Pap.* **2009**, *63*, 598–607.

(23) MacKerell, A. D. Jr.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(24) MacKerell, A. D., Jr.; Brooks, C., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; John Wiley & Sons: Chichester, 1998; pp 271–277.

(25) MacKerell, A. D. Jr.; Feig, M.; Brooks, C. L. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(26) Brooks, B. R.; Bruccoleri, R.; Olafson, D.; States, D.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(27) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(28) Guvench, O.; Hatcher, E. R.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D. Jr. *J. Chem. Theory Comput.* **2009**, *5*, 2353–2370.

(29) Sherwood, P.; et al. *J. Mol. Struct. (THEOCHEM)* **2003**, *632*, 1–28.

(30) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

(31) Frisch, M. J.; et al. *Gaussian03*, revision D.01; Gaussian, Inc.: Wallingford, CT, 2004.

(32) Thiel, W. *Program MNDO2005*, version 7.0; Max-Planck-Institut für Kohlenforschung: Mülheim, 2005.

(33) Slater, J. C. *Phys. Rev.* **1951**, *81*, 385–390.

(34) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(35) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(36) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(37) Perdew, J. P. *Phys. Rev. B* **1986**, *34*, 7406.

(38) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(39) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(40) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(41) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.

(42) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(43) Frauenheim, T.; Seifert, G.; Elstner, M.; Niehaus, T.; Köhler, C.; Amkreutz, M.; Sternberg, M.; Hajnal, Z.; Carlo, A. D.; Suhai, S. *J. Phys.: Condens. Matter* **2002**, *14*, 3015–3047.

(44) Smith, W. *J. Mol. Graphics* **1996**, *14*, 136–141.

(45) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580–10594.

(46) de Vries, A. H.; Sherwood, P.; Collins, S. J.; Rigby, A. M.; Rigutto, M.; Kramer, G. J. *J. Phys. Chem. B* **1999**, *103*, 6133–6141.

(47) Sherwood, P.; de Vires, A.; Collins, S.; Greatbanks, S.; Burton, N.; Vincent, M.; Hillier, I. *Faraday Discuss.* **1997**, *106*, 79–92.

(48) Nocedal, J. *Math. Comp.* **1980**, *35*, 773–782.

(49) Liu, D.; Nocedal, J. *Math. Programming* **1989**, *45*, 503–528.

(50) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52–57.

(51) Baker, J. *J. Comput. Chem.* **1986**, *7*, 385–395.

(52) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177–2186.

(53) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

(54) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.

(55) Eichkorn, K.; Treutler, O.; Ohm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289.

(56) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

(57) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(58) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.

(59) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735–746.

(60) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

(61) Ly, H. D.; Lougheed, B.; Wakarchuk, W. W.; Withers, S. G. *Biochemistry* **2002**, *41*, 5075–5085.

(62) Goedl, C.; Nidetzky, B. *ChemBioChem* **2009**, *10*, 2333–2337.

(63) Boix, E.; Swaminathan, G. J.; Zhang, Y.; Natesh, R.; Brew, K.; Acharya, K. R. *J. Biol. Chem.* **2001**, *276*, 48608–48614.

(64) André, I.; Tvaroška, I.; Carver, J. P. *Carbohydr. Res.* **2003**, *338*, 865–877.

Supporting Information for

Retaining glycosyltransferase mechanism studied by QM/MM methods: Lipopolysaccharyl-α-1,4-galactosyltransferase C transfers α-galactose via an oxocarbenium ion-like transition state.

Hansel Gómez, [§,‡,†] Iakov Polyak, [§] Walter Thiel, [§] José M. Lluch[‡,†] and Laura Masgrau[‡,*]

[§]Max-Planck-Institut für Kohlenforschung, D-45470 Mülheim an der Ruhr, Germany.
[‡]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.
[†]Departament de Química, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.

Table of contents:

1. Computational details

1.1 System preparation

PROTONATION STATES OF THE TITRATABLE RESIDUES:

All Lys and Arg side chains are in the ammonium or guanidinium forms, respectively, and all Asp and Glu side chains are in the carboxylate form. The chosen protonation states for all histidines are listed below:

HSD: 26 34 167 225 244 253

HSE: 78

ACTIVE RESIDUES INCLUDED IN QM/MM GEOMETRY OPTIMIZATIONS:

TYR11, PRO73, LEU74, ASN75, ILE76, ARG77, HSE78, ILE79, SER80, THR82, THR83, TYR84, ARG86, LEU87, TYR101, LEU102, ASP103, ILE104, ASP105, VAL106, LEU125, ALA127, SER128, ILE129, ASP130, LEU131, PHE132, VAL133, GLU134, GLN136, TYR139, ILE143, TYR151, PHE152, ASN153, ALA154, GLY155, VAL156, LEU157, MET184, GLN185, TYR 186, GLN187, ASP188, GLN189, ASP190, ILE191, LEU192, ASN193, GLY194, ASN208, MET210, PRO211, THR212, ASN213, HSD244, TYR245, CYS246, GLY247, PRO248, ALA249, LYS250, TRP252.

1.2 Details of the electronic structure methods used for the QM/M calculations

Geometry optimizations were generally carried out at the QM(BP86/SVP)/CHARMM level using the BP86 functional and the SVP basis set in combination with the resolution-of-the-identity (RI) approximation. Alternative QM treatments (B3LYP/SVP and SCC-DFTB) were tested for geometry optimization in reaction path calculations, and it was found that single-point QM/MM energy profiles are essentially the same when evaluated at QM/CHARMM geometries optimized with different QM treatments. For example, single-point QM(B3LYP/SVP)/CHARMM profiles are essentially identical when calculated at QM/CHARMM geometries optimized with the use of BP86/SVP or B3LYP/SVP (Figure S1). Therefore, we adopted BP86/SVP as standard QM treatment for geometry optimizations and transition state (TS) searches, even though we also used B3LYP/SVP for TS optimization in the wild-type enzyme. For more accurate energy evaluations, we performed single-point energy calculations with other functionals and larger basis sets. The M05-2X functional describes carbohydrate chemistry well, including the dissociation of the glycosidic bond in sugar phosphates (ref 23 of main text and Csonka, G.I.; French, A. D.; Johnson, G.P.; Stortz, C.A. *J. Chem. Theory Comp.* **2009**, *5*, 679-692), and it has therefore been applied along with the standard B3LYP functional. The energy barriers calculated with B3LYP or M05-2X are commonly somewhat higher than those obtained with BP86 because of the admixture of Hartree-Fock exchange in these two hybrid functionals. In summary, the SVP and TZVP basis set were used for single-point energy calculations with the functionals BP86, B3LYP and M05-2X; the def2-TZVPP(d) basis set was also considered in the case of BP86 and B3LYP. Finally, empirical dispersion corrections for DFT (DFT-D) were applied in combination with the B3LYP/TZVP treatment.

1.3 Umbrella sampling molecular dynamics calculations

Umbrella sampling at the SCC-DFTB/CHARMM22 level was performed to compute the potential of mean force (PMF) and the free energy profile for the front-side attack mechanism using the dynamics module within ChemShell. The reaction coordinate defined as *d(O3B-C1')– d(O4-C1')* was scanned in steps of 0.1 Å, with a force constant of 237 kcal mol^{-1} Å^{-2}. The same active region as in QM/MM optimizations (see above) was subjected to NVT dynamics with the Nosé-Hoover thermostat (Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511-519 and Hoover, W. *Phys. Rev. A* **1985**, *31*, 1695-1697). The SHAKE procedure (Ryckaert, J.; Ciccotti, G.; Berendsen, H. *J. Comp. Phys.* **1977**, *23,* 327-341) was applied at every step for the O-H bonds in the water molecules. A 20 ps molecular dynamics (MD) run for equilibration was followed by 5 ps of data collection for every sampling window. Both the umbrella integration analysis (Kästner, J.; Thiel, W. *J. Chem. Phys.* **2005**, *123*, 144104 and Kästner, J.; Thiel, W. *J. Chem. Phys.* **2006**, *123*, 234106) and the weighted histogram analysis method (WHAM) (Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011-1021 and Souaille, M. *Comput. Phys. Commun.* **2001**, *135*, 40-57) were used to compute the free energy profile. Results are plotted in Figure S3.

2. QM/MM results

Comparison of the results obtained by the different methods used (Table 1, Table S2 and Figure S1) shows that the energy barriers obtained range between 8.1 and 16.6 kcal/mol (BP86/TZVP and M05-2X/SVP single-point energies, respectively). The computed reaction energies vary less (from 2.8 kcal/mol at the B3LYP-D/TZVP level to 6.1 kcal/mol at the B3LYP/TZVP and B3LYP/def2-TZVPP(d) levels). Similar geometries for R, TS and P are obtained at the BP86/SVP and B3LYP/SVP levels, with some small differences for the TS: at the B3LYP/SVP level the breaking C1'-O3B bond is slightly longer, and the forming O4-C1' bond is slightly shorter. Strong deviations are found when the QM region is treated at the SCC-DFTB level, which gives a very high energy barrier of 32.7 kcal/mol and a reaction energy of 9.7 kcal/mol. The transition state at this level of theory is still of dissociative character but SCC-DFTB/CHARMM yields a shorter *d(C1'-O3B)* distance (2.63 Å) and a longer *d(O4-C1')* distance (2.44 Å) than the other two methods used for optimization in this work (Table S1). These discrepancies are most likely related to the difficulty of the SCC-DFTB method to correctly describe chemical processes involving phosphorous. When comparing the different DFT methods for a given basis set, the barrier heights increase in the order BP86 < B3LYP < M05-2X as has been observed before because of the Hartree-Fock exchange that is present in B3LYP and M05-2X but not in BP86. The M05-2X relative energies are higher at the TS and also along the reaction pathway (Figure S1), in agreement with the tendency observed between different DFT methods when describing the breakage of the glycosidic bond in sugar phosphates (see ref. 23 of the main paper). For a given method, the extension from the SVP to the TZVP basis set reduces the energy barrier, whereas the inclusion of more polarization and of diffuse functions in the def2-TVZPP(d) basis set increases the barrier height again.

Table S1. Bond distances (in Å) in the optimized reactants, transition state, and products for the proposed S_Ni mechanism at the SCC-DFTB level of theory.

	React	TS	Prod
d(C1'-O3B)	1.51	2.63	3.23
d(O4-C1')	3.13	2.44	1.51
d(O4-HO4)	0.99	1.15	1.69
d(HO4-O3B)	1.78	1.27	0.99
d(OE1-C1')	3.38	2.78	3.08
d(C1'-O5)	1.44	1.28	1.44

Table S2. QM/MM potential energy barriers and reaction energies for the wild type enzyme at several levels of theory calculated at stationary points optimized at the QM(BP86/SVP)/CHARMM level. Energies are given in kcal/mol.

	BP86			B3LYP			B3LYP-D	M05-2X	
	SVP	TZVP	def2-TZVPP(d)	SVP	TZVP	def2-TZVPP(d)	TZVP	SVP	TZVP
R*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TS	9.18	8.13	8.81	12.87	11.82	12.52	9.50	16.61	14.61
P	3.28	4.63	4.59	4.60	6.05	6.13	2.79	3.62	4.63

*absolute energies (a.u): **BP86**: -3784.16478, -3787.66218, -3787.94300; **B3LYP**: -3782.26151, -3785.77506, -3786.06026; **B3LYP-D**: -3785.95579; **M05-2X**: -3783.48006, -3787.10053.

Table S3. QM/MM potential energy barriers and reaction energies for the wild type enzyme at several levels of theory calculated at stationary points optimized at the QM(B3LYP/SVP)/CHARMM level. Energies are given in kcal/mol.

	B3LYP			B3LYP-D	M05-2X	
	SVP	TZVP	def2-TZVPP(d)	TZVP	SVP	TZVP
R*	0.00	0.00	0.00	0.00	0.00	0.00
TS	11.14	10.09	10.96	7.23	14.46	12.44
P	4.46	6.02	6.14	2.49	3.56	4.63

*absolute energies (a.u): **B3LYP**: -3782.26833, -3785.78459 ,-3786.07594; **B3LYP-D:** -3785.96469; **M05-2X**: -3783.49413, -3787.11671.

Table S4. QM/MM potential energy barriers and reaction energies for the wild type enzyme with 2'-deoxygalactose calculated at stationary points optimized at the QM(BP86/SVP)/CHARMM level. Energies are given in kcal/mol.

	BP86			B3LYP			B3LYP-D	M05-2X	
	SVP	TZVP	def2-TZVPP(d)	SVP	TZVP	def2-TZVPP(d)	TZVP	SVP	TZVP
R*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TS	12.71	10.77	11.19	17.02	15.06	15.49	12.41	19.69	16.91
P	7.83	8.46	7.98	9.40	10.21	9.83	7.19	8.44	9.06

*absolute energies (a.u): **BP86**: -3708.98620, -3712.4003, -3712.67705; **B3LYP**: -3707.12263, -3710.55270, -3710.83386; **B3LYP-D**: -3710.72719; **M05-2X**: -3708.31648, -3711.85067.

Table S5. QM/MM potential energy barriers and reaction energies for the WT enzyme with 3-deoxylactose calculated at stationary points optimized at the QM(BP86/SVP)/CHARMM level. Energies are given in kcal/mol.

	BP86			**B3LYP**			**B3LYP-D**	**M05-2X**	
	SVP	TZVP	def2-TZVPP(d)	SVP	TZVP	def2-TZVPP(d)	TZVP	SVP	TZVP
R*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TS	13.77	11.20	12.13	17.51	14.97	15.82	12.84	21.15	17.71
P	2.46	2.60	2.51	3.45	3.81	3.78	0.28	2.58	2.46

*absolute energies (a.u): **BP86**: -3708.99351, -3712.40560, -3712.68209; **B3LYP**: -3707.13001, -3710.55810, -3710.83887; **B3LYP-D**: -3710.73282; **M05-2X**: -3708.32418, -3711.85662.

Table S6. QM/MM potential energy barriers and reaction energies for the Q189A mutant enzyme calculated at stationary points optimized at the QM(BP86/SVP)/CHARMM level. Energies are given in kcal/mol.

	BP86			**B3LYP**			**B3LYP-D**	**M05-2X**	
	SVP	TZVP	def2-TZVPP(d)	SVP	TZVP	def2-TZVPP(d)	TZVP	SVP	TZVP
R*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TS	9.26	7.40	8.06	12.05	10.05	10.68	9.58	17.22	14.26
P	-0.96	0.12	0.11	-0.42	0.77	0.85	-0.94	0.02	0.68

*absolute energies (a.u): **BP86**: -3576.24241, -3579.51524, -3579.78493; **B3LYP**: -3574.46567, -3577.75357, -3578.02721; **B3LYP-D**: -3577.91773; **M05-2X**: -3575.60240, -3578.9868.

Table S7. QM/MM potential energy barriers and reaction energies for the second step of a double displacement mechanism in the Q189E mutant enzyme. The stationary points were optimized at the SCC-DFTB/CHARMM level. Energies are given in kcal/mol.

	SCC-DFTB	**B3LYP**			**B3LYP-D**	**M05-2X**	
		SVP	TZVP	def2-TZVPP(d)	TZVP	SVP	TZVP
R*	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TS	28.57	36.16	33.86	34.25	32.13	41.74	39.47
P	14.87	27.71	26.38	25.12	24.48	30.65	30.18

*absolute energies (a.u): **B3LYP**: -3801.23378, -3804.79205, -3805.05941; **B3LYP-D**: -3804.97238; **M05-2X**: -3802.44688; -3806.11344.

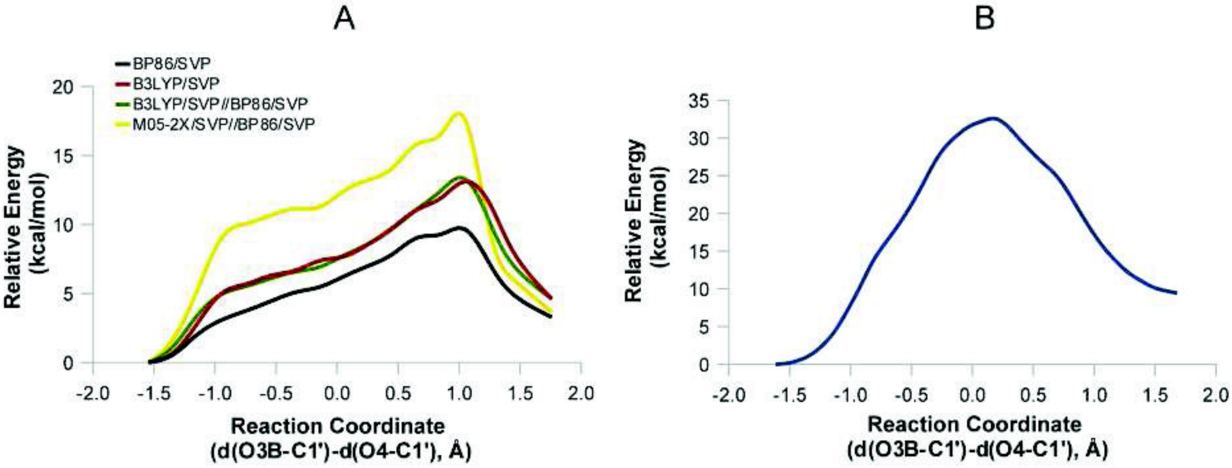


Figure S1. QM/CHARMM potential energy profiles for the front-side attack (S_Ni) mechanism at different QM levels of theory: (A) BP86/SVP, B3LYP/SVP, B3LYP/SVP//BP86/SVP, M05-2X/SVP//BP86/SVP; and (B) SCC-DFTB.

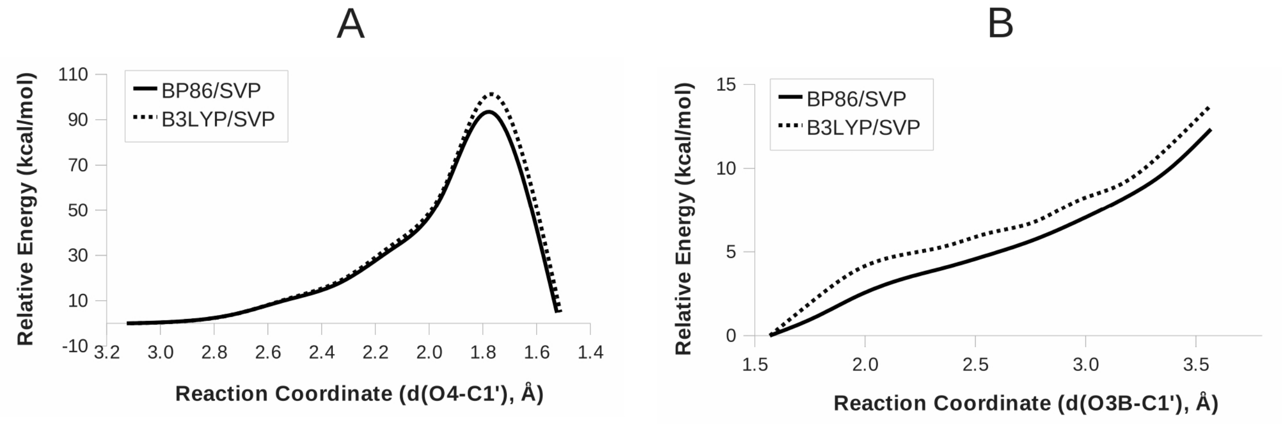


Figure S2. QM(BP86/SVP)/CHARMM energy profile for the front-side attack using as reaction coordinate (A) the *d(O4-C1')* distance and (B) the *d(O3B-C1)* distance. The use of the *d(O4-C1')* distance results in a energy profile with a very high barrier and a late exit of the leaving group (Figure S2A). On another side, by only forcing C1'-O3B bond cleavage, the reaction does not occur: the energy increases up to values similar to the ones obtained with the combined reaction coordinate (Figure S1), but there is no stabilization after breaking the C1'-O3B since proton transfer to O3B does not occur and, as a consequence, C1'-O4 bond formation is not achieved (Figure S2B). The QM(B3LYP/SVP)/CHARMM single-point energies along these paths are also given.

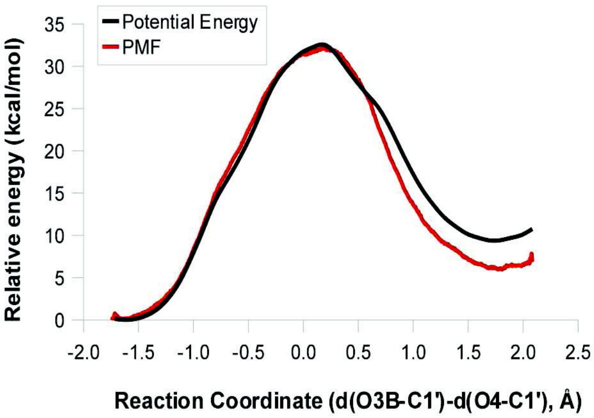


Figure S3. SCC-DFB/CHARMM potential energy profile and potential of mean force (PMF) for the front-side attack mechanism.

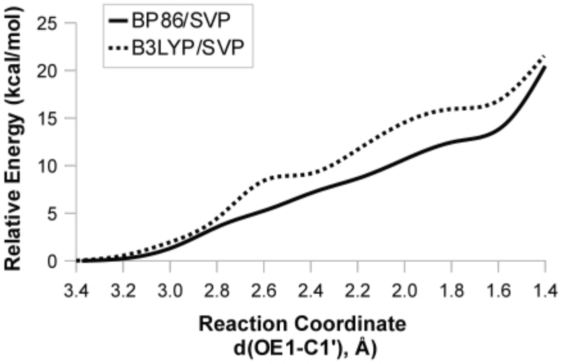


Figure S4. QM(BP86/SVP, B3LYP/SVP)/CHARMM energy profiles for the first step of the double displacement mechanism in the wild type enzyme.

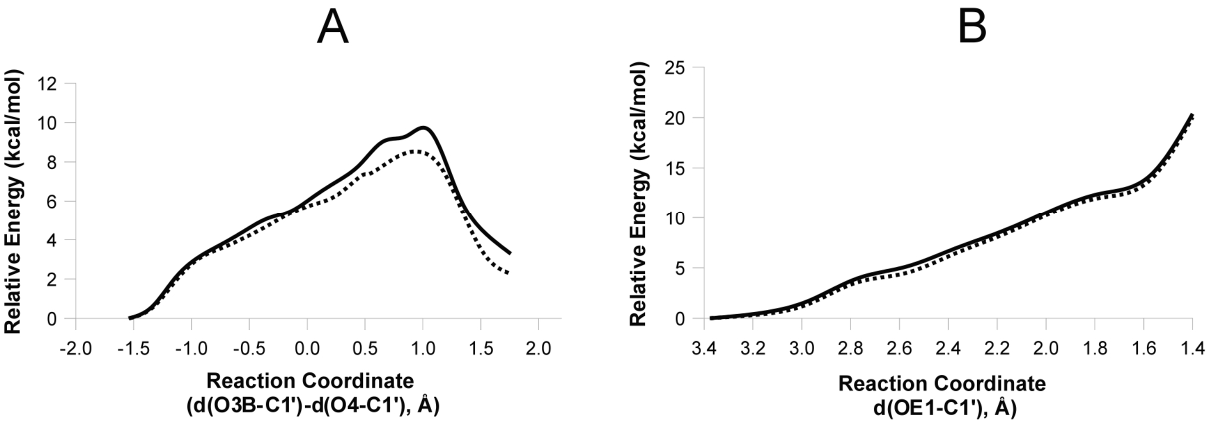


Figure S5. QM(BP86/SVP)/CHARMM potential energy profiles for (A) the front-side attack (S_Ni) mechanism and (B) the first step of the double displacement mechanism with Mg^{2+} and Mn^{2+} as divalent cation (solid and dotted lines, respectively).

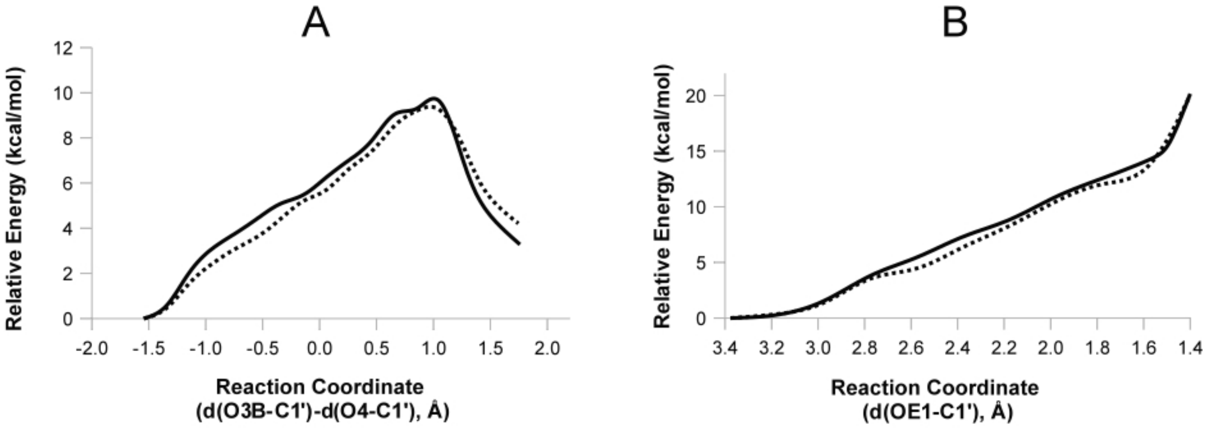


Figure S6. QM(BP86/SVP)/CHARMM potential energy profiles for (A) the front-side attack (S_Ni) mechanism and (B) the first step of the double displacement mechanism in the wild type enzyme, including or not including the residues ASP188 and LYS250 in the QM region (dashed and solid lines, respectively).

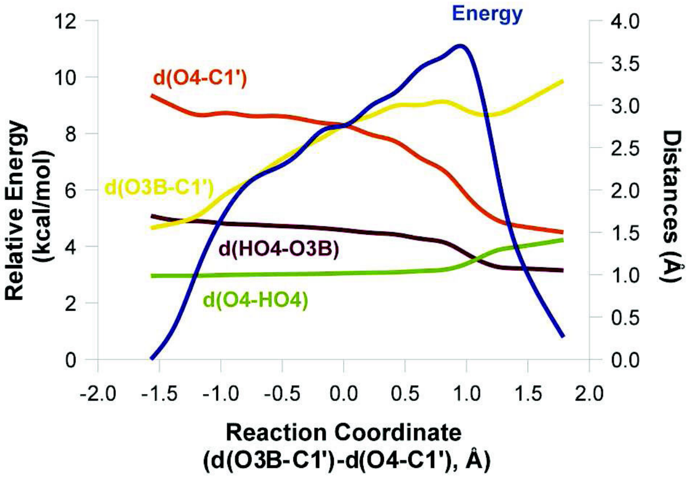


Figure S7. QM(B3LYP/TZVP)/CHARMM energy profile for the proposed S_Ni mechanism calculated at QM(BP86/SVP)/CHARMM optimized geometries, for the Q189A mutant with unmodified substrates. The variation of selected interatomic distances involved in the reactive process is also shown.

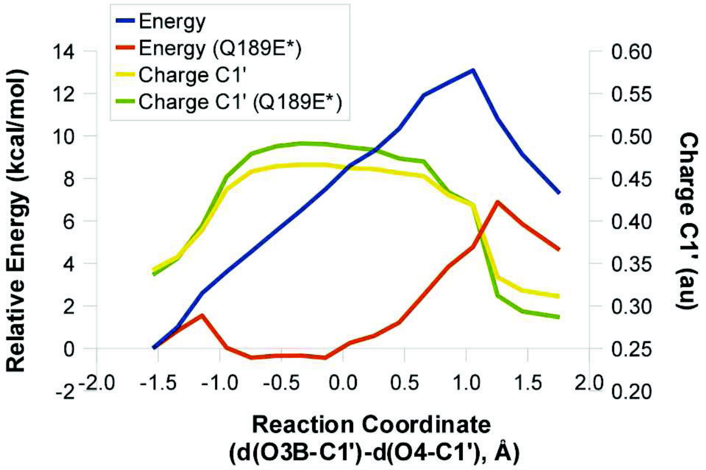


Figure S8. QM(BP86/SVP)/CHARMM potential energy profile for the proposed S_Ni mechanism in the wild type and the *pseudo* Q189E* mutant enzymes. In both cases, residue 189 is described by the MM force field. The *pseudo* Q189E* mutant is built by simply assigning glutamate charges to Gln189. Its energy profile is calculated at the optimized geometries of the wild type enzyme. QM(BP86/TZVP)/CHARMM charge evolution at C1' calculated at QM(BP86/SVP)/CHARMM optimized geometries is also depicted.

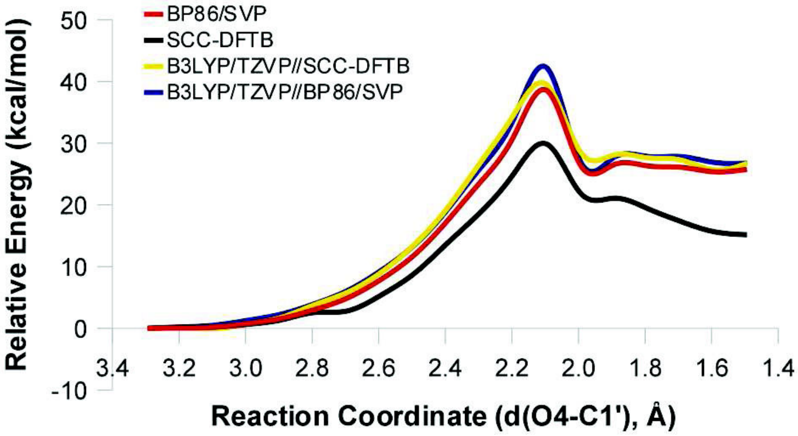


Figure S9. QM/CHARMM potential energy profiles computed at different QM levels of theory, for the second step of the double displacement mechanism in the Q189E mutant. Energies are given in kcal/mol, distances in Å.

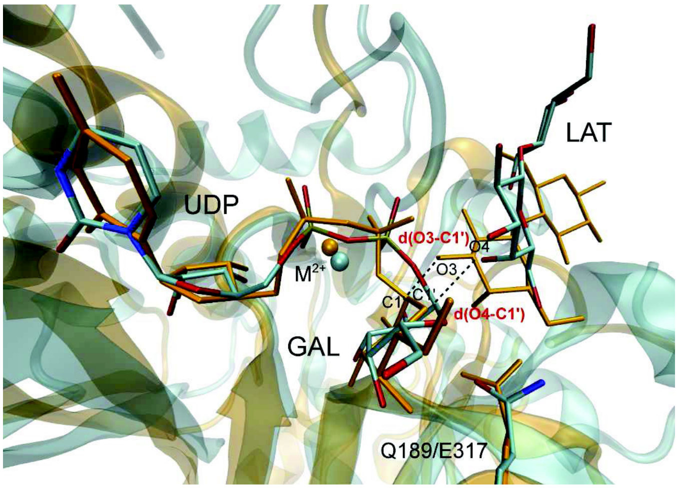


Figure S10. Comparison of LgtC (cyan; PDB code: 1GA8) and α3GalT (orange; PDB codes: 2VS5, protein and UDP-GAL; 1O7O, LAT) active sites with substrates or substrate analogues bound (using the crystallographic data).

3. Complete references 24, 28, 30 and 32 of the main text.

[24] MacKerell, Jr., A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr., 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, III, W.E.; Roux, B.; Schlenkrich, M.; Smith, J.C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, 102, 3586-3616.

[28] Brooks, B. R.; Brooks, C. L. III.; MacKerell, A. D. Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545-1614.

[30] (a) 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; Sjovoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *J. Mol. Struct. (Theochem)* **2003**, *632*, 1-28.
(b) ChemShell is a modular QM/MM program developed in the European QUASI project under the coordination of P. Sherwood (see http://www.chemshell.org).

[32] Frisch, M. T., G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Montgomery, J.; Vreven, T.; Kudin, K.; Burant, J.; Millam, J.; Iyengar, S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J.; Hratchian, H.; Cross, J.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.; Yazyev, O.; Austin, A.; Cammi, R.; Pomelli, C.; Ochterski, J.; Ayala, P.; Morokuma, K.; Voth, G.; Salvador, P.; Dannenberg, J.; Zakrzewski, V.; Dapprich, S.; Daniels, A.; Strain, M.; Farkas, O.; Malick, D.; Rabuck, A.; Raghavachari, K.; Foresman, J.; Ortiz, J.; Cui, Q.;Baboul, A.; Clifford, S.; Cioslowski, J.; Stefanov, B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R.; Fox, D.; Keith, T.; Al-Laham, M.; Peng, C.; Nanayakkara, A.; Challacombe, M.; Gill, P.; Johnson, B.; Chen, W.; Wong, M.; Gonzalez, C.; Pople, J. *Gaussian 03*, revision D.01; 2004.

A Microiterative Intrinsic Reaction Coordinate method for large QM/MM systems

I. Polyak, E. Boulanger, K. Sen, W. Thiel

Phys. Chem. Chem. Phys. (2013),

DOI: 10.1039/c3cp51669e

PCCP

RSC Publishing

PAPER

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

Cite this: DOI: 10.1039/c3cp51669e

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

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.

Received 18th April 2013,
Accepted 30th May 2013

DOI: 10.1039/c3cp51669e

www.rsc.org/pccp

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:

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

where x denotes the mass-weighted Cartesian coordinates of the nuclei, s is the arc length along the IRC, and g is the mass-weighted gradient at 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

---

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

however become increasingly accurate with the growth of the 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,[25] 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:

$$x_{k+1} = x_k - \Delta s \frac{g(x_k)}{|g(x_k)|}. \tag{2}$$

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]

$$x_{k+1} = x_k + A(t)g(x_k), \tag{3}$$

with

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

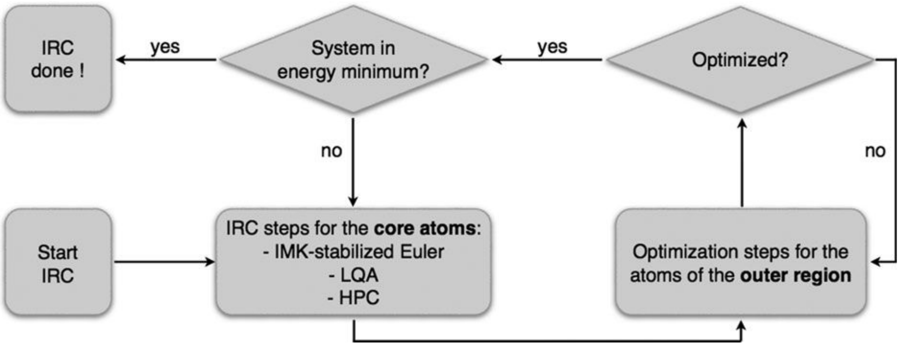


Fig. 1 Scheme of the microiterative IRC procedure as implemented in the HDLCopt program.

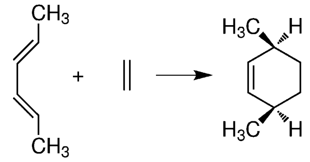


Fig. 2 Diels–Alder cycloaddition reaction of 2,4-hexadiene and ethene.

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:

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

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

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \sqrt{\sum_i g_i'(x_k)^2 \mathrm{e}^{-2\lambda_{ii}t}}, \tag{6}$$

where $g'(x_k) = U_k^{\dagger} g(x_k)$. 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 (Fig. 2).

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 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[26] that was reoptimized at the B3LYP/SVP[27–33] level using ChemShell in combination with the Gaussian 09 program.

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 $0.15\sqrt{\mathrm{amu}}$ 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 $0.05\sqrt{\mathrm{amu}}$ 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 $0.15\sqrt{\mathrm{amu}}$ 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 $0.05\sqrt{\mathrm{amu}}$ 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 $0.01\sqrt{\mathrm{amu}}$ 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 $0.15\sqrt{\mathrm{amu}}$ 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.

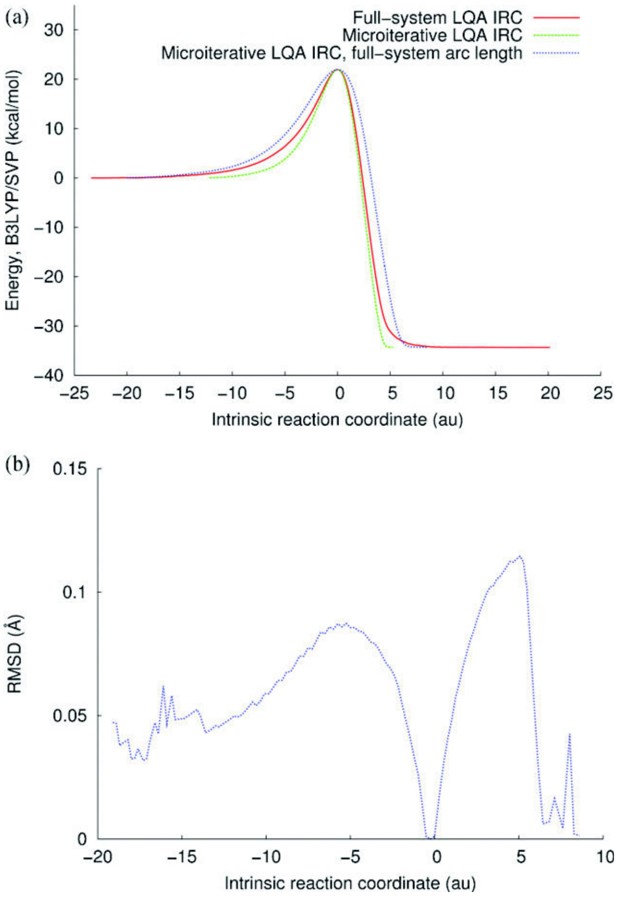


Fig. 3 IRC results for the Diels–Alder reaction obtained using the LQA method and a step size of $0.15\sqrt{\text{amu}}$ bohr. (a) Comparison between full-system and microiterative IRC energy profiles, see the text. (b) RMS deviations between the geometries along the full-system and microiterative IRC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

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 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

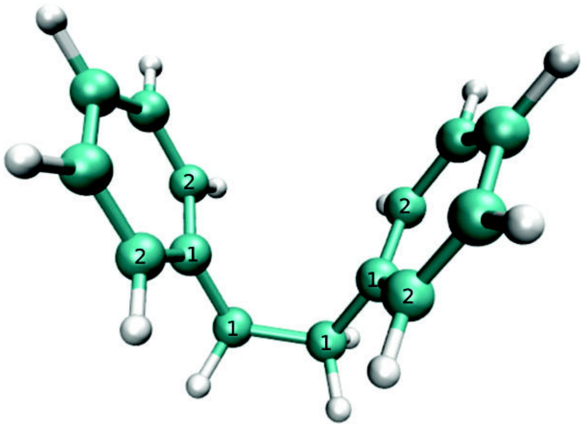


Fig. 4 Transition state for internal rotation in 1,2-diphenylethane, with assignment of carbon atoms to core regions (see text).

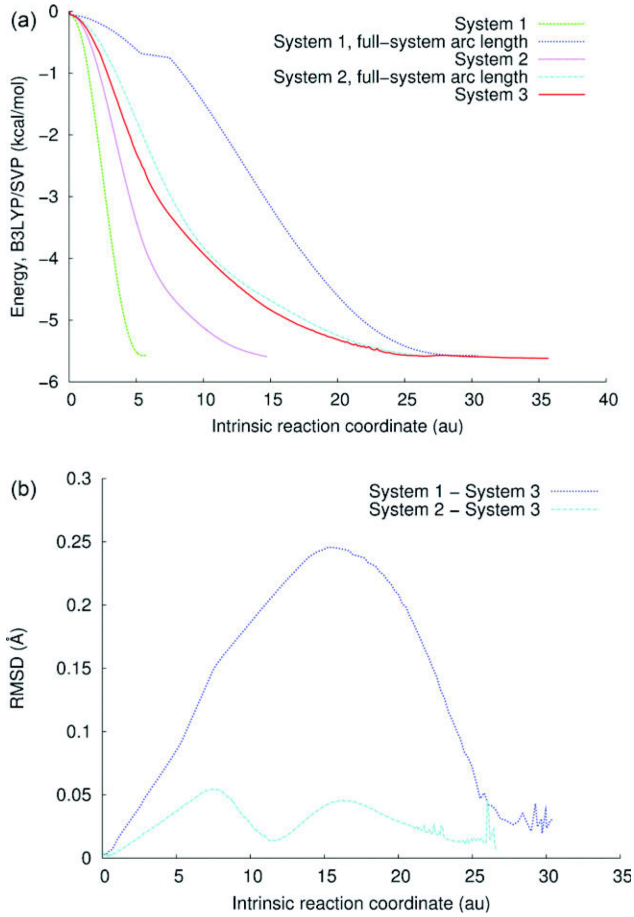


Fig. 5 IRC results for the internal rotation in 1,2-diphenylethane from the microiterative procedure for systems 1 and 2 (see text) and from the standard treatment for the full system 3. LQA step sizes: $0.15\sqrt{\text{amu}}$ bohr for 3 and $0.05\sqrt{\text{amu}}$ bohr for 1 and 2. (a) IRC energy profiles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

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 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.[34]

Using the QM/MM approach we treated the substrate (24 atoms) using the semiempirical OM3 method[35,36] and the rest of the system (including the protein and the solvent shell, 13 421 atoms in total) using the CHARMM22 force field.[37] The initial preparation of the system has been described elsewhere.[38] A snapshot from the previous classical molecular dynamics (MD) simulations[38] was selected and subjected to another MD run in the NVT ensemble using the CHARMM33b1 program.[39,40] 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

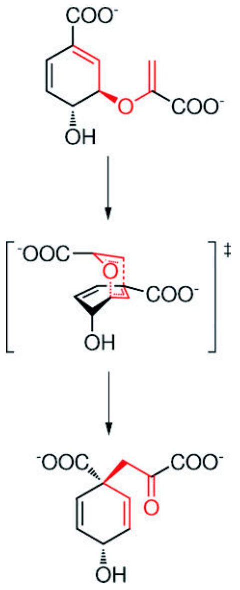


Fig. 6 Claisen rearrangement of chorismate to prephenate in BsCM.

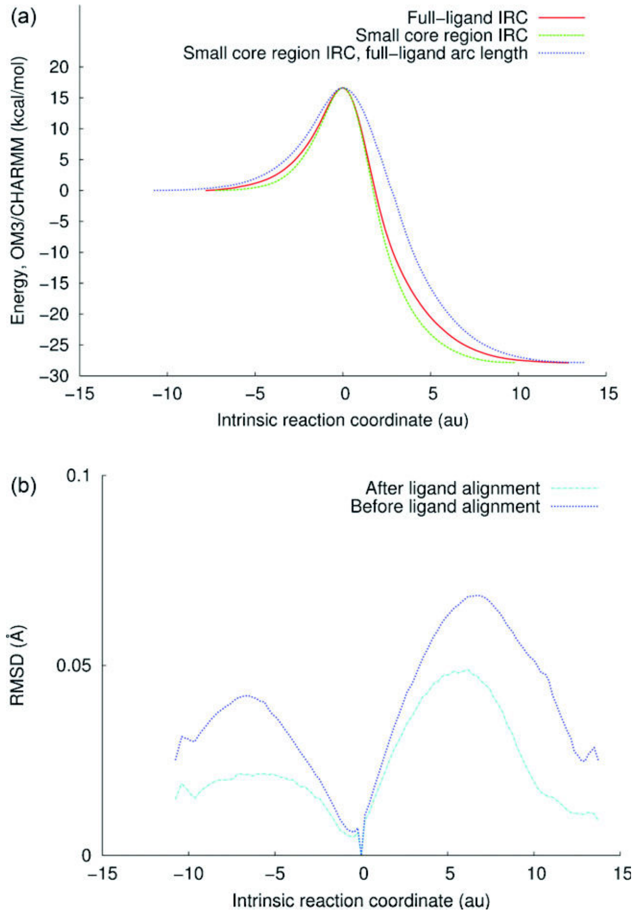


Fig. 7 Microiterative IRC results for the chorismate–prephenate conversion catalyzed by BsCM. LQA step size: $0.15\sqrt{\text{amu}}$ bohr. (a) Energy profiles from full-ligand and small-core IRC calculations (plotted in the latter case with different arc length definitions). (b) RMS deviations between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

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.

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).

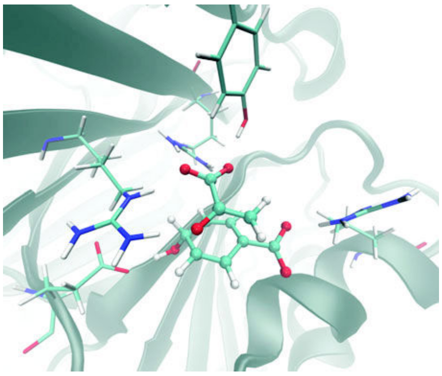


Fig. 8 Active center of BsCM. Shown is the substrate in its transition state and the five hydrogen-bonded active-site residues.

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 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.[41] 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.[42] The substrate and the isoalloxazine ring of the FADHOOH with the attached hydroperoxide group (48 atoms) were included in the QM region and treated using the semi-empirical AM1 method,[43] while the rest of the system was described by the CHARMM22 force field. A randomly chosen snapshot from those used in ref. 42 was subjected to a restrained potential energy scan along the reaction coordinate defined in ref. 42. The structure with the highest energy on this scan served as a

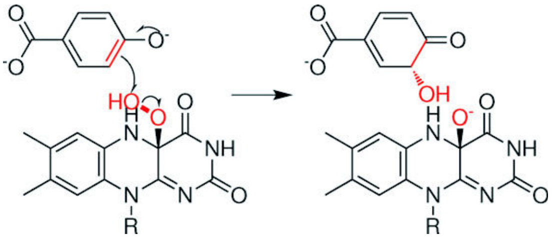


Fig. 9 Hydroxylation reaction catalyzed by *p*-hydroxybenzoate hydroxylase. R denotes the ribityl side chain of the FADHOOH cofactor.

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 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).

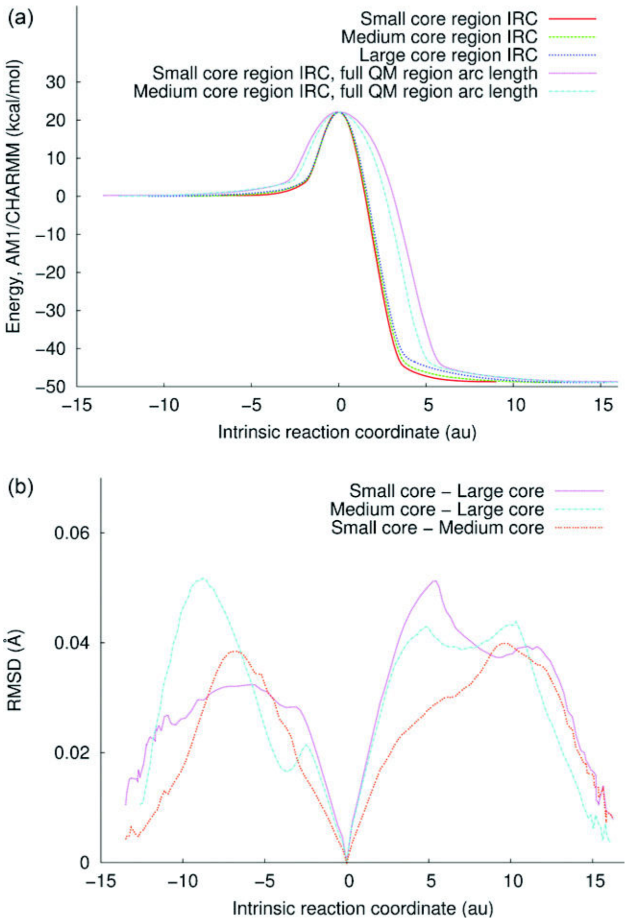


Fig. 10 Microiterative IRC results for the hydroxylation reaction conversion catalyzed by PHBH. LQA step size: $0.15\sqrt{\text{amu}}$ bohr. (a) Energy profiles from the IRC calculations with three different core regions (see text); those for the two smaller core regions are plotted with different arc 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. 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 H. P. Hratchian and H. B. Schlegel, *J. Chem. Theory Comput.*, 2005, **1**, 61.
26 S. M. Bachrach and P. B. White, *THEOCHEM*, 2007, **819**, 72.
27 J. C. Slater, *Phys. Rev.*, 1953, **91**, 528.
28 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200.
29 A. D. Becke, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1988, **38**, 3098.
30 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648.
31 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623.
32 C. T. Lee, W. T. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785.
33 A. Schäfer, H. Horn and R. Ahlrichs, *J. Chem. Phys.*, 1992, **97**, 2571.
34 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.
35 M. Scholten, PhD thesis, Universität Düsseldorf, 2003.
36 N. Otte, M. Scholten and W. Thiel, *J. Phys. Chem. A*, 2007, **111**, 5751.
37 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.
38 H. M. Senn, J. Kästner, J. Breidung and W. Thiel, *Can. J. Chem.*, 2009, **87**, 1322.
39 B. Brooks, R. Bruccoleri, D. Olafson, D. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187.
40 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.
41 H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.*, 2009, **48**, 1198.
42 T. Benighaus and W. Thiel, *J. Chem. Theory Comput.*, 2011, 7, 238.
43 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902.

QM/MM Dual Hamiltonian Free Energy Perturbation

I. Polyak, T. Benighaus, E. Boulanger, W. Thiel

J. Chem. Phys., revised version (2013)

QM/MM 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

(Dated: 16 June 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.

---

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

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,[6–12] and also employed in a modified form by Ryde,[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 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 coworkers,[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 → MM perturbation along the reaction coordinate and for estimating the vertical MM → 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 the following sections, 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:

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

In the QM/MM-FE method introduced by Yang et al.,[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 the ”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

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

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:

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

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*, equation 8 can then be reformulated accordingly:

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

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 section 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.

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 $A_1(\xi) = f(\xi) + \ln(k(\xi))/2\beta + const.$

We used the same parameters as in our previous work.[30] In atomic units, the barrier is

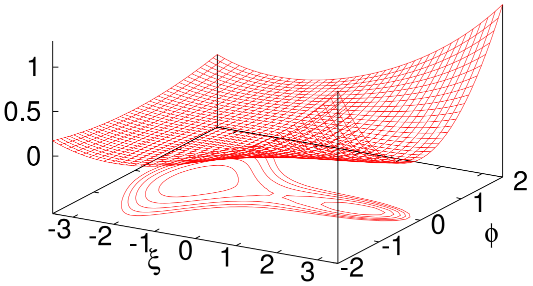


FIG. 1. 3D plot and contour plot of the analytic potential $E_1(\xi, \phi)$ with a contour spacing of 0.005. All values in atomic units.

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. 50000 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.

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

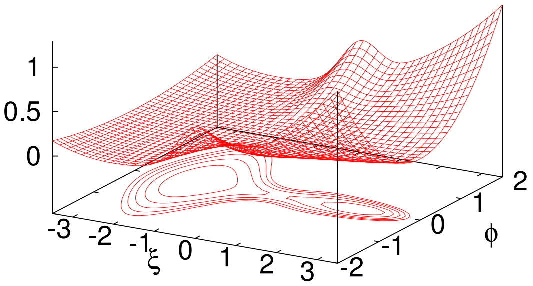


FIG. 2. 3D plot and contour plot of the analytic potential $E_{2a}(\xi, \phi)$ with a contour spacing of 0.005. All values in atomic units.

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.

The drastic rise of the error in the activation free energy confirms the importance of having

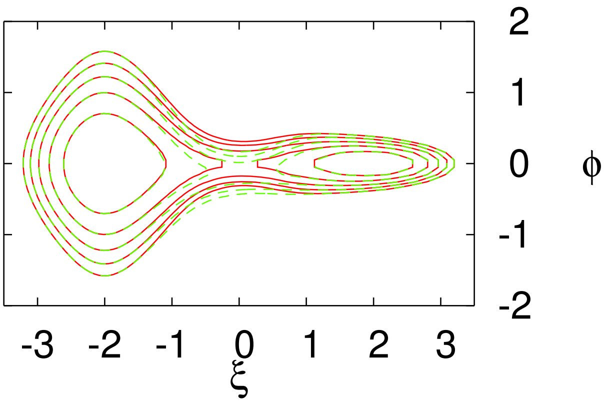


FIG. 3. Contour plot of the analytic potentials $E_{2a}(\xi, \phi)$ (solid lines) and $E_{2b}(\xi, \phi)$ (dashed lines) with a contour spacing of 0.005 and $\Delta = 0.2$. All values in atomic units.

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 $T = 300K$, 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 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

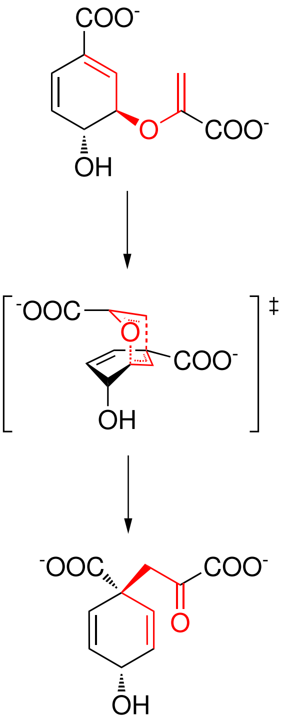


FIG. 4. Claisen rearrangement of chorismate to prephenate in chorismate mutase. The two parallel red dashed lines in the transition state indicate the forming and the breaking bonds. The difference between the corresponding distances is the reaction coordinate.

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 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,

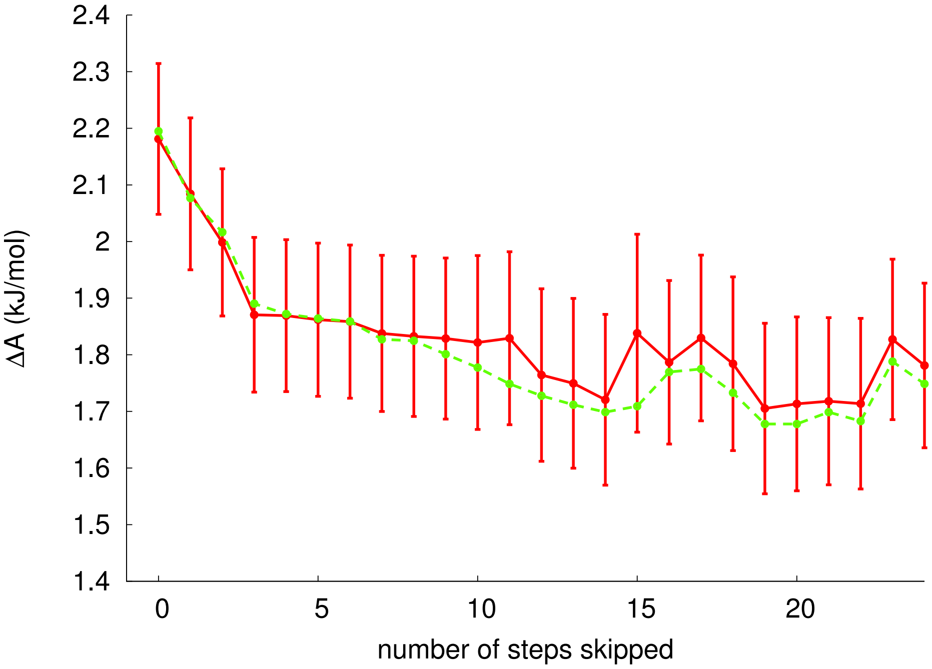


FIG. 5. Free energy difference between two windows calculated with a different number of steps skipped between the ΔE_{pert} evaluations, with the overall number of these evaluations fixed to 1000. The values in red (solid line) were obtained after subjecting the data to statistical tests for lack of 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).

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 Supporting Information[53]).

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

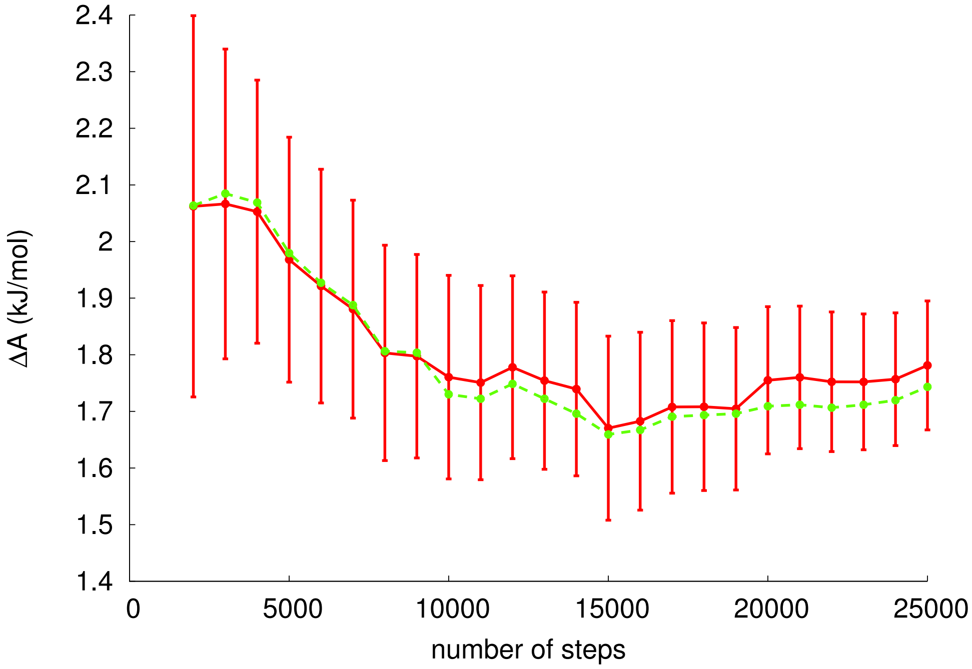


FIG. 6. Free energy difference between two windows calculated with 14 steps skipped between two ΔE_{pert} evaluations, with the overall number of these evaluations being varied. The values in red (solid line) were obtained after subjecting the data to statistical tests for lack of trend and decorrelation. 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).

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

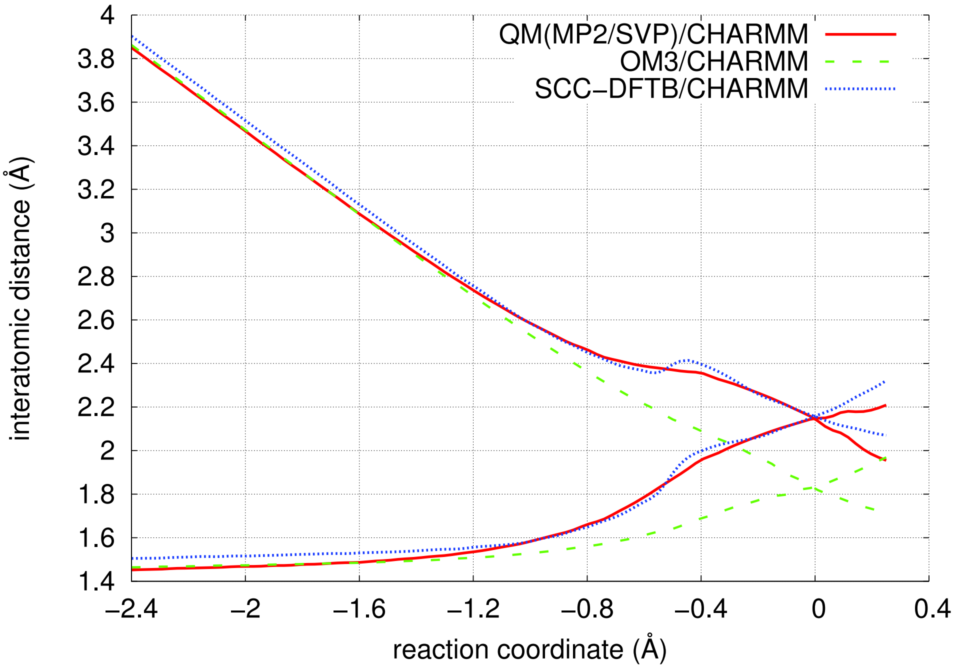


FIG. 7. Optimized C-C and C-O distances along the RC for the three different QM methods.

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 snaphot 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 Supporting Information[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 the previous sub-section 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 an 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$\}$).

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 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 Supporting Information[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

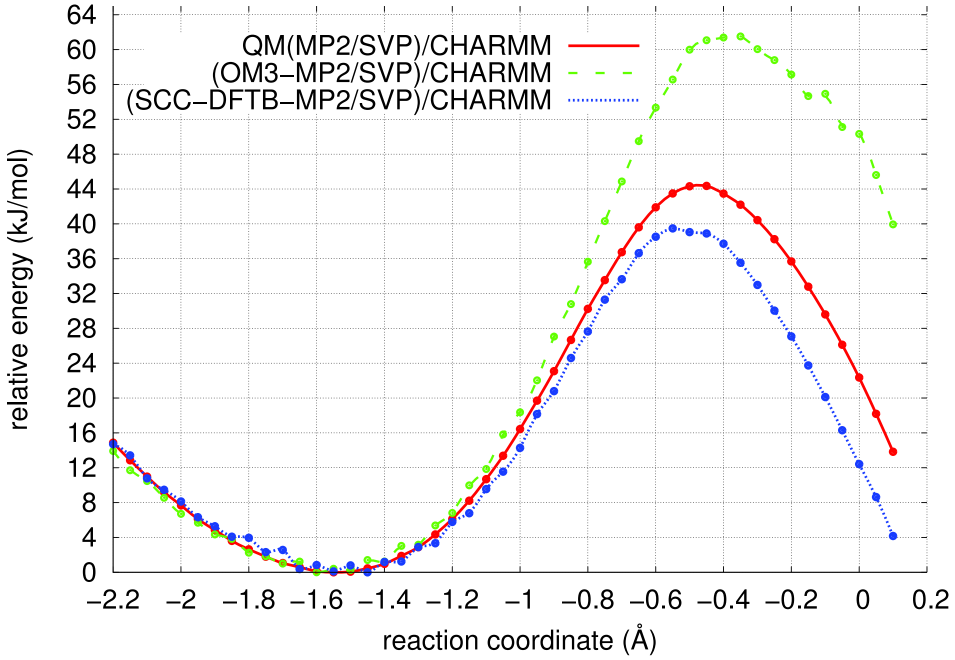


FIG. 8. Potential energy profile computed at the QM(MP2/SVP)/CHARMM level and QM(MP2/SVP)/CHARMM single-point energies at the optimized OM3 and SCC-DFTB structures along the reaction path.

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= $\{-2.4\ \text{Å}, -1.25\ \text{Å}\}$) with OM3/MM and using SCC-DFTB/MM for the second part (RC= $\{-1.25\ \text{Å}, 0.1\ \text{Å}\}$). 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 (dash-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

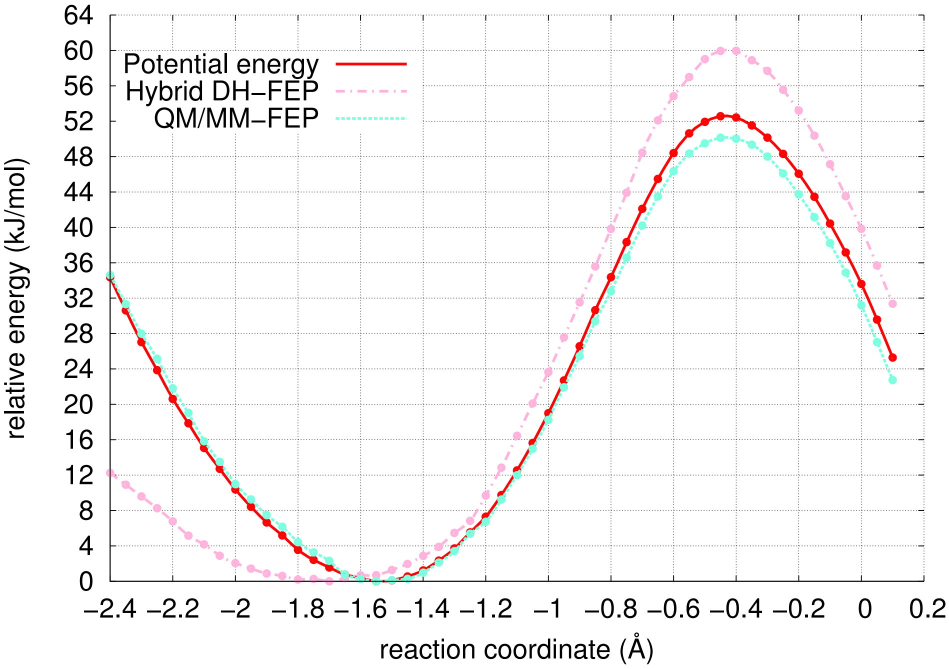


FIG. 9. Potential energy, DH-FEP, and QM/MM-FE profiles obtained for snapshot 6. The potential energy was computed at the QM(MP2/SVP)/CHARMM theory level. The DH-FEP profile was determined with a hybrid approach, in which the first part of the reaction path was sampled with OM3/CHARMM, and the second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

-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

TABLE I. Free energy and potential energy barriers and entropic contributions to the barrier of the BsCM-catalyzed reaction for the six snapshots considerd. 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

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 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

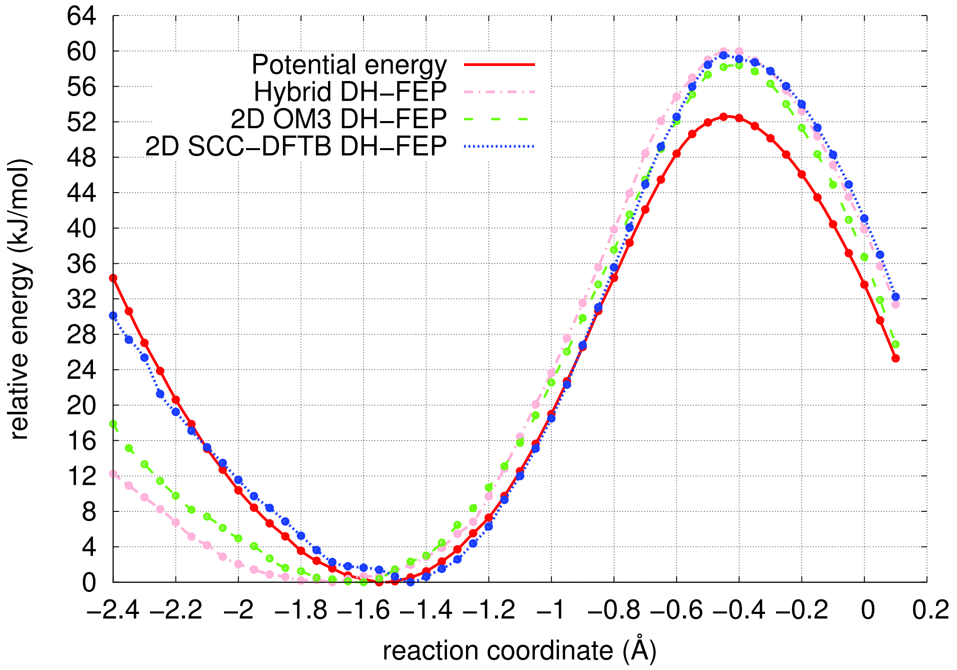


FIG. 10. Potential energy profile from QM(MP2/SVP)/CHARMM calculations and three free energy profiles computed for snapshot 6. Hybrid DH-FEP profile, sampling with OM3/CHARMM and SCC-DFTB/CHARMM for the first and second part of the reaction path, respectively (see text); 2D DH-FEP profiles, evaluated 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).

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). 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 dual Hamiltonian free energy perturbation (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 cooordinate (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.

REFERENCES

[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. Letters **294**, 143 (1998).

[43]W. Smith and T. Forester, J. Mol. Graph. **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 Letters **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 [URL will be inserted by AIP] for additional evaluation of the effect of sampling time on the free energy convergence and OM3/CHARMM and SCC-DFTB/CHARMM DH-FEP profiles.

[54] I. Polyak, E. Boulanger, K. Sen, and W. Thiel, "Microiterative intrinsic reaction coordinate method for large qm/mm systems," (2013), *online*, DOI:10.1039/c3cp51669e.

Supporting Information for:

QM/MM 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

Contents

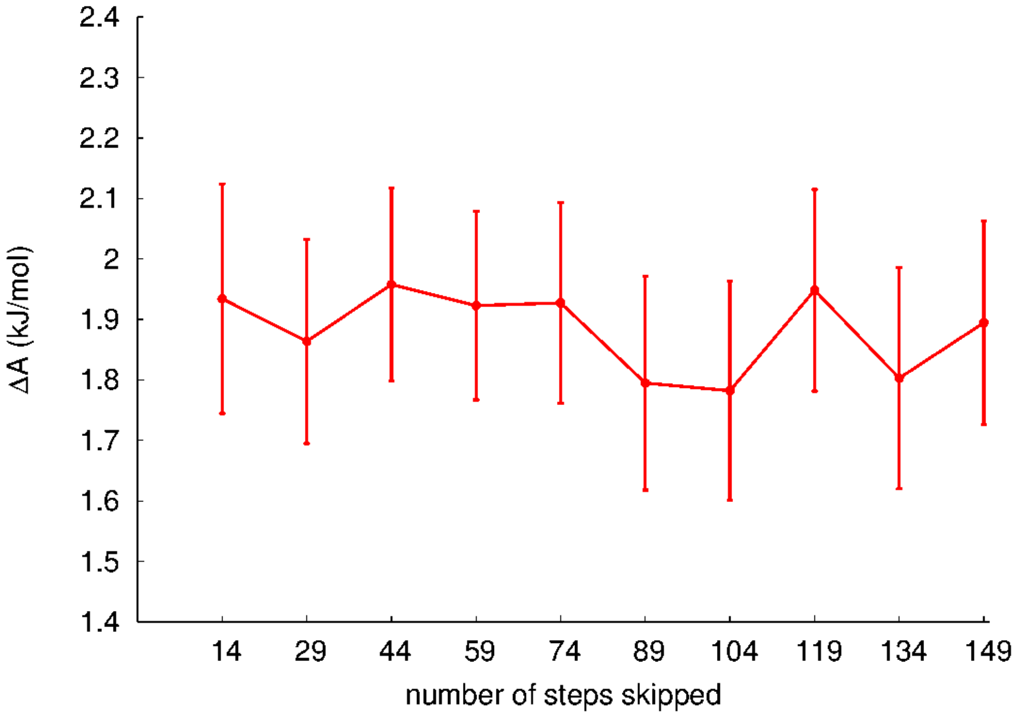
---

[a] Permanent address: Lanxess Deutschland GmbH, 51369 Leverkusen, Germany

[b] Electronic mail: thiel@mpi-muelheim.mpg.de

Fig. S1. Free energy differences between two windows of snapshot 4.

Fig. S1a.

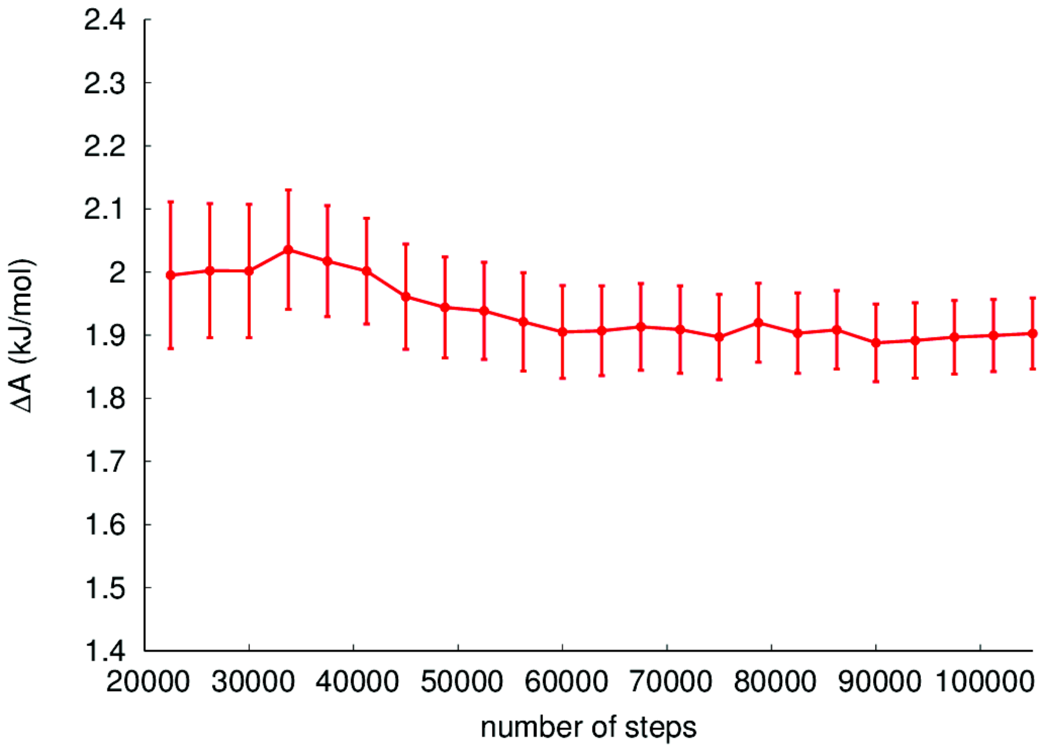


Free energy differences between two windows calculated with different numbers of steps skipped between the ΔE_{pert} evaluations, with the overall number of evaluations fixed to 700. Data taken starting from the end of a 105 ps OM3/CHARMM MD sampling run.

The computed free energy differences vary between 1.78 and 1.96 kJ/mol, without any obvious trend, when increasing the number of steps skipped between evaluations from 14 to 149. They remain within the range covered by the individual error bars, which are typically around 0.35 kJ/mol.

Comparison with Figure 5 of the main paper for snapshot 6: Fluctuations of the computed free energy differences persist at a roughly constant level, within the range of the error bars, when extending the time between ΔE_{pert} evaluations from 15 to 150 fs (thus decreasing the correlation between successive evaluations).

Fig. S1b.

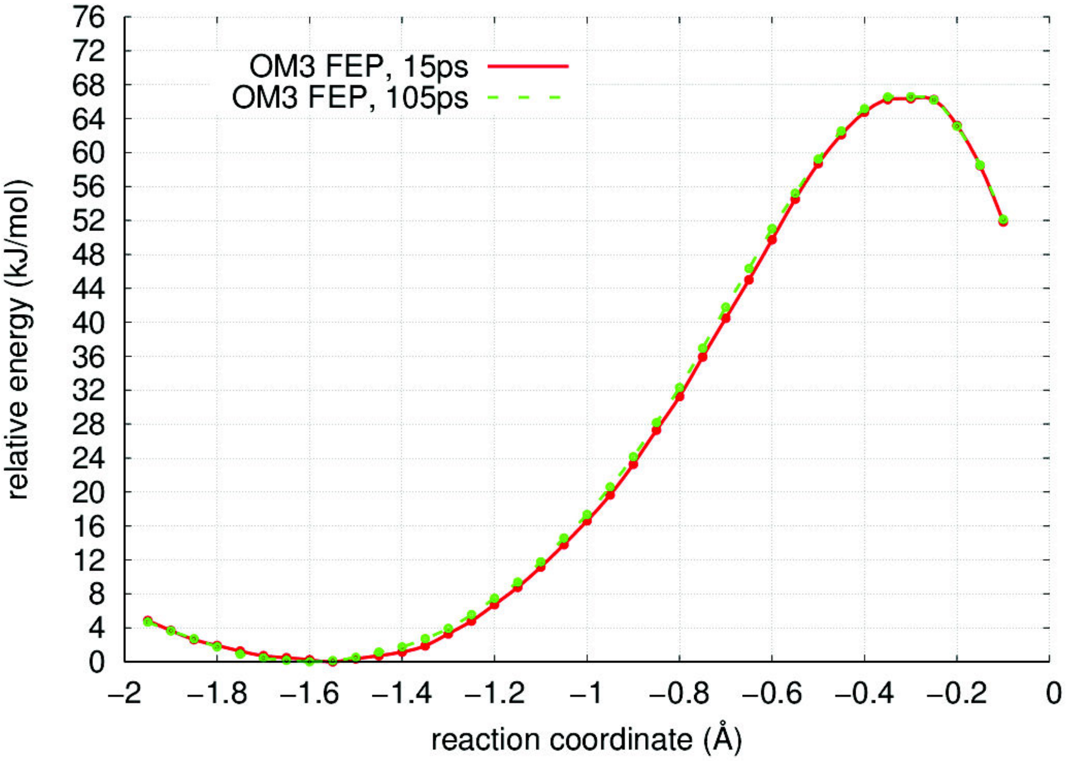


Free energy differences between two windows calculated with 14 steps skipped between two ΔE_{pert} evaluations, as a function of the overall number of evaluations. Data taken starting from the beginning of a 105 ps OM3/CHARMM MD sampling run.

The computed free energy differences vary between 1.89 and 2.04 kJ/mol, when increasing the sampling time from 22.5 to 105 ps. Initially they decrease slightly, before reaching an essentially constant plateau at around 1.90 kJ/mol. The individual error bars also decrease slightly with increasing sampling time, from about 0.2 to 0.1 kJ/mol.

Comparison with Figure 6 of the main paper for snapshot 6: The computed free energy differences and the associated error bars show reasonable convergence when extending the sampling time to 105 ps. The results obtained with the standard sampling time of 10 ps (see the main paper) differ only slightly from the converged value (which remains within the error bar of the standard value).

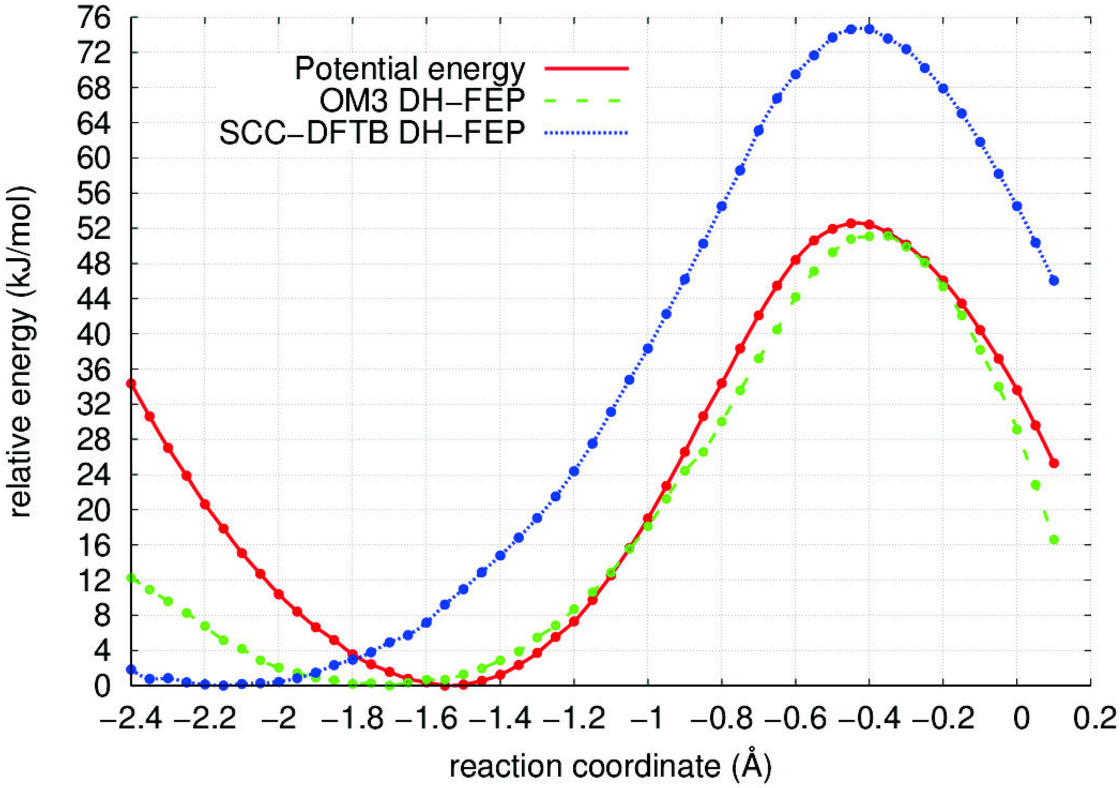
Fig. S2. FEP profiles obtained for snapshot 4 with different sampling times.



In both cases, OM3/CHARMM MD sampling was performed after 25 ps of equilibration, with ΔE_{pert} being evaluated at every 15th step. In the first case (red solid line) the sampling time for each window was 15 ps. In the second case (green dash-dotted line) it was prolonged to 105 ps.

The resulting activation free energies agree to within 0.2 kJ/mol. The corresponding uncertainties are 0.7 kJ/mol and 0.3 kJ/mol for the 15 ps and 105 ps simulations, respectively.

Fig. S3. Potential energy and DH-FEP profiles for snapshot 6.



The potential energy was computed at the QM(MP2/SVP)/CHARMM level. The DH-FEP profiles were computed with sampling done at the OM3/CHARMM or SCC-DFTB/CHARMM level, while ΔE_{pert} was evaluated using QM(MP2/SVP)/CHARMM.