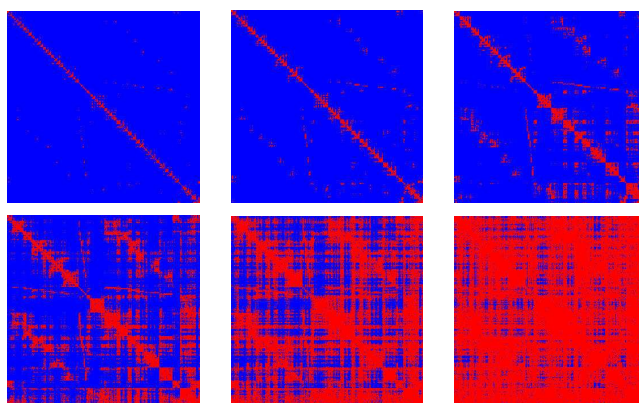


Study of the reaction mechanism in Mandelate Racemase enzyme: reaction path and dynamical sampling approaches

Xavier Prat Resina



Bellaterra, February 2004

JOSEP MARIA LLUCH I LÓPEZ I ÀNGELS GONZÀLEZ I LAFONT,
PROFESSOR CATEDRÀTIC I TITULAR, RESPECTIVAMENT, DEL
DEPARTAMENT DE QUÍMICA DE LA UNIVERSITAT AUTÒNOMA
DE BARCELONA, CERTIFIQUEM QUE

En Xavier Prat i Resina, llicenciat en Química per la Universitat de Barcelona,
ha realitzat sota la nostra direcció, en el Departament de Química de la Uni-
versitat Autònoma de Barcelona, el treball d'investigació titulat

**”Study of the reaction mechanism in Mandelate Racemase
enzyme: reaction path and dynamical sampling approaches”**

que es presenta en aquesta memòria per optar al grau de Doctor en Química.
I perquè consti als efectes escaients, signem aquest certificat a Bellaterra 26
de febrer de 2004:

Dra. Àngels Gonzàlez i Lafont

Dr. Josep Maria Lluch i López

*... tot i que mil i una tesi em caldrien
Als meus pares
i a la memòria dels meus avis*

*No hay un solo hombre que no sea un descubridor.
Empieza descubriendo lo amargo, lo salado, lo cóncavo, lo liso, lo áspero,
los siete colores del arco y las ventitantas letras del alfabeto;
pasa por los rostros, los mapas, los animales y los astros;
concluye por la duda o por la fe
y por la certidumbre casi total de su propia ignorancia.*

Contents

Preface	ix
1 Introduction to theoretical chemistry	1
1.1 Overview of the methods	2
1.2 Introduction: Potential Energy	5
1.2.1 Quantum Mechanics	5
1.2.1.1 Basic equations for a molecular system	5
1.2.1.2 Born-Oppenheimer approximation: Potential Energy Surface	6
1.2.1.3 Quantum nuclear motion	8
1.2.1.4 Electronic problem: Hartree-Fock	9
1.2.1.5 Semiempirical approximations	15
1.2.1.6 Density Functional Theory	18
1.2.1.7 More accurate solutions to the electronic problem	20
1.2.2 Molecular Mechanics	21
1.2.2.1 Energy terms: bonded and non-bonded	21
1.2.2.2 Long range interactions	23
1.2.2.3 Force fields	26
1.2.3 Hybrid methods	26
1.2.3.1 Polarized QM/MM	27
1.2.3.2 IMOMM-ONIOM	33
1.2.3.3 EVB/MM	34
1.2.4 Derivatives of the potential energy	35
1.3 Introduction: Optimization Methods	37
1.3.1 Common issues: convergence criteria and step length	37
1.3.2 Non derivative methods	38
1.3.3 First derivatives methods	39
1.3.4 Second derivative methods	39

1.3.4.1	Newton Raphson and quasi-Newton methods	39
1.3.4.2	Rational Function Optimization	40
1.3.4.3	Direct Inversion of Iterative Space (DIIS) . .	41
1.3.4.4	Update expressions and initial Hessians . . .	42
1.3.5	Reaction path	43
1.3.5.1	Coordinate scan	44
1.3.5.2	Intrinsic reaction coordinate	45
1.3.5.3	Interpolation between reactants and products	45
1.3.5.4	Chain methods	46
1.3.6	Cartesian, internal and redundant coordinates	46
1.3.7	Second order methods for large systems	47
1.3.7.1	Limited memory: L-BFGS	48
1.3.7.2	ABNR	49
1.3.7.3	Truncated Newton	50
1.3.7.4	Coupled or micro-iterative method	51
1.3.7.5	Internal vs Cartesian coordinates	52
1.4	Introduction: Molecular Dynamics	53
1.4.1	Basic equations and algorithms	54
1.4.2	Thermostats and barostats	54
1.4.3	Constraints	56
1.4.4	Langevin Dynamics and Stochastic Dynamics	57
1.5	Introduction: Statistical Mechanics	60
1.5.1	Free energy calculations	60
1.5.2	Potential of mean force	62
1.5.3	Chemical kinetics: Transition state theory	65
1.6	Enzymatic reaction simulations	67
2	Mandelate Racemase enzyme	69
2.1	Introduction: experimental results	69
2.1.1	Racemases and the aim of their study	70
2.1.2	Presentation of Mandelate Racemase enzyme	71
2.2	Introduction: Previous theoretical studies	77
2.3	Modulation of substrate activity	81
2.3.1	Methods and strategies	81
2.3.2	Reaction mechanism of propargylglycolate substrate .	85
2.3.3	Reaction mechanism of mandelate substrate	90
2.3.4	Comparison with experimental kinetics	93

2.3.5	Inhibition by propargylglycolate substrate	94
2.3.6	Discussion	96
2.3.7	Conclusions	99
2.3.8	Tables of results	99
2.4	Gas phase calculations	103
2.4.1	Semiempirical calculations	105
2.4.2	DFT calculations	108
2.4.3	Some short conclusions and perspectives	108
3	Optimization of big systems	111
3.1	Optimization in QM/MM surfaces	114
3.1.1	Equations and its implementation	115
3.1.1.1	RFO and updates	115
3.1.1.2	Initial Hessian onminimization and TS search	117
3.1.1.3	External minimizers: BFGS and L-BFGS	118
3.1.1.4	Implementation	119
3.1.2	Tests on model systems	121
3.1.2.1	Description of systems	121
3.1.2.2	Results and discussion	124
3.1.3	Conclusions	129
3.2	Micro-iterative method	130
3.2.1	Strategies and its implementation	131
3.2.1.1	Possible optionsin the micro-iterative method	131
3.2.1.2	Implementation	134
3.2.2	Tests on Mandelate Racemase	135
3.2.2.1	Results on the core size	137
3.2.2.2	Results on the frequency of the iteratedprocesses	140
3.2.2.3	Results on the interaction between QM and MM zones	142
3.2.3	Conclusions	143
3.3	How important is an accurate optimization	145
3.3.1	Procedure	146
3.3.2	Comparison between Mandelate Racemase mechanism structures	147
3.3.2.1	Energetic comparison	147
3.3.2.2	Geometric comparison	148
3.3.3	Conclusions	156
3.4	Avoiding the memory problem	158
3.4.1	Initial hessians, sparsity and storage	159

3.4.2	General scheme for the strategy	163
3.4.3	Results and discussion	167
4	Molecular Dynamics and Free Energy	169
4.1	Model and setup	170
4.1.1	Potential Energy Surface	171
4.1.2	Molecular Dynamics	173
4.1.3	Potential of Mean Force	175
4.2	PMF on different reaction coordinates	178
4.2.1	Selection of a reaction coordinate	178
4.2.2	Combining two bond distances	180
4.2.3	Combining four bond distances	184
4.2.4	Reactivity for the mutant N197A	188
4.3	Discussion and conclusions	191
5	Final Conclusions	193
A	Computer resources and informatics	197
A.1	An example of numerical computation	198
B	Source Code	201
	Acronyms	203
	Bibliography	205

Preface

I started the writing of this thesis not only to obtain a doctoral degree, but also to compile in a particular way all the work that I have done during all this time. The articles published during these years can only give a short overview of my research task. They may be found in the given references [1, 2, 3, 4, 5, 6, 7]. I decided to give my own perspective of the things I have learnt and the results I have obtained. Some sections are directly the published articles, but some other are not and contain a significative amount of unpublished data.

As you probably have already noticed, English is not my mother language. You may find lots of grammatical mistakes all along this thesis, but I thought that it was preferable to write science in bad english that many people could understand, rather than to write it in my beloved Catalan that unfortunately only few researchers in the scientific community read.

This thesis has four main chapters:

Chapter 1 In this chapter we can find an introduction not only to the methods used in the thesis, but there are also some sections devoted to some other methods. The number of strategies in theoretical chemistry is so big that a wide perspective helps to understand the choice we have done. Moreover, sometimes when a particular method has been modified it is difficult to split the theoretical framework between the standard methodology and our own contribution. This is why some redundancy can be found between the first chapter and the rest of the thesis.

Chapter 2 Here we can find an initial study of Mandelate Racemase enzymatic catalysis with QM/MM methods. An interesting insight is given to the chemistry of this enzyme. However, the complexity of the different reaction mechanisms with different substrates emphasizes the

need for additional methods to study Mandelate Racemase reactivity and enzymatic catalysis in general. The work of this section has been published in reference [1].

Chapter 3 This is the main body of the thesis. In this chapter a progressive development of a method to locate stationary points in big systems is carried out. In the first section the equations for minima and transition state search are developed, implemented and tested in small systems. This part corresponds mainly to the reference [2].

The second and third sections are devoted to the development and application respectively of a micro-iterative method. Here the equations implemented in the first section are now coupled to a minimizer in order to find real transition states in enzymatic systems [3], [4].

The work of the last section 3.4 contains unpublished results. We propose, implement and test a strategy to avoid the computational problems that arise in the optimization process when dealing with very big matrices.

All this chapter would have not been possible without the collaboration of Dr. Gérald Monard and Dr. Josep Maria Bofill. I am indebted to them for their help, their time and for all the things they have taught me.

Chapter 4 This last chapter of results contains the calculation of free energy profile corresponding to the Mandelate Racemase reaction. In particular, the potential of mean force is calculated including a wide discussion about the choice of the crucial reaction coordinate according to the previous results obtained in the transition state optimization. The paper corresponding to this section is under preparation[7].

I must thank Dr. Jiali Gao for his help and hospitality during a three months stay at the University of Minnesota. I learnt there the fundamental part of the simulation techniques applied in this chapter.

Appendix Two appendices conclude the thesis. The first is a brief exposition of several important aspects referring our daily tool, the computer. The second is a route map of the source code that my collaborators and I had to write to carry out the simulations.

This thesis has two main subjects studied explicitly, namely, the applied study of Mandelate Racemase reactivity (chapter 2 and 4) from one part, and the development of methods for optimization of structures in big systems (chapter 3) from the other.

However, both subjects are interrelated. The variety and complexity of Mandelate Racemase mechanisms studied in chapter 2 is the motivation to develop some new tools and it serves as an appropriate system to test the optimization methods of chapter 3. And the knowledge of the optimized structures is very valuable for the posterior study of Mandelate Racemase reaction by molecular dynamics in chapter 4. Therefore, despite some new tools have been developed and they can be exported to other molecular systems, I prefer to include Mandelate Racemase enzyme and its particular chemical problem as the central motivation of my investigation.

Finally I must thank to my directors Dr. Angels González and Dr. Josep Maria Lluh for their patience and the given opportunity to develop my PhD. I thank also Dr. Mireia Garcia, who has been my unofficial third supervisor during an important part of my four years work.

Xavier Prat Resina
Bellaterra, Febrer 2004

Chapter 1

Introduction to theoretical chemistry

In this chapter a revision of the methods that had to be studied during this thesis it is found. Some of them had to be modified or developed, some others only used as already implemented algorithms in program packages, and finally we also explain some methods that are not used at all. One may find this chapter a too long exposition of methods, or an unnecessary task since these are standard methods. However, during a research task the theoretical chemist must know the wide variety of methods available in the literature in order to proceed with his own work. Moreover, in my opinion the learning labor is one of the most important tasks during the initial formation of a scientist.

Even when using very standard techniques, theoretical chemistry is constantly progressing and sometimes there is not a clear separation between the chemist that develops or implements the tools (theoretical chemist) and the chemist that applies standard techniques to solve relevant chemical problems (computational chemist).

The first section contains a general but likely particular overview of techniques and strategies in theoretical chemistry. In the second, third, fourth and fifth section a deeper explanation of the methods used in this thesis is given. Since most of these methods are well known there will not be a full coverage of these fields. A lot of good books and articles already exist and the original references may be found in there. In particular, for the Quantum Chemistry section I prefer the Pauling's[8], Pilar's[9] and

Daudel's [10] books for its fundamental exposition of the problem, and the Jensen's [11] book for a very complete and updated review.

For the optimization section in my opinion there is no book covering all the wide range of methods very deeply. They may be found in Fletcher's [12], Leach's [13] and Schlick's [14] book. In the molecular dynamics and statistical simulation section, the Leach's [13] covers a recent overview of standard techniques and while Allen-Tildesley's [15] covers the basis of simulation techniques, McQuarrie's [16] stands for a deeper conceptual and fundamental background on the whole field.

1.1 Few historical remarks and overview of the methods

Theoretical chemistry was born in the context of theoretical physics, when physicists at the beginning of the XXth century tried to apply the newborn quantum mechanics to molecular systems. The understanding of the chemical bond, the electromagnetic spectra and the explanation of the stability of small molecular structures were some of the greatest achievements. A lot of effort has been done since then to improve the initial works made by Slater, Mulliken, Pauling, London and many others. This is why historically quantum chemistry has almost taken the whole task in theoretical chemistry [17].

After the second world war physicists were more interested in the sub-atom world, so the task to find an accurate description of molecular entities was left to a new class of scientists, the quantum chemists or theoretical chemists.

During the fifties and the sixties the appearance of computers made possible the first semi-empirical calculations and the establishment of the basis in *ab initio* techniques. The application of quantum mechanics to molecular systems was so challenging and difficult that theoretical chemists have sometimes forgotten the real experimental results and have focused only on the numerical solution of the equations. The development, implementation and computation of the equations is such a hard and uncertain task that it must be contemplated as a *computational experiment* by itself.

At the same time, during the fifties the first computers permitted also the first computation of thermodynamic magnitudes using condensed phase

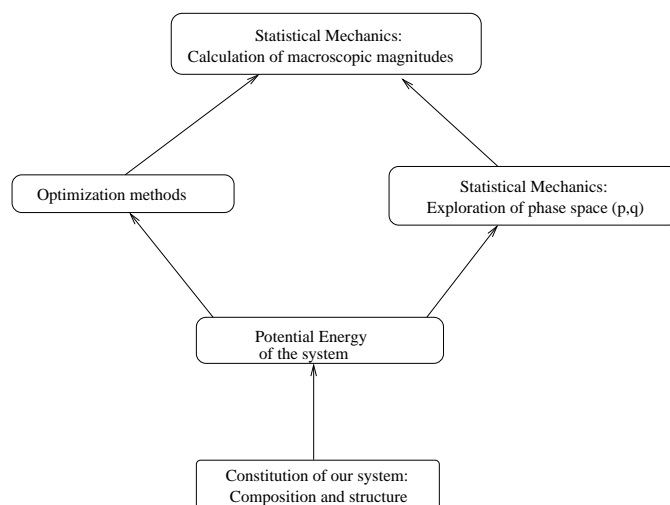


Figure 1.1: General scheme for a theoretical study of a chemical reaction

simulations. Metropolis with the Monte Carlo method in 1953, Alder and Mainwright using molecular dynamics in 1957 to calculate the phase transition for a hard sphere system, and Rahman in 1964 with the first simulation of a Lennard-Jones liquid were the first attempts in the new area of simulation.

Both condensed phase simulations and quantum chemistry calculations have evolved separately until recently, when the computational power and some new theoretical developments have paved the way for a joint and complete overview of techniques. In this sense a theoretical chemist should contemplate his research task in a wider and interdisciplinary research.¹ Actually, recent text-books already contain a larger and interdisciplinary view of theoretical chemistry. For example, twenty years ago Levine's Quantum Chemistry [18] and Allen-Tildesley's Liquids simulation book [15] covered very different areas. On the contrary, nowadays Leach's [13], Cramer's [19] or Jensen's book [11] include the molecular orbital theory along with some simulation techniques.

My particular unified view of a theoretical study of a chemical reaction can be drawn in figure 1.1. The scheme is not intended to be exhaustive, there are many fields in theoretical chemistry that fall out of this classifi-

¹Lots of names can be used to define the theoretical chemist research, *e.g.* biophysics, molecular modeling, nano-materials, drug design, solid state physics, computational structural biology ... *etc.*

cation (spectroscopy, structural chemistry, structure-activity relationships *etc*). But in my opinion every study of a chemical reaction should be contemplated by the different stages displayed in the figure. Every stage in figure 1.1 will be a section of this chapter. An important exception is the initial stage of knowing the structure and composition of the system. This question is not a problem in small sized systems, but in condensed phase systems the problem remains. The solvation of a macromolecule, protein folding or ligand binding are addressed to solve this problem in biopolymers, and the modeling of defects and impurities in solid state is also far from being solved.

In general optimization techniques and the exploration of phase space are usually applied to small sized systems and condensed phase systems respectively. Their common objective is the knowledge of the shape of the potential energy surface. However, one of the main arguments in this thesis is that optimization techniques can also be applied to condensed phase systems even when the exploration of phase space is feasible. We will see how both techniques can be applied to Mandelate Racemase reaction and how they complement each other.

1.2 Theoretical methods used in this thesis: Potential Energy

The simulation of chemical reactions should be ideally carried out by the tools provided by Quantum Mechanics. In practice, very small systems can be solved only by means of quantum theory. The potential energy that gives title to this section is the potential that molecular nuclei feel under the kinetic and potential energy influence of electrons. This electrons/nuclei separation comes from the Born-Oppenheimer approximation (section 1.2.1.2) which is a mathematical strategy and a conceptual milestone that gives meaning to the molecular entities and escapes from the no intuitive quantum conception of matter.

Big molecules such as enzymes and other condensed phase systems are too big to be computationally affordable with quantum methods, even to compute only their potential energy. To overcome these limitations these last years the so-called Quantum Mechanics / Molecular Mechanics (QM/MM) methods have been one of the most successful applications.

In order to explain the QM/MM methods used in this thesis, it is important to review the currently available methods in Quantum Mechanics and Molecular Mechanics, its advantages and its failures. In this way, we can justify the appearance and the success of hybrid methods in general and of the QM/MM strategies in particular.

1.2.1 Quantum Mechanics

The proposed solutions of a molecular system by means of Quantum Mechanics are not exact. We will not deduce rigorously all the equations, but rather illustrate the most important conclusions. The aim of this section is to show how accurate and how expensive are the current techniques to solve the so-called electronic problem.

1.2.1.1 Basic equations for a molecular system

The equations that rule a non-relativistic and unperturbed molecule are given by the Quantum Mechanics formulation.

From the fifth postulate in atomic units we have the time dependent

Schrödinger equation that describes the evolution of a quantum system.

$$- \frac{\hbar}{i} \frac{\delta |\Psi\rangle}{\delta t} = \hat{H} |\Psi\rangle \quad (1.1)$$

Where $|\Psi\rangle$ is the vector of the space that contains all the information of the system and \hat{H} is the Hamiltonian operator (the sum of kinetic and potential energy). Under the representation of positions the vector becomes the wave function and for a molecular system this Hamiltonian has the following form

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V}_{NN} + \hat{V}_{Ne} + \hat{V}_{ee} \quad (1.2)$$

or in a more detailed description

$$\begin{aligned} \hat{H} = & -\frac{1}{2} \sum_K^{nuclei} \frac{\Delta_K}{m_K} - \frac{1}{2} \sum_i^{electrons} \Delta_i + \sum_K^{nuclei} \sum_{K>L}^{nuclei} \frac{Z_K Z_L}{R_{KL}} \\ & - \sum_i^{electrons} \sum_K^{nuclei} \frac{Z_K}{r_{iK}} + \sum_i^{electrons} \sum_{i>j}^{electrons} \frac{1}{r_{ij}} \end{aligned} \quad (1.3)$$

Under the consideration that $|\Psi\rangle$ is a stationary state we get the time independent Schrödinger equation

$$\hat{H} |\Psi\rangle = E |\Psi\rangle \quad (1.4)$$

The value E is the energy eigenvalue of the Hamiltonian operator, a scalar that offers the spectrum of the operator.

The equation 1.4 cannot be solved exactly for a molecular system. The term \hat{V}_{Ne} in equation 1.2 does not permit to solve the independent Schrödinger equation by splitting the problem into a nuclear part separated from an electronic part. Consequently we will need to solve the equation 1.4 by different stages. This two-stage solution is provided by the Born-Oppenheimer approximation.

1.2.1.2 Born-Oppenheimer approximation: Potential Energy Surface

Empirical observations of molecular spectroscopy show that the total energy of a molecule can be viewed as the sum of several approximately non-interacting parts. Born-Oppenheimer approximation shows how the elec-

tronic motions can be approximately separated from the nuclear motions.

Let us define a molecule by a geometry structure determined by the nuclear positions. If the nuclei have fixed positions, the nuclear kinetic term vanishes $\hat{T}_N = 0$ and the nuclear repulsion term \hat{V}_{NN} becomes a constant. The Hamiltonian expression of equation 1.2 has a shortened form that we label as electronic Hamiltonian

$$\hat{H}^{elec} = \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee} \quad (1.5)$$

The solutions of the electronic Hamiltonian are the electronic wavefunctions that will have to be solved for every nuclear configuration R_K

$$\hat{H}^{elec}|\Psi_{R_K}^{elec}\rangle = E_{R_K}^{elec}|\Psi_{R_K}^{elec}\rangle \quad (1.6)$$

The \hat{V}_{nn} term is not usually included in the electronic Hamiltonian since it is only a constant at a given nuclear configuration. However we can define the potential energy adding the term \hat{V}_{nn} to the electronic energy.

$$U_{R_K} = E_{R_K}^{elec} + \hat{V}_{NN} \quad (1.7)$$

The solutions $\{|\Psi_{R_K}^{elec}\rangle\}$ is a complete set of functions of the n-electrons space. So, the total wavefunction of the system should belong to a full space created from the tensorial product between the nuclear space and electronic space: $\mathbb{E}_{tot} = \mathbb{E}_{elec} \otimes \mathbb{E}_{nucl}$.

$$|\Psi\rangle = \sum_p \sum_q C_{pq} |\chi^{nucl}(R_K)_p\rangle |\Psi^{elec}(R_K)_q\rangle \quad (1.8)$$

where $|\chi^{nucl}(R_K)_p\rangle$ is the wavefunction that describes the nuclear motion.

Although it is not a basis of the whole space, in the Born-Oppenheimer expansion the crossed terms are avoided and the wavefunction has the form

$$|\Psi\rangle = \sum_p C_p |\chi^{nucl}(R_K)_p\rangle |\Psi^{elec}(R_K)_p\rangle \quad (1.9)$$

If the total wavefunction has the form of equation 1.9 when it is introduced in the total Schrödinger equation 1.4 we have

$$(\hat{H}^{elec} + \hat{V}_{NN} + \hat{T}_N) \sum_p C_p |\chi_p^{nucl}\rangle |\Psi_p^{elec}\rangle = E \sum_p C_p |\chi_p^{nucl}\rangle |\Psi_p^{elec}\rangle \quad (1.10)$$

where the subscripts R_K are omitted for clarity. Multiplying this equation by $\langle \Psi_q^{elec} |$ we find the following coupled differential equations

$$\sum_p C_p \left\{ U_p \delta_{pq} + \langle \Psi_q^{elec} | \hat{T}_N | \Psi_p^{elec} \rangle \right\} |\chi_p^{nucl}\rangle = E |\chi_q^{nucl}\rangle \quad (1.11)$$

it means that the electronic states are coupled through the nuclear kinetic operator. Several steps are needed to proceed. The chain rule must be applied to the expression $\hat{T}_N |\Psi_p^{elec}\rangle |\chi_p^{nucl}\rangle$ and the stationary state must be assumed. Finally, to uncouple the above equations we have to assume that the nuclear kinetic operator is diagonal under the electronic representation

$$\langle \Psi_q^{elec} | \hat{T}_N | \Psi_p^{elec} \rangle = \delta_{qp} \langle \Psi_q^{elec} | \hat{T}_N | \Psi_p^{elec} \rangle \quad (1.12)$$

it means that the nuclear motion only involves one electronic state. This is the Born-Oppenheimer adiabatic approximation. Sometimes this approximation is not good and the nuclear motion has to include the participation of several electronic states[20, 21]. In any case, the assumptions considered so far permit to write the whole Schrödinger equation as

$$(\hat{T}_N + U_q^{(c)}) |\chi^{nucl}\rangle_q = E |\chi^{nucl}\rangle_q \quad (1.13)$$

where

$$U_q^{(c)} = U_q + \langle \Psi_q^{elec} | \hat{T}_N | \Psi_q^{elec} \rangle \quad (1.14)$$

The final approximation is to consider that even the diagonal correction can be neglected and therefore $U^{(c)} \approx U$. This last step is usually valid [22].

In conclusion, the nuclei move on a *Potential Energy Surface* (PES) U , where U comes from the solution of electronic Schrödinger equation 1.6, and E is the total energy of the system. The PES is a concept that will be used all along this thesis and we must keep in mind the level of approximations we have assumed to achieve such concept.

1.2.1.3 Quantum nuclear motion

Assuming that we can solve the electronic Schrödinger equation 1.6 the next step would be to solve the quantum nuclear motion through equation 1.13. In many textbooks we can find an analytical solution to this equation when the PES is a quadratic term. This is the harmonic oscillator and the radial

part of the nuclear wavefunctions contains the Hermite polynomial series. This treatment is also valid for the molecular vibrations of the molecule when every normal mode of vibration is assumed to move in a harmonic potential[23].

However, in many cases the harmonic oscillator is far from being adequate and some accurate treatments are needed to solve the nuclear Schrödinger equation and even to contemplate its evolution with time (equation 1.1). See the references [24, 25, 26, 27] for quantum and semi-classical methods used to solve this problem.

In this thesis the nuclei will be considered classical. It means that the nuclear kinetic operator will have the classical form

$$\hat{T}_N = \frac{1}{2} \sum_i^{nucl} m_i v_i^2 \quad (1.15)$$

and the nuclear wavefunction $|\chi^{nucl}\rangle$ will have no uncertainty in its momenta nor its position. It means that all nuclei can be represented by points of mass m_i and velocity v_i which move, according to equation 1.13, in a potential energy surface. This is usually a good approximation for heavy atoms and high temperatures. When this is not the case the quantum behavior of nuclei should be considered to reproduce purely quantum effects such as tunneling and reflection.

1.2.1.4 Electronic problem: Hartree-Fock

In this subsection we will try to show the main topics about the task that has occupied theoretical chemists during more than seventy years, the solution of electronic Schrödinger equation 1.6 for molecules. Obviously no analytical solution for this equation exists. Actually the many-body interacting problem is a very common problem that exists in many fields of physics, from the subparticles world to the astronomy.

The usual strategy is to consider a non-interacting system whose solution is usually known and then try to solve the real problem by means of perturbation theory or variational theory.

Independent particle model: Hydrogen atom:

The hydrogen atom was solved analytically by Schrödinger in 1926. Its

solution serves as a basis to consider the molecular case[8].

$$\hat{h}\Phi = \epsilon\Phi \quad \text{with} \quad \Phi(r\theta\phi) = R_{nl}(r)Y_{lm}(\theta\phi) \quad (1.16)$$

Where Φ is the atomic orbital (radial and angular) and \hat{h} is the mono-electronic Hamiltonian

$$\hat{h} = -\frac{1}{2}\Delta - \frac{1}{r} \quad (1.17)$$

Variational Principle:

The variational principle states that any wavefunction of the space accomplishes that its expectation value of the energy is an upper bound to the ground state energy of the system E_0 .

$$E_0 \leq \frac{\langle \Psi | \hat{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \quad (1.18)$$

For the sake of clarity the superscript *electronic* will be avoided from now on.

A test function: Slater determinant

An intuitive initial wavefunction would be considering the independent particle model. That is, a molecule constituted by non-interacting electrons so that every electron will be in an atomic spin-orbital (from here on ϕ_i for atoms) or in a molecular spin-orbital (Φ for molecules). This choice characterizes the molecular orbital theory. An alternative to this choice are the Valence Bond Methods where the test function is a combination of configurations of atomic orbitals. This strategy was named the Heitler-London-Pauling-Slater [8] and during the thirties it was a good alternative to the molecular orbital theory. However, despite valence bond methods give a chemical vision of the wavefunction easier to interpret, the numerical convergence is low and quantitative results are expensive.

The operator of non-interacting electrons is a sum of independent one-particle operators and its eigenfunction is the product of the one-particle wavefunctions, this is the so-called Hartree product

$$|\Psi^{hartree}\rangle = \prod_i^n \Phi(i) \quad (1.19)$$

Since the electrons are fermions they must be antisymmetric when interchanged. To *antisymmetrize* the Hartree product we will introduce the Slater

determinant

$$|\Psi^{slater}\rangle = \frac{1}{\sqrt{n!}} \begin{vmatrix} \Phi_1(1) & \Phi_2(1) & \dots & \Phi_n(1) \\ \Phi_1(2) & \Phi_2(2) & \dots & \Phi_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_1(n) & \Phi_2(n) & \dots & \Phi_n(n) \end{vmatrix} \quad (1.20)$$

This can be a good test function, even though it is important to note that we are dealing with only one function of the electronic space (equation 1.16 has infinite solutions), that is, the exact wavefunction for the real system is found in the basis that considers all electrons in all orbitals. It is defined a Slater determinant as a unique *configuration*. Then the exact solution cannot come from here but when we consider all possible configurations. In section 1.2.1.7 this option is commented as Configuration Interaction.

In any case we will see that a Slater determinant is not that bad. So, let us put this Slater determinant in equation 1.18. Considering the orthonormality of the molecular orbitals we obtain the following expressions known as Slater rules[28].

$$\langle \Psi^{slater} | \hat{H} | \Psi^{slater} \rangle = \sum_i^n h_{ii} + \sum_i^n \sum_{j>i}^n (J_{ij} - K_{ij}) \quad (1.21)$$

where

$$\begin{aligned} h_{ii} &= \langle \Phi_i(1) | -\frac{1}{2}\Delta_i - \sum_K^N \frac{Z_K}{R_{iK}} | \Phi_i(1) \rangle \\ &= \langle \Phi_i(1) | \hat{h}_1 | \Phi_i(1) \rangle \end{aligned} \quad (1.22)$$

$$J_{ij} = \langle \Phi_i(1)\Phi_j(2) | \frac{1}{r_{12}} | \Phi_i(1)\Phi_j(2) \rangle \quad (1.23)$$

$$K_{ij} = \langle \Phi_i(1)\Phi_j(2) | \frac{1}{r_{12}} | \Phi_j(1)\Phi_i(2) \rangle \quad (1.24)$$

The coulombic integrals J_{ij} and the exchange integrals K_{ij} can be rewritten in an operator form that will be useful later.

$$\hat{J}_i | \Phi_j(2) \rangle = \langle \Phi_i(1) | \frac{1}{r_{12}} | \Phi_i(1) \rangle | \Phi_j(2) \rangle \quad (1.25)$$

$$\hat{K}_i | \Phi_j(2) \rangle = \langle \Phi_i(1) | \frac{1}{r_{12}} | \Phi_j(1) \rangle | \Phi_i(2) \rangle \quad (1.26)$$

In this way, it can be seen the coulombic J_{ij} and exchange integrals K_{ij} as

the matrix elements representation of operators \hat{J}_i and \hat{K}_i in the molecular orbitals basis. Obviously, the \hat{J}_i and \hat{K}_i depend explicitly on the molecular orbitals themselves.

Minimization of the energy:

In order to obtain the best molecular orbitals that give the lowest energy we have to minimize the energy with respect to the orbitals, provided that these orbitals cannot be zero or linear dependent, that is, constraining the orbitals to be orthogonal and to have a norm of one.

As in many other situations, the optimization of a functional with constraints is carried out by the Lagrange Multipliers method. Building the Lagrangian \mathcal{L} , differentiating with respect to the molecular orbitals and imposing the stationary conditions ($\delta\mathcal{L} = 0$) we could see that the molecular orbitals can be obtained by solving the final Hartree-Fock (HF) equations:

$$(\hat{h}_i + \sum_j^n (\hat{J}_j - \hat{K}_j))|\Phi_i\rangle = \epsilon_i|\Phi_i\rangle \quad (1.27)$$

or defining the Fock operator as

$$\hat{F} = \hat{h}_i + \sum_j^n (\hat{J}_j - \hat{K}_j) \quad (1.28)$$

we have the condensed form

$$\hat{F}|\Phi_i\rangle = \epsilon_i|\Phi_i\rangle \quad ; \quad \epsilon_i = \langle\Phi_i|\hat{F}|\Phi_i\rangle \quad (1.29)$$

where ϵ_i are the Lagrange multipliers and can be interpreted somewhat as the energy of the *ith* molecular orbital (Koopmans theorem gives an alternative and sometimes useful definition[9]). However we must recall that the Hartree-Fock equations are just an strategy to obtain the optimized molecular orbitals and that the energy comes from the equation 1.21.

Equation 1.29 is a pseudo-eigenvalue equation because the Fock operator cannot be known unless we know the molecular orbitals, and molecular orbitals are obtained by the Fock operator. This dependence (exemplified in equations 1.25 and 1.26) forces that the Hartree-Fock equations must be solved iteratively until self-consistency. This procedure is called Self-Consistent-Field (SCF).

Introduction of a basis: Roothan-Hall equations:

Hartree-Fock equations can only be solved numerically (mapping the orbitals on a set of grid points) for highly symmetric systems (mostly atoms). In molecular systems we must introduce a known basis set $\{|\phi_j\rangle\}$ to span our molecular orbitals as a combination of these m basis functions.

$$|\Phi_i\rangle = \sum_r^m c_{ri} |\phi_r\rangle \quad (1.30)$$

This is the Linear Combination of Atomic Orbitals method (LCAO). The term atomic is used because it has been seen that taking atomic-like orbitals centered on every nucleus as a basis set the numerical results are optimal. The implementation of the LCAO method (equation 1.30) to the HF equations (equation 1.29) leads the so-called Roothan-Hall equations[10]

$$\hat{F} \left| \sum_r^m c_{ri} |\phi_r\rangle \right\rangle = \epsilon_i \sum_r^m c_{ri} |\phi_r\rangle \quad (1.31)$$

Multiplying equation 1.31 by $\langle\phi_s|$ with $s = 1, \dots, m$ and writing the equation in matrix form

$$\mathbf{FC} = \mathbf{SC}\epsilon \quad (1.32)$$

where \mathbf{S} is the overlap matrix and is not diagonal since the basis set are not orthogonal. An orthogonalization of the basis can lead to an eigenvalue equation $\mathbf{F}'\mathbf{C}' = \mathbf{C}'\epsilon$. Remember that the fock operator \hat{F} depends on the molecular orbitals, so if we substitute equation 1.30 in 1.28 we have an expression for the elements of fock matrix F_{pq} in the basis set representation

$$\begin{aligned} F_{pq} &= \langle\phi_p|\hat{F}|\phi_q\rangle \\ &= \langle\phi_p|\hat{h}|\phi_q\rangle + \sum_i^n \langle\phi_p|\hat{J}_i - \hat{K}_i|\phi_q\rangle \end{aligned} \quad (1.33)$$

$$\begin{aligned} &= \langle\phi_p|\hat{h}|\phi_q\rangle \\ &+ \sum_i^n \sum_r^m \sum_s^m c_{ri} c_{si} (\langle\phi_p\phi_r|\frac{1}{r}|\phi_q\phi_s\rangle - \langle\phi_p\phi_r|\frac{1}{r}|\phi_s\phi_q\rangle) \end{aligned} \quad (1.34)$$

$$F_{pq} = h_{pq} + \sum_{r,s} P_{rs} G_{pqrs} \quad (1.35)$$

The term h_{pq} are the core integrals and G_{pqrs} the two-electron integrals.

The two electron integrals nomenclature is usually simplified as $\langle pr|sq\rangle$. Depending on the atom where the four different functions are centered it will be a two-electron integral of one, two, three or four centers. The three and four centers are the most common integrals and then the most expensive to evaluate. The matrix $P_{rs} \equiv \sum_i^n c_{ri}c_{si}$ is usually called the density matrix. Its elements will be the coefficients to optimize in order to have a diagonal representation of the Fock matrix and therefore the pursued coefficients to obtain the molecular orbitals.

The method described above is the Restricted Hartree-Fock (RHF), every spin-orbital is a molecular orbital *occupied* by two electrons with spin function α and β so that the whole expectation value of spin operator \hat{S}^2 is zero. For open shell systems there are the Restricted Open Shell HF (ROHF) or the more adequate Unrestricted HF method (UHF). The corresponding Roothan-Hall equations for the UHF case are the Pople-Nesbet equations.

Few comments about the basis set:

The analytical solution of Hydrogen atom of equation 1.16 has a radial part represented by the Laguerre polynomials and an angular part represented by the Spherical Harmonics which are basically the associate Legendre Polynomials [8, 9]. The basis that we usually introduce in the Roothan-Hall equations will have a radial and an angular part. The angular is almost never commented because it is always the same (s,p,d,f...). It is the radial part which decides how far or how close is the electron with respect to the nucleus. The radial function can be represented by the Slater Type Orbitals (STO). However, despite of the adequacy of the STO basis, the integrals are not analytical using STO. Then to avoid expensive numerical integrals it is more common to use Gaussian functions (GTO) which integrals are analytical or easier to evaluate. In any case some modern methods still employ the STO functions. Furthermore, the semiempirical methods used in this thesis are carried out by STO as well. Many textbooks can be found covering the basis set, the different types, their efficiency and the computational requirements[11].

The SCF process:

We can summarize all the HF process through the Roothan-Hall equations as follows:

1. Input data: composition (Z_k) and geometry (X_K) of the system and functions basis set $\{\phi_r\}$

2. Integral computing: One electron (h_{pq} and S_{pq}) and two electron (G_{pqrs})
3. Orthogonalize the basis set: $\mathbf{C} = \mathbf{L}\mathbf{C}'$ so that $\mathbf{S} \rightarrow \mathbf{1}$
4. Give an initial guess to obtain a tentative molecular orbitals
5. Build the Fock matrix as core integrals + density matrix \times two-integrals
6. Transform the Fock matrix to the orthogonal basis set $\mathbf{F}' = \mathbf{L}^T\mathbf{F}\mathbf{L}$
7. Diagonalize the Fock matrix and obtain the coefficients $\mathbf{C}' \rightarrow \mathbf{C}$ and the new molecular orbitals
8. Compute the HF energy and check for convergence.
9. If it is not converged compute the new density matrix and go back to point 5.

This procedure has some bottlenecks that could make the whole process not feasible. The computation and the storage of integrals, several diagonalization and matrix transformations combined with a slow convergence process. A discussion on how expensive is every step and the different possibilities to overcome these problems may be found in the literature. Particularly the linear scaling techniques[29] deal with the problem of converting the whole computational requirements into a process that scales linearly with the size and/or with the basis set.

It is important to note that the HF method is the best method for a one-configuration wavefunction because we have used the exact electronic Hamiltonian to develop the equations. However there are many improvements to the HF method. Some of them will be outlined very briefly in section 1.2.1.7

1.2.1.5 Semiempirical approximations

The SCF process as explained above is a too expensive process for our purpose in enzymatic systems. For both the size of our system and the high number of energy evaluations we will need a cheaper method. The so-called semiempirical methods simplify the HF-SCF equations in such a way that the molecular orbitals and the energy are obtained faster. In this subsection

we will briefly mention the level of approximations just to be aware on how cheap and how accurate these methods are.

Only valence electrons with minimal basis set:

First of all only the valence electrons are considered explicitly, that is, there will be only basis functions for them, so the nucleus in the electronic Hamiltonian will have a lowered effective charge Z' due to the screening by the core-electrons². Moreover, the basis set used for the valence electrons will be minimal. It means, for example, one function for hydrogen and four functions, representing the s and three p, for the second and third row elements.

Discarding integrals: ZDO approximation

Most of the semiempirical strategies are based on the Zero Differential Overlap approximation (ZDO). That is to consider zero the product (not the integral) between two basis functions.

$$\mu_A \cdot \nu_B = 0 \quad (1.36)$$

Note that we changed the notation, now the basis function are denoted as μ, ν, λ and σ and the subscript indicates the atom where they are centered.

In particular, the Neglect of Diatomic Differential Overlap (NDDO) method sets to zero only those products which functions are centered on different atoms A and B that depend on the same electron coordinates.

$$\mu_A(i) \cdot \nu_B(i) = 0 \quad A \neq B \quad (1.37)$$

As a consequence of the NDDO approximation the overlap matrix becomes the identity and the three and four center electrons are vanished. In consequence the Fock matrix elements become

$$F_{\mu\nu} = h_{\mu\nu} + \sum_{\lambda,\sigma}^{AO} P_{\lambda\sigma} [\langle \mu\nu | \lambda\sigma \rangle - \langle \mu\lambda | \nu\sigma \rangle] \quad (1.38)$$

where for the monoelectronic elements $h_{\mu\nu}$:

$$\langle \mu_A | \hat{h} | \nu_A \rangle = \langle \mu_A | -\frac{1}{2}\Delta - V_A | \mu_A \rangle - \sum_{a \neq A} \langle \mu_A | V_a | \nu_A \rangle \quad (1.39)$$

$$\langle \mu_A | \hat{h} | \nu_B \rangle = \langle \mu_A | -\frac{1}{2}\Delta - V_A - V_B | \mu_B \rangle \quad (1.40)$$

²Sometimes a screening or repulsive function is used

and bielectronic

$$\langle \mu_{A\nu_B} | \lambda_C \sigma_D \rangle = \langle \mu_{A\nu_B} | \lambda_A \sigma_B \rangle \delta_{AC} \delta_{BD} \quad (1.41)$$

Since many terms are discarded the rest of integrals must be parameterized by experimental data, in such a way that we include in the surviving terms a correction to give the adequate results.

Further approximated methods exist, for example, the Completely Neglect Differential Overlap (CNDO) and the Intermediate Neglect Differential Overlap (INDO) Hamiltonians, which discard even more elements of the Fock matrix. Although these methods had their importance in the sixties and seventies they are not used any more. The methods currently used and still improved such as the Modified Neglect of Diatomic Overlap (MNDO) [30], the Austin Model 1 (AM1) [31] and the Parameterized Model 3 (PM3) [32, 33] are based on the NDDO Hamiltonian. They only differ in the way the core-core repulsion is treated and how the atomic parameters are assigned. There are other new methods such as MNDO/d, SAM1, ZINDO and PM5³. We will not comment the differences between these methods, check the reference [34] for an updated review in semiempirical methods.

Some earlier versions of semiempirical approximations prior to the ZDO-type existed for the treatment of the π -bonds. The Hückel method, the Pariser-Parr-Pople are some examples of such over simplification. However no useful results for the molecular structure came out until the explicit treatment of σ -bond with Pople's CNDO [35].

The most frequently used Hamiltonians in QM/MM enzymatic studies are AM1 and PM3. Both of them have their own recognized failures. A clear example is the low rotation barrier in amide bond which is usually corrected by an additional molecular mechanics term. Another required improvement is the adequate description of hydrogen bonds [36]. While AM1 has been parameterized by hand in a chemical intuition fashion, PM3 parameters have been obtained in a relatively automated way using a wider range of chemical data. However its accuracy is very similar, so before the choice a previous comparison with *ab initio* results is always recommended.

³The parameters of PM5 are not published and therefore unknown, these kind of commercial initiatives does not help very much to the progress in computational chemistry

1.2.1.6 Density Functional Theory

Density functional theory (DFT) has a long story on the treatment of electronic problems. Mainly after the Hohenberg and Kohn theorems that state that the exact energy of the ground electronic state of a system can be obtained in a biunivocal relationship from the electron density ρ expression [37]. This theorem shows that there exists an expression (a functional of the electron density) that connects electron density and the exact energy. The problem is that the theorem does not specify how this functional looks like.

The application of the DFT to the computational chemistry was not found useful until the late eighties. After that time, the DFT techniques have had an explosion of successful applications to chemical problems due to the usage of more appropriate functionals for these kind of systems [38, 39].

The DFT applied to a molecular system may be solved in such a way that keeps many parallelism with the already explained Hartree-Fock wave mechanics. However, in this case the electron density distribution plays a central role rather than the many-electron wavefunction. In defense of the DFT it is said that while in some cases the wave mechanics must take into account thousands of determinants into the wavefunction, the DFT gives an easier understanding by its only three-space coordinates electron density.

Kohn-Sham equations:

The equivalent of the Hartree-Fock equations in DFT are the self consistent Kohn-Sham equations (KS)[40]. Although this is not the only way to consider the DFT equations, it is the most frequently used in computational chemistry, so we will focus on this particular aspect of the DFT, the rest of applications and theories may be found in the literature[41].

This strategy can be seen as modeling an independent particle model that is able to reproduce, in theory, the full interacting system. The energy is obtained through a sum of separated functionals of the electronic density ρ .

$$E_{DFT}[\rho] = T[\rho] + E_{Ne}[\rho] + V[\rho] \quad (1.42)$$

Both the kinetic term $T[\rho]$ and the electron-electron term $V[\rho]$ are unknown functionals. The $V[\rho]$ can be derived as a sum of the correlated coulombic interaction $J[\rho]$ and another functional that would account for the exchange found in HF.

$$V[\rho] = J[\rho] + K[\rho] \quad (1.43)$$

For the kinetic term $T[\rho]$ Kohn and Sham proposed to define an independent electron system described by a set of orbitals $\{\Phi_i\}$ that give a density ρ_0

$$\rho_0(\vec{r}) = \sum_i^n |\Phi_i(\vec{r})|^2 \quad (1.44)$$

The central assumption of the Kohn-Sham strategy is to assume that the above fictitious system give the same electron density ρ_0 than the real system ρ

$$\rho_0(\vec{r}) = \rho(\vec{r}) \quad (1.45)$$

The kinetic term corresponding to this hypothetical system $T_0[\rho]$, although it is not the exact, is known and is the one that is used for HF.

$$T_0[\rho] = \sum_i^n \langle \Phi_i | -\frac{1}{2} \Delta_i | \Phi_i \rangle \quad (1.46)$$

$$E_{DFT}[\rho] = T_0[\rho] + E_{Ne}[\rho] + J'[\rho] + E_{xc}[\rho] \quad (1.47)$$

Where the uncorrelated coulombic functional $J'[\rho]$ has the known expression from the HF. And the functional $E_{xc}[\rho]$ must now include the correction to the Kinetic energy as well

$$E_{xc}[\rho] = (T[\rho] - T_0[\rho]) + (V[\rho] - J'[\rho]) \quad (1.48)$$

This unknown exchange-correlation functional is where all the recent developments in DFT theory have achieved its major success. $E_{xc}[\rho]$ can be designed by local (LDA), gradient corrected or generalized gradient corrected (GGA) (*e.g.* BLYP[42, 43]) and hybrid approximations (B3LYP [44]). The hybrid functional B3LYP has produced excellent results when it is compared with other strategies of similar computational cost. However this field is still in progress, and new functionals of the energy are being proposed to cover a wider variety of chemical systems [45].

Despite of the new nomenclature it is easy to see that equation 1.47 has similar functionality that the Hartree-Fock expression. So if the functional $E_{xc}[\rho]$ is only the exchange part \hat{K}_j found in the HF scheme (equation 1.26) we recover the same HF expression. In this sense, the implementation of HF equations might be seen as a particular case of the DFT-KS equations. However, if the functional $E_{xc}[\rho]$ was known we would obtain the exact

energy including all the electron correlation.

In the same manner the functional $E_{DFT}[\rho]$ is minimized with the constraint of the orthonormality of the Kohn-Sham orbitals. The Kohn-Sham orbitals can be spanned as a combination of basis functions centered on every atom leading to the Kohn-Sham equations that have the same scheme of the Roothan-Hall procedure.

The computational cost of the KS equations is very similar to the Roothan-Hall-HF. However some differences exist, it could depend strongly on the expression used for the $E_{xc}[\rho]$ functional. Usually some integrals due to the exchange-correlation expression have to be evaluated numerically in a grid of points.

$$\langle \phi_r | V_{xc}[\rho(r), \nabla \rho(r)] | \phi_s \rangle \approx \sum_i^{\text{points}} V_{xc}[\rho(r_i), \nabla \rho(r_i)] \phi_r(r_i) \phi_s(r_i) \Delta v_i \quad (1.49)$$

The number of points, the size and the shape of the grid is crucial to reproduce a computational result.

1.2.1.7 More accurate solutions to the electronic problem

At this point we must mention that there exist more accurate approximations to the electronic problem. These are usually referred as the post-Hartree-Fock methods or electron correlation methods. Starting from the HF equations we can include more configurations (determinants) to the wavefunction. The solution is progressively improved variationally or by perturbation theory. Mainly there are three categories in the post-HF methods, the Configuration Interaction (CI or MCSCF), the Many-Body perturbation theory and Coupled Cluster theory. An alternative to the post-HF methods, although very expensive, are the promising Quantum Monte Carlo methods[46] which give the most accurate results in small molecular systems.

Unfortunately these more sophisticated tools are still too expensive to be applied to condensed phase reactions where the systems are too big and many energy and gradient evaluations are required. Although this field occupies a significative part of theoretical chemistry they will not be commented further. Related to this topic some very good reviews and books have been published [11, 28, 47].

1.2.2 Molecular Mechanics

In the last section we have seen how hard is to obtain the energy of a molecular system. Many integral computations, many diagonalizations and transformations are needed to obtain, for example, the relative stability of a molecular conformation.

This is why, parallel to the methods of Quantum Chemistry there has always been the thinking that we could obtain similar results if we introduce all the qualitative chemical knowledge about molecular structure into a parametric function. That is, the strength of a chemical bond between two atoms, the steric hindrance, dispersion forces, hydrogen bonds, electrostatics ... All these interactions could be put together as a sum of analytical functions that give as a result a parametric energy function of the nuclear coordinates. This energy function is called an empirical force field and the strategy is the Molecular Mechanics (MM).

Historically there are several analytical functions that describe the inter and intramolecular interactions. Some examples of the oldest expressions are the Morse potential for the chemical bond, the Lennard-Jones or the Buckingham potential for the intermolecular interactions, the LEPS procedure ... and many others. However it is not until the end of sixties and the beginning of seventies with the work of Lifson, Allinger and Scheraga, and with the help of the emerging computers that some useful results are obtained. These Molecular Mechanics parametric functions need to be parameterized according to experimental results or *ab initio* calculations. We can find many different force fields in the literature. A force field will be characterized by the number and functional type of the energy terms and by the way the parameters are obtained.

1.2.2.1 Energy terms: bonded and non-bonded

We are going to assume that the energy of the system is separable in different terms. The usual separation is the following

$$E_{bonded} = E_{stretching} + E_{bending} + E_{torsion} + E_{improper} \quad (1.50)$$

$$E_{non-bonded} = E_{vanderWaals} + E_{electrostatic} \quad (1.51)$$

$$E_{total} = E_{bonded} + E_{non-bonded} \quad (1.52)$$

There are many expressions for every term, some force fields incorporate crossing terms to account for the coupling between two different interaction types. However, we are interested in those force fields that are specialized in the treatment of biomolecules. In this case the common energy terms are

$$E_{stretching} = \sum_i^{bonds} k_{ri}(r - r_{eq})^2 \quad (1.53)$$

$$E_{bending} = \sum_i^{angles} k_{\theta i}(\theta - \theta_{eq})^2 \quad (1.54)$$

$$E_{torsion} = \sum_i^{dihedr} \frac{V_i}{2} [1 + \cos(n\phi - \gamma)] \quad (1.55)$$

$$E_{improper} = \sum_i^{impropers} k_{\gamma i}(\gamma - \gamma_{eq})^2 \quad (1.56)$$

where n in equation 1.55 is the multiplicity of the conformation. And for the non-bonded interactions

$$E_{vanderWaals} = \sum_i^{atoms} \sum_{j>i}^{atoms} \epsilon_{ij} \left[\left(\frac{r_{min_{ij}}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{min_{ij}}}{r_{ij}} \right)^6 \right] w_{ij}$$

with $r_{min} = \sigma 2^{1/6}$

$$= \sum_i^{atoms} \sum_{j>i}^{atoms} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] w_{ij}$$

or simply

$$= \sum_i^{atoms} \sum_{j>i}^{atoms} \left[\left(\frac{A_{ij}}{r_{ij}} \right)^{12} - \left(\frac{B_{ij}}{r_{ij}} \right)^6 \right] w_{ij} \quad (1.57)$$

$$E_{electrostatic} = \sum_i^{atoms} \sum_{j>i}^{atoms} \frac{q_i q_j}{r_{ij}} w_{ij} \quad (1.58)$$

the weighting function w_{ij} is typically set to zero for atoms i and j connected by a bond or angle and w_{ij} ranges from ~ 0.4 to 1 for 1-4 interactions.

The general combination rules for Van der Waals parameters between two atoms i and j are

$$\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \quad ; \quad \sigma_{ij} = \frac{\sigma_i \sigma_j}{2} \quad (1.59)$$

Assuming this separation of the energy the set of parameters k_{ri} , $k_{\theta i}$, V_i , q_i , ϵ_i define a specific force field. If every different atom had its own set of parameters the amount of parameters to optimize would be huge. To simplify this task the concept of atom type is used. The atom type idea is based on the common chemical intuition of transferability. For example, all the sp³ carbon atoms in an alkylic chain will have the same set of parameters because its chemical behavior is approximately the same in different length chains. However every force field defines its own set of atom types, so the parameters for a certain atom in a force field are not usually transferable to another force field.

1.2.2.2 Long range interactions

When size increases the most time-consuming process is the computation of non-bonded interactions shown in equation 1.57 and 1.58. For a pairwise model the number of non-bonded evaluations increase as the square of the number of atoms. However sometimes such computational effort is not necessary, for example, the van der Waals interactions reproduced by a Lennard-Jones potential decays very rapidly (at a distance of 2.5σ there is only 1% of the interaction at σ).

Non-bonded cutoff:

A very popular strategy is to set a cutoff distance. In this method each atom or group of atoms interacts with all the atoms that are closer than the considered cutoff. The rest of the atoms that fall outside this *sphere of interaction* do not contribute to the energy nor the gradient of the considered atom. It has been seen that the simple truncation can be too rough, actually in this case the energy is not conserved. Sometimes a cubic switching function is used at a certain distance in order to obtain a smooth decay. At a certain distance r_{on} the switching function is activated so that the potential is reduced until it vanishes at distance r_{off} . For the switch approximation the following $S(r)$ function is multiplied to the potential energy

$$S_{switch}(r) = \begin{cases} 1, & r \leq r_{on} \\ \frac{(r_{off}^2 - r^2)^2 (r_{off}^2 + 2r^2 - 3r_{on}^2)}{(r_{off}^2 - r_{on}^2)^3}, & r_{on} < r \leq r_{off} \\ 0, & r > r_{off} \end{cases} \quad (1.60)$$

As an alternative, the whole interaction can be shifted and smoothed

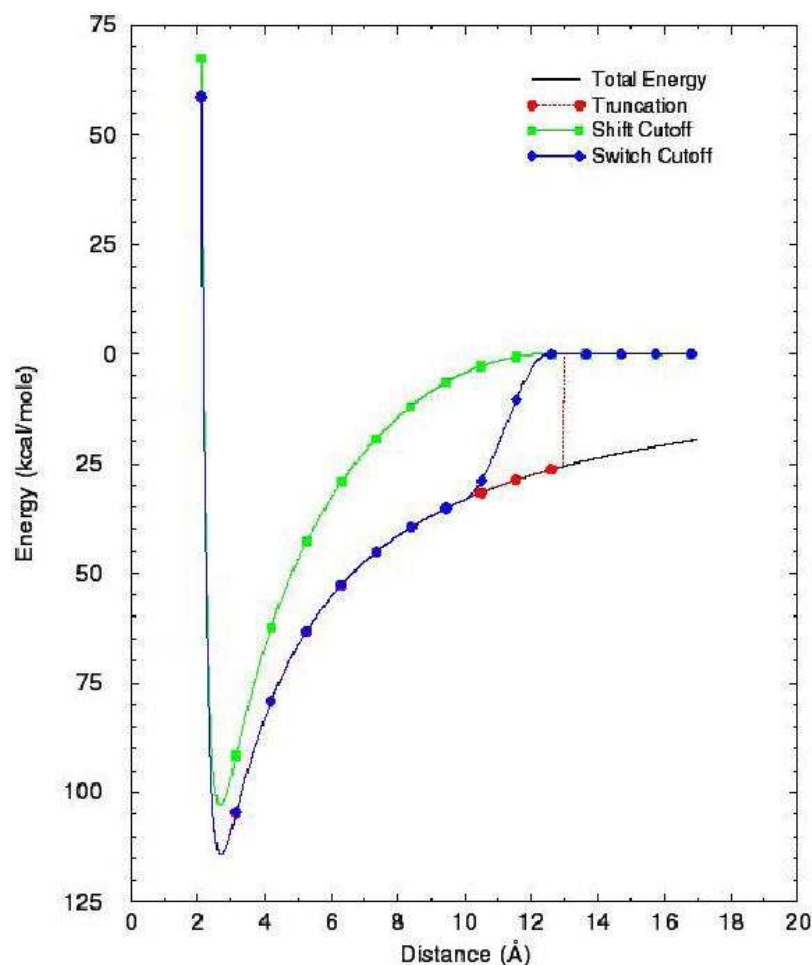


Figure 1.2: Potential energy profile corresponding to the different cutoff approximations

to avoid the abrupt truncation[48]. In this case the $S(r)$ function has the following expression

$$S_{shift}(r) = \begin{cases} (1 - (r/r_{off})^2)^2, & r \leq r_{on} \\ 0, & r > r_{off} \end{cases} \quad (1.61)$$

See figure 1.2 for an approximated shape of the cutoff potentials.

Related to the non-bonded cutoff strategy, an additional problem arises. We must compute all the distances between each atom to decide whether its contribution is calculated or not, that is, every atom has a list of the nearest interacting atoms. If we had to calculate this list at every step of a Molecular Dynamics run it would be too expensive. So this non-bonded

pair-list is built atom-wise or group-wise and it is updated every certain frequency.

We have seen that there are many parameters that we have to consider. A cutoff distance for the electrostatics and for the van der Waals, a cutoff for the smooth decay, a frequency for the non-bonded pair-list update. Despite of all these variables some experience on these kind of simulations has been accumulated, and the different possibilities and the extent of the error has been reported [48, 49].

Minimum image and periodic boundary:

The simulations in condensed phase usually have to reproduce an infinite bulk system such as the solvent that surrounds our solute. If this solvent is reproduced by a single box of solvent the edge of the box will have an artificial surface tension. In this case the Periodic Boundary Conditions is the most suitable strategy. The box is reproduced throughout the space to form an infinite lattice. In the course of a simulation, as a molecule moves in the original box, each periodic image in the neighboring boxes moves in exactly the same way.

Obviously all the possible interactions in an infinite lattice would be infinite. An adequate approximation is the minimum image convention where every atom in the original box interacts with the nearest elements that fall within the given cutoff. These interacting atoms may be in the original box or in the neighboring, however, in order to avoid the interaction of a molecule with itself the cutoff must be smaller than the half of the box.

Long range electrostatic effects:

We may find the truncation schemes described above a brute-force strategy. Mainly for electrostatics there are cases where the interaction does not decay that fast and the cutoff approximation is not adequate. The Ewald lattice-sum, Fast Multipole Method (FMM) and the Reaction Field approximations are the possible solutions [50]. These methods are appropriate for highly charged systems and for the calculations of some delicate magnitudes such as the dielectric constant. Even though, the Ewald summation, which calculates the true electrostatic interactions for an infinitely repeating system, imposes long-range correlations on the system that may be undesirable, and it is computationally more expensive than spherical-cutoff methods.

1.2.2.3 Force fields

The AMBER [51, 52], the CHARMM [53, 54], the OPLS [55] and the GRO-MOS force field [56] are the most widely used set of parameters applied to the simulation of biomolecules.

There are studies that perform benchmarks to compare between some of these force fields [57]. The first thing that one must realize is that their set of parameters are different. This is by no means strange because the parameterization process, the atom types and the reference data is different for all of them. So, for example, the atomic partial charges do differ in some magnitude between two force fields. In this sense, we must not give much chemical interpretation to the particular parameters, but they must be seen rather as a set of variables that must adjust to give a global good fitting. This is why all these force fields tend to give similar global results. It means for example, that after a Molecular Dynamics simulation some properties such as the conformation distribution is similar in any force field. It is also true that some specific failures of a force field in certain systems have been reported [54].

1.2.3 Hybrid methods

Methods for modeling big systems, mainly solute/solvent or in a more general scheme core/environment interactions, can be divided in two big groups. The first one and the earliest is the inclusion of a continuum model characterized by a bulk dielectric constant.

These methods are widely used in *ab initio* techniques[58] as well as coupled to a Molecular Mechanics force field [59], and they provide valuable information. However they do not give specific interactions between the solute and the bulk. In the case of enzymatic reactivity it is this specificity in the interaction which makes the enzyme so efficient. Sometimes the system is partitioned at different shells, a core and a first sphere of the environment are modeled explicitly while the outer environment is represented by a continuum model[60]. Although there are several methods to model solvent effects in biomolecules [61] with implicit models they will not be commented further.

The second group is the family of methods where the environment is modeled explicitly at a lower computational cost. These methods are applied successfully to enzymatic reactions [62], to organometallic catalytic processes

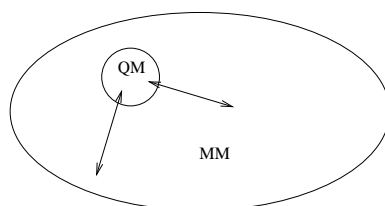


Figure 1.3: QM/MM methods are based on a splitting of the system in a reactive(QM) part and an environment(MM) part

[63, 64] and to solid state materials such as zeolites [65]. Since these discrete environment methods are used all along this thesis, they will be commented in more detail.

1.2.3.1 Polarized QM/MM

The earliest hybrid method modeling the environment explicitly is the so-called combined Quantum Mechanics/Molecular Mechanics (QM/MM). In QM/MM [66, 67, 68] methodology a small reactive part of a chemical system is described by quantum mechanics (QM) whereas the remaining large non-reactive part is described by molecular mechanics (MM). Although the first paper is attributed to Warshel and Levitt [66] some publications make reference of a preceding work by Warshel and Karplus [69] where the electronic structure of a polyene is split in a π system represented by a semiempirical PPP method and the σ by an empirical force field. In any case, the method that is used today to address the reactivity of biochemical systems is still very similar to that used in 1976 by Warshel and Levitt [66]. There is a significant amount of information reviewing the current status of QM/MM methods in the literature [70, 71, 72, 5].

The main advantage of QM/MM methods is its easy implementation in computational codes while giving rather good chemical results. Its main disadvantage, especially in enzymatic systems, is to go beyond qualitative results and thus to obtain quantitative numbers out of QM/MM computations. This problem is mainly due to three factors: i) the need for good *ab initio* description for the QM part whereas the usual size of the QM part mostly only allow for semiempirical calculations; ii) the need for accessing free energy numbers through extensive sampling which is too computationally expensive; iii) the difficult calibration of the interaction between the

QM part and the MM part, especially in biochemical systems as mentioned there after.

The first two factors are related to actual computational bottlenecks and should be overpassed in the near future.

Splitting the System:

In most reactive systems, the number of atoms involved in a chemical reaction is fairly limited (*i.e.*, whose electronic properties are changed during the reaction), the rest of the atoms may have a strong influence on the reaction but this is usually limited to short and long range non-bonded interactions that can be represented through both electrostatic and van der Waals interactions. The main idea of QM/MM methodology is to split the chemical system into two parts: the first part is small and described by quantum mechanics, it is called the *quantum part*. The second part is the rest of the system and is described by molecular mechanics, it is called the *classical part*. The full Hamiltonian can therefore be expressed as:

$$\hat{H} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}} \quad (1.62)$$

where \hat{H}_{QM} is the electronic Hamiltonian describing the quantum atoms in the Born-Oppenheimer approximation. It has the same expression that in equations 1.5 and 1.7 in page 7

In equation 1.62, the Hamiltonian \hat{H}_{MM} describes the classical part. As we defined in the preceding section, a set of atoms interacting in this part can be seen as a set of point charges $\{Q_c\}$ in space interacting through a defined force field. The usual energy expression for every term of the force field has already been commented in section 1.2.2.1 (page 21).

The last term $\hat{H}_{\text{QM/MM}}$ in equation 1.62 stands for the explicit interaction between the quantum and the classical part. It can be represented as the sum of two terms: a van der Waals term describing the non-electrostatic interactions between quantum and classical atoms and an electrostatic term describing the interaction between a classical point charge $\{Q_c\}$ and the electrons and nuclei of the quantum part:

$$\hat{H}_{\text{QM/MM}} = V_{\text{QM/MM}}^{\text{van der Waals}} - \sum_i^{\text{electrons classical}} \sum_C \frac{Q_C}{r_{iC}} + \sum_K^{\text{nuclei classical}} \sum_C \frac{Z_K Q_C}{R_{KC}} \quad (1.63)$$

Note that the second term in equation 1.63 that describes the electrons-

classical charge interaction depends on the coordinates of electrons r_{iC} . Therefore it must be incorporated into the core Hamiltonian of the quantum subsystem and obtained through a SCF process. Indeed this term is the responsible of the polarization on the QM part due to the presence of the MM charges. The other two terms in equation 1.63 depend on particles with fixed positions and can be computed analytically as the rest of MM terms.

QM/MM interactions:

The Hamiltonian $\hat{H}_{\text{QM/MM}}$ should reproduce quantitatively the interaction between the QM and MM parts as if the system was computed fully quantum mechanically [73]. The quantitative reproduction of the QM/MM interactions depends on three points:

- 1) the choice of the set of non-polarizable point charges $\{Q_C\}$ or more generally of the MM force field;
- 2) the choice of the van der Waals parameters to describe $V_{\text{QM/MM}}^{\text{van der Waals}}$;
- 3) the way the classical charges polarize the quantum subsystem.

It is usually a good approximation to take the charge definition from an empirical force field and to incorporate those charges into the core Hamiltonian of the quantum subsystem, being this one the procedure used in this thesis. This gives quite reliable results because the charges in molecular mechanics are defined in order to reproduce electrostatic potential properly [74]. However, it is well-known that between different force fields the charge definition on atoms can change dramatically, even between different generations of the same force field. These differences between different set of charges should have a non-negligible impact on the quality of a QM/MM study.

The choice of van der Waals parameters is another crucial point in the production of a good QM/MM interaction. In theory, a new set of parameters should be defined for each atom in the system (QM + MM) to exclusively compute $V_{\text{QM/MM}}^{\text{van der Waals}}$ [75, 76]. It is well known that the parameters from a forcefield are optimized in a consistent fashion. In consequence the new van der Waals parameters should be compatible with the forcefield used. In practice, one uses the van der Waals parameters from the current force field definition for each atom in the MM part and, when possible, define a new set of van der Waals parameters for the QM atoms [77, 78].

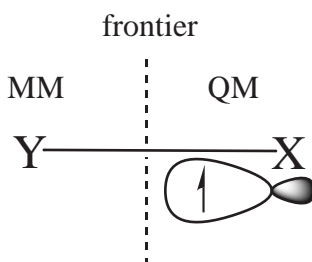
A third point needing to be clarified in the description of QM/MM interactions is the way the classical set of charges $\{Q_C\}$ polarize the electronic wavefunction from the QM subsystem. This is usually done by adding to the core Hamiltonian of the QM part a perturbation describing the interaction between the QM electrons and the classical charges. An element of the core Hamiltonian matrix is expressed as:

$$\begin{aligned} h_{pq}^{\text{core}} &= \langle \phi_p | \hat{h}^{\text{core}} | \phi_q \rangle \\ &= h_{pq}^{\text{core}} - \sum_i^{\text{electrons}} \sum_C^{\text{classical}} \langle \phi_p | \frac{Q_C}{r_{iC}} | \phi_q \rangle \end{aligned} \quad (1.64)$$

However with NDDO semiempirical Hamiltonians it has been shown [75] that electron-classical charges and nuclei-classical charges interactions should not be treated in the same way as electron-nuclei and nuclei-nuclei interactions.

Cutting covalent bonds:

In the study of chemical reactions in solution with QM/MM methods the solute usually belongs to the quantum part and the solvent to the MM part, so that the two zones are perfectly differentiated. However, in enzymatic reactions the reactive part treated with QM usually includes the substrate and some residues that participate actively in the chemical reaction. In consequence, the frontier between the two zones, mainly in lateral chains of the active aminoacidic residues resides on a covalent bond.



In this case the way to link the QM and MM part is not straightforward. A problem occurs at this frontier because the electron of X (see figure above) involved in the covalent bond with Y is not paired with any other electron because in molecular mechanics the electrons of Y are not explicitly represented. This unpaired electron would give a radical character in the QM part that would change all the chemistry.

General Strategy	Name and Reference
	Link atom (blind) [68]
	Dummy groups[79]
	Link atom (partial interaction) [80, 81]
Adding an Atom	Adjusted Connection atoms [82]
	Pseudohalogens[83]
	Pseudobond approach [84]
	Double link atom [85]
	Quantum capping potentials [86]
	AddRemove [87]
Frontier Bond Orbital	Proposed bonding hybrid orbital [66]
	Local SCF (LSCF) [88, 89, 90]
	Generalized Hybrid Orbital (GHO) [91, 92, 93, 94]
	Frozen orbital [95]

Table 1.1: A brief classification of the methods addressed to treat the covalent frontier in QM/MM methods

The way this frontier between the two zones is treated has been the objective of many publications. Actually, many laboratories that developed their tools in QM/MM schemes have their own method to treat the frontier

The different solutions can be divided in two main categories. Those that add an atom or pseudoatom to fill the valencies on the quantum frontier, and those which deal specifically with the frontier bond orbital by trying to compute directly its main characteristics from known parameters. In table 1.1 some of the different approaches published so far with its original references is given.

In this thesis we have used the Link atom approximation in chapter 2 and the Generalized Hybrid Orbital (GHO) in chapter 4. The Link atom method is the simplest implemented and widest used method. It consists in adding a monovalent atom, the so-called *link* or *dummy atom*, along the X—Y bond to fill the valency of the quantum frontier atom X. Usually, this link atom is an hydrogen [68], but some implementations use an halogen like fluorine or chlorine [96]. This method has two major problems still in debate: The interaction of this link atom with the classical part and its additional degree of freedom. See reference [5] for a deeper discussion.

The GHO formalism is an extension of the LSCF method which at the same time is derived from the original work from Warshel and Levitt [66].

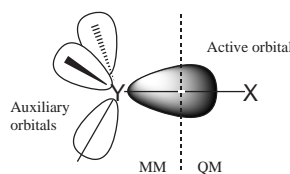


Figure 1.4: Schematic representation of GHO model for the QM/MM frontier

In GHO the classical frontier atom is described by a set of orbitals divided into two sets of auxiliary and active orbitals. The latter set is included in the SCF calculation, while the former generates an effective core potential for the frontier atom.

As we can see, it seems that there are many variations in QM/MM methods. These methods are still under development and there is not yet a unique and standardized procedure to be used as a black box. Actually, one must face some numerical and methodological problems for every system under study.

More accurate potentials:

Most of publications using QM/MM methods have chosen semiempirical Hamiltonians in the QM part. This is mainly because an exploration of the configuration space requires thousands of energy and gradient evaluations that so far only semiempirical methods are fast enough to be permitted. An interesting initiative is the coupling of a semiempirical DFT method (SCC-DFTB) with CHARMM package [97], which is a good compromise between accuracy and computational cost [98]. However, many groups have interfaced *ab initio* packages to MM programs to perform QM/MM at a higher level of accuracy. Programs such as Gaussian98, GAMESS-US, GAMESS-UK, CADPAC, HONDO, DALTON, JAGUAR, deMon, Turbomol are used to carry out Hartree-Fock/MM [99, 100], DFT/MM [101, 102, 90, 103, 104], or even Coupled Cluster/MM [105].

Finally we must not forget the Carr-Parrinello approximations [106] that has been interfaced with GROMOS force field to perform *ab initio* simulations of biological systems [107]. Even though, as we have seen, QM/MM methods are far from being accurate. Therefore, in addition to an insufficient sampling, the frontier, the arbitrary polarization charges are sources of error in such a way that the increase of accuracy in the QM part may be in some cases wasted.

1.2.3.2 IMOMM-ONIOM

The Integrated Molecular Orbital Molecular Mechanics (IMOMM) method [108] is relatively young compared to the Warshel and Levitt work in 1976. It was designed to study organometallic reactions where bulky ligands were important but too expensive to employ the same *ab initio* method that the metal requires. Since these big ligands usually only play a steric role, in the IMOMM hybrid method the polarization of the MM charges to the one-electron Hamiltonian is not included and only van der Waals interactions are considered between the two regions. As any other hybrid method the energy expression can be defined as

$$E_{TOT} = E_{QM} + E_{MM} + E_{QM/MM} \quad (1.65)$$

Where the third term is only the van der Waals between the two zones

$$E_{QM/MM} = V_{QM/MM}^{vanderWaals}$$

Or in a very simplified expression

$$E_{TOT} = E_{real}^{low} - E_{model}^{low} + E_{model}^{high} \quad (1.66)$$

The formulation of equation 1.66 is too simplified because does not indicate explicitly the crossed terms between the model (core) and the environment (real - model) region. In the original publication by Maseras and Morokuma [108] there is a discussion referring to the topic of double counting the interaction at the low and high level and how the frontier is treated.

The "Our-own N-layered Integrated molecular Orbital and Molecular mechanics" (ONIOM) strategy [109] is essentially an extension to the usage of more than two layers of theory and the consideration that the low level can be a quantum mechanical method such as semiempirical (IMOMO). This possibility involves that the low level energy cannot be decomposed in energy terms like molecular mechanics. In this case the expressions in equations 1.65 and 1.66 are not equivalent, and the total energy is an addition-subtraction of energy as displayed in equation 1.66. The problems and solutions of this issue will be found in the original papers [109].

Some authors think that IMOMM-ONIOM are not properly hybrids methods because they do not include explicitly the polarization that the

environment provokes on the core region. My personal opinion is that a hybrid method is any strategy that combines two methods in two regions of a system avoiding the full high-level treatment. So IMOMM-like methods should belong to this hybrid family as a full member.

Besides, this method that was designed to study organometallic chemistry it has been used in enzymatic systems with successful results [110, 111].

1.2.3.3 EVB/MM

In the Empirical Valence Bond method (EVB) [112, 113] from Warshel and co-workers a chemical reaction is described using a valence bond approach. The system wavefunction is represented by a linear combination of the most important ionic and covalent resonance forms and the potential energy is found by solving the related secular equation. The electronic interaction Hamiltonian is built using parameter terms extracted from empirical values and/or *ab initio* surfaces. So, this fact must be emphasized, EVB approach is built as a sum of parametric functions very similar to a MM force field without any explicit treatment of the electrons. This makes the method usually tens of times faster than the QM(SCF)/MM method.

The main advantage of the EVB approach is its ability to give at low computational cost very good quantitative results in comparison with experiment as long as the incorporated empirical terms are carefully chosen [114, 115]. This is mainly accomplished by first calibrating free energy surfaces from reference reactions in solution before incorporating the enzyme effects. However, the main disadvantage is that choosing the valence bond forms (*i.e.*, the most prevalent ionic and covalent forms), one implicitly directs how the chemical reaction should be held. Therefore the EVB method cannot allow unusual reaction pathways that could occur in reactive chemical systems or a chemical reaction not previously defined in the valence bond forms.

Other Alternatives:

Hybrid methods are continually in expansion. Some methods are based on valence bond approximations such as the MOVb from Mo and Gao [116], where all configurations are calculated *ab initio*. On the same valence bond framework SEVB approximation is designed to correct a PES previously calculated [117]. Another option is the Warshel's frozen or constrained DFT (CDFT) [118] which basic idea is to treat the entire protein solvent system

quantum mechanically while freezing (or constraining) the electron density of the environment

The QM/MM framework has had many different implementations. An example of them are the overlapping mechanically embedded [119], the effective charge operators [120], the effective group potential [121] or the effective fragment potential [122]. Other interesting but yet not fruitful strategies are devoted to couple two wavefunctions at two different QM levels polarizing each other in a self consistent fashion [123, 124, 125, 126].

In my opinion these wide variety of QM/MM methods that address to the same problems must converge to a unique and standardized procedure. Hopefully, linear scaling techniques will permit one day to study big systems without artificial partitions and ambiguous embeddings.

1.2.4 Derivatives of the potential energy

The derivative of the potential energy respect to its geometrical (nuclear) coordinates is a crucial issue needed in optimization and molecular dynamics methods.

It is obvious that depending on the energy expression the derivative will be an affordable task or not. There are available methods to calculate the first, the second and sometimes higher order derivatives of the energy.

In general there are two ways to calculate the energy derivatives, numerically and analytically.

- **Numerical derivatives**

Numerical derivatives will be a more expensive task but easier to implement. In this case the derivative is approximated by small finite displacements of the geometry and performing energy evaluations at the displaced structures.

The *ith* gradient component $g_i(\mathbf{q})$ of a molecular system with a geometry \mathbf{q} can be approximated as:

$$g_i(\mathbf{q}) = \frac{\delta E}{\delta q_i} \approx \frac{E(\mathbf{q} + \mathbf{e}_i \Delta q) - E(\mathbf{q})}{\Delta q} \quad (1.67)$$

where *ith* coordinate has been displaced Δq . The \mathbf{e}_i vector has zero in its components but 1 in the *ith* component.

An improvement to the equation 1.67 is displacing the *ith* coordinate

forward and backward.⁴

$$\frac{\delta E}{\delta q_i} \approx \frac{E(\mathbf{q} + \mathbf{e}_i \Delta q) - E(\mathbf{q} - \mathbf{e}_i \Delta q)}{2\Delta q} \quad (1.68)$$

The same argument can be used for second derivatives. In this case we have diagonal H_{ii} and off-diagonal H_{ij} elements of the Hessian matrix:

$$\begin{aligned} H_{ii} &= \frac{\delta^2 E}{\delta q_i^2} \approx \frac{g_i(\mathbf{q} + \mathbf{e}_i \Delta q) - g_i(\mathbf{q})}{\Delta q} \\ H_{ij} &= \frac{\delta^2 E}{\delta q_i \delta q_j} \approx \frac{g_j(\mathbf{q} + \mathbf{e}_i \Delta q) - g_j(\mathbf{q})}{\Delta q} \end{aligned} \quad (1.69)$$

And with the forward-backward improvement.

$$\begin{aligned} H_{ii} &= \frac{\delta^2 E}{\delta q_i^2} \approx \frac{g_i(\mathbf{q} + \mathbf{e}_i \Delta q) - g_i(\mathbf{q} - \mathbf{e}_i \Delta q)}{2\Delta q} \\ H_{ij} &= \frac{\delta^2 E}{\delta q_i \delta q_j} \approx \frac{g_j(\mathbf{q} + \mathbf{e}_i \Delta q) - g_j(\mathbf{q} - \mathbf{e}_i \Delta q)}{2\Delta q} \end{aligned} \quad (1.70)$$

The quality of the derivative depends on the magnitude of Δq and the accuracy of potential energy $E(\mathbf{q})$. (See section 3.1.1.2 for the numerical quantities of these magnitudes and some practical recommendations to decrease the high computational cost of numerical derivatives)

- **Analytical derivatives**

Analytical derivatives have a more complicated expression than the equations above. It can be found in the bibliography first and second derivatives for many of the QM methods (based on the Coupled Perturbed Hartree Fock equations [10, 28] or on Hellman-Feynman theorem). There are also derivatives in Cartesian coordinates for the most common terms in a Molecular Mechanics force field. Recently Cui and Karplus published the analytical second derivatives for a QM/MM potential energy expression [127].

⁴It can be shown that while expression 1.67 is a first order approximation in a Taylor expansion, equation 1.68 is the second order one.

1.3 Theoretical methods used in this thesis: Optimization Methods

In this section we will describe the methods that optimize the potential energy as a function of the nuclear coordinates. That is, those methods that permit moving the nuclear coordinates of a molecular structure to find stationary points, mainly minima and saddle points, on the Potential Energy Surface. The stationary points may explain the chemistry of the molecule. It is expected that a minimum energy structures will be representative of a stable chemical species, as well as the energy and the structure of a saddle point will describe the mechanism and the kinetics of the considered reaction.

It is important to note that all the stationary points found by these methods are local, they are not absolute minima. Perhaps an important exception is the non-derivative algorithms, but their application is not usually addressed to reactivity. The theoretical prediction of a global minimum in a molecular system is still an unsolved problem [128]. A lot of work is devoted to this area, a particular example is the mechanism of protein folding [129].

The optimization methods explained here can be classified in several ways. We classify the methods depending on the need of first or second derivatives which is related with its efficiency and computational cost. It could be classified depending on the kind of stationary point we are looking for, that is, methods to locate minima and methods to locate saddle points. However, any of these classifications would not cover all the possibilities.

1.3.1 Common issues: convergence criteria and step length

Since the PES is not an analytic function the optimization and the search of the local roots must be performed numerically in an iterative fashion. In the sections that follow a wide landscape of techniques will be described. But all of them have two points in common:

convergence criteria: we start from an initial structure but since the mathematically exact minimum will never be reached some convergence criteria must be adopted.

step length: all the techniques have an algorithm to predict a geometry step to get progressively closer to the stationary point. But while most strategies provide a displacement vector that points to a certain

direction, few of them are able to predict with accuracy the length of the vector along this direction. In this case a step length must be determined.

The **convergence criteria** is not unique. The most common criterion is to consider that the search has converged to a point when the norm of the gradient is lower than a threshold (actually this is the only requirement for a stationary point)⁵. Sometimes the maximum component of the gradient vector is required to be under a threshold as well. Other convergence criteria are the change in energy between the previous and the current structure. Otherwise the RMS of the displacement predicted by the algorithm for the next step search.

The **step length** prediction is not unique either. The trust radius approximation considers a fixed step length to the direction of optimization. Line search technique is a very useful strategy where an interpolated one-dimensional polynomial function describes the profile in the displacement direction. In this case the length of the displacement is the length to reach the minimum in the polynomial function.

Line search usually needs additional energy evaluations for such interpolation, but some techniques exist that avoid this waste [12]. Other techniques, such as the Rational Function Optimization method used in this thesis, and outlined in section 1.3.4.2, have an implicit step size determination.

1.3.2 Non derivative methods

Optimization methods that do not require any derivative of the function are not usually applied to stationary points in molecular systems. Although they are generally easy to implement, their convergence properties are rather poor. They may work well in special cases when the function is quite random in character or the variables are essentially uncorrelated. Some examples of these methods are the Simplex, Genetic Algorithms, Neural Networks and Simulated Annealing.

⁵Since the numerical value of the norm depends on the dimension of the vector, the root mean square (RMS) is a more adequate magnitude

$$RMS = \sqrt{\frac{\sum_i^N x^2}{N}}$$

1.3.3 First derivatives methods

The methods that require up to first derivatives of the energy with respect to the nuclear coordinates are mainly the steepest descent and the conjugate gradient family methods [12].

Since the magnitude of the gradient indicates the steepness of the local slope, the energy of the system can be lowered by moving each atom in response to the force acting on it. This is the basis of the steepest descent, where the displacement of the geometry at iteration k $\Delta \mathbf{q}$ may be obtained from the gradient \mathbf{g}_k at the current geometry

$$\Delta \mathbf{q}_k = -\alpha_k \frac{\mathbf{g}_k}{|\mathbf{g}_k|} \quad (1.71)$$

Where α_k is the step length determined by trust radius or line search.

In conjugate gradient method the displacement is computed from the gradient at the current point plus the scaled previous displacement

$$\Delta q_k = \alpha_k \left[-\frac{\mathbf{g}_k}{|\mathbf{g}_k|} + \gamma_k \Delta q_{k-1} \right] \quad (1.72)$$

where the scaling factor γ_k is computed using the previous gradient vectors. There are several expressions for this factor, the easiest form is the Fletcher-Reeves

$$\gamma_k = \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\mathbf{g}_{k-1} \cdot \mathbf{g}_{k-1}} \quad (1.73)$$

1.3.4 Second derivative methods

1.3.4.1 Newton Raphson and quasi-Newton methods

The simplest second derivative method is *Newton-Raphson* (NR). In a system involving N degrees of freedom a quadratic Taylor expansion of the potential energy about the point \mathbf{q}_k is made, where the subscript k stands for the step number along the optimization.

$$E(\mathbf{q}_k + \Delta \mathbf{q}_k) = E(\mathbf{q}_k) + \mathbf{g}_k^T \Delta \mathbf{q}_k + \frac{1}{2} \Delta \mathbf{q}_k^T \mathbf{H}_k \Delta \mathbf{q}_k \quad (1.74)$$

The vector $\Delta \mathbf{q}_k = (\mathbf{q}_{k+1} - \mathbf{q}_k)$ describes the displacement from the reference geometry \mathbf{q}_k to the desired new geometry \mathbf{q}_{k+1} , \mathbf{g}_k is the first derivative vector (gradient) at the point \mathbf{q}_k and \mathbf{H}_k is the second derivative matrix (Hessian) at the same geometry. Under the approximation of a purely quadratic

PES, and imposing the condition of a stationary point $\mathbf{g}_k = 0$ we have the Newton-Raphson equation that predicts the displacement that has to be performed to reach the stationary point.

$$\Delta\mathbf{q}_k = -\mathbf{H}_k^{-1}\mathbf{g}_k \quad (1.75)$$

Because the real PES are not quadratic, in practice an iterative process has to be done to reach the stationary point, and several steps will be required. In this case the Hessian should be calculated at every step which is high computationally demanding. A variation on the Newton-Raphson method is the family of quasi-Newton-Raphson methods (qNR), where an approximated Hessian matrix \mathbf{B}_k (or its inverse) is gradually updated using the gradient and displacement vectors of the previous steps. In section 1.3.4.4 we will summarize these methods to update the Hessian.

1.3.4.2 Rational Function Optimization

While standard Newton-Raphson is based on the optimization on a quadratic model, by replacing this quadratic model by a rational function approximation we obtain the RFO method [130, 131].⁶

$$\Delta E = E(\mathbf{q}_k + \Delta\mathbf{q}_k) - E(\mathbf{q}_k) \cong \frac{\frac{1}{2}(1 \quad \Delta\mathbf{q}_k^T) \begin{pmatrix} 0 & \mathbf{g}_k^T \\ \mathbf{g}_k & \mathbf{B}_k \end{pmatrix} \begin{pmatrix} 1 \\ \Delta\mathbf{q}_k \end{pmatrix}}{(1 \quad \Delta\mathbf{q}_k^T) \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{S}_k \end{pmatrix} \begin{pmatrix} 1 \\ \Delta\mathbf{q}_k \end{pmatrix}} \quad (1.76)$$

The numerator in equation 1.76 is the quadratic model of equation 1.74. The matrix in this numerator is the so called Augmented Hessian (AH). \mathbf{B}_k is the Hessian (analytic or approximated). The \mathbf{S}_k matrix is a symmetric matrix that has to be specified but normally is taken as the unit matrix \mathbf{I} . The solution of RFO equation, that is, the displacement vector $\Delta\mathbf{q}$ that extremalizes ΔE (*i.e.* $\nabla_{\mathbf{q}}(\Delta E) = 0$) is obtained by diagonalization of the Augmented Hessian matrix solving the $(N + 1)$ -dimensional eigenvalue equation 1.77

$$\begin{pmatrix} 0 & \mathbf{g}_k^T \\ \mathbf{g}_k & \mathbf{B}_k \end{pmatrix} \mathbf{v}_\theta^{(k)} = \lambda_\theta^{(k)} \mathbf{v}_\theta^{(k)} \quad \forall \theta = 1, \dots, N + 1 \quad (1.77)$$

⁶In the original publication this method was designed for wavefunction optimization rather than molecular geometries

and then the displacement vector $\Delta \mathbf{q}_k$ for the k_{th} step is evaluated as

$$\Delta \mathbf{q}_k = \frac{1}{v_{1,\theta}^{(k)}} \mathbf{v}'_{\theta}{}^{(k)} \quad (1.78)$$

where

$$(\mathbf{v}'_{\theta}{}^{(k)})^T = (v_{2,\theta}^{(k)}, \dots, v_{N+1,\theta}^{(k)}) \quad (1.79)$$

In equation 1.79, if one is interested in locating a minimum then $\theta = 1$, and for a transition structure $\theta = 2$. As the optimization process converges, $v_{1,\theta}^{(k)}$ tends to 1 and $\lambda_{\theta}^{(k)}$ to 0.

1.3.4.3 Direct Inversion of Iterative Space (DIIS)

The DIIS method [132] was firstly applied to SCF convergence problems and then addressed to geometry optimizations [133]. The method is suitable at the vicinity of the stationary point, and it is based on a linear interpolation/extrapolation of the available structures so as to minimize the length of an error vector.

The DIIS method minimizing the gradient (GDIIS) has been implemented in section 3.1 in the RFO framework, therefore it will be briefly described.

We want to obtain a corrected gradient $\bar{\mathbf{g}}$ as a combination of the previous m gradient vectors $\{\mathbf{g}_i\}$

$$\bar{\mathbf{g}} = \sum_{i=1}^m c_i \cdot \mathbf{g}_i \quad (1.80)$$

The error function to minimize is $\bar{\mathbf{g}} \cdot \bar{\mathbf{g}}$ with the condition $\mathbf{1}^T \mathbf{c} = 1$. Then, building the corresponding Lagrangian in the matrix form

$$L(\mathbf{c}, \lambda) = \frac{1}{2} \mathbf{c}^T \mathbf{G} \mathbf{c} - \lambda (\mathbf{c}^T \mathbf{1} - 1) \quad (1.81)$$

Where $\mathbf{G}_{ij} = (\mathbf{g}_i \cdot \mathbf{g}_j)$ is the matrix containing the scalar products between the gradients of the last m steps. \mathbf{c} is the coefficient vector containing as much components as iterations. The dimension of matrix \mathbf{G} and vector \mathbf{c} is equal to the number of iterations. Derivating 1.81 with respect to λ and to

\mathbf{c} and imposing the stationary condition

$$\nabla_{\mathbf{c}}L = \mathbf{G}\mathbf{c} - \mathbf{1}\lambda = \mathbf{0} \quad (1.82)$$

$$\frac{\delta L}{\delta \lambda} = -(\mathbf{1}^T \mathbf{c} - 1) = 0 \quad (1.83)$$

Joining both derivative conditions in matrix form

$$\begin{pmatrix} \mathbf{G} & -\mathbf{1} \\ -\mathbf{1}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -1 \end{pmatrix} \quad (1.84)$$

and then

$$\begin{pmatrix} \mathbf{c} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{G} & -\mathbf{1} \\ -\mathbf{1}^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ -1 \end{pmatrix} \quad (1.85)$$

To obtain the coefficients \mathbf{c} only requires the inversion of a matrix as large as the number of iterations + 1. From the previous system of equations we can have the improved gradient and then build up the improved Augmented Hessian

$$\begin{pmatrix} 0 & \bar{\mathbf{g}}^T \\ \bar{\mathbf{g}} & \mathbf{B} \end{pmatrix} \quad \text{instead of} \quad \begin{pmatrix} 0 & \mathbf{g}^T \\ \mathbf{g} & \mathbf{B} \end{pmatrix} \quad (1.86)$$

1.3.4.4 Update expressions and initial Hessians

The Hessian is not usually calculated explicitly but is obtained by updating an initial estimate. The calculation of an expensive initial Hessian can be also avoided by suitable estimates obtained from empirical, molecular mechanics or semiempirical methods[134].

In any case the initial Hessian must have the same number of negative eigenvalues than the stationary point we want to find.

For quasi-Newton-Raphson and RFO methods, at every step, the approximated Hessian matrix is updated from the information of previous steps.

$$\mathbf{B}_{k+1} = \mathbf{B}_0 + \sum_{i=0}^k [\mathbf{j}_i \mathbf{u}_i^T + \mathbf{u}_i \mathbf{j}_i^T - (\mathbf{j}_i^T \Delta \mathbf{q}_i) \mathbf{u}_i \mathbf{u}_i^T] \quad k = 0, 1, \dots \quad (1.87)$$

Where $\mathbf{j}_i = \mathbf{D}_i - \mathbf{A}_i$, $\mathbf{D}_i = \mathbf{g}_{i+1} - \mathbf{g}_i$, $\mathbf{A}_i = \mathbf{B}_i \Delta \mathbf{q}_i$, $\mathbf{u}_i = \mathbf{M}_i \Delta \mathbf{q}_i / (\Delta \mathbf{q}_i^T \mathbf{M}_i \Delta \mathbf{q}_i)$. Different election of the \mathbf{M}_i matrix leads to different update Hessian matrix formula [135]. In particular, for the Broyden-Fletcher-Goldfarb-Shanno

(BFGS) update $\mathbf{M}_i = a_i \mathbf{B}_{i+1} + b_i \mathbf{B}_i$ for some selected positive definite scalars a_i and b_i . For the Powell update case the matrix \mathbf{M}_i is equal to unit matrix \mathbf{I} .

These expressions give an approximation to the proper Hessian matrix. However, in NR equation (equation 1.75) we need the inverse rather than the Hessian. In this case we can use the update expression for its inverse. But we must never forget that while working with the Hessian matrix we can know and control the number of positive eigenvalues and then, the order of stationary point reached. When the inverse of the Hessian is used this information is lost. This is why we will prefer RFO technique rather than pure NR because an explicit control of eigenvalues is performed without being necessary a full diagonalization of a Hessian or Augmented Hessian matrix.

The BFGS update is not acceptable to locate first-order saddle points because this update is positive definite. While several update formulae are used to locate minimum energy structures, the Powell formula is a more suitable update to find transition state structure [136].

1.3.5 Reaction path

The reaction path will connect the several chemical species that evolve during the reaction process. The word "connected" needs to be clarified. The most adequate definition of such connection is the trajectories that an ensemble of molecular entities travel through the PES at a given temperature until the reaction is completed. This is the most general definition of reaction paths and must be studied by statistical mechanics [137]. However, it can be studied in a more simplified definition by optimization methods [138, 139].

Optimization methods provide very valuable information and may give a chemical insight where statistical methods are far from being applicable. Therefore, restricting our study to one molecule, one stationary point and one valley of reaction without the effect of temperature, we can associate the reaction path as the path by which stationary points are connected through the lowest energy valley in the PES. This path is called the minimum energy path (MEP).

In some cases, such as small and/or gas phase systems, the MEP is a very good approximation to the reaction paths. Strictly, this MEP would be, without considering the zero point energy correction, the reaction path

that would follow an hypothetical system at zero Kelvin.

In this section some methods to find the MEP are commented. In addition, some of these methods are useful to find a geometry guess to start a saddle point *free-search*.

1.3.5.1 Coordinate scan

The most intuitive strategy to draw a MEP is to identify an internal coordinate (bond distance, angle, dihedral ...) or any combination of them as a reaction coordinate, and then to perform several restrained energy minimizations at different values of this coordinate kept frozen. At every restrained minimization this coordinate is modified, going from reactant to product, in order to have a discontinuous representation of the supposed reaction path. The way this coordinate is fixed is usually applying a harmonic potential with a force constant big enough to keep unmoved the atoms involved in internal coordinate.

$$V_{total} = V_{system} + k(x_{a,fix} - x_a)^2 \quad (1.88)$$

$x_{a,fix}$ is the intended fixed value at each restrained minimization, x_a is the current value of the reaction coordinate along the simulation. The above expression can be generalized to a combination of variables [140].

$$V_{total} = V_{system} + k((x_{a,fix} - x_a) + (x_{b,fix} - x_b) + \dots)^2 \quad (1.89)$$

This method is also useful when we want to discriminate between a concerted or stepwise mechanism, where, in this case, a and b are those coordinates governing the two reaction steps. This last option is the best since only one degree of freedom is kept unmoved while contemplating both distances variation.

The point of maximum potential energy along the reaction coordinate can be taken as a first approach to the transition state structure. However, it is not always so easy or intuitive to identify an internal coordinate as the reaction coordinate. If the coordinate is not appropriate we cannot be sure of visiting the saddle point region. In any case, even when a coordinate seems to be intuitive, it should be always checked if the Hessian matrix has an unique negative eigenvalue that will be associated to the transition eigenvector.

When this strategy is applied to condensed phase systems many parallel reaction paths may exist. If the minimization process is brusque during the scanning, the system may fall down into a parallel valley of the reaction lower in energy. This would provoke a discontinuity in the energy profile meaning that reactants and products are not actually connected. To avoid this hysteresis Some authors calculate the reaction path from reactants to products and back to reactants many times [99] until the obtained energy profile is unique.

1.3.5.2 Intrinsic reaction coordinate

It is important to note that while the stationary points are invariant to any coordinate transformation, the MEP is not. It means that it will depend, to some extent, on the coordinate system. The MEP located in mass-weighted coordinates is unique and it is called the Intrinsic Reaction Coordinate (IRC) [141]⁷. There are several methods that follow the IRC from the saddle point down to the reactants and down to the products [139]. All of them try to solve the steepest descent reaction path as defined in the following differential equation

$$\frac{d\mathbf{x}(s)}{ds} = -\frac{\mathbf{g}(s)}{|\mathbf{g}(s)|} \quad (1.90)$$

1.3.5.3 Interpolation between reactants and products

An alternative strategy to the free search methods that start from the saddle point and determine the path downhill is interpolating between reactants and products. In this case, the search takes as reference points the reactants and in some cases the products. There are methods that go gradually uphill looking for the saddle point and others that progressively refine the interpolation between reactants and products.

The coordinate scan explained above would be a particular case of these methods. Examples of other more sophisticated strategies are the synchronous transit methods (LST and QST), following gradient extremals, Reduced Gradient Following and walking up valleys (see a recent review by Schlegel and the references therein [138]).

⁷Some authors consider that the MEP is only when the path is computed in mass-weighted coordinates, that is, MEP and IRC are synonyms

1.3.5.4 Chain methods

The chain methods are designed to optimize the entire path between reactants and products. In a paper by Henkelman *et al.* the different chain methods are reviewed [142]. The most popular techniques are based on the progressive optimization of a chain of points that connect reactants and products. Since they are designed to be applied in big systems the usage of a big Hessian matrix is always avoided. This makes that the efficiency is relatively low. In some cases they have been applied to biomolecular systems such as the Nudged Elastic Band (NEB) [143, 144, 145], the Conjugate Peak Refinement [146, 147] and the Replica Path [148].

1.3.6 Cartesian, internal and redundant coordinates

It is well known that in any optimization problem, the choice of coordinates can have an important influence on the efficiency of the optimization. Cartesian coordinates are easy to define and are used for calculating the molecular energy and its derivatives. However, the potential energy surface has very strong coupling between coordinates when represented in Cartesians.

On the other hand internal coordinates (bond lengths, valence angles and torsions) are more appropriate coordinates to describe the behavior of molecules. Because they express the natural connectivity of chemical structures, there is much less coupling between these internal coordinates.

There are different types of redundant internal coordinate systems (primitive, natural, delocalized) [149, 150], and all work better than Cartesians or non-redundant internals (*e.g.* Z-matrix coordinates).

The transformation of Cartesian coordinates and displacements to internals is straightforward, but the transformation of the gradients requires a generalized inverse of the transformation Wilson B matrix [23]

$$\mathbf{B} = \frac{\delta \mathbf{q}}{\delta \mathbf{x}} \quad (1.91)$$

$$\mathbf{q} = \mathbf{B} \cdot \mathbf{x}, \quad \mathbf{g}_q = \mathbf{B}^{-1} \cdot \mathbf{g}_x, \quad \mathbf{H}_q = \mathbf{B}^{-T} (\mathbf{H}_x - \delta \mathbf{B} / \delta \mathbf{x} \cdot \mathbf{g}_q) \mathbf{B}^{-1} \quad (1.92)$$

where \mathbf{x} are the Cartesian coordinates, \mathbf{q} are the (redundant) internal coordinates. See reference [149] for the meaning of \mathbf{B}^{-T} matrix and the

derivatives of \mathbf{B} . Here, we will only mention that because of the redundancy and the curvilinear nature of the internal coordinates, the transformation back to Cartesians involves an iterative process as well as a generalized inverse [149].

The calculation of a generalized inverse scales cubically. For small molecules, this is not a problem. However, for larger systems these transformations can become a bottleneck.

It has been suggested that a good initial Hessian in a second derivative optimization can decrease the coupling in Cartesian coordinates and obtain a convergence as fast as in redundant coordinates [151, 2].

1.3.7 Second order methods for large systems

The first derivatives methods such as steepest descent and conjugate gradients are widely used to minimize molecular structures of thousands of atoms. However most of times the potential energy that describes these big systems is cheap. It means that the computational bottleneck is the storage of big matrices rather than the potential energy evaluation and its first derivative. Conversely, when potential energy is not that cheap we need the efficiency of a second derivatives method in order to save as many energy evaluations as possible. Moreover, when we are looking for transition state structures the information of the PES curvature provided by the Hessian matrix is essential for the success of the search.

Although some chain methods described in the last section are specially designed to be applied on large systems they are not able to locate stationary points by a *free search*. In addition, none of the chain methods uses either approximate or exact second derivatives⁸.

The four principal strategies described in this section are optimizers that, in some extent, uses the information of the second derivatives to increase the efficiency or to keep track of the PES curvature. Avoiding at the same time the computational cost that the classical Newton-like methods usually require.

When trying to apply Newton-like methods to systems bigger than very few hundreds of atoms with the nowadays computers some computational problems arise. Three main bottle-necks exist

⁸ The conjugate peak refinement [146] represents the Hessian matrix in its equations by the identity matrix

- Computation of an accurate initial Hessian scales with $O(N^2)$ when calculated numerically
- Diagonalization process that scales with $O(N^3)$
- Hessian storage scales with $O(N^2)$
- Internal-Cartesian coordinates interconversion scales with $O(N^3)$

The scaling factor must not be directly related to the computational cost. These four problems appear at different size and situations (see section 3.4 for a particular benchmark). For example depending on the level of theory of the energy, the second derivatives calculation can be the main computational demanding task. A full diagonalization will be problematic when both the system has hundreds of dimensions and the process must be performed at every step of a long optimization. The storage will not be a problem until the system has about tens of thousand atoms⁹ Finally, the internal-Cartesian coordinates interconversion scales as the diagonalization but in this thesis only Cartesian coordinates will be used.

1.3.7.1 Limited memory: L-BFGS

The limited memory strategy avoids two of the bottlenecks specified above in the Newton-Raphson algorithm when applied to big systems. The storage and the inversion of big matrices. Here although a quasi-Newton-Raphson algorithm is used, the inverse of the Hessian matrix is never built up, but directly the product of the inverse of the Hessian by the gradient. Then, no Hessian diagonalization is required.

$$\mathbf{B}_{k+1}\mathbf{v} = \mathbf{B}_0\mathbf{v} + \sum_{i=k-m}^k [\mathbf{j}_i\mathbf{u}_i^T\mathbf{v} + \mathbf{u}_i(\mathbf{j}_i^T\mathbf{v} - (\mathbf{j}_i^T\Delta\mathbf{q}_i)\mathbf{u}_i^T\mathbf{v})] \quad (1.93)$$

For instance, in equation 1.93, if the vector \mathbf{v} is the gradient of the current geometry and \mathbf{B}_0 the inverse of the initial Hessian we obtain directly by a vector-matrix product the displacement of Newton-Raphson equation for the $k + 1$ step (equation 1.75). A diagonal matrix must be used for \mathbf{B}_0

⁹ A computer of 512 MB of free memory can store $1024*1024*512=536\ 870\ 912$ bytes $\sim 67\ 108\ 864$ double precision real numbers which in triangular form $(n(n+1)/2)$ can build a matrix of dimension 11584.

unless we want to face an expensive full inversion process for a full Hessian. Unit matrix is usually a good choice.

What makes this method powerful is that in order to update this matrix product only information of last m steps is used. In this way only the geometry and gradient of the last m steps have to be stored. When a BFGS update formula is used this procedure is called L-BFGS. This method was developed by Jorge Nocedal [152, 153]. The source code can be obtained free of charge from the web. Recently, Nocedal and co-workers [154] have combined the LBFGS with a Hessian free Newton method that improves the efficiency in the minimization process.

This useful method for minima, as it will be explained later, cannot be applied to transition state search. Note that there is no control on the number of positive eigenvalues and, as we already said, the BFGS formula is not suitable for TS.

Other strategies based on the idea of limited memory can be adopted when another update scheme is more adequate. This is, for example, the limited memory Powell [155, 2] that will be explained and tested in section 3.4.

1.3.7.2 ABNR

The adopted basis Newton-Raphson method (ABNR) was developed originally by M. Karplus and D.J. States. It has been used widely for the optimization of biomolecules since its original implementation in CHARMM package [53]. Even though, there has never been a paper describing the method and the corresponding implementation. Only in a very recent paper by B. R. Brooks and co-workers [144] we can find some equations that picture the most important aspects of the method.

In the ABNR minimization the Newton-Raphson scheme is only applied in a small subspace of the molecule. So the whole displacement of the geometry is a combination of a steepest descent(SD) step plus a small contribution of Newton-Raphson(NR).

$$\Delta\mathbf{q}_k = \Delta\mathbf{q}_k^{SD} + \Delta\mathbf{q}_k^{NR} \quad (1.94)$$

At the beginning of the minimization only the steepest descent component is employed ($\Delta\mathbf{q}_k^{NR} = 0$). After several SD steps, the last m geometry displacements can be used as a basis $\{\Delta\mathbf{q}_{lk}\}_m$ of dimension m to

obtain the NR step. So at step k the last m geometries are used \mathbf{q}_l ; $l = k - 1, k - 2, \dots, k - m$

$$\Delta \mathbf{q}_k^{NR} = \sum_{l=1}^m \Delta \mathbf{q}_{lk} c_{lk} \quad \text{where} \quad \Delta \mathbf{q}_{lk} = \mathbf{q}_{k-l} - \mathbf{q}_k \quad (1.95)$$

This equation can be approximated in a Taylor expansion with respect to \mathbf{q}_k and then the Newton-Raphson equation becomes

$$\sum_{l'=1}^m \sum_{l=1}^m \Delta \mathbf{q}_{kl'} \cdot [\mathbf{g}(\mathbf{q}_{k-l}) - \mathbf{g}(\mathbf{q}_k)] c_{lk} = - \sum_{l'=1}^m \Delta \mathbf{q}_{kl'} \cdot \mathbf{g}(\mathbf{q}_k) \quad (1.96)$$

Equation 1.96 is a set of m equations that can be solved diagonalizing the small $m \times m$ matrix. With the coefficients $\{c_{lk}\}$ we can obtain the geometry displacement component $\Delta \mathbf{q}_k^{NR}$. But what incorporates the Hessian character in ABNR method is the particular steepest descent step that is computed in the following way

$$\Delta \mathbf{q}_k^{SD} = \frac{\mathbf{F}_k}{|\mathbf{F}_k|}, \quad \mathbf{F}_k = -\mathbf{g}(\mathbf{q}_k + \Delta \mathbf{q}_k^{NR}) \quad (1.97)$$

Where this gradient vector at point $\mathbf{q}_k + \Delta \mathbf{q}_k^{NR}$ is estimated as

$$\mathbf{g}(\mathbf{q}_j + \Delta \mathbf{q}_k^{NR}) \approx \mathbf{g}(\mathbf{q}_k) + \sum_{l=1}^m [\mathbf{g}(\mathbf{q}_{k-l}) - \mathbf{g}(\mathbf{q}_k)] c_{lk} \quad (1.98)$$

This additional dimension added to steepest descents steps stands for an estimation update of the local Hessian. If any eigenvalue of equation 1.96 is negative the NR step is deactivated and only SD steps are performed. As in L-BFGS a subspace dimension of $m = 5$ is usually enough for a good performance.

1.3.7.3 Truncated Newton

Newton Raphson equation (Eq. 1.75) can be rewritten as:

$$\mathbf{H}_k \Delta \mathbf{q}_k = -\mathbf{g}_k \quad (1.99)$$

The truncated Newton-Raphson method [156, 157, 14] applied to the minimization of big molecular systems has been developed by the Tamar Schlick

group. It finds iteratively an approximation to the implicit solution $\Delta\mathbf{q}_k$ in equation 1.99 using the combination of a Truncated Newton-Raphson optimizer with a conjugate gradient technique.

The TN package is the software that contains most of implementations described in the publications and can be obtained free of charge from the web servers of Schlick group.

1.3.7.4 Coupled or micro-iterative method

The methods presented so far for big systems do not permit the explicit control on the number of positive eigenvalues of the Hessian. This feature can be avoided for minima but it is crucial when locating first-order saddle points. In addition, even in the case of minima, when energy and gradient evaluations are expensive an efficient method is required. In this case we need a Hessian matrix provided that its size permits an easy manipulation. The first option would be to freeze all the atoms that may not be important for our saddle point location, and then build up a Hessian for a core region not bigger than the size computationally affordable. The Newton-like search is performed only on this small core zone, while the environment is neither permitted to relax nor contribute to the *reaction* vector. This intuitive approximation has been applied in enzymatic reactions (see reference [158] for an example).

Since systems such as enzymes are rather flexible, a movement of an atom, group or a side chain provokes in turn a coupled movement of the interacting atoms. This means that the above approximation of a frozen environment may not be adequate as a definitive strategy. The logical solution would be to permit the environment atoms to relax during the search in the core. This is the so-called micro-iterative method.

The micro-iterative method is a strategy first used by Maseras and Morokuma [108] in the IMOMM scheme applied to organometallic systems. Few years later, the GRACE package [159, 160] permitted the location of TS structures and IRC pathways in enzymatic systems of thousands of atoms. Several groups have applied micro-iterative method to enzymatic systems [81, 161, 84, 95, 120, 162] and zeolites [65]. This method splits the system in two parts, a core zone where an accurate second order search is done, and an environment that is kept minimized with a cheap first order method. Both processes are carried out until consistency. This separation makes that the

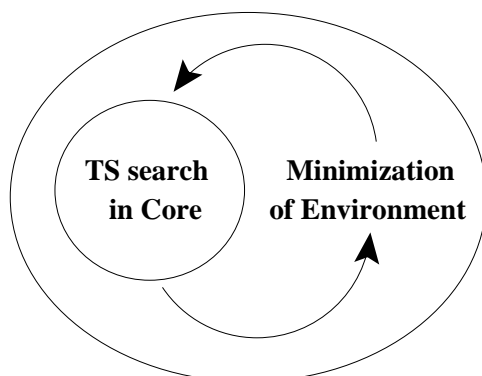


Figure 1.5: Micro-iterative scheme. A quadratic search is performed in the core while a linear minimization keeps the environment relaxed

sum of the expenses of the two processes is considerably lower than a single global search.

The computational requirements of a second order search are only needed to optimize the small core zone, while the big part of the system is moved according to a cheaper method.

This is maybe the only strategy that can locate real saddle points in big systems with the direct usage of second derivatives information. Obviously the control and information given by the eigenpair will be only referred to the core zone where the main relevant movements of the reaction are expected.

Since this method is a central part of this thesis, it will be explained in more detail in the third chapter where we develop, implement and test some crucial aspects of the micro-iterative strategy.

1.3.7.5 Internal vs Cartesian coordinates

As we commented, the usage of internal coordinates implies some matrix transformations between internal and Cartesian coordinates that become a computational bottleneck in big systems. Some work has been devoted to overcome this computational drawback through linear scaling techniques [163, 164, 161, 165]. However we must take into account that even in internal coordinates, the optimization of a big system will require a large amount of steps to converge. Moreover, in flat energy surfaces such as biopolymers, the improvement on the usage of internal coordinates may be concealed under the long iterative process and the trade-off may not be favorable.

1.4 Theoretical methods used in this thesis: Molecular Dynamics Methods

Classical molecular dynamics (MD) can be used to propagate in time the nuclear coordinates of molecular system using the classical equations of motion¹⁰

$$-\frac{dV}{dq} = m\frac{d^2q}{dt^2} \quad (1.100)$$

This is an excellent approximation for many particles, however, when we consider light particles such as hydrogen the quantum nature of the molecules must be considered [168] and we must recover the equations of nuclear quantum motions (section 1.2.1.3).

Equation 1.100 must be solved numerically propagating a trajectory at small time-steps. While a typical time-step is about one femtosecond (10^{-15}) most of chemical interesting events take place at time scales several orders of magnitude higher (micro or millisecond). Therefore the MD equations should be propagated until 10^9 - 10^{12} steps to observe a reactive event (rare event). Despite of the recent acceleration techniques [169] this task is yet too expensive to be performed in the nowadays computers. This gap of time scales makes that so far Molecular Dynamics are rarely used to obtain a real picture of the thermally activated chemical processes. Therefore, in this thesis MD will be employed as a technique that may be used to compute equilibrium as well as kinetic properties of a many-body system. Most of these properties can also be computed with Monte Carlo (MC) techniques. Although MC covers a vast area of techniques they will not be commented here, and it may be found in the literature [13, 15, 50].

In this section, the main topics of molecular dynamics techniques are presented. However it must be taken in consideration that MD is a huge field of research and we will only mention those methods that we will apply to model our enzymatic system. In this sense, a volume of the *Account of Chemical Research* was dedicated to review the state of the art in Molecular Dynamics simulations of biomolecules [170].

¹⁰Other formulations of classical mechanics alternative to the Newton's are also used [166, 167]

1.4.1 Basic equations and algorithms

The position of a set of particles \mathbf{q} after a time step displacement Δt can be obtained by a Taylor expansion

$$\mathbf{q}_{i+1} = \mathbf{q}_i + \frac{d\mathbf{q}}{dt}\Delta t + \frac{1}{2}\frac{d^2\mathbf{q}}{dt^2}\Delta t^2 + \frac{1}{6}\frac{d^3\mathbf{q}}{dt^3}\Delta t^3 \quad (1.101)$$

where in the second term in the right-hand of equation 1.101 appears the velocity $\mathbf{v} = \frac{d\mathbf{q}}{dt}$ and in the third term the acceleration $\mathbf{a} = \frac{d^2\mathbf{q}}{dt^2} = -\frac{1}{m}\frac{dV}{d\mathbf{q}}$

Developing the same expansion at a previous step and added to equation 1.101, with both equations truncated at third order, give place to the Verlet algorithm which is the basis of the current molecular dynamics simulations techniques

$$\mathbf{q}_{i+1} = (2\mathbf{q}_i - \mathbf{q}_{i-1}) + \mathbf{a}_i(\Delta t)^2 \quad (1.102)$$

The acceleration is taken from the derivatives of the potential energy and at the initialization point \mathbf{q}_0 the previous positions can be estimated as $\mathbf{q}_{i-1} = \mathbf{q}_0 - \mathbf{v}_0\Delta t$. The Verlet algorithm offers several numerical problems. In addition, the fact that the velocity does not appear explicitly is a problem when generating ensembles at constant temperature.

There are several improvements to the Verlet algorithm. The Leap-Frog algorithm includes the velocity in its equations

$$\mathbf{q}_{i+1} = \mathbf{q}_i + \mathbf{v}_{i+\frac{1}{2}}\Delta t \quad (1.103)$$

where the velocity is

$$\mathbf{v}_{i+\frac{1}{2}} = \mathbf{v}_{i-\frac{1}{2}} + \mathbf{a}_i\Delta t \quad (1.104)$$

velocity Verlet or the higher order predictor-corrector are other alternative integration methods. Discussion about their adequacy, the numerical stability, the energy conservation and the time-reversible character must be found in the extensive bibliography [15, 50, 13].

1.4.2 Thermostats and barostats

Using the equations of motion specified above the total energy is a constant of motion. In this case the time averages obtained in this MD simulation are equivalent to ensemble averages in microcanonical ensemble (NVE).

In order to run MD simulation at other non-NVE statistical ensembles

we must introduce a thermostat and/or barostat.

Constant Temperature:

The temperature of a particle system is related to the time average of the velocity of the particles

$$\left\langle \sum_i^n \frac{1}{2} m_i v_i^2 \right\rangle = \frac{3}{2} n k_B T \quad (1.105)$$

The initial velocities can be given from a Maxwell-Boltzmann distribution at the desired temperature. And the most intuitive strategy to keep constant the temperature would be to multiply at every step the velocity of the particles by a scaling factor $\lambda = \sqrt{T_{required}/T_{current}}$.

Other less rough approaches exist, Andersen introduced the stochastic collisions method[171] and Berendsen and co-workers [172] introduced a coupling parameter τ to an external bath

$$\lambda^2 = 1 + \frac{\Delta t}{\tau} \left(\frac{T_{required}}{T_{current}} - 1 \right) \quad (1.106)$$

However, the most popular strategy is the extended Lagrangian method that contains additional, artificial coordinates and velocities. It was introduced by Andersen, Nosé[173] and reformulated by Hoover[174]. The reservoir is represented by an additional degree of freedom s with a fictitious mass Q and with potential energy $(n+1)k_B T_{req} \ln(s)$. The parameter Q determines the coupling and the energy flow between the reservoir and the real system and therefore the energy oscillations. Some non-ergodicity problems have been reported with Nosé-Hoover method in this case a series of Nosé-Hoover chains are introduced [175, 176].

Constant Pressure:

A macroscopic system maintains constant pressure changing its volume. The calculation of the pressure from a particle system is not as direct as the temperature (equation 1.105). The Virial theorem can be used although it can give problems for Periodic Boundary systems.

$$P = \frac{2}{3V} \left\langle \sum_i^n \frac{1}{2} m_i v_i^2 \right\rangle - \left\langle \frac{dV}{d\mathbf{q}} \mathbf{q} \right\rangle \quad (1.107)$$

The amount of volume fluctuation is related to the isothermal compressibil-

ity

$$\kappa = \frac{1}{k_B T} \frac{\langle V^2 \rangle - \langle V \rangle^2}{\langle V \rangle} \quad (1.108)$$

Many of the methods used for the control of pressure are analogous to those used for the constant temperature. The volume may be changed scaling the positions $\mathbf{q}'_i = \lambda^{1/3} \mathbf{q}_i$ of the particles through a coupling parameter τ

$$\lambda = 1 - \kappa \frac{\Delta t}{3\tau} (P_{required} - P_{current}) \quad (1.109)$$

Otherwise, the Nosé-Hoover scheme can be applied using an external piston as additional degree of freedom [174, 177].

1.4.3 Constraints

The inclusion of constraints to the fastest movements that are not of great interest in themselves (*e.g.* bond vibration) permit increasing the time step of the MD simulation. As a consequence of a larger time step the simulation becomes computationally cheaper.

The most commonly used method for applying holonomic constraints is the SHAKE procedure [178]. The procedure is based on the determination of Lagrange multipliers (l_k) imposed as a constriction to the equations of motion. For n constraints with d_k as the corresponding constrain distance we have

$$m_i \frac{d^2 q_i(t)}{dt^2} = -\frac{\delta}{\delta q_i} [V(\mathbf{q}) + \sum_k^n l_k(t)(q_k^2 - d_k^2)] \quad (1.110)$$

Here we will only mention that for solving the above equations in SHAKE procedure the Lagrange multipliers are determined iteratively and therefore they depend on a threshold. In small systems the procedure can be carried out by a matrix inversion [179].

While SHAKE works in Cartesian coordinates Tobias and Brooks generalized this to an arbitrary internal coordinate [180].

In addition, constraints in Molecular Dynamics simulations can be applied to other interesting areas. It may be used, for example in Potential of Mean Force calculations or in Ligand Binding techniques [181, 182].

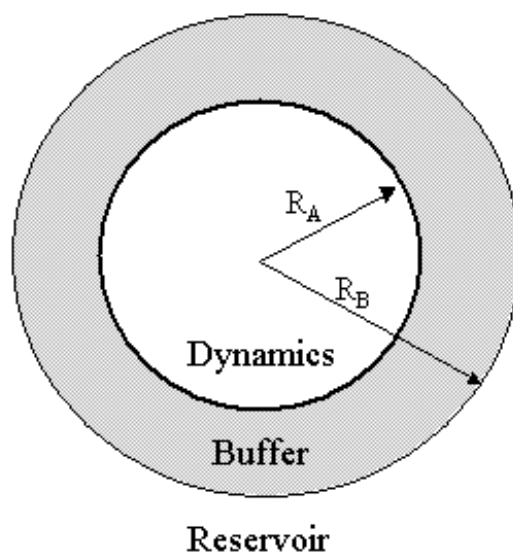


Figure 1.6: General scheme for the partition of the system in Stochastic Boundary MD method

1.4.4 Langevin Dynamics and Stochastic Dynamics

There are several formulations for Langevin and Stochastic equations of motion that may be used to perform Non-equilibrium, Brownian and Hydrodynamics simulations [15].

However, here we will only cover a very small part of this area that is particularly useful for simulating the localized dynamics of a portion of a protein. This method is very suitable in many enzymatic reactions where the chemical process is expected to take place in a localized region of the enzyme, in this case the expensive Periodic Boundary Conditions may not be necessary and therefore it can be avoided. This is the Stochastic Boundary Molecular Dynamics method (SBMD) [183, 184, 185, 186].

Here we will explain in some detail the SBMD method as it is used in simulations of enzymatic systems [187, 188, 189] as well as in chapter 4. The SBMD setup has three main characteristics, the partition of the system in three zones, the forces applied to certain zones of the system and the particular equations of motion.

Partitioning the system:

The system is divided in three main regions: dynamics, buffer and reservoir (see figure 1.6).

Dynamics or reaction region: it consists of atoms within a sphere of radius R_A centered in the active site. The selection is done by residue, in other words, the entire residue is considered to be in this region if any atom of the residue falls in the criterion. This is to avoid disjunct *freely floating* atoms in the system.

Buffer region: which contains the rest of residues surrounding the dynamics region up to a certain distance R_B ($R_B > R_A$).

Reservoir region: is the rest of the system, namely, any residue that has all its atoms beyond R_B distance.

The atom-labels partitioning the protein are kept fixed all along the simulation while the solvent molecules are allowed to diffuse between the regions (even between the subpartitions in the buffer region).

Evaluating boundary and stochastic forces:

The atoms that belong to the reservoir region are eliminated¹¹. Therefore, atoms in buffer region must feel an interaction that reproduces the forces from inexistent reservoir region. These forces may be divided in **average** (boundary) forces and **stochastic** forces.

While reactions in solution the solvent molecules undergo translational diffusion, in proteins the atomic motions have a more localized nature. In this sense the atoms in buffer regions feel a boundary force represented by linear and isotropic restoring forces derived from the atomic mean-square fluctuations. The effective force constant is

$$m_i \Omega_i^2 = \frac{3k_B T}{\langle \Delta \mathbf{r}_i^2 \rangle} \quad (1.111)$$

Atoms near the frontier buffer/reservoir region should not feel the same boundary forces than those in buffer/dynamics frontier. For this reason the force constant is progressively scaled by a screening function $S(r)$ that equals from zero in the dynamics region to a value of one half in the reservoir.

$$S(r) = \begin{cases} 0 & r \leq R_A \\ \frac{(r-R_A)^2(3R_B-R_A-2r)}{2(R_B-R_A)^3} & R_A < r \leq R_B \\ 0.5 & r > R_B \end{cases} \quad (1.112)$$

The boundary force constants are derived from the X-ray temperature fac-

¹¹It has been seen that inclusion of the mean electrostatic field without properly accounting for fluctuations in this field leads to distortions in the simulation [186]

tors (see the publications [186, 187] for more details). Since solvent molecules are allowed to diffuse no boundary forces are assigned to them.

In addition to the boundary forces it is necessary to mimic the thermal fluctuations, the energy flow between the buffer and reservoir region. The simplest way to do it is coupling the buffer region to a stochastic heat bath. This done by ascribing friction coefficients β to the atoms and treating them as interacting Langevin atoms. The friction coefficient should reproduce the velocity relaxation, and they are computed from the velocity correlation function. The same friction coefficient is assigned for all the buffer atoms of the protein but considering again the same screening function as considered in the boundary force. On the contrary, solvent molecules have a constant scaling factor 0.5. Note that this stochastic heat bath will be used as a thermostat in isothermal ensemble simulations.

Equations of motion:

Atoms that belong to the dynamics region evolve according to Newton dynamics using a Verlet-type algorithm. On the other hand, a stochastic dynamics based on the Langevin equation is used for buffer region.

$$m_i\ddot{q}(t) = F_i - m_i\Omega_i^2[q_i(t) - q_i^{ref}] - m_i\beta_i\dot{q}(t) + f_i(t) \quad (1.113)$$

where F_i is the natural force coming from the potential energy, Ω_i is the boundary frequency and the Langevin parameters β_i and f_i are the friction coefficient and random force respectively.

The Stochastic Boundary method was designed for reducing the total number of solvent particles included in simulations of localized processes. This reduction must be designed avoiding the spurious edge effects. However, the method does not include the lowest frequency collective motions, this means that the dynamics and buffer regions must be small enough to avoid the necessity of a more sophisticated heat-bath model.

1.5 Theoretical methods used in this thesis: Statistical Mechanics

The statistical mechanics can be used to transform our detailed information at microscopic level obtained during the simulation to thermodynamic magnitudes [16, 190, 191].

There are many macroscopic properties that can be calculated from computer simulation of molecular systems *e.g.* heat capacity, radial distribution functions or non-equilibrium properties such as diffusion coefficients. However, here we will only give a short perspective to the free energy calculations. Reaction free energy (Helmholtz or Gibbs functions) is the magnitude that describes the spontaneity of thermic processes in NVT and NPT ensembles, respectively, that is, the tendency of molecular systems to associate and/or react. In addition, under the framework of Transition State theory, it will describe the kinetics of such process.

1.5.1 Free energy calculations

The computation of free energy can be applied to a variety of chemical and biochemical phenomena [192]. Solvation processes, molecular stability, molecular association such as ligand binding ... *etc.* We are interested in free energy differences between reactants, products and transition states. While in gas phase and small molecular systems free energy differences can be computed by means of analytical expressions, in condensed phase we have to compute the Potential of Mean Force (PMF) along a distinguished coordinate using sampling techniques.

Free energy calculations for small molecules:

Before describing the PMF calculation in condensed phase systems, it can be useful to indicate the particular case of free energy calculations in small molecular systems. In the canonical ensemble (NVT) the Helmholtz free energy F is computed from the canonical partition function Z

$$F(NVT) = -k_B T \ln(Z) \quad (1.114)$$

For N independent and identical particles in the classical limit, Z is computed from the partition function of a single molecule z .

$$Z = \frac{z^N}{N!} \quad ; \quad z = \sum_i^{\text{states}} e^{E_{tot}i/k_B T} \quad (1.115)$$

The total energy of every molecular state $E_{tot}i$ can be approximated as a sum involving translational, rotational, vibrational and electronic states. This assumption implies that the molecule partition function z_{tot} can be written as a product of terms

$$\begin{aligned} E_{tot} &= E_{trans} + E_{rot} + E_{vib} + E_{elec} \\ z_{tot} &= z_{trans} \cdot z_{rot} \cdot z_{vib} \cdot z_{elec} \end{aligned} \quad (1.116)$$

In small molecular systems the partition functions can be approximated by analytical expressions. The term z_{trans} is computed with the free particle model, z_{rot} as the rigid rotor and the z_{vib} is described as a factorization of normal modes of vibration within the harmonic oscillator. Some improvements exist for anharmonic and hindered rotor [193] models but in any case an exploration of the phase space is never performed.

In non-rigid condensed phase systems, the independent particle assumption $Z = z^N/N!$ is not valid, and the translational partition function z_{trans} cannot be modeled as a free particle. Therefore, we will have to compute the whole partition function through the phase space integral which in the NVT ensemble takes into account all the degrees of freedom of the configuration space.

$$Z = \frac{1}{h^{3N}} \int_{\mathbf{q}} \int_{\mathbf{p}} d\mathbf{q} d\mathbf{p} e^{-\frac{H(\mathbf{q}, \mathbf{p})}{k_B T}} \quad (1.117)$$

Note that if we consider that our particle system is made up of atoms (nuclei), the vibrational and translational will be considered altogether.

The usual Hamiltonian expressions permits to separate the atomic momenta from the potential energy.

$$H(\mathbf{p}, \mathbf{q}) = \sum_i^N \frac{\mathbf{p}_i^2}{2m_i} + V(\mathbf{q}) \quad (1.118)$$

Therefore, the integral over the phase space (p, q) becomes the configura-

tional integral multiplied by a constant

$$Z \propto \int_{\mathbf{q}} d\mathbf{q} e^{-\frac{V(\mathbf{q})}{k_B T}} \quad (1.119)$$

We could consider our condensed phase system (solution or biomolecule) as a supermolecule and apply the same formula than equations 1.116. However, when computing the vibrational motion of this supermolecule the local and harmonic approximation is not valid since there are many minima separated by barriers lower than the $k_B T$ factor. In consequence we still need to compute the internal nuclear motion by MD or MC techniques until convergence of the integral of configuration.

After the assumption of Born-Oppenheimer approximation the nuclei have a classical behavior and it can be entirely reproduced by classical statistical mechanics. However, it has been seen that the absence of the zero point energy and the effects of quantized vibrational motion not contemplated in classical statistical mechanics can be a source of error, mainly when computing activation free energies in reactions involving hydrogen transfer [194].

1.5.2 Potential of mean force

The Potential of Mean Force (Kirkwood 1935) of a system with N molecules is strictly the potential that gives the average force over all the configurations of all the $n+1 \dots N$ molecules acting on a particle j at any fixed configuration keeping fixed a set of molecules $1 \dots n$ [16].

$$-\nabla_j w^{(n)} = \frac{\int e^{-\beta V} (-\nabla_j V) d\mathbf{q}_{n+1} \dots d\mathbf{q}_N}{\int e^{-\beta V} d\mathbf{q}_{n+1} \dots d\mathbf{q}_N} \quad j = 1, 2, \dots, n \quad (1.120)$$

Where $\beta = 1/k_B T$, the $\nabla_j w^{(n)}$ is the average force and therefore $w^{(n)}$ is the so-called Potential of Mean Force (PMF). A particular example would be $w^{(2)}(r_{12})$ that describes the interaction between two molecules held a fixed distance r when the remaining $N - 2$ molecules are canonically averaged over all configurations.

In a more practical way, the PMF can be used to know how the free energy changes as a function of a coordinate of the system. It can be a geometrical coordinate or a more general energetic (solvent) coordinate. Unlike the mutations used frequently in free energy perturbation calculations which

are often along non-physical pathways, the PMF is usually calculated for a physical achievable process. In particular it is useful for predicting the rates in chemical solution and for elucidating the reaction mechanism of condensed phase reaction such as enzymatic processes.

The PMF $w(\chi)$ along some coordinate χ is defined from the average distribution function $\langle\rho(\chi)\rangle$

$$w(\chi) = w(\chi^*) - k_B T \ln \left[\frac{\langle\rho(\chi)\rangle}{\langle\rho(\chi^*)\rangle} \right] \quad (1.121)$$

where $w(\chi^*)$ and $\langle\rho(\chi^*)\rangle$ are arbitrary reference values. The average distribution function along the coordinate is obtained from a Boltzmann weighted average.

$$\langle\rho(\chi)\rangle = \frac{\int d\mathbf{q} \delta(\chi'(\mathbf{q}) - \chi) e^{-V(\mathbf{q})/k_B T}}{\int d\mathbf{q} e^{-V(\mathbf{q})/k_B T}} \quad (1.122)$$

Where $\delta(\chi'(\mathbf{q}) - \chi)$ is the Dirac delta function for the coordinate χ . We are assuming that the chosen coordinate χ is a geometrical coordinate $\chi(\mathbf{q})$, but χ can have any other functionality. For processes with an activation barrier higher than $k_B T$ the distribution function $\langle\rho(\chi)\rangle$ cannot be computed by a straight molecular dynamics simulation. Such computation would not converge due to low sampling in higher-energy configurations. Special sampling techniques (non-Boltzmann sampling) have been developed to obtain a PMF along a coordinate χ . Although PMF of enzymatic reactions can be calculated using free energy perturbation [112, 188, 195], the method used in this thesis and explained here is the Umbrella Sampling technique [196, 197].

Umbrella Sampling:

In this method, the microscopic system is simulated in the presence of an artificial biasing window potential $V_b(\chi)$ that is added to the potential energy $V(\mathbf{q})$.

$$V' = V(\mathbf{q}) + V_b(\chi) \quad (1.123)$$

This forces the system to compute an ensemble average over a non-Boltzmann distribution within a small interval of a prescribed value of χ .

Unless the entire range of χ coordinate is spanned in a single simulation, multiple simulations (windows) are performed with different biasing umbrella potentials $V_b(\chi)_i$ that center the sampling in different overlapping regions or windows of χ .

A reasonable choice to produce the biased ensembles, though not the unique¹², is to use for every window i an harmonic function of the form

$$V_b(\chi)_i = \frac{1}{2}k(\chi - \chi_i)^2 \quad (1.124)$$

At every window i the distribution function is efficiently converged $\langle \rho(\chi) \rangle_i^b$ through the expression of the distribution function in equation 1.122 substituting $V(\mathbf{q})$ for $V'(\mathbf{q})$. The superindices b and u indicate *biased* and *unbiased* respectively. The distribution functions from the various windows need to be unbiased $\langle \rho(\chi) \rangle_i^u$ (the non-Boltzmann factor is removed) and then recombined together to obtain the final estimate PMF $w(\chi)$. Otherwise, obtaining the unbiased PMF for the i th window

$$w(\chi)_i^u = w(\chi^*) - k_B T \ln \left[\frac{\langle \rho(\chi) \rangle_i^b}{\langle \rho(\chi^*) \rangle} \right] - V_b(\chi)_i + F_i \quad (1.125)$$

where F_i are also undetermined constant that represents the free energy associated with introducing the window potential.

$$e^{-F_i/k_B T} = \langle e^{-V_b(\chi)_i/k_B T} \rangle \quad (1.126)$$

These undetermined constants are obtained by adjusting the various adjacent windows $w(\chi)_i$. The process of unbiased and recombine the different simulation windows is the main difficulty in Umbrella Sampling. There are different strategies to do it, the most intuitive is the adjusting the various adjacent windows $w(\chi)_i^u$ (manually or automatically by least-squares) in the region in which they overlap until they match.

A more sophisticated strategy is the Weighted Histogram Analysis Method (WHAM) [198] that makes usage of all the information in the umbrella sampling, and does not discard the overlapping regions. In addition, WHAM does not require a significant amount of overlap and it can be easily extended to multi-dimensional PMF [199, 200].

In particular the WHAM technique computes the total unbiased distribution function $\langle \rho(\chi) \rangle^u$ as a weighted sum of the unbiased distribution functions $\langle \rho(\chi) \rangle_i^u$. This weighting function can be expressed in terms of the

¹²For example, in chapter 4 we will use an additional polynomial function

known biased distribution functions $\langle \rho(\chi) \rangle_i^b$

$$\langle \rho(\chi) \rangle^u = \sum_{i=1}^N n_i \langle \rho(\chi) \rangle_i^b \times \left[\sum_{j=1}^N n_j e^{-(V_b(\chi)_j - F_j)/k_B T} \right]^{-1} \quad (1.127)$$

where N is the number of windows and n_i is the number of data points (number of steps in a MD) in the window. The free energy constants F_i need to be obtained from an optimal estimate of the total distribution function

$$e^{-F_i/k_B T} = \int d\chi e^{V_b(\chi)_i/k_B T} \langle \rho(\chi) \rangle^u \quad (1.128)$$

Equations 1.127 and 1.128 need to be solved iteratively until self-consistence. A common procedure is giving a guess to the set of free energy constants $\{F_i\}$, obtain the total distribution function by equation 1.127 and use it to compute $\{F_i\}$ in equation 1.128. The procedure is stopped when the difference between the constants in two consecutive iterations is below a threshold.

See reference [201] for a comparative study between the different techniques to unbias and recombine the data extracted from Umbrella Sampling.

1.5.3 Chemical kinetics: Transition state theory

Transition state theory (TST) is one of most successful theories in theoretical chemistry [202]. It gives the framework of chemical reaction rate theory and today it is the general name for many theories based in whole or in part on the fundamental assumption of the existence of a hypersurface (transition state) in the phase space¹³ that divides reactants and products. Three properties are assumed:

- i)* Reactant states are in local equilibrium along a progress coordinate, which is the reaction coordinate.
- ii)* Trajectories that cross the transition state hypersurface do not recross it before becoming thermalized on the reactant or product side
- iii)* The reaction coordinate degree of freedom can be separated from the rest and it is treated by classical mechanics.

¹³Note that consider a saddle point in the configuration space as the transition state may not always be correct

The rate constant at a given temperature T is

$$k(T) = \gamma(T) \frac{1}{\beta h} e^{-\beta \Delta G^\ddagger(T)} \quad (1.129)$$

The magnitude $\Delta G^\ddagger(T)$ is the free energy difference between the reactants and the transition state. The transmission coefficient $\gamma(T)$ is a correction term that stands for all the approximations assumed in the TST.

$$\gamma(T) = g(T) \kappa(T) \Gamma(T) \quad (1.130)$$

Given in the same order as the approximations these three correction terms are:

- i)* $g(T)$ accounts for the deviations from equilibrium distribution in phase space. It can be either less than or greater than 1.
- ii)* $\Gamma(T)$ arises from dynamical recrossing. It will be 1 or less than 1.
- iii)* $\kappa(T)$ is the contribution from quantum mechanical tunneling therefore almost always this correction is greater or equals to 1.

For a review of the current status of the TST theory see reference [203], and [62] for TST applied to enzymatic systems.

However, the most expensive and problematic task is the computation of the free energy difference $\Delta G^\ddagger(T)$, which must be calculated along a predefined reaction coordinate. In some condensed phase reactions the mechanism should be intuitively predicted, such as proton transfers, and therefore the reaction coordinate is adequately chosen. On the other hand, many other reactions have a complicated or unpredictable mechanism and therefore the choice of a predefined reaction coordinate is difficult. A paradigmatic example is the reaction of autoionization of water [204]. In this sense there has been recent improvements in the computation of reaction rates without the knowledge of the reaction mechanism through transition path sampling method [137], but the computational cost is still too high and some alternatives are needed to model enzymatic reactions adequately.

1.6 State of the art in enzymatic reaction simulations

The collection of methods presented so far have been shown in terms of the methodology context. In this section, we will go over the same methods but emphasizing its important contributions to the study of enzymes.

Every method has its particular area of applicability. From pure quantum calculations of few tens of atoms through systems with thousands of atoms. From single point energy calculation to statistical sampling of the configuration space and prediction of rate coefficients. Some very good reviews exist in the literature [114, 205].

A lot of work has been done with MM potentials with successful calculations on ligand binding, conductivity through channels and protein folding. But since these methods do not consider explicitly the chemical reaction mechanism they are not mentioned here.

- When potential energy requires to be very accurate (DFT or post-HF) and a strict control of the wavefunction is needed, the enzymatic system is reduced and only the active site is modeled. This is the biomimetic approximation and it is mainly applied to metallo-enzymes systems [206, 207]. Usually, only minima and saddle point structures are located, in some cases the zero point energy is also computed.
- The inclusion of more and more atoms in the systems make the quantum chemistry calculations not feasible. Linear scaling techniques can be applied to accelerate this calculations [208] although they are still not fast enough to study the reactivity problem.
- An accurate but still cheap potential energy is the Carr-Parrinello molecular dynamics coupled to a Molecular Mechanics potential [107]. Potential of Mean Force calculations are possible [209] although in many cases the method is not fast enough to permit an extensive sampling over the phase space.
- The usage of QM(SCF)/MM potentials when QM is *ab initio* [103] is an alternative to the biomimetic models where reliable structures can be obtained by optimization techniques, without oversimplifying the system or constraining any atom.

- QM(SCF)/MM when QM is semiempirical or EVB permits long MD and full thermodynamic properties calculations. In this case rate constant calculations of the chemical step are also possible. However, the accuracy of semiempirical or EVB potentials is too low to go beyond the qualitative results.
- The methods specified above can be used to elucidate the origin of enzymatic catalysis. The enzymatic simulations are carefully analyzed, the different contributions to the rate acceleration are studied in order to know how enzymes work. This last item is currently a very active area [62, 210, 211, 6].

Chapter 2

Mandelate Racemase enzyme

Chapter overview

This first chapter of results is a first attempt to investigate by standard QM/MM methods the Mandelate Racemase reactivity. Some very useful conclusions will be deduced for the Mandelate Racemase reaction mechanisms. However, some other problems will arise. For example, the need of an accurate technique to locate transition states and the need to include the essential temperature effects will be the motivation to proceed with the investigation in the following chapters.

In the first section 2.1 there is a wide survey of the experimental results that are useful for our theoretical study. After a review in section 2.2 of almost the unique previous theoretical calculation[212] on Mandelate Racemase, in section 2.3 we will study the racemization reaction of propargylglycolate and mandelate substrate by the enzyme Mandelate Racemase. The chapter will conclude with a study of a gas phase model of Mandelate Racemase reaction.

2.1 Introduction: experimental results

A considerable number of experimental studies have been published about Mandelate Racemase. Some of these studies focus on the substrate binding, the chemical mechanism or the design of alternative substrates and inhibitors. The general aim of these works is to shed some light on the origin of the enzymatic catalysis or to look for the possible industrial application of racemases. Although we are not directly interested in these challenging objectives the experimental studies will give us essential information for our

theoretical study.

2.1.1 Racemases and the aim of their study

Asymmetric but not stereoselective:

The racemases enzyme family catalyze the interconversion of both substrate enantiomers ¹. Although most enzymes are famous for being exquisitely asymmetric, by definition racemases process both enantiomers and have equilibrium constants equal to unity. So the question arises, how can these enzymes, which are inherently asymmetric, deal with both enantiomers with at least approximately equal facility? The logical answer to this question is that racemases must have evolved a functional and structural *pseudosymmetry* in their active sites.

Deprotonation of a high pK_a hydrogen:

Another question, in this case not exclusive of racemases, is the rapid proton exchange involving carbon-hydrogen bond cleavage of carbon acids with relatively high pK_a . In the case of mandelic acid its α -hydrogen has an estimated pK_a in solution of ~ 22 and ~ 29 for its anion mandelate[213].

While in 0.40 M NaOD the α -proton of sodium mandelate undergoes exchange very slowly even at 100°C , in contrast, Mandelate Racemase provokes the same exchange reaction with a turnover number of $\sim 1000\text{ s}^{-1}$ at 25°C even at pH 7.5[214].² This puzzling situation is a general problem that appears in many other enzymatic reactions, for example, triose-phosphate isomerase, Δ^5 -ketosteroid isomerase, citrate synthase, enolase, aconitase and fumarase[215].

Bearne and Wolfenden[216] published a work where the non-enzymatic mandelic racemization reaction is studied in the presence of many different acid-base catalysts at different concentrations and different pH. A comparison between the reaction kinetics of enzymatic and non-enzymatic processes has concluded that Mandelate Racemases produces a rate enhancement under neutral conditions at 25°C of $1.7 \cdot 10^{15}$ -fold.

¹The epimerases family are closely related since they interconvert the two diastereoisomers

² Turnover number is the k_{cat} often applied to enzyme catalyzed reactions. It represents the maximum number of substrate molecules (when the substrate concentration is very high) which can be converted to products per molecule of enzyme per unit time

2.1.2 Presentation of Mandelate Racemase enzyme

Biological context:

Mandelate Racemase (E.C. 5.1.2.2.; MR)³ from the soil bacteria *Pseudomonas putida* catalyzes the interconversion of the enantiomers of mandelic acid[217] (see figure 2.1).

Experimentally it has been studied as a paradigm for enzymes which catalyze rapid carbon-hydrogen bond cleavage. Early studies suggested that the mandelate pathway of *P. putida* consists of five enzymes that facilitate the conversion of both (R) and (S)-mandelate to benzoate, which is subsequently converted to acetyl coenzyme A (CoA) and succinyl-CoA.

Why Mandelate Racemase?:

Among all the racemases family MR is the most studied enzyme. The following reasons justify the choice:

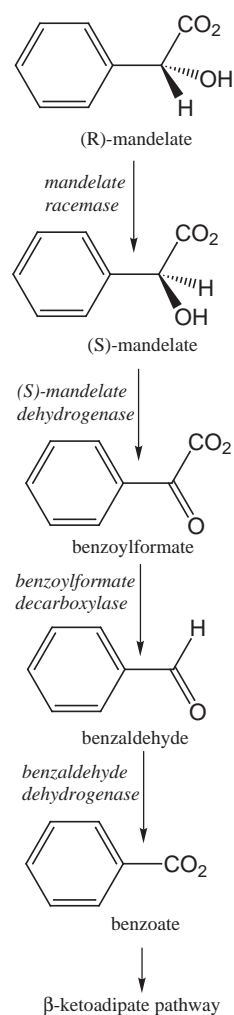
1. It is a cofactor independent inducible enzyme that can be obtained in large amounts by fermentation from *Pseudomonas Putida*
2. Immobilization leads to an enhanced activity and facilitates its recovery
3. The catalytic efficiency of this enzyme is exceptional (turnover frequency $\sim 1000 \text{ s}^{-1}$)
4. It promotes a reaction that is almost impossible by chemical means

These are probably the reasons why MR was the first of the racemases to have an elucidated structure by X-ray spectroscopy.

X-ray:

The Mandelate Racemase X-ray structure and several of its mutational variants are available at reasonable high resolution. This is an essential experimental data that gives us the opportunity to start a theoretical study.

In its crystal structure MR is an octamer of 422 symmetry. Every subunit is composed of two major structural domains. An N-terminal $\alpha + \beta$ domain and a central parallel α/β barrel, there is also a third, smaller, irregular



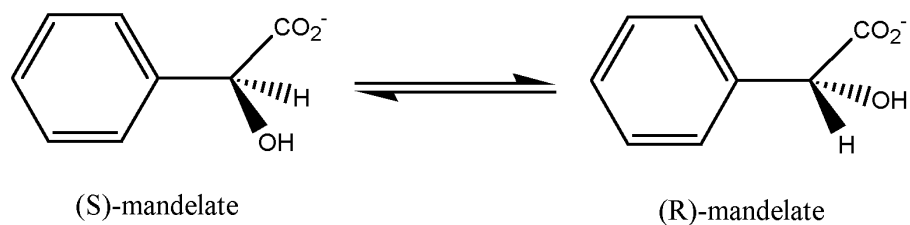
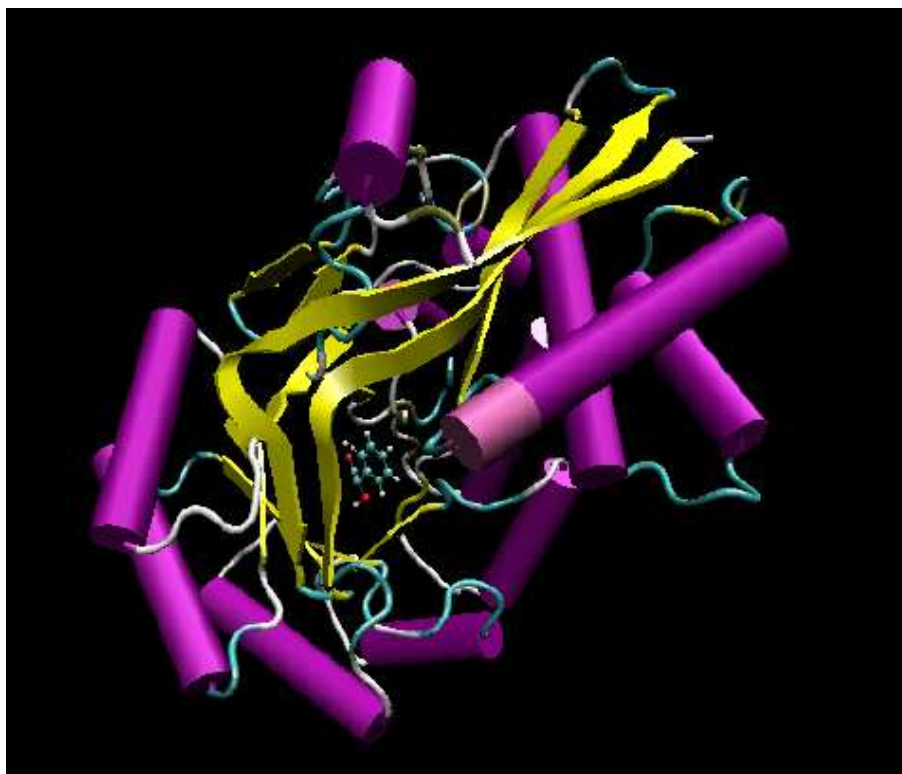


Figure 2.1: Scheme of Mandelate Racemase natural reaction

Figure 2.2: Stereoview of a ribbon diagram showing the three dimensional structure of MR backbone. In *ball & stick* the mandelate substrate is bound in the active site. This figure was prepared by VMD [218]

C-terminal domain (see figure 2.2).

Common evolution ancestors:

The molecular weight of enzyme's subunit is 38 570 and its secondary, tertiary and quaternary structures are strikingly similar to those of muconate lactonizing enzyme and galactonate dehydratase. This similarity indicates a common evolution ancestor to perform the necessary chemical task of abstracting protons α to a carboxylate that have relatively high pK_a values.

Ion dependent:

Wild Mandelate Racemase is Mg^{2+} dependent. Some studies suggest other less effective divalent metal ions such as Co^{2+} , Ni^{2+} , Mn^{2+} and Fe^{+2} [217]. This required divalent metal ion was found to be tightly bound to MR and close to the bound mandelate, suggesting that the metal ion helps to the deprotonation reaction withdrawing the excess of electron density. In figure 2.3 we can see the ligand sphere of the divalent cation.

Many substrates, inhibitors and inactivators:

Many possible substrates can bind the active site, or even racemize through Mandelate Racemase. Besides its natural substrate, the enzymatic racemization of other α -hydroxy carbonyl compounds can take place. In table 2.1 a brief description of the most studied substrates is given.

The design of substrates gives mechanistic information:

The design of alternative substrates has helped to find out experimentally the important residues in the active site for the binding process. For example, the α -OH group seems to be crucial for the binding and ortho-substituted phenyl ring provokes a remarkable steric hindrance[224]. Since the amide derivative of mandelate, although with a significant lower rate, can racemize [221], the presence of a carboxyl or negative charged oxygen on the substrate does not seem to be essential for the binding⁴.

This substrate spectrum also gives information about the possible mechanism, for example, electron-donating phenyl substituents position enhance the enzyme activity, which means that a negative charged stabilization is needed for the racemization[226]. While phenyl substituents on para and meta positions bind the active site, ortho-analogues do not bind due to steric limitations. An aromatic system must be present in β position being the

³Enzyme Commission (EC) classifies MR in the main class 5 of isomerases

⁴This new experimental result seems to be against the hypothesis that a formation of a strong and short hydrogen bond between mandelate and Glu317 is essential for the catalysis effect[225]

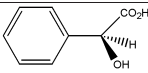
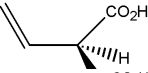

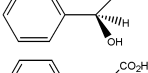
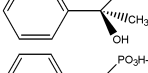
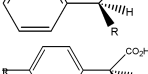
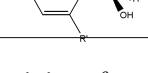
Substrate		Activity
Mandelic acid		natural substrate $k_{cat} = 350s^{-1}$ [214]
Vinyl-glycolate		active substrate $k_{cat} = 250s^{-1}$ [219]
Propargyl-glycolate		active and partially inactivator $k_{cat} = 79$ [220]
Mandelic acid amide		active substrate (15% of mandelate) [221]
(R)- α -PhenylGlycidate		inhibitor (used in Xray) [222]
Benzylphosphonate		inhibitor [223]
(Hetero)-aryl-substituted		active substrates [224]

Table 2.1: Different enzymatic activity of some substrates that bind to Mandelate Racemase active site

vinyl-glycolate the minimal conjugated system. In the absence of π -electrons in this position, such as for lactate, no racemization occurs.

Vinylglycolate has also been found to be an excellent substrate of Mandelate Racemase, with a value of k_{cat} somewhat lower, but comparable to that of mandelate. In addition, propargylglycolate has been determined to be a moderately good substrate for racemization, with a k_{cat} value of about 10% relative to mandelate. The case of propargylglycolate has been found to be specially interesting because it is also an irreversible inhibitor, with a partition ratio of racemization/inactivation of about 17 000 [220]. These two alternative substrates along with mandelate have been used in this thesis (section 2.3, page 81) to study its racemization process. This study will provide a mechanistic insight that could explain the tendency of the reaction kinetics.

The active site:

In figure 2.3 a schematic picture of the active site deduced from x-ray spectroscopy it is shown. Other residues are not displayed in figure 2.3 for clarity reasons. Aspartic270 has been proposed to form a dyad with Histidine297 to increase the proton-donor capacity. Asparagine197 seems to interact with the α -hydroxyl group to stabilize the transition state racemization. Both mutagenesis substitution experiments (D270N and N197A) provoke a reduc-

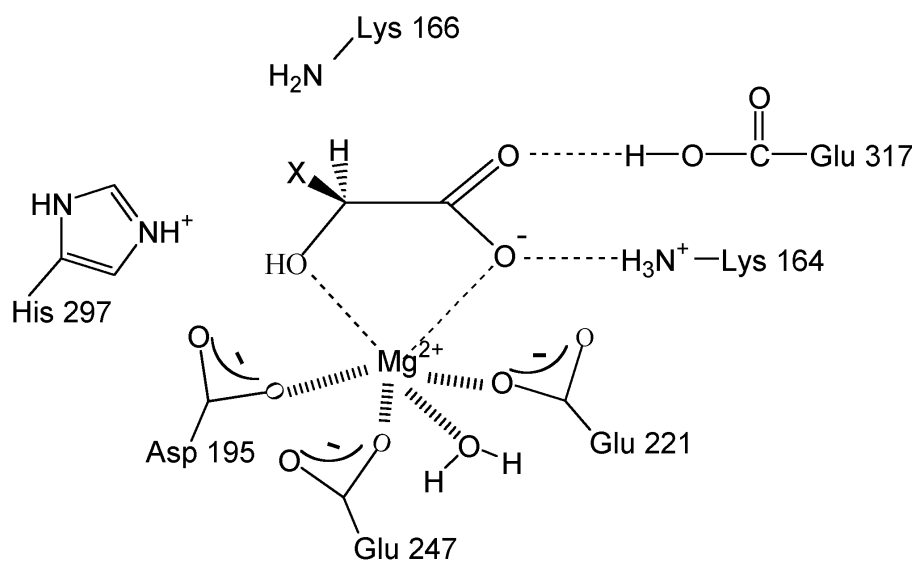


Figure 2.3: Schematic representation of the active site of Mandelate Racemase. When X=phenyl we have the natural substrate in S configuration

tion in catalytic efficiency of the enzyme.

Proposed mechanism and structure-function relationships:

Many experimental studies have been published to elucidate the Mandelate Racemase mechanism.

- Kinetic measurements (reference [214])
- Isotopic effects (references [214, 222])
- Site-directed mutagenesis experiments: K166R [227], H297N [228], E317Q [214], D270N [229] and N197A [223]

The main conclusions of these experiments are:

Lys166 and His297 role:

Mandelate Racemase proceeds by a two-base mechanism. The ϵ -amino group of Lys166 is the general base catalyst that abstracts the α -proton from (S)-mandelate. Whereas the doubly protonated imidazole group of His297 acts as the general base catalyst that removes the corresponding α -proton from (R)-mandelate. The conjugate acids of Lys166 and His297 serve as the proton donors in the formation of (S) and (R)-mandelate, respectively.

Stabilization by Lys164, Glu317 and Mg: Lysine164, Glutamic317 and magnesium cation help to the ligand binding at the first stage of the reaction

and to the electron density withdrawing to stabilize the enolic intermediate when proton abstraction takes place.

Intermediate:

The experimental results suggest that the mechanism takes place through an enolic intermediate. The term *enolic intermediate* is preferred rather than carbanion, enolate or enol to avoid specifying the extent to which the proton is transferred from the general acid catalyst (Glu317 or Lys164) to the oxygen atom of the intermediate. In any case, any isolation of this intermediate has not been reported. At the end of the next chapter we will propose that such intermediate is a transition state rather than a stable species.

Moreover, Glutamic317 has been suggested to form a so called Low Barrier Hydrogen Bond (LBHB)[230] to explain the overstabilization of the intermediate by the enolate anion formation.[215, 225] However Guthrie and Kluger[231] suggest that the electrostatic stabilization in combination with a reduction in medium polarity may be sufficient to stabilize the unstable species formed during the catalysis. Actually, if the mandelic acid amide, as we said above, was racemized at an acceptable rate[221] we can conclude that the formation of a LBHB, more difficult in the amide derivative, is not essential for the chemical step.

The possible explanation of pseudosymmetry: The ϵ -ammonium group of Lys166 has a pK_a that is evidently lowered by about 4 pK [217] units by electrostatic effects of the active site, while the pK_a of the His297 is a more nearly normal value for a histidine imidazole group. This lowering of Lys166 makes both pK_a closer to each other and this can account for the "pseudosymmetry" in the racemases reaction pointed out at the beginning of this section.

2.2 Introduction: Previous theoretical studies

Mandelate Racemase has already been studied in our group [212]. Before that, only a previous study by Alagona, Ghio and Kollman [232] covered the Mandelate Racemase reactivity. Other studies such as that from Maurice and Bearne [223] perform geometry optimizations and electrostatic map potentials of several substrates to elucidate which of the interacting groups of the substrate play an important role in the active site. But they give no insight into the chemistry of the enzymatic reaction.

Kollman and co-workers

The work of Kollman *et al.* [232] is a very preliminary gas phase study where a small model of active site is computed at HF/(STO-3G,3-21G and 6-31G) level. This *in vacuo* study of the enzymatic reaction has some drawbacks such as the exclusion of important residues in the model system. This is probably the reason why they observe no isomerization reaction. They had important difficulties to obtain an optimized geometry that could reproduce the X-ray active site, so they had to perform restricted geometry optimizations in order to avoid artificial geometries and interactions. This fact gave also a too rigid system unable to follow all the charge redistribution along the reaction path. In section 2.4 we will perform similar gas-phase calculations and we will observe that when the active site is represented with few atoms the model is not representative and it is unable to reproduce the racemization reaction.

Garcia-Viloca *et al.*

We will briefly comment on the work carried out by Mireia Garcia-Viloca *et al.* some time ago [212]. This paper was done just before the author of this thesis started his work, so this can be taken as a starting point of the whole thesis.

This is a QM/MM study of the racemization of vinylglycolate by Mandelate Racemase. In table 2.1 vinylglycolate has already been shown as a possible substrate. The reactivity is elucidated finding the possible intermediates by minimization of the energy function. A coordinate scan between the intermediates is performed to have an approximation to the energy barrier of the reaction step.

We do not want to summarize all the study, but some conclusions stated there will be used in our results section 2.3. This is the case for the reaction mechanisms encountered for vinylglycolate racemization. It will be also

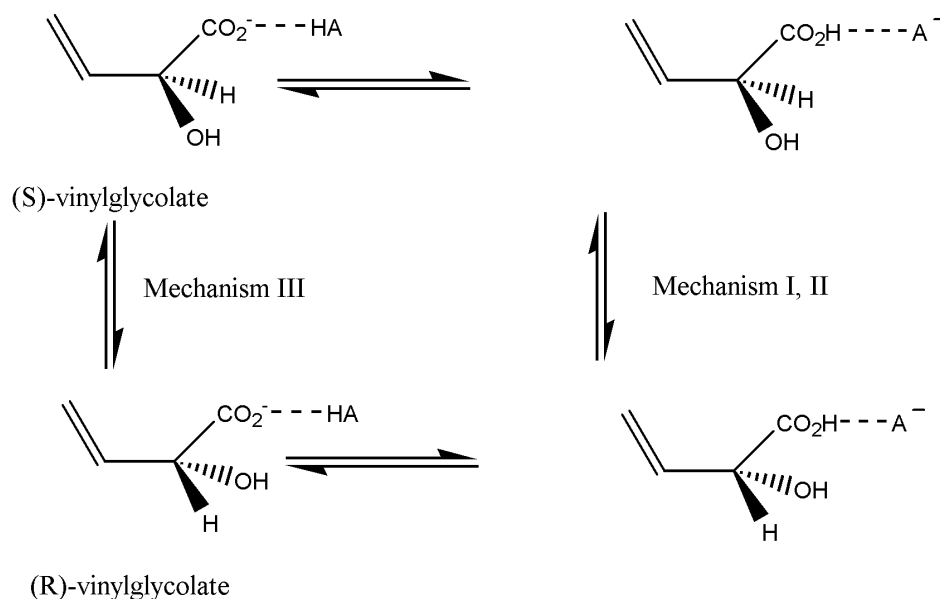


Figure 2.4: A very simplified representation of the three mechanisms encountered in vinylglycolate racemization.

important to know how the semiempirical Hamiltonian and the number of QM atoms were chosen.

Brief description of the reaction mechanisms

Three possible mechanisms are found. Mechanism I and mechanism II involve six steps through six transition states and five intermediates each. Both mechanisms require a proton transfer (the first step) from Lys164 or Glu317, respectively, to a carboxylate oxygen of vinylglycolate through the corresponding hydrogen bond. That proton transfer precedes the abstraction of the α -proton (the second step), so avoiding the accumulation of more than one negative net charge in the substrate. In contrast, a two-step mechanism (mechanism III) that takes place through a dianionic intermediate is also possible. In this case, no previous proton transfer is required, the conjugate acid of His297 now plays a stabilizing role together with the hydrogen bonds from both Lys164 and Glu317. The three parallel mechanisms are competitive at room temperature, mec I being slower than the other two.

Choice of the QM Hamiltonian

Small models of the reaction were designed to select the semiempirical Hamiltonian. They had compared the AM1 and PM3 performance with DFT results. The energy profile for abstraction of the α -proton of vinylgly-

Lys164	3	His297	5
Lys166	4	Glu317	4
Asp195	4	Substrate	7
Glu221	4	Mg	1
Glu247	5	H ₂ O	1

Table 2.2: Number of heavy atoms represented quantum mechanically in each residue of the active center of the complex MR-substrate

colate by ethylamine that models the residue Lys166 in the active site has been calculated at the different levels. The comparison of geometries and energies gives the PM3 as the best semiempirical choice.

The X-ray crystal structure of the enzyme⁵ indicates that the distance between the hexa-coordinated Mg ion and the oxygen atoms of the ligands is within the range 2.0-2.5 Å (see figure 2.3 in page 75). However, the preliminary results using PM3 gave shorter distances with deviations of 0.2-0.6 Å from the experimental. This is probably because PM3 parameters for magnesium have been obtained mainly from magnesium halides and other small inorganic compounds data. Then, Hutter *et al.*[234] developed new AM1 parameters for magnesium including a wide variety of biologically relevant molecules that contain magnesium atoms with different coordinations. The usage of AM1-SRP⁶ parameters for Mg and the rest of atoms in PM3 improves significantly the results (the calculated distances are in the experimental range). These same parameters have been applied successfully to other magnesium dependent enzymatic systems[189, 158].

Selection of the QM model

The final QM subsystem had been chosen after testing several sizes. They looked for the model that minimized the polarization of the frontier link atoms. In the end it included the substrate, the general acid based catalyst Lys166 and His297, the charge stabilization residues Lys164, Glu317 and the magnesium cation with its additional four ligands. The picture of the final QM subsystem is plotted in figure 2.3 in page 75.

Besides, some residues have QM lateral chains larger than the ones usually adopted in simulations (see table 2.2). That is, glutamic and aspartic are usually represented by ethanoic and lysine by ethylamine. But the au-

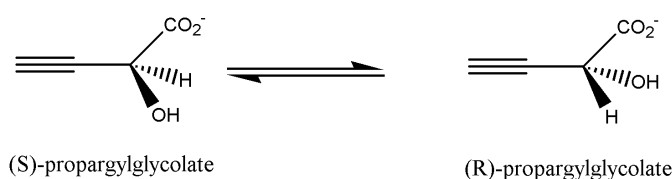
⁵PDB code 1MNS[233]

⁶Although the AM1 parameters optimized for magnesium can be used in a wide variety of molecules, since they are not the standard ones we use the term SRP to indicate it.

thors found that some residues could reproduce better the *ab initio* results by enlarging the QM lateral chain. This is the case of Lys166 that is represented by propilamine, and Glu247 by propionate.

2.3 Modulation of substrate activity

At this point the first results obtained in this thesis are presented. The aim of this section is to carry out a realistic QM/MM study of the racemization of mandelate and propargylglycolate by Mandelate Racemase. The results will be compared with those previously obtained for the racemization of vinylglycolate [212]. This comparison will enable us to get a deeper insight into the mechanistic details of the reaction and, in particular, into the basis for substrate activity.



All the numerical results (reaction distances and energetics) are shown in tables 2.4, 2.5 and 2.6 in page 100. The results originally calculated in this work (propargylglycolate and mandelate substrate) are shown along with the vinylglycolate substrate results for an easier comparison.

2.3.1 Methods and strategies

Methods:

The reaction has been modeled using a QM/MM potential as implemented in AMBER 5.0[52] (Roar-cp[235] module). Roar-cp module is the result of coupling together SANDER (the standard AMBER minimizer) with mopac 7.0. This module has been developed in the Merz laboratory during the past decade. A modified version of this program has been used.

The QM/MM scheme used here is displayed in equation 1.62 in page 28. Hydrogen link atoms have been employed to cap exposed valence sites due to bonds which cross the QM-MM boundary. The link atom model used in this work is the same that is used in the original reference [68]. It means that the hydrogen link atom does not feel any interaction from the MM part at all. It has been observed that the model is better reproduced when the link atoms can interact with all the MM atoms excluding the MM host[80]. Doing so we would have avoided the artificial polarization that we eventually had found. In addition, the link atom is left free as an additional degree of freedom. It is permitted to move during optimizations and its coordinates

are saved as any other atom. Despite other approaches freeze at 1.0\AA the distance of the link atom, we think that in the optimization process this approach would cause problems on the convergence criterion.

As in the previous work on vinylglycolate[212], the QM part of the system has been represented at the PM3 semiempirical molecular orbital level, but for the magnesium ion the new AM1 parameters developed by Hutter *et al.*[234] have been used. Several recent theoretical studies on biological systems have successfully used PM3[32, 33] and AM1 [31] methods (see reference [72] for a review of the applications).

As for the MM part of the system and the QM/MM van der Waals and electrostatic interactions we have utilized the AMBER force field by Weiner *et al.*[51]. The full QM/MM energy computation has been carried out taking into account all the interactions, that is, no cutoff was applied for the non-bonded interactions during the most of the simulations (with the exception of the molecular dynamics run).

For the sake of comparison with the case of vinylglycolate, in the present work we have chosen the same QM/MM partition used in the Garcia-Viloca *et al.* study (see figure 2.3 page 75 and table 2.2 page 79). This model has a QM subsystem with zero total electronic net charge, and 80 or 88 atoms, link atoms included, for propargylglycolate or mandelate complexes, respectively. This means that a total of 3963 atoms will constitute the whole QM/MM model for the mandelate case.

In QM/MM scheme, when cutting a covalent bond, the MM part of the cut residue can have atomic charges which sum is not an integer. In order to preserve integral charge in the MM region, in each partially QM modeled residue, the MM charges of the lateral chain atoms have been changed[236]⁷.

At the moment of writing this thesis the method used in this chapter (carried out more than three years ago) may appear too rough and rather antiquated. This is a sign that fortunately this field and the computational power is changing very rapidly. However, we believe that the chemical insight that these calculations give is still useful.

Setup of the model:

The 2.0\AA resolution structure of the complex of *Pseudomonas putida* Man-

⁷This strategy is not necessary when a force field has an integer charge for every small group of atoms (*e.g.* CHARMM force field) or when the QM/MM frontier already contemplates the correction (GHO frontier, for example). This method will not be employed in chapter 4

delate Racemase enzyme inactivated with (R)- α -phenylglycidate (Protein Data Bank code 1MNS[233]) has been taken as the starting point for the calculations. From this crystallographic structure we have taken the Cartesian coordinates of the 2700 atoms that belong to enzyme residues and we have discarded the coordinates corresponding to the 209 crystallographic waters, with the exception of the water directly bound to the metal Mg ion in the active site. Instead of keeping the crystallographic waters, we have solvated the active site with a sphere of TIP3P[237] water molecules of radius 20 Å, centered on the magnesium atom. This sphere includes 202 water molecules that are submitted to a soft harmonic potential to avoid their moving away from the active center.

We have substituted the (R)- α -PGA inhibitor found in the X-ray structure by the different substrates studied here, that is, (S)-propargylglycolate and (S)-mandelate, superimposing as many atoms as possible. Since no hydrogen atoms were determined by the crystallographic technique, they have been added using the EDIT module of AMBER[51]. However, we have not represented explicitly all the hydrogen atoms because we have used AMBER united atom model for the enzymatic residues that belong to the MM part of the system.

The protein complex has been neutralized by placing four Na^+ in positions of large negative electrostatic potential and far enough from the active site.

A 15 Å sphere was defined around the active site magnesium atom, and only residues within this sphere as well as the water molecules were allowed to move during the simulations, that results in 1299 moving atoms for the mandelate substrate case.

We have also carried out QM/MM molecular dynamics simulations in the NVT ensemble to find new possible structures that could represent the reactant and the product complexes of propargylglycolate and mandelate with the enzyme. Starting from the minimized structures, a QM/MM molecular dynamics simulation was performed with a 1 fs time step to heat the systems from 0 to 300 K over an interval of 6 ps, with atom velocities assigned from a Gaussian distribution every 2 ps in 100 K increments. The systems were equilibrated for an additional 10 ps run. The temperature was maintained by coupling to two thermostat chains (one for the MM region and the other for the QM region) within the Nose-Hoover chain temperature

scheme [174, 176]. We have used three thermostats for each chain and the program has automatically set the mass values of each of them. In this case the non-bond cutoff distance was 10 Å for the MM atoms whereas we used no cutoff for the QM atoms. The non-bond pair list was updated every 25 steps and the SHAKE algorithm[178] was used to constrain bond distances that imply hydrogen atoms. The structures obtained after the molecular dynamics simulation turned out to be very similar to those corresponding to the minimized complexes. This ensures that both structures, that is, before and after the molecular dynamics, are a good starting point for the study of the reaction mechanisms.

Procedure for the energy profiles:

Starting from the model structure of the reactive complex for each different substrate, the QM/MM potential energy was minimized until the RMS gradient fell below 0.001 kcal/(mol·Å), by means of the L-BFGS method[153] (see section 1.3.7.1 in page 48 for a description of the method). The resultant minima were taken as the reference structure that models the reactant of the S→R reactions. From each reference structure we have calculated the energy profiles of different possible mechanisms of the enzymatic reactions.

We have minimized the QM/MM potential energy along a suitable reaction coordinate for each of the reactions, with the convergence criterion described above. When the step consists basically in a proton transfer the distance between the acceptor atom and the hydrogen that is being transferred is taken as the reaction coordinate. When the reaction is a conformational change (inhibition of propargylglycolate in section 2.3.5) the appropriate dihedral angle is scanned. In any case, a penalty harmonic potential is applied to this reaction coordinate: $U = K(r - r_{eq})^2$, where r_{eq} defines every point of the corresponding energy profile, and $K = 10\,000$ kcal/(mol Å²), that permits the scanning of the reaction coordinate in small increments of 0.1 Å. In this section, the structure of maximum classical potential energy at every step of the mechanism will be considered as the transition state of that step.

The steps studied here for propargylglycolate and mandelate substrates are those described in the previous work for vinylglycolate. For a full treatment of a chemical reaction, whether in a gas or condensed phase, the dynamics of the process should be included. However, studies of the potential energy surfaces[160, 84, 161] governing enzymatic reactions can in them-

selves be mechanistically revealing, and are an essential precursor to more extensive dynamic studies.

First, we will present the results corresponding to propargylglycolate. Secondly, the case of mandelate will be analyzed. All reactions have been studied from the (S)-enantiomer to the (R)-enantiomer. We recall that the relevant bond distances along with the energetics of the different intermediates are displayed in the tables of section 2.3.8 page 99.

2.3.2 Reaction mechanism of propargylglycolate substrate

As mentioned previously three different reaction mechanisms were found for vinylglycolate in the previous paper[212]. Mec I and mec II start from the same structure, which we have called structure S. However, mec III starts from a different structure denoted S2. The corresponding structures for propargylglycolate, S and S2, have also been found here (see figure 2.5) by substituting vinylglycolate by propargylglycolate in each case and optimizing the moving part of the system.

The labels used in figure 2.5 to identify some atoms will be used throughout the chapter and correspond to atoms located in eight positions that are able to be occupied or unoccupied by means of proton transfer processes. It should be taken into account that positions 1 and 7 are really the same, but when the hydrogen is attached to C1 or C7 it means that the proton comes from the pro-(S) or from the pro-(R) side, respectively.

(S) structure:

A comparison of these structures with the corresponding S and S2 geometries of vinylglycolate does not reveal any significant differences. In all the structures Lys166 residue is closer to the substrate than His297. This fact is consistent with the basic catalytic role attributed to Lys166 residue in the S-to-R direction. His297 residue is nearly 1.0 Å closer to the substrate in S2 than in S (the distances from the hydrogen atom attached to N2 to C1 are 2.95 Å and 3.85 Å, respectively). This is a very important geometric difference that will have mechanistic consequences, as will be described later.

There are two other structural differences related to the location of Lys166 and His297 residues in the active site of the minimum energy structures of S and S2. First, whereas in S2 Lys166 forms a hydrogen bond with one of the ligands of magnesium ion (Asp195), in S Lys166 and Asp195

residues are quite far away. Secondly, a hydrogen bond between His297 and Glu247 exists in S but it does not form in S2. On the other hand, the hydrogen bond between propargylglycolate and Lys164 is present in the two minimized structures S and S2. In both cases, the distance (2.70 Å in S and 2.69 Å in S2) between the heavy atoms (O4–N3) involved in this hydrogen bond, compares well with the X-ray experimental value of 2.76 Å [233].

Finally, a hydrogen bond between the substrate and Glu317 residue appears in S as well as in S2. However, the O–O distance in this hydrogen bond is slightly longer for S2 than for S. These two bond distances are in good agreement with the experimental result of 2.68 Å [233].

(R) structure:

The minimization of the QM/MM potential energy starting from the minimum energy structure of (R)-vinylglycolate and Mandelate Racemase enzyme but substituting vinylglycolate by propargylglycolate, leads to the minimum energy structure partially represented in 2.5 (structure R). Contrary to structures S and S2, the deprotonated ϵ -nitrogen of His297 is oriented to the α -proton attached to the C7 carbon forming with it a hydrogen bond of 1.76 Å. This is the expected orientation of His297 if it is supposed to be the basic catalyst in the R→S direction. In this R structure His297 does not interact with Glu247. In addition, we can observe how the conjugate acid of Lys166 has moved further from the α -carbon with respect to its position in structure S, and it is interacting (as in S2) with a magnesium ligand (Asp195) by hydrogen bonding. The two other hydrogen bonds between propargylglycolate and the two residues Lys164 and Glu317 are still maintained in the active site of the (R)-enantiomer.

Mechanism I and II for propargylglycolate:

For propargylglycolate mec I and mec II start from the same structure S and consist of six steps that are qualitatively very similar to the case of vinylglycolate. Figure 2.6 and 2.7 show the energy profile of mechanism I and II, respectively. In these two mechanisms, a proton transfer (which we will denote “catalytic” from now on) precedes the proton transfer corresponding to the racemization process itself (which we will denote “reactive” from now on). The role of those catalytic steps is to diminish the negative charge on the substrate along the reaction. In mec I the catalytic proton transfer takes place from the ϵ -ammonium group of Lys164 to the O4 of the substrate. In mec II it is through the hydrogen bond between Glu317 and one of the

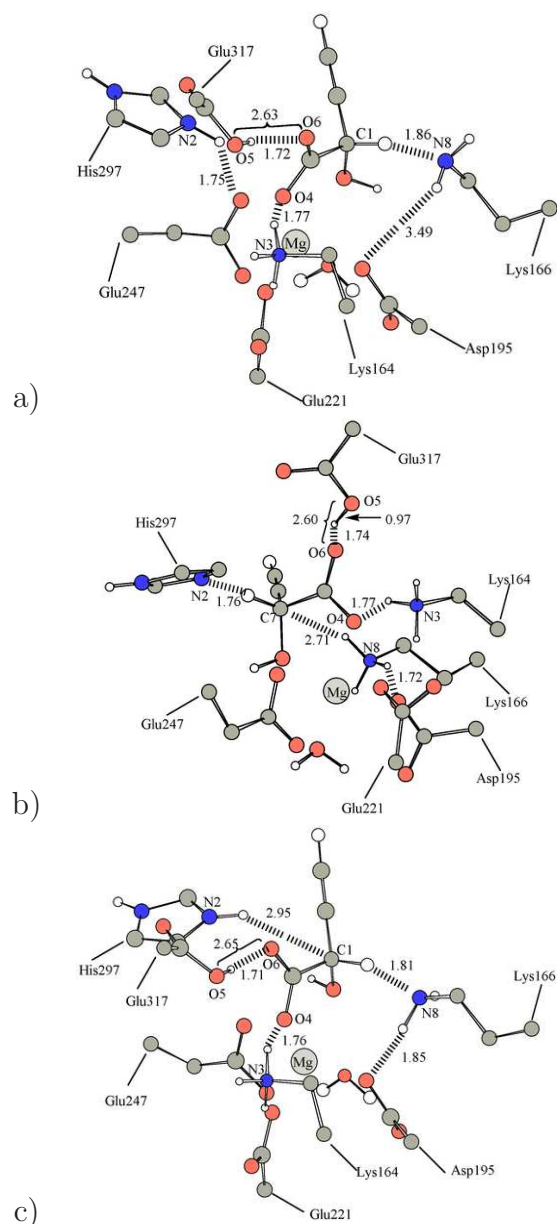


Figure 2.5: **(a)** Structure of the active site at the stationary point S for propargylglycolate. The labels used correspond to the following atoms: C1, α -carbon of (S)-propargylglycolate enantiomer; N2, ϵ -nitrogen of His297; N3, amino nitrogen of Lys164; O4, oxygen of the carboxylate group of the substrate; O5, oxygen of the carboxylic group of Glu317; O6, oxygen of the carboxylate group of the substrate; N8, amino nitrogen of Lys166. **(b)** Structure of the active site at the stationary point R for propargylglycolate. The labels used to identify some atoms are the same as in (a), with the exception of C7, that corresponds to the α -carbon of (R)-propargylglycolate enantiomer. **(c)** Structure of the active site at the stationary point S2 for propargylglycolate. Distances are given in Å.

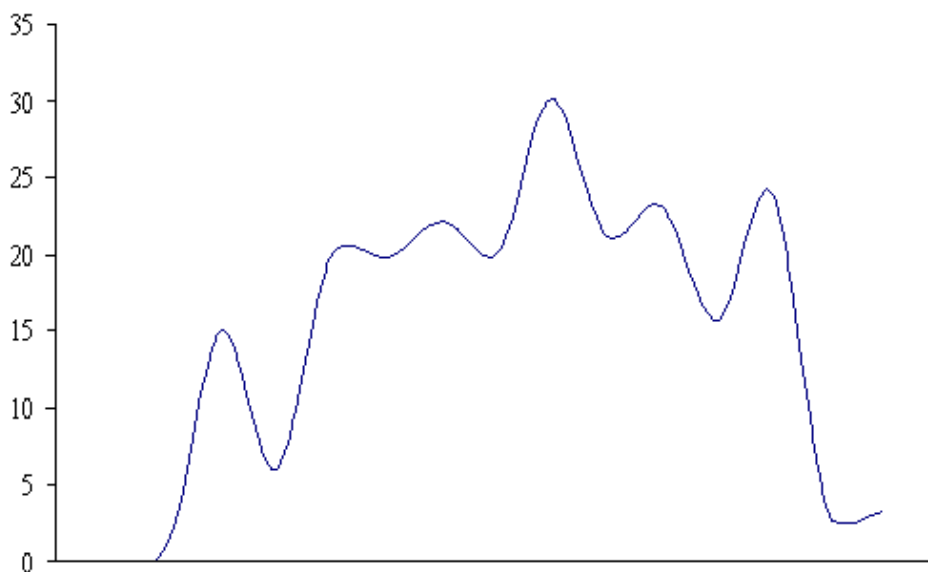


Figure 2.6: Potential energy profile for the mechanism I of propargylglycolate substrate racemization

carboxylic oxygens (O6) of the substrate. The result of this step ($S \rightarrow I1$) in both mechanisms is the neutral substrate, that is propargyl acid within our model. The catalytic proton transfer in mec I presents a higher potential energy barrier (15.9 kcal/mol) than in mec II (11.8 kcal/mol). In addition, the product of this proton transfer, intermediate I1, is around 6 kcal/mol more endoergic than S in mec I, whereas S and I1 are nearly isoergic in mec II. The catalytic proton transfers do not provoke significant changes in the bond distances involved in the reactive proton transfers of mec I and mec II.

Step 2 in both mechanisms corresponds to the abstraction of the α -proton attached to C1 by Lys166, that is the first reactive proton transfer of the racemization process, and leads in each case to the anionic structures I2. Parallel to this proton transfer, residue Lys166 forms a new hydrogen bond with Asp195, one of the ligands of the magnesium ion. The other reactive bond distance N2-H remains practically unaltered from I1 to I2 whereas the reactive bond distance C7-H shows a slight decrease, indicating a small approximation of residue His297 to the substrate. In any case, at I2 His297 (the residue that is the acid catalyst in the $S \rightarrow R$ direction) is still quite far from propargylglycolate. Steps 3 and 4 do not correspond to any proton-

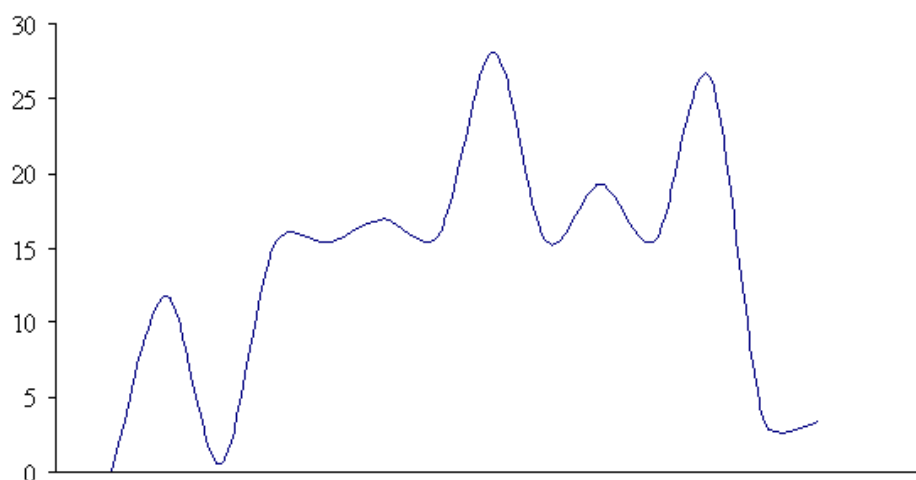


Figure 2.7: Potential energy profile for the mechanism II of propargylglycolate substrate racemization

transfer process. The energy barriers of these two steps are the result of some important changes in the active site. First, residue His297 approaches the substrate and the C7-H distance clearly diminishes (from 3.70 Å at I2 to 1.71 Å at I4 in mec I and from 3.67 Å at I2 to 1.80 Å at I4 in mec II). Secondly, parallel to its approximation to propargylglycolate, residue His297 loses its hydrogen-bond interaction with Glu247 and, in consequence, the N2-H...Glu247 bond distance increases to 2.94 Å and to 2.96 Å at I4 in mec I and mec II, respectively. Thirdly, the configuration change of C1: from a configuration closer to that of the reactant S to a configuration similar to that of R, through an sp^2 hybridization. That is, the change from C1 to C7 in our code. In both mechanisms this configuration change takes place in step 4 and involves the energy maxima of the reaction paths. This change is produced when His297 is close enough to the anionic intermediate to form an ionic pair with it.

Step 5 is the second reactive step that consists of the proton transfer from N2 to C7, through structure TS5. From I4 this step presents a smaller barrier than the first reactive proton transfer and is exoergic in mec I or nearly isoergic for mec II. The final product R is obtained in both mechanisms after the catalytic proton transfer during step 6, that returns the catalytic positions to their initial state.

Mechanism III for propargylglycolate:

As described in the previous paper of the group [212] mec III for vinylglycolate is clearly different from mec I and mec II. This alternative mechanism has also been found for propargylglycolate in MR as a third feasible reaction path of racemization. Mec III starts from the structure denoted S2 (see figure 2.5).

In the case of vinylglycolate this mechanism, much simpler than the two previous reaction paths, was shown to take place, in only two steps (without any catalytic proton transfer) via a dianionic intermediate I⁸. The stabilization of this intermediate is achieved in mec III by the residue His297, which from the beginning (in S2) is situated closer to the substrate than in S. The two steps of this mechanism consist in part in the two reactive proton transfers. The first between positions C1 and N8 (Lys166) through TS1 and the second between positions C7 and N2 (His297) through TS2. In a concerted manner with the α -proton abstraction by Lys166, the configuration change of C1 and the approach of His297 take place. Therefore, the so-called intermediate between TS1 and TS2 is a product-like structure, with a C7 clearly in R configuration rather than the planar sp² hybridisation structure.

In contrast to the results for vinylglycolate, where the two-step mechanism was found, mec III for propargylglycolate + MR is a one-step mechanism with only one clear energetic barrier of 21.9 kcal/mol at TS1. The reaction coordinate corresponding to this mec III from S2 to R was recalculated with a smaller step trying to find a minimum energy structure for the dianionic intermediate I of propargylglycolate. Even so and although the reaction energy profile shows a plateau in between 1.65 Å and 1.58 Å for the C7-H bond, no minimum was located with the L-BFGS minimizer.

2.3.3 Reaction mechanism of mandelate substrate

Next, the results obtained in the study of mandelate racemization by Mandelate Racemase are presented. Concerning mec I and mec II, the first remarkable difference between mandelate and the other two substrates is found in the minimum energy structure of the (S)-enantiomer (see figure 2.8). In this minimum energy structure (which will be called S'), the His297 residue is somewhat closer to the substrate (C7-H distance of 3.21 Å). Despite the approach of His297 to the substrate, the hydrogen bond between this residue and Glu247 is maintained in S'.

⁸ It is clearly a dianionic since no proton transfer occurs from Lys164 or Glu317

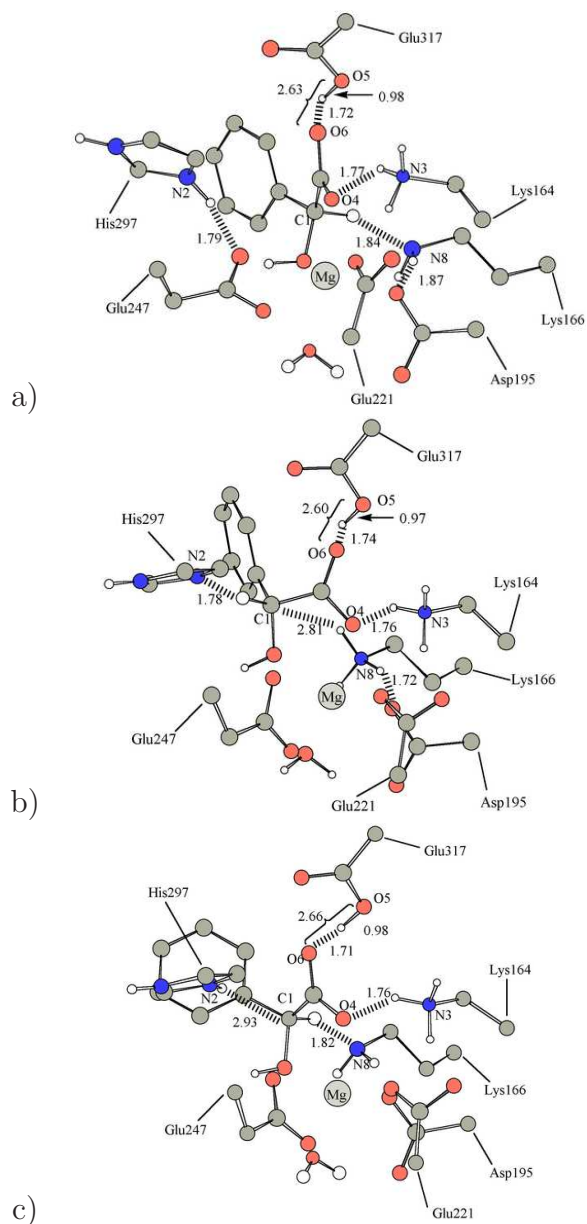


Figure 2.8: **(a)** Structure of the active site at the stationary point S' for mandelate. The labels used to identify some atoms are the same as in figure 2.5. **(b)** Structure of the active site at the stationary point R for mandelate. **(c)** Structure of the active site at the stationary point S2 for mandelate. Distances are given in Å.

Mechanism I for mandelate:

Mec I for mandelate consists of the same six steps with the same stationary points as in the corresponding mechanisms for vinylglycolate and propargylglycolate (see table 2.4 and 2.5). However, the last step of this mechanism (from I5 to R) is around 4 kcal/mol above the corresponding stationary points on mec I for the other two substrates. Consequently, the highest energy point on mec I for mandelate corresponds to the transition state of the last proton transfer, that is, from the substrate and along the catalytic hydrogen bond back to Lys164.

Mechanism II for mandelate:

Mec II for mandelate presents more significant differences when compared with the same mechanism for the two other substrates. Mec II for mandelate consists of only five steps. The change of hybridization of the α -carbon takes place in a concerted way with the α -proton donation from His297. This fragment of the reaction coordinate, from I3 to I5, was recalculated using a smaller step, trying to locate the I4 intermediate. However, this stationary point was not found on the reaction path of mec II for mandelate.

Other significant differences arise in mec II of mandelate: first, the barrier for the catalytic proton transfer between positions O5 and O6 is appreciably higher than for the two other substrates and the product of this proton transfer (intermediate I1) is substantially more destabilized. In contrast, step 3 (from I2 to I3) is lower in energy than for vinylglycolate and propargylglycolate, specially I3, that is around 8 kcal/mol more stabilized than the other two substrates.

The highest energy point in mec II of mandelate corresponds to the proton transfer from the substrate and along the catalytic hydrogen bond back to Glu317. The global processes (mec I and mec II) are more endoergic than for vinylglycolate and propargylglycolate.

Mechanism III for mandelate:

The minimum energy stationary point that we have called S2 for the other two substrates also exists for mandelate. In S2 for mandelate His297 is 0.28 Å closer to the substrate than in S' (see figure 2.8) and consequently this residue does not form a hydrogen bond with Glu247, similar to what happens in the S2 structure for vinylglycolate and propargylglycolate complexes. The S2 structure connects with an (R)-enantiomer (R') *via* mec III. This structure R' is 2.10 kcal/mol more stable than the (R)-enantiomer for

	k_{cat}/s^{-1}	ΔG^\ddagger (<i>kcal/mol</i>)	ΔV^\ddagger		
			I	II	III
(S)-Propargyl	79[220]	14.96	30.1	28.1	21.9
(S)-Vinyl	250±20[219]	14.27	26.7	20.8	20.8
(S)-Mandelate	350±5[214]	14.07	27.2	28.1	19.3
(R)-Propargyl	37[220]	15.41	26.8	24.8	18.7
(R)-Vinyl	240±30[219]	14.29	24.1	18.2	17.7
(R)-Mandelate	500±10[214]	13.85	20.5	21.4	15.4

Table 2.3: Comparison between the k_{cat} determined experimentally and the Gibbs free energy barriers, ΔG^\ddagger calculated from them, and the theoretical potential energy barriers, ΔV^\ddagger for the three mechanisms I, II, III

mec I and mec II, although there are no significant geometrical differences in the active site⁹.

Mec III for mandelate, similar to vinylglycolate, is a two-step energy profile rather asymmetric, with a higher barrier in the S-to-R direction that accounts for the abstraction of the α -proton by Lys166 and the hybridization change of the α -carbon. The values of the C-C1(C7)-O(OH)-C mandelate dihedral angle are 51.5°(S2), -23.7°(TS1), -44.9°(I), -48.4°(TS2) and -56.1°(R'). The lowest energetic barrier corresponds to the reprotonation of C7 by His297. A very shallow minimum appears in between the two barriers. The two steps of this mec III correspond to the experimentally proposed mechanism for this reaction. However, contrary to the experimental conclusions, it is a rather asymmetric mechanism.

2.3.4 Comparison with experimental kinetics

From an energetic point of view, the Gibbs free energy barriers, ΔG^\ddagger , corresponding to a one-step process that would proceed with the experimental rate constant k_{cat} , and our theoretical potential energy barriers, ΔV^\ddagger , taken as the highest energy along the QM/MM potential energy profiles, are shown in table 2.3.

Our results indicate that the reaction proceeds at least through three different mechanisms, in such a way that the effective potential energy barrier attributed to the overall reaction would indeed be lower than the particular

⁹ This is a very common situation. Two minima can differ in some units of kcal/mol without any significant geometry difference in the active site.

potential energy barriers associated to each mechanism. It can be seen that in both S→R and R→S directions mec II and mec III are faster than mec I for vinylglycolate and propargylglycolate. For mandelate, mec III and mec I are more favorable than mec II.

In spite of the approximations used, there is a good qualitative agreement when comparing the Gibbs free energy barriers deduced from the experimental k_{cat} for the (S)-enantiomers (from 14.96 kcal/mol to 14.07 kcal/mol) with the value of the potential energy maximum of the most favorable mechanism (mec III) that goes from 21.9 kcal/mol to 19.3 kcal/mol. In addition, mec III is the only one that gives the experimental trend of k_{cat} with respect to the substrate. That is, mandelate, which is the substrate that undergoes racemization by the enzyme at the highest rate, has the lowest potential energy barrier of the three substrates in the two directions (S→R and R→S). Vinylglycolate has been found to be an excellent substrate of Mandelate Racemase with kinetic parameters comparable to those of mandelate. In the second column of table 2.3 it can be verified that (S)-vinylglycolate and (S)-mandelate are racemized by MR enzyme at a similar rate although vinylglycolate is somewhat slower. In agreement, the global potential energy barriers for mandelate are around 1 kcal/mol lower than for vinylglycolate in the forward and reverse reactions.

Propargylglycolate is the substrate that undergoes racemization at the lowest rate and it is also the one that presents the highest potential energy barriers in both the S-to-R and the R-to-S directions. On the other hand, the reaction with mandelate turns out to be experimentally faster from R to S than from S to R, but when the substrates are (R)-vinylglycolate or (R)-propargylglycolate the interconversion is slower than in the S to R direction. In agreement with the experimental results, we obtain a smaller potential energy barrier for the reaction with (R)-mandelate than with (S)-mandelate. However, in disagreement with the experimental values, (R)-vinylglycolate and (R)-propargylglycolate present lower potential energy barriers than the corresponding (S)-enantiomers.

2.3.5 Inhibition by propargylglycolate substrate

In addition to being a substrate of mandelate racemase, propargylglycolate has been determined to be an inactivator of the enzyme. The process of inactivation is consistent with an enzyme-catalyzed rearrangement of the

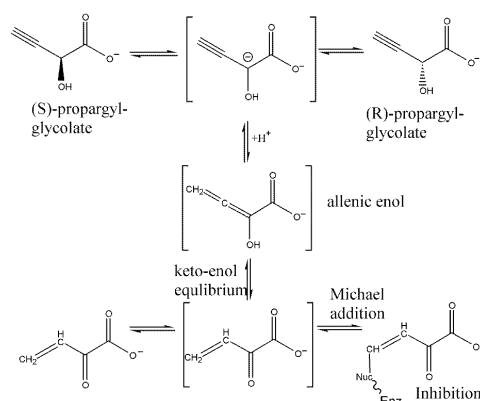


Figure 2.9: Scheme of the reaction of inhibition of the mandelate racemase through a michael addition reaction to the propargylglycolate derivate 2-keto-3-butenate

acetylenic substrate to an allenic-enol (see figure 2.9). This affords 2-keto-3-butenate as the ultimate electrophile that can then react with an active site nucleophile resulting in inactivation of the racemase.

In order to check the viability of this mechanism for propargylglycolate, we have calculated a reaction path from stationary point I2 (the anionic intermediate in the S to R direction) of mec II to the allenic intermediate.

This mechanism consists of three steps. First, the protonated Lys166 residue rotates to form a hydrogen bond with the Asp195 residue. This rotation step has a potential energy barrier of 2.82 kcal/mol and the formation of the new hydrogen bond stabilizes the system by 3.24 kcal/mol. In a second step, the OH group of propargylglycolate rotates from a pro-S position (with the OH group pointing to Lys166) to a pro-R position (with the OH group pointing to His297). This rotation stabilizes the system by 5.52 kcal/mol and passes over a potential energy barrier of 1.99 kcal/mol.

At this conformation of the substrate and of Lys166 (denoted I2' in figure 2.11), this same residue (the general base catalyst when the substrate is (S)-propargylglycolate) is able to protonate (S)-propargylglycolate at the C terminal position to give the allenic-enol intermediate. This proton-transfer process implies a potential energy barrier of 15.33 kcal/mol and the allenic product is stabilized by 2.77 kcal/mol with respect to I2' and 11.53 kcal/mol with respect to I2 (see figure 2.11).

The highest energy point of this QM/MM potential energy profile for the inactivation mechanism lies 17.24 kcal/mol above (S)-propargylglycolate.

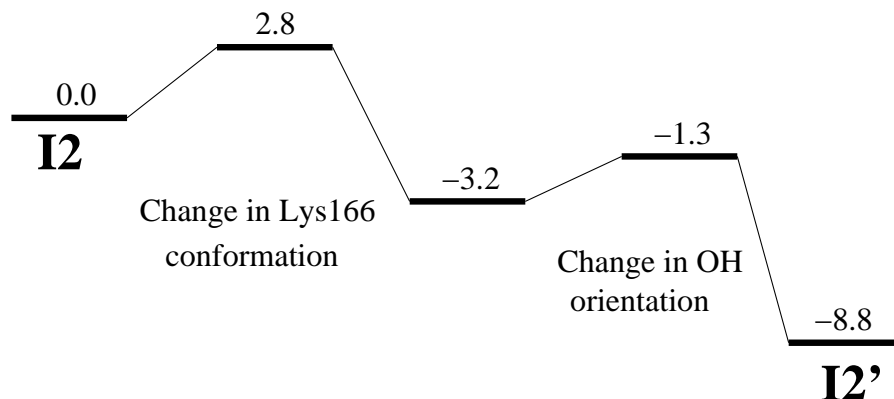


Figure 2.10: Energy profile for the geometric rearrangement in the active site before the inhibition by propargyl-glycolate takes place.

This energy barrier is lower than the highest energy point encountered by the system along the racemization mechanism *mec II*. However, it has to be remembered that we have only calculated one of the possible reaction paths from (S)-propargylglycolate to the allenic intermediate. The experimental proposal is that the formation of the allene must be faster than the rate of inactivation of the enzyme and this final covalent modification of racemase implies several more steps.

First, there is a facile ketonization of the allenic-enol intermediate to form 2-keto-3-buteonate. On our QM/MM potential energy surface, the different isomers of this buteonate molecule lie around 30 kcal/mol below the allenic intermediate. Secondly, what is properly the inactivation process consists in the attack of the electrophile molecule, *via* a Michaelis-type conjugate addition, to an active site nucleophile and this is the slowest part of the whole inactivation mechanism, which explains the partition ratio for racemization/ inactivation of $\sim 17\,000$ mentioned in the first section of this chapter.

2.3.6 Discussion

The racemization reaction of propargylglycolate and mandelate by the enzyme Mandelate Racemase takes place through at least three mechanisms (*mec I*, *mec II* and *mec III*), that are analogous to those previously found for vinylglycolate[212] as a substrate.

The stabilization of the anionic intermediates seems to be the clue to understand the catalytic role of the enzyme. This stabilization is achieved,

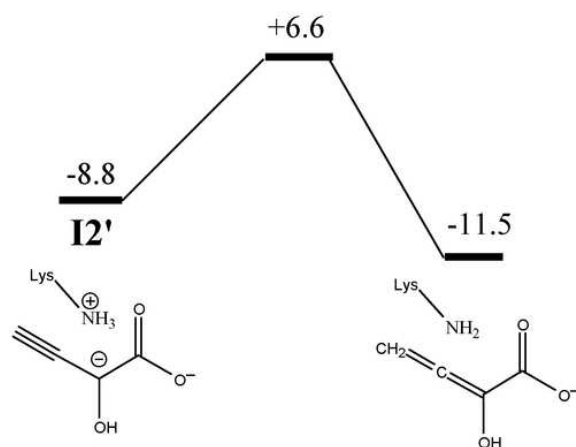


Figure 2.11: Formation of the allenic-enol intermediate. Energies (in kcal/mol) are given with respect to stationary point I2 for mec II of propargylglycolate.

on the one hand, by the interaction of the substrate with the active site residues and, on the other hand, by the particular substrate structure. As a consequence, in good agreement with the kinetic experimental data, we have shown that mandelate is racemized faster than vinylglycolate, which, in turn, is isomerized faster than propargylglycolate.

The first step of mec I and mec II consists of a catalytic proton transfer from Lys164 or Glu317, respectively, to a carboxylate oxygen of the substrate through the corresponding hydrogen bond. That proton transfer precedes the abstraction of the α -proton (the second step), so avoiding the accumulation of more than one negative charge on the substrate.

The proton transfer along the hydrogen bond between Glu317 and each (S)-enantiomer (mec II) “catalyzes” more efficiently the α -proton abstraction by Lys166. This fact is in agreement with the experimental proposal by Gerlt and Gassman[238, 215] confirmed also by mutagenesis experiments[214], about the importance of Glu317 as a putative general acid catalyst in the Mandelate Racemase pathway and about the role of this residue in reducing the pKa of the α -proton.

In contrast, Lys164 has always been described as a charged residue contributing with its ϵ -ammonium group to the positive electrostatic environment of the active site[229]. However, Gerlt and coworkers[228] also indicate that it is likely that Lys164 participates by donating a proton to the substrate carboxyl group but that the confirmation of the proposed binding

mode would require additional experimental verification.

Due to the stabilization of the α -proton abstraction transition state, we have seen that the highest energy point on the reaction path of mec I and mec II for (S)-propargylglycolate and (S)-mandelate corresponds to different steps of the mechanism: the hybridization change of the substrate and the final proton transfer from the substrate back to Lys164 or Glu317, respectively.

In mec III (the fastest) the stabilization of the intermediate is achieved mainly by the His297 residue. Moreover, it is in this mechanism where the structure of the substrate plays a more decisive role in stabilizing the stationary points found along the path. Without the catalytic proton transfers, more negative charge accumulates on the substrate and its efficiency to delocalize this extra charge is crucial for the enzymatic process. This is the reason why, among the three substrates, mandelate presents the lowest potential energy barrier in mec III.

In addition, we have found minimum energy structures corresponding to formal dianionic intermediates only on the QM/MM potential energy surfaces for vinylglycolate + MR complex (previous paper[212]) and for mandelate + MR complex in this paper. The viability of mec III and the existence of those intermediates (even though they are only slightly stabilized on our semiempirical QM/MM potential energy surfaces with respect to the second-step transition state) are in qualitative agreement with the experimental proposal of a stepwise racemization mechanism via a transiently stable intermediate[214]. The formation of this intermediate is required to explain several experimental isotope effects although its concentration has not yet been determined[217, 227].

Finally we must remark that when exploring an enzymatic mechanism with optimization methods some differences from a gas-phase study must be taken into account. An enzyme presents many parallel reaction pathways on a PES which might be very smooth in different regions¹⁰. In this sense, a change between two parallel pathways or two minima is very easy during a bad step of an optimization. Keeping this in mind we must accept the following consequences.

- An accurate optimization method is needed to avoid bad steps which make the system escape from the local valley

¹⁰ A normal mode analysis gives a high number of low frequencies.

- Despite of the last point, we will find small energetic differences between two minima energy structures without any significant geometry differences. The energetic difference is distributed among the high number of degrees of freedom.
- Sometimes the geometric differences between two intermediates will be not relevant to the mechanism.

In conclusion, we need a more accurate optimizer, mainly for the TS structure. So, before any inclusion of the dynamics and temperature effects (chapter 4) an appropriate method to locate TS in big systems will be developed in chapter 3.

2.3.7 Conclusions

In this QM/MM study we have shown how and why the rate of racemization of β - γ -unsaturated-hydroxycarboxylates by Mandelate Racemase can be modulated as a function of the particular electronic properties of the relatively similar substrates. Understanding the main factors that determine that modulation helps us to learn how more efficient substrates or inhibitors can be designed in general for enzymatic reactions.

2.3.8 Tables of results

mec I	S	TS1	I1	TS2	I2	TS3	I3	TS4	I4	TS5	I5	TS6	R
propargyl glycolate													
C1-H	1.15	1.16	1.16	1.50	1.57	1.58	1.60	2.69	2.56	2.66	2.79	2.76	2.71
N8-H	1.86	1.84	1.83	1.19	1.14	1.13	1.12	1.01	1.01	1.01	1.01	1.01	1.01
C7-H	3.85	3.81	3.86	3.72	3.70	3.31	2.96	1.61	1.71	1.51	1.19	1.19	1.17
N2-H	1.03	1.03	1.03	1.03	1.03	1.00	0.99	1.10	1.07	1.16	1.73	1.74	1.76
V	0	15.90	5.98	19.88	19.73	22.18	20.15	30.12	21.23	23.16	15.66	23.89	3.34
vinyl glycolate													
C1-H	1.15	1.15	1.16	1.56	1.58	1.66	1.62	2.43	2.56	2.67	2.79	2.75	2.71
N8-H	1.86	1.87	1.83	1.24	1.13	1.10	1.11	1.01	1.00	1.00	1.00	1.00	1.00
C7-H	3.79	3.79	3.84	3.68	3.69	2.69	2.97	1.89	1.74	1.49	1.19	1.18	1.17
N2-H	1.03	1.03	1.03	1.03	1.03	1.02	0.99	1.03	1.06	1.18	1.74	1.75	1.77
V	0	18.01	6.13	20.27	19.27	24.24	19.71	26.72	19.76	22.48	14.79	23.23	2.61
mandelate													
C1-H	1.15	1.16	1.16	1.48	1.66	1.67	1.71	2.45	2.59	2.76	2.91	2.85	2.81
N8-H	1.84	1.81	1.78	1.21	1.09	1.09	1.08	1.01	1.01	1.00	1.01	1.00	1.01
C7-H	3.21	3.28	3.32	3.45	3.36	3.12	3.01	1.99	1.77	1.49	1.19	1.18	1.17
N2-H	1.02	1.02	1.02	1.02	1.02	1.00	1.00	1.02	1.06	1.18	1.75	1.76	1.78
V	0	17.77	7.82	19.52	18.34	25.15	19.06	20.04	22.35	22.54	18.60	27.22	6.74

Table 2.4: Distances (in Å) between the atoms that participate in the reactive proton-transfer processes and potential energy (in kcal/mol) at the different stationary points found along mechanism I for the different substrates. The labels used for atoms are explained in the caption of figure 2.5.

mec II	S	TS1	I1	TS2	I2	TS3	I3	TS4	I4	TS5	I5	TS6	R
propargyl glycolate													
C1-H	1.15	1.16	1.16	1.51	1.53	1.55	1.57	2.81	2.49	2.72	2.82	2.80	2.71
N8-H	1.86	1.84	1.83	1.19	1.17	1.15	1.15	1.00	1.01	1.00	1.01	1.01	1.01
C7-H	3.85	3.82	3.82	3.67	3.67	3.41	3.14	1.60	1.80	1.40	1.20	1.20	1.17
N2-H	1.03	1.02	1.02	1.02	1.02	1.00	0.99	1.09	1.05	1.52	1.69	1.71	1.76
V	0	11.83	0.51	15.33	15.32	16.96	15.89	28.11	15.41	19.23	15.51	26.45	3.34
vinyl glycolate													
C1-H	1.15	1.16	1.16	1.55	1.55	1.62	1.58	2.43	2.45	2.70	2.84	2.82	2.71
N8-H	1.86	1.86	1.84	1.24	1.16	1.12	1.14	1.00	1.00	1.00	1.00	1.00	1.00
C7-H	3.79	3.80	3.82	3.64	3.66	2.89	3.13	1.89	1.84	1.39	1.20	1.18	1.17
N2-H	1.03	1.02	1.02	1.02	1.02	1.02	0.99	1.03	1.04	1.54	1.79	1.75	1.77
V	0	10.40	-0.50	15.54	14.89	18.08	15.58	20.81	13.63	17.8	14.41	19.41	2.61
mandelate													
C1-H	1.15	1.16	1.17	1.47	1.65	1.66	1.69			2.68	2.82	2.70	2.81
N8-H	1.84	1.78	1.76	1.25	1.10	1.09	1.08			1.00	1.01	1.01	1.01
C7-H	3.21	3.19	3.23	3.35	3.26	3.15	2.96			1.40	1.20	1.18	1.17
N2-H	1.02	1.02	1.02	1.02	1.02	1.01	0.99			1.49	1.69	1.74	1.78
V	0	17.69	5.14	14.77	13.74	14.55	8.30			23.57	18.67	28.14	6.74

Table 2.5: Distances (in Å) between the atoms that participate in the reactive proton-transfer processes and potential energy (in kcal/mol) at the different stationary points found along mechanism II for the different substrates. The labels used for atoms are explained in the caption of figure 2.5.

mecIII	S2	TS1	I	TS2	R
propargyl glycolate					
C1-H	1.16	1.76			2.71
N8-H	1.81	1.07			1.01
C7-H	2.95	2.06			1.17
N2-H	1.00	1.02			1.76
V	0.16	22.05			3.34
vinyl glycolate					
C1-H	1.16	2.02	2.62	2.62	2.75
N8-H	1.81	1.03	1.01	1.01	1.01
C7-H	2.94	1.76	1.55	1.55	1.17
N2-H	0.99	1.07	1.16	1.17	1.77
V	-0.48	20.28	15.03	15.04	2.61
mandelate					
C1-H	1.16	2.10	2.69	2.74	2.81
N8-H	1.82	1.03	1.01	1.01	1.01
C7-H	2.93	1.79	1.59	1.50	1.17
N2-H	1.00	1.06	1.13	1.22	1.78
V	0.89	20.19	16.68	16.78	4.63

Table 2.6: Distances (in Å) between the atoms that participate in the reactive proton-transfer processes and potential energy (in kcal/mol) at the different stationary points found along mechanism III for the different substrates. The labels used for atoms are explained in the caption of figure 2.5

2.4 Quantum calculations on mandelate mimic systems

An easy question and a difficult answer:

This thesis will have a significant part devoted to the location of stationary points in big enzymatic systems. However, looking to the current publications in theoretical chemistry and mainly quantum chemistry, we find many studies where the enzyme is simplified to a few tens of atoms [206]. Then, an obvious question emerges. Will we obtain the same results if we model the enzyme by reducing the system to a small representation of the active site?

This strategy when applied to enzymes is widely used, the model is usually referred as biomimetic, cluster or simply gas phase model. However the reduction of the studied system from several thousands to a few tens of atoms has some important drawbacks that must be accepted from the beginning:

Energy: The absence of the whole environment excludes the important long range interactions. Some authors solve this problem with a continuum model.

Mechanism: The oversimplification of the active site may exclude some *a priori* unknown residues that have an important role in the mechanism ¹¹.

Artificial constraints: In order to keep an adequate structure of the active site these calculations usually need to constrain some selected degrees of freedom. Commonly, some atoms are frozen at the position found in the PDB structure. However, the selection of atoms to freeze and their fixed coordinates may change along the reaction steps of a mechanism.

Unreal structures: Despite of the constriction of some atoms we may obtain unreal interactions between two residues due to the oversimplified structure.

Obviously, the two last points will depend on how flexible is our system. In any case, gas phase modelizations are still useful in some chemical systems

¹¹This problem exists also in QM/MM when we have to treat some residues with MM potentials.

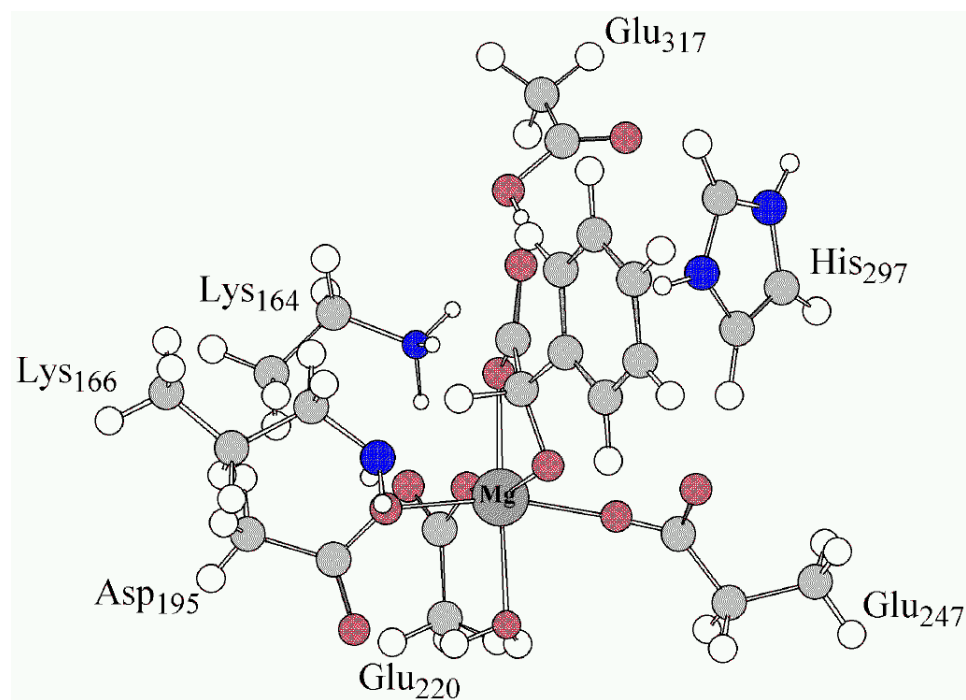


Figure 2.12: Gas phase model 1 of the active site of Mandelate Racemase for Quantum Mechanics calculations.

where accurate potential energies and a strict control of electronic state are essential (mainly in metalloproteins) [206, 207].

In this section we model the racemization reaction of mandelate substrate already studied in section 2.3.3. Two gas phase models have been chosen. The first one (model 1) is the quantum part selected in the previous QM/MM study. In figure 2.12 we show again the 88 atoms represented in model 1.

In the second model (model 2; 60 atoms) we excluded the magnesium atom and the corresponding ligands. As it is shown in figure 2.13, we only include the mandelate substrate, the two general acid-base residues Lys166 and His297 and the catalytic residues Lys164 and Glu317.

We have performed calculations at the PM3-SRP(Mg) semiempirical level and at the B3LYP/6-31G* density functional level of theory with the two different models. The calculations have been carried out by a modified version of Gaussian 98 package[239] to incorporate the SRP parameters for Mg atom in PM3 (model 1).

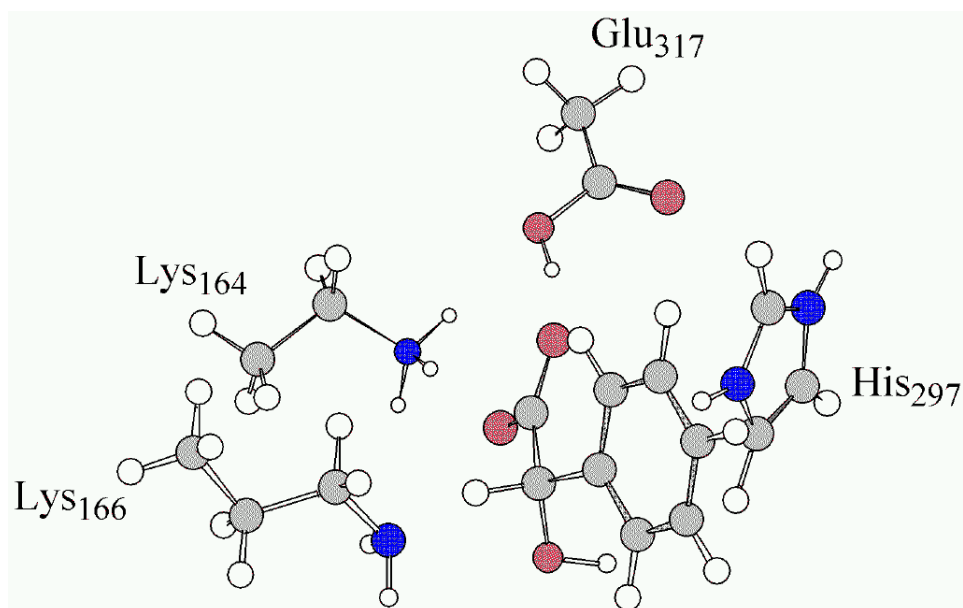


Figure 2.13: Gas phase model **2** of the active site of Mandelate Racemase for Quantum Mechanics calculations.

2.4.1 Semiempirical calculations

We tried to calculate all the intermediates starting from the S2 structure (see section 2.3). We have frozen the position of the hydrogen atoms which in the QM/MM model were link atoms. The initial coordinates are taken from the QM/MM S2 minimized structure, and there will be a total of 7 atoms with its position fixed all along the mechanism. The convergence criteria used here is the default in Gaussian98. In general, it took hundreds of steps to reach every stationary point. It can be seen in figure 2.12 that the gas phase model is built by a very weak interactions, then its energy surface is very flat and its stationary points are very difficult to optimize.

Some problems encountered:

Our results obtained using gas phase models are difficult to interpret. We obtain different results when fixing the last hydrogen atom or the last heavy atom (to avoid false rotations). Many reactants and products are found, for example, for the (S)-mandelate reactant the residue His297 may interact with OH group of mandelate or with Glu317. These different structures are obtained during the process of minimization without any other external constraint. A small rotation of some hydrocarbon chain can differentiate

two found minima. In comparison with the QM/MM model where we found also several structures, here these structures have significant structural differences. Contrary to the expected, we obtained more significantly different structures in this gas phase model than in the QM/MM model.

The reason of the difficulties may be explained as follows: During the racemization step, the acid-base residues Lys166 and His297, at their respective steps, have to move about 2 Å to be closer to the substrate. In the QM/MM model, for the (S)-mandelate and (R)-mandelate structure the non active residue (His297 and Lys166 respectively) is quite far away from the substrate and not strongly coordinated to any other functional group. This weak interaction situation is very hard to reproduce in a gas phase model. Besides, His297 is too rigid and when it must approximate to the pro-R face of the substrate provokes a movement in the backbone of the protein that is impossible to reproduce freezing the position of an atom in the gas phase model. The acetate that models Glu317 has the same behavior. If we leave free Glu317 we obtain false structures. If we fixe one of its hydrogens we get a too rigid residue whose distance from the substrate should change during the different steps of the mechanism.

Despite of these difficulties we have been able to depict an energy profile for the three mechanisms. The indirect mechanisms I and II reported in section 2.3 have been found also here. We started from a (S2)-like structure, that is, His297 is not coordinated to Glu247 because this interaction is not essential for any of the mechanisms in the racemization.

The results are presented in table 2.7(mec I) and 2.8 (mec II). As we said above, some likely artificial structures have been found and these results have been excluded. However, even among the several intermediates shown in the corresponding tables, the connection cannot be ensured.

The results for the mechanism III are shown in table 2.9. The shallow intermediate found in the QM/MM model after the central TS does not exist in gas phase, however a flat region around the same energy profile zone exist. It has been found another shallow intermediate (II in table 2.9) but in this case before the TS of configuration inversion. This structure not reported in the previous QM/MM model will be found by means of the TS search algorithm in the QM/MM model in the following chapter.

model 2:

The model 2 in figure 2.13 has been used to calculate the mechanism III

Structure	ΔE	N8-H	C1-H	C7-H	N2-H
S	0	1.77	1.16	2.91	1.00
TS1	20.69	1.75	1.16	2.97	1.00
I1	13.59	1.74	1.17	3.01	1.00
TS2	21.40	1.29	1.42	2.87	1.00
I2	21.85	1.08	1.66	2.45	1.00
TS3	23.79	1.02	1.91	1.98	1.02
I3	18.20	1.01	2.82	1.57	1.14
TS4	18.24	1.01	2.87	1.51	1.19
I4	6.22	1.01	3.25	1.17	1.75
TS5	13.36	1.01	3.19	1.16	1.77
R	-5.64	1.01	2.94	1.16	1.79

Table 2.7: Energies (kcal/mol) and distances (\AA) corresponding to the Mechanism I of model 1

Structure	ΔE	N8-H	C1-H	C7-H	N2-H
S	-1.30	1.77	1.16	3.02	1.03
TS1	19.36	1.75	1.16	3.01	1.00
I1	8.53	1.74	1.17	3.14	1.00
TS2	16.75	1.28	1.42	2.88	1.00
I2	14.76	1.08	1.65	2.70	1.00
TS3	18.38	1.02	2.01	1.97	1.02
I3	14.91	1.01	2.82	1.67	1.08
TS4	16.35	1.01	2.98	1.46	1.26
I4	6.69	1.01	3.20	1.17	1.74
TS5	15.19	1.01	3.15	1.16	1.76
R	-4.52	1.01	2.99	1.16	1.79

Table 2.8: Energies (kcal/mol) and distances (\AA) corresponding to the Mechanism II of model 1

Structure	ΔE	N8-H	C1-H	C7-H	N2-H
S	3.01	1.77	1.16	2.85	1.00
TS1	15.17	1.20	1.48	2.53	1.00
I1	15.10	1.14	1.55	2.49	1.00
TS2	18.57	1.04	1.85	1.93	1.03
R	-6.24	1.01	3.00	1.16	1.79

Table 2.9: Energies (kcal/mol) and distances (\AA) corresponding to the Mechanism III of model 1

with semiempirical methods. This smaller model has also been selected to use it in the DFT calculations and it will be used as a quantum part in the QM/MM free energy calculations of chapter 4.

In model 2 the absence of the Mg cation coordinated to the substrate removes some rigidity to the structure. As a consequence, this shortened selection of the active site is even more flexible and both Lys166 and His297 interact in the (S) and (R) structures, respectively, with the carboxylic group of mandelate. Despite of this fact, using PM3 semiempirical Hamiltonian we have located the central transition state corresponding to the carbon inversion of configuration of mechanism III with very similar characteristics in comparison with the model 1. This fact means that the essential chemistry is already contained in this smaller model and it enables us to use it in the forthcoming chapter 4.

2.4.2 DFT calculations

DFT calculations have been carried out using gas phase model 2. Starting from the structures optimized at the semiempirical level a B3LYP functional is used with the 6-31G* basis set. We got no results from the DFT calculations. The problem was both the 726 basis functions and the flat surface that it took so many steps to converge that the location of stationary points was too expensive. In the end we were unable to converge any meaningful geometry of the mechanism. So we discarded to insist in such expensive calculations.

2.4.3 Some short conclusions and perspectives

This section is intended to show the problems that may arise when studying very flexible gas phase models.

We will not comment the differences between QM/MM results in section 2.3 and the results obtained here. Both results are not comparable because we cannot ensure the connection of the intermediates under a unique MEP. The transition states found in the previous QM/MM model are also found here, but in general we can hardly connect the TS and the several intermediates due to the tricky behavior of the optimizations. Different structures have been found and all of them are possible candidates to the same intermediate.

In the QM/MM study of Mandelate Racemase we also found different minima in small energetic range, but, as we already pointed out, all the minima had the same structure in the active site. This is not the case in these gas phase calculations. There are some interactions that we did not find in the QM/MM model and we attribute it to the intrinsic problem of constriction of some coordinates.

Probably the problems encountered in the work of Alagona *et. al.*[232] and already commented in page 77 were due to these same problems, the difficult combination of the needed constraints and the flexible active site of Mandelate Racemase.

This is why we think that, in this case, QM/MM methods are not only an improvement to the solute/solvent interaction but an easier way to obtain reliable structures. For these reasons we insist that the exploration of big-dimensioned QM/MM surfaces with optimization methods is a valuable tool. In the next chapter we will present a method to carry out such exploration.

Chapter 3

Optimization of systems constituted by thousands of atoms

State of the art in the study of the reaction paths:

These last years theoretical chemistry has focused its efforts on a deeper understanding of reactive systems constituted by thousands of atoms. Condensed phase reactions, enzymatic reactions and solid state catalysis are some of the most successful areas. When so many degrees of freedom are taken into account, in addition to the eternal problem of having an adequate potential, an appropriate exploration of the configurational space is crucial to calculate thermodynamic, kinetic magnitudes and to understand the mechanism of the considered process.

Some work has been done on the acceleration of this exploration by molecular dynamics for example[169, 240]. Another important issue has to do with the need of building up a path connecting reactant to products prior to any dynamics calculation. That is, no matters how fast is our molecular dynamics if we explore a wrong region that is far away from any of the reaction pathways.

As we stated in the introduction section, some strategies have been proposed to elucidate the reaction pathway. Going from the classical minimum energy path (MEP)[141, 138], to consider the effect of the temperature on this MEP[241, 145], or even taking into account dynamical effects on an ensemble of transition paths[137].

Location of stationary points:

The procedures mentioned above need of the availability of a very cheap potential energy. However, the state of the art in enzymatic catalysis makes usage of the QM/MM methods [72]. Although QM/MM potentials were already designed to obtain a fast energy evaluation in big systems, its computational cost is still too big to afford the full exploration of the configurational space or even to locate a temperature-depending path. So the location of one of the local saddle points of a certain reaction step in the PES is still a good tool to verify the reliability of the reaction pathway. Moreover, when a QM/MM potential can be so expensive that a posterior dynamics is prohibitive, mainly in QM(*ab initio*)/MM, the exploration of the stationary points in a local valley is the only way out to obtain reliable information about the reactivity of our chemical system.

The location of stationary points in chemical systems involving up to few tens of atoms is a very well established field. The most effective algorithms are based on the Newton-Raphson equation that employs second derivatives of the energy. However, when hundreds of atoms need to be moved some problems emerge due to the manipulation of these very big matrices, mainly its computation, diagonalization and storage.

Coming from the reduced dimension of systems usually studied in *ab initio* quantum chemistry, the problem is not only to overcome the computational problems but it is also a conceptual problem. There will be many available parallel reaction paths at a given finite temperature. Actually in systems such as enzymes the potential energy surface becomes so flat that special care must be taken to perform always the search in the same local valley. So we need efficient and tight algorithms. Even more, when expensive potentials such as the QM/MM ones are used the process needs to spend as few steps as possible.

The available software is not adequate:

Most of program packages designed to study biological systems have several standard minimizers. These algorithms have been conceived for cases in which potential energy, usually molecular mechanics, is computationally cheap and the main computer demand is the storage of enormous vector arrays and the slow matrix diagonalizations. An example of these methods have already been outlined, they are steepest descent, conjugate gradient, or the more accurate Adopted Basis Newton Raphson (ABNR)[53, 144], Trun-

cated Newton[156, 157] and Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)[153] (see reference [242] for a recent comparison between some of these methods). Most of times they are applied to minimize an experimental Protein Data Bank (PDB) structure before running a molecular dynamics or a normal mode analysis. However none of those methods were originally designed to search for TS structures. Then we need efficient methods to locate directly TS structures in enzymatic catalysis using the information provided by the second derivatives. In addition these methods have to be developed to use the quite expensive QM/MM potentials, but just demanding a reasonable computer cost. This is why we found the necessity to implement a method capable to find TS with QM/MM potentials in big systems. Although there are some packages such as GRACE [159, 160] specially designed to perform geometry optimizations in enzymes that may be downloaded from the web free of charge ¹, an accurate design and test of the algorithms involved in the optimization process is still needed.

Overview of the chapter:

In the first section 3.1 we propose a methodology to locate stationary points on a QM/MM potential energy surface. This algorithm is based on a suitable approximation to an initial full Hessian matrix, either a modified Broyden-Fletcher-Goldfarg-Shanno (m-BFGS) developed here or a Powell update formula for the location of, respectively, a minimum or a transition state, and the so-called Rational Function Optimization. Although this second order method is tested on systems of small and medium size, it is an essential stage before we move to the study of big systems. The work of this section has already been published as indicated in reference [2].

In section 3.2 a systematic analysis of the micro-iterative method described in section 1.3.7.4 (page 51) is presented. The accurate RFO method developed and tested in section 3.1 is coupled to a cheap minimizer such as L-BFGS. We decided that the micro-iterative procedure is the most suitable strategy to locate stationary points in our systems. So we formulated, implemented and tested several options of this method. Features such as the two regions (core/environment) size, the interaction between these two regions and the alternating frequency between the two corresponding optimization processes are analyzed. The methods are tested on two different and rep-

¹<http://www.bath.ac.uk/~chsihw/grace/grace.html> A similar package [161] is implemented in the Carr-Parrinello package: <http://www.cpmid.org>

representative steps of Mandelate Racemase mechanism. This work coincides with our recent publication [3].

In section 3.3 we will investigate how important the accurate location of transition states is. Sometimes, when the appropriate reaction coordinate is more complicated than just a unique distance or an angle, the direct location of the stationary point making use of second derivatives is a good strategy. This will be the case to prevent from wrong conclusions when both MEP analysis or posterior free energy calculation along the reaction path are made. In particular, all the TS found in section 2.3 have been refined and the differences have been evaluated. The reference [4] is the published work that contains the data described in this section.

The last section 3.4 is a description of a method that we designed to locate stationary points on very big systems without splitting the system in different parts like micro-iterative method does. This objective has been a challenge for many researchers, that is, the search with a second-order algorithm with a strict control of positive eigenvalues without the storage and diagonalization problem. We propose a limited-memory update combined with an iterative diagonalization method similar to those used in multiconfigurational electronic problems[243, 244]. We advance that this method did not work as we desired and we will describe its advantages and its failures.

3.1 An algorithm to locate stationary points on QM/MM surfaces

We want to design a second order optimization method applied to QM/MM energy surfaces. We want also to test some important features: The shape of the initial Hessian, the update Hessian formula, the usage of Cartesian coordinates for algorithms where internal coordinates are frequently used, the efficiency of the method to converge at high and low gradient norms. All these aspects are interrelated and will influence altogether to the efficiency of the method.

The comparison of performance with other methods such as BFGS and L-BFGS will be a reference test to know the efficiency of our method.

Besides, the implementation task is by itself a tedious work that must be carefully checked before we proceed to apply the method to big size systems.

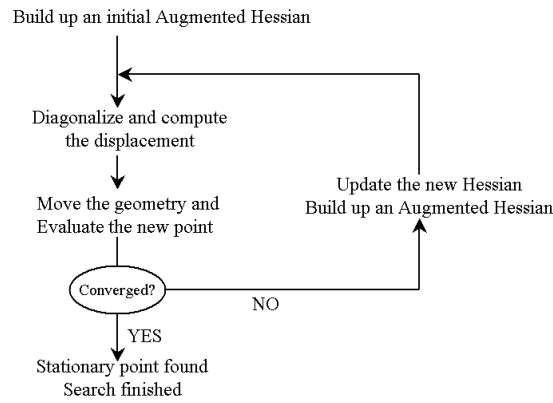
3.1.1 Equations and its implementation

3.1.1.1 RFO and updates

As we said in section 1.3 Rational Function Optimization is a second order method[130, 131, 245]. It is basically a Newton-Raphson procedure, a Hessian eigenmode following algorithm that avoids the inversion of the Hessian matrix and provides an implicit determination of step length.

Although we already described the mathematics of RFO method in the introduction section 1.3.4.2, now we will give a handful explanation of the method that will stand out the advantages and the difficulties of its practical usage.

RFO is an iterative process that can be outlined as follows:



We first build the Augmented Hessian(AH), which is the Hessian matrix (\mathbf{B}_k) with an additional row and an additional column that contains the gradient.

$$AH = \begin{pmatrix} 0 & \mathbf{g}_k^T \\ \mathbf{g}_k & \mathbf{B}_k \end{pmatrix} \quad (3.1)$$

At every iteration k the $(N+1)$ eigenvalue system must be solved

$$\begin{pmatrix} 0 & \mathbf{g}_k^T \\ \mathbf{g}_k & \mathbf{B}_k \end{pmatrix} \mathbf{v}_\theta^{(k)} = \lambda_\theta^{(k)} \mathbf{v}_\theta^{(k)} \quad \forall \theta = 1, \dots, N+1 \quad (3.2)$$

The displacement vector $\Delta \mathbf{q}_k$, is obtained dividing the first ($\theta = 1$) or second ($\theta = 2$) eigenvector (depending if we look for a minimum or a saddle point

respectively) by its first component

$$\Delta \mathbf{q}_k = \frac{1}{v_{1,\theta}^{(k)}} \mathbf{v}'_{\theta}{}^{(k)} \quad (3.3)$$

The quadratic variation energy $Q(\Delta \mathbf{q}_k)$ is evaluated as

$$Q(\Delta \mathbf{q}_k) = \frac{1}{2} \frac{\lambda_{\theta}^{(k)}}{v_{1,\theta}^{(k)}} \quad (3.4)$$

It can be shown that as the optimization process converges $v_{1,\theta}^{(k)}$ tends to 1 and $\lambda_{\theta}^{(k)}$ to zero. This criteria is an additional test of convergence aside from the gradient norm and the change in energy (section 1.3.1).

The $(N + 1)$ eigenvalue system (equation 3.2) may be solved completely if we want to check at every step the correct concavity of the PES. If this is not the case, a partial diagonalization that only gives the two lowest eigenpairs will be enough to calculate the displacement (equation 3.3) and proceed with the search².

Updating formula for the Hessian matrix:

When we are trying to locate a minimum with RFO, the Hessian matrix is updated using a modified form of the BFGS formula[12]. The most general rank-two update Hessian matrix formula is

$$\mathbf{B}_{k+1} = \mathbf{B}_0 + \sum_{i=0}^k [\mathbf{j}_i \mathbf{u}_i^T + \mathbf{u}_i \mathbf{j}_i^T - (\mathbf{j}_i^T \Delta \mathbf{q}_i) \mathbf{u}_i \mathbf{u}_i^T] \quad k = 0, 1, \dots \quad (3.5)$$

Where $\mathbf{j}_i = \mathbf{D}_i - \mathbf{A}_i$, $\mathbf{D}_i = \mathbf{g}_{i+1} - \mathbf{g}_i$, $\mathbf{A}_i = \mathbf{B}_i \Delta \mathbf{q}_i$, $\mathbf{u}_i = \mathbf{M}_i \Delta \mathbf{q}_i / (\Delta \mathbf{q}_i^T \mathbf{M}_i \Delta \mathbf{q}_i)$ and \mathbf{B}_{k+1} is the approximated Hessian matrix. Different election of the \mathbf{M}_i matrix leads to different update Hessian matrix formula. In particular, for the BFGS update $\mathbf{M}_i = a_i \mathbf{B}_{i+1} + b_i \mathbf{B}_i$ for some selected positive definite scalars a_i and b_i [136].

The proposed modified form of the BFGS expression only differs with respect to the normal BFGS in the calculation of the two scalars a_i and b_i .

² In LAPACK library there are several algorithms that permit this more economic diagonalization[246]. We will use the term partial for those diagonalization methods that extract a limited number of eigenpairs but require the whole matrix in memory. The label iterative will be used for the methods studied in section 3.4 where the memory storage of the matrix is avoided

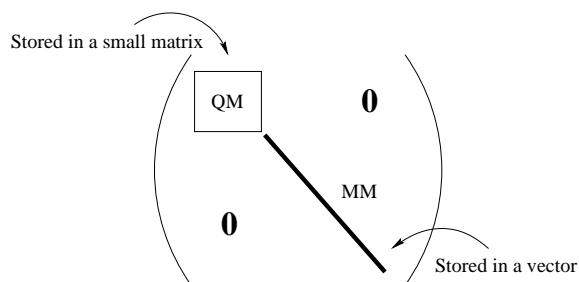


Figure 3.1: Approximated initial Hessian designed to optimize large systems

In this modified form these two scalars are evaluated as

$$a_i = \frac{[\mathbf{A}_i^T \mathbf{D}_i]^2}{(\mathbf{A}_i^T \mathbf{A}_i)(\mathbf{D}_i^T \mathbf{D}_i)} \quad (3.6)$$

$$b_i = 1 - a_i \quad (3.7)$$

Note that both a_i and b_i are positive quantities. The resulting \mathbf{B}_{k+1} updated Hessian will be positive-definite if both the \mathbf{B}_k matrix and the scalar $\Delta \mathbf{q}_k^T \mathbf{D}_k$ are positive-definite as occurs in the normal BFGS formula.

For location of first-order saddle points the Powell formula is used, where in this case the matrix \mathbf{M}_i contained in equation 3.5 is equal to unit matrix \mathbf{I} .

Acceleration of optimization process: DIIS

We implemented the Direct Inversion of Iterative Space to the gradient vector (GDIIS) as a process to accelerate the convergence in the vicinity of the stationary point. As described in section 1.3.4.3 a combination of the previous gradient vectors is used so as to minimize the length of the current gradient vector.

The GDIIS procedure has only been used in a development stage of the source code. We saw that the quality of a good initial Hessian could improve in a more efficient way the convergence process. So it is not used in the test systems nor in the forthcoming sections of application to enzymatic systems.

3.1.1.2 Initial Hessian on minimization and TS search

We tested two forms of initial Hessian matrix. The first is the identity matrix. The second is a numerical Hessian. The initial numerical Hessian matrix used in this section has the shape shown in figure 3.1 This form has

been proposed in reference [155] in order to avoid the storage of a big matrix. The systems tested in this section are not that big, but we wanted to know the behavior of such initial Hessian in order to apply it to really big systems (section 3.4).

The numerical Hessian components have been calculated by forward-backward expression (equation 1.70 in page 36) of the gradient. We have seen that the numerical value has an important dependence on the convergence criteria in the SCF calculation and on the finite displacement of the geometry. So, after some benchmarks tests we determined the SCF threshold to $10^{-8} kcal/mol$ and the finite displacements reduced to 10^{-7}Å .

We have also tested some strategies to build faster the numerical Hessian matrix:

- 1SCF/MM for MM atoms: When calculating the derivatives corresponding to MM atoms only one cycle in the SCF process is done
- pure MM for MM atoms: discard the QM and QM/MM part of the Hamiltonian for the MM atoms far away from the quantum zone.

However, we have noticed a strong dependence of the number steps required to converge on the quality of the approximated Hessian. So, hereafter QM(*full SCF*)/MM interaction is used for the numerical derivatives.

3.1.1.3 External minimizers: BFGS and L-BFGS

The RFO method will be compared with the quasi Newton process BFGS[12, 136] method and with the limited memory L-BFGS[153].

$$\Delta \mathbf{q}_k = -\mathbf{B}_k^{-1} \mathbf{g}_k \quad (3.8)$$

Few comments will be devoted to these methods here. They have already been described in section 1.3. We will only point out that the BFGS implemented here starts the iterative process with the identity matrix as initial Hessian. As the optimization proceeds, an update of the inverse of the Hessian is build up following the BFGS formula (equation 3.6). So no matrix diagonalization is required and, therefore, there is no test on the suitable curvature of the surface.

The limited memory BFGS, the so-called L-BFGS procedure has also been described in section 1.3.7.1. It is important to note that the unique

difference, as we have implemented here, between BFGS and LBFGS is that while the former builds the inverse of the Hessian matrix at every iteration and stores it in memory, the latter never stores the whole Hessian, but only the gradient and the position of a certain number of previous steps in order to perform *on the fly*, the matrix-vector product $\mathbf{B}_k^{-1}\mathbf{g}_k$ of equation 3.8.

3.1.1.4 Implementation

To obtain the QM/MM potential energy every molecular system was partitioned into a reactive part (the core³) treated quantum mechanically with the PM3 [32] or the AM1 [31] semiempirical Hamiltonians and a nonreactive part (the environment) treated by means of molecular mechanics with the AMBER force field[51].

The QM atoms are influenced by the partial charges of the MM atoms, and, in addition, bonding and van der Waals interactions between the two regions are included consistently. All the non-bonded interaction parameters used for $H_{QM/MM}$ (see section 1.2.3.1) are taken from AMBER force field. We used hydrogen link atoms to cap exposed valence sites due to bonds which cross the QM/MM boundary. The AMBER 5.0 (Roar-cp module)[235] program was used to carry out the QM/MM calculations. This potential energy method is the same as the one used in the Mandelate Race-mase QM/MM study in section 2.3, and its equations are described in the introduction section 1.2.3.1.

The search algorithm works in the following way (the name of the subroutine that performs the task is indicated in *italics*. The implemented source code is described in the appendix, page 201):

1. An initial geometry \mathbf{q}_0 in Cartesian coordinates is chosen. The QM region, the MM region and the corresponding link atoms are defined. Then, three environment atoms are always fixed in order to get rid of the translations and rotations of the whole molecular system. (*subr. rdigpr pqqorder*)
2. The QM/MM energy and the gradient at \mathbf{q}_0 are calculated using the Roar-cp module. (*subr. energy-qmmm*)

³the nomenclature core/environment used here will appear in the micro-iterations scheme. But in this case it only indicates the partition in a QM and MM region. So, the whole set of atoms are moved in a unique process

3. The guess \mathbf{B}_0 Hessian matrix is built up according to figure 3.1 (*subr. calchess*). To this aim, we implemented numerical second derivatives in the Roar-cp module as required here. Diagonalization of \mathbf{B}_0 allows us to test whether \mathbf{q}_0 lies on the suitable quadratic region of the PES (*i.e.*, zero or one negative eigenvalue for searching a minimum or a transition state, respectively). If this is not the case, the \mathbf{B}_0 matrix is forced to have the convenient number of negative eigenvalues. (*subr. corr_eigval*)
4. The RFO method plus a procedure to solve the corresponding secular equations through the full diagonalization of the AH matrix are used in order to obtain the displacement vector at each iteration. (*subr. ts_rfo: build_ah, diagonalize, choose_eig*)

This vector is scaled by a factor according to the procedure described in section 3.1.1.1 (*subr. calc_step*). Then, the new geometry \mathbf{q}_{k+1} is obtained as a result of the current k iteration.

5. If the RMS of the gradient at \mathbf{q}_{k+1} is smaller than a suitable convergence criterion, it is considered that the corresponding stationary point has been reached and the search ends (*subr. eval_step*). Otherwise, the algorithm proceeds to step 6.
6. For minima, the modified BFGS formula (m-BFGS) given in equations 3.5, 3.6 and 3.7 is used to update the Hessian matrix. For transition states the Powell formula is employed (equation 3.5). Owing to the update the QM and the MM parts of the resulting \mathbf{B}_{k+1} Hessian matrix become coupled and the MM part is no longer diagonal (as displayed in the initial Hessian of figure 3.1). (*subr. update_hess*)

Diagonalization of \mathbf{B}_{k+1} allows us to test whether \mathbf{q}_{k+1} lies on the suitable quadratic region of the PES. If this is not the case, the \mathbf{B}_{k+1} matrix is forced to have the convenient number of negative eigenvalues (*subr. corr_eigval*). Then the algorithm proceeds to the step 4 to start a new iteration cycle.

All the calculations, except the evaluation of energies and gradients, were carried out with the algorithm just described. Hereafter this algorithm is called RFO-m-BFGS or RFO-Powell when a minimum or a transition state, respectively, is looked for.

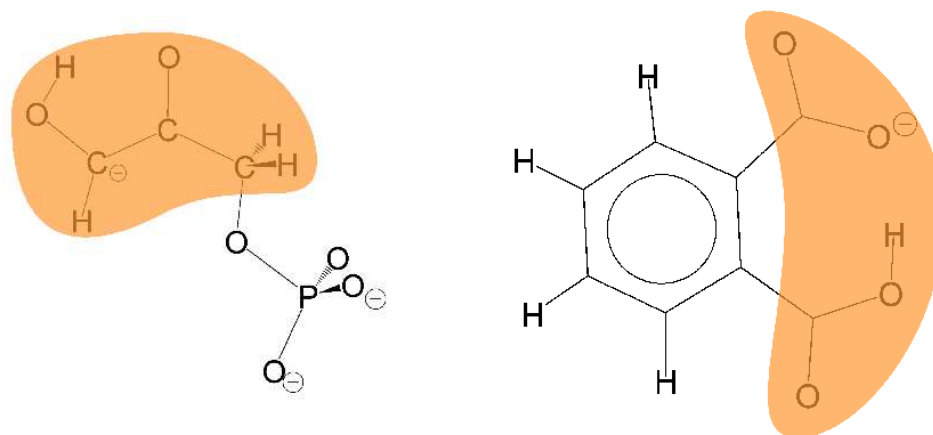


Figure 3.2: Schematic description of the studied molecular systems. The *shaded zone* in each picture corresponds to the quantum mechanical region. The left molecule is labeled **DHAP** and the right model is **PHTAL**.

The BFGS and L-BFGS procedure were implemented in a development version of ROAR[235] by Gérald Monard and Mireia Garcia-Viloca.

3.1.2 Tests on model systems

3.1.2.1 Description of systems

To test our algorithm we have chosen several chemical or biochemical systems, taken from published works, from small to medium size. We have run both geometry minimization and transition state search on every molecular system. In what follows we will describe (see figures 3.2 and 3.3) the molecular systems we have chosen as a test.

The so called **DHAP** system is the deprotonated dihydroxyacetone phosphate (the dihydroxyacetone phosphate is the substrate of triosephosphate isomerase studied by Kollman and co-workers [247]). The reaction tested is a proton transfer between the hydroxy and the ketone group. DHAP has 14 atoms and we have partitioned it in a reactive part of 9 atoms treated with the PM3 Hamiltonian and a non-reactive part of 5 atoms treated with a molecular mechanics potential. Note that in all systems pictured in Figure 3.2 and 3.3 the shaded zone corresponds to the QM region. A link atom is required for each covalent bond that joins any atom belong-

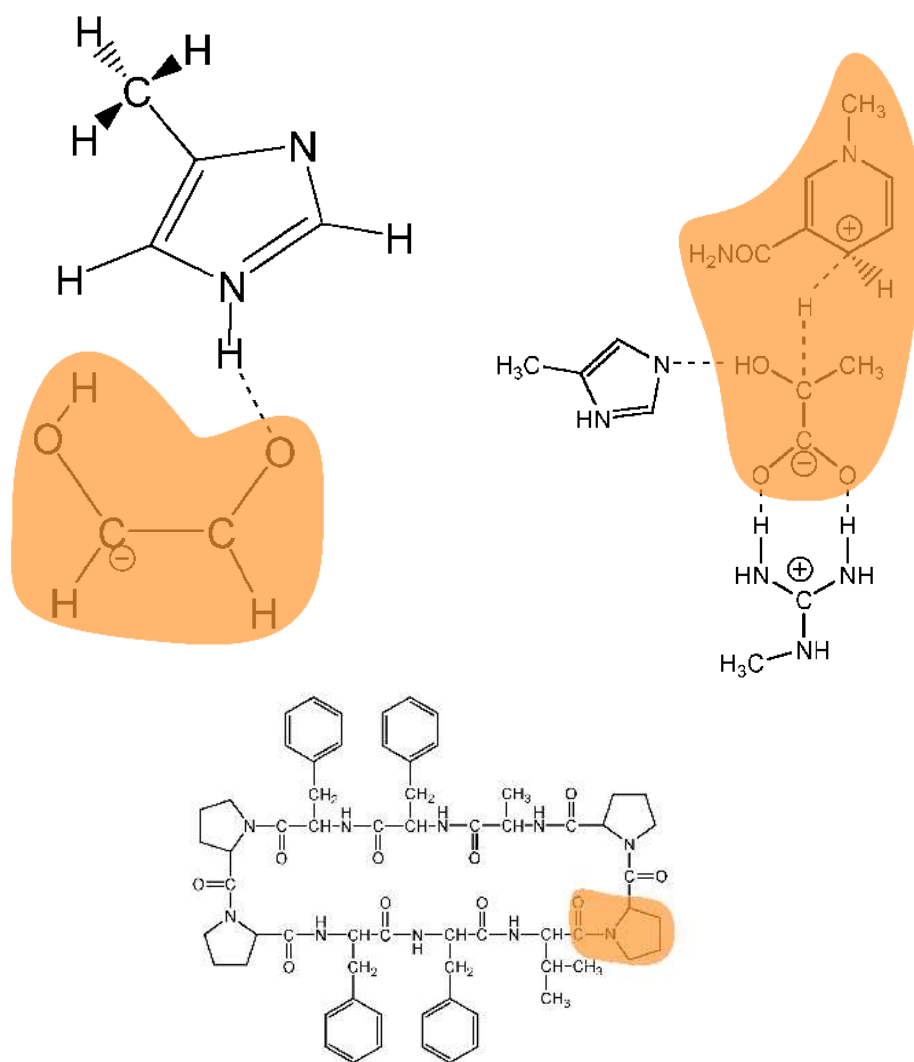


Figure 3.3: Schematic description of the studied molecular systems. The *shaded zone* in each picture The upper left picture is called **TIM**, the upper right **LDH** and the central one **ANTA**.

ing to the QM region with any atom lying in the MM (non-reactive) region. Then only one link atom has to be included in this case.

The **PHTAL** system is the phtalate anion. The reaction consists of a proton transfer between the two carboxylic groups. The whole system has 17 atoms. Seven of them, including both carboxylic groups, are treated with the AM1 semiempirical Hamiltonian, and the 10 atoms of the phenyl ring are in the MM part. Two link atoms are added.

The **TIM** system is again a model of the active site of triosephosphate isomerase studied by Cui and Karplus[127]. It is constituted by an enediolate emulating the deprotonated dihydroxyacetone phosphate. In all 19 atoms are included. 7 atoms are in the quantum part treated with the PM3 Hamiltonian, and a singly protonated imidazole ring modeling an histidine, that is 12 atoms, defines the MM part. No link atoms are required since there is no covalent bond crossing the QM/MM frontier. The reaction studied is a proton transfer between the two oxygen atoms, similarly to that studied in DHAP. In table 3.1 TIM1 and TIM2 stand for the reactant and the product of the reaction, respectively.

LDH is a model of the active site of lactate dehydrogenase enzyme studied by Andrés and co-workers [248]. The whole system is medium-sized and constituted by a total of 55 atoms. It includes a pyruvate and a nicotinamide ring involving 30 atoms in the reactive part treated with PM3, and a guanidino group and a methyl imidazole including 25 atoms for the MM part. As in the TIM system, no link atoms are required. The reaction studied is the transformation of pyruvate to lactate due to a hydride transfer between pyruvate and nicotinamide. In order to reproduce the system studied in the reference we have also performed an full QM calculation, including all 55 atoms in the quantum part.

The biggest system tested here is **ANTA** which is a decapeptide called antamanide, studied by Fischer and Karplus[146]. In that work the reaction studied was a conformational change in a proline ring. The fact that no bond is broken during this reaction enables us to study the whole system with a molecular mechanics force field. With the AMBER united atom model a total of 90 atoms are handled. When the whole system is treated classically the initial Hessian matrix has not the form shown in figure 3.1, but the full Hessian matrix is built.

In addition, a QM/MM partition has also been done on the **ANTA**

system. The QM part includes the 14 atoms from the all-atom proline ring treated with the PM3 Hamiltonian (this partition takes into account that the frontier cannot be in the peptide bond due to the fact that it has some double bond character), whereas the molecular mechanics part is the rest of the decapeptide involving 83 united atoms. Two link atoms are also included.

3.1.2.2 Results and discussion

Location of the minima:

Although nowadays several cheap and efficient minimization algorithms exist, we present here the comparative results between our algorithm and two others that are widely used. To this aim we carried out geometry minimizations with our algorithm (RFO-m-BFGS) and two quasi-Newton-Raphson algorithms. These two quasi-Newton-Raphson algorithms use the BFGS and the L-BFGS update Hessian matrix formula. These two algorithms, which were also coupled to the Roar-cp module, are labeled qNR-BFGS and qNR-L-BFGS. The results corresponding to the minimizations with the L-BFGS update formula were obtained using information from the five previous iterations. After several tests this number of iterations was proven to be the best compromise between the efficiency of the method and memory requirements.

System	QM+MM coordinates	$ g_0 ^a$	ΔE^b	qNR-L-BFGS ^c	qNR-BFGS ^c	RFO-m-BFGS(HI) ^c	RFO-m-BFGS(HF) ^c
DHAP	30 + 15	10.811	11.2	96/104	34/77	46/77	10/11
PHTAL	27 + 30	11.739	3.48	35/41	39/83	63/108	28/43
TIM1	21 + 36	5.191	0.55	121/129	68/146	81/128	37/48
TIM2	21 + 36	5.429	0.71	123/130	56/117	90/142	57/66
LDH	90 + 75	15.39	76.04	4101/4196	479/965	1852/1935	1838/1959
LDH	165 + 0	18.47	26.86	972/1014	215/433	716/809	282/302
ANTA	0 + 270	8.044	10.2	403/423	268/539	321/544	58/60
ANTA	48 + 249	124.61	815.0	3215/3340	625/1258	1949/2054	1459/1467

^a Initial gradient norm in kcal/(mol·Å)

^b Energy difference between the initial geometry and the minimum in kcal/mol

^c Number of steps/number of gradient and energy evaluations

Table 3.1: Comparative results corresponding to the location of the minima for the different systems studied

The comparative results for the search for the minima are presented in table 3.1. We are able to choose the minimization algorithm between qNR-BFGS, qNR-L-BFGS and RFO-m-BFGS. As we said the unity matrix is taken as the initial guess Hessian matrix for qNR-BFGS and qNR-L-BFGS. For the sake of comparison we used two different initial guess Hessian matrices for RFO-m-BFGS. So, RFO-m-BFGS(HI) stands for the RFO-m-BFGS algorithm with the unity matrix as an initial guess, whereas the initial Hessian matrix for the RFO-m-BFGS(HF) algorithm was calculated numerically according to equation 1.70 in page 36 building up a matrix of the form shown in figure 3.1 (except for the ANTA system when the whole system is treated classically).

In each system the minimum reached by all of the algorithms is always the same. So we just specify the energy difference between the starting point and the minimum reached. A final numerical Hessian calculation was done in order to characterize the stationary point. The convergence criterion on the root mean square (RMS) of the gradient is 10^{-3} kcal/(moli·Å), except for ANTA, which is 10^{-4} kcal/(mol·Å). We also present the number of steps and the number of energy and gradient evaluations (the energy and gradient calculations required to build up the numerical initial guess Hessian matrix are not counted). This information will give us the efficiency of every step. Note that the energy and the gradient are calculated only once each step unless the displacement vector needs to be corrected. This is why the number of steps is always smaller than the number of energy and gradient calculations, as seen in table 3.1.

No global conclusions about the compared efficiency between the different algorithms can be drawn. We can just note a general tendency because the behavior of an optimization depends not only on the algorithm but also on the intrinsic characteristics of the system (size, starting point, fixed atoms and convergence criteria). Nevertheless it can be seen that qNR-L-BFGS tends to need more steps than the other algorithms. This is due to the fact that it only works with the information of the last five preceding steps.

Comparing the results for the columns corresponding to the RFO-m-BFGS(HI) and RFO-m-BFGS(HF) algorithms, there is an evident conclusion. When an initial Hessian matrix is calculated numerically the number of steps and the energy and gradient evaluations required decrease compared to when the starting Hessian matrix is a unity matrix. It can be

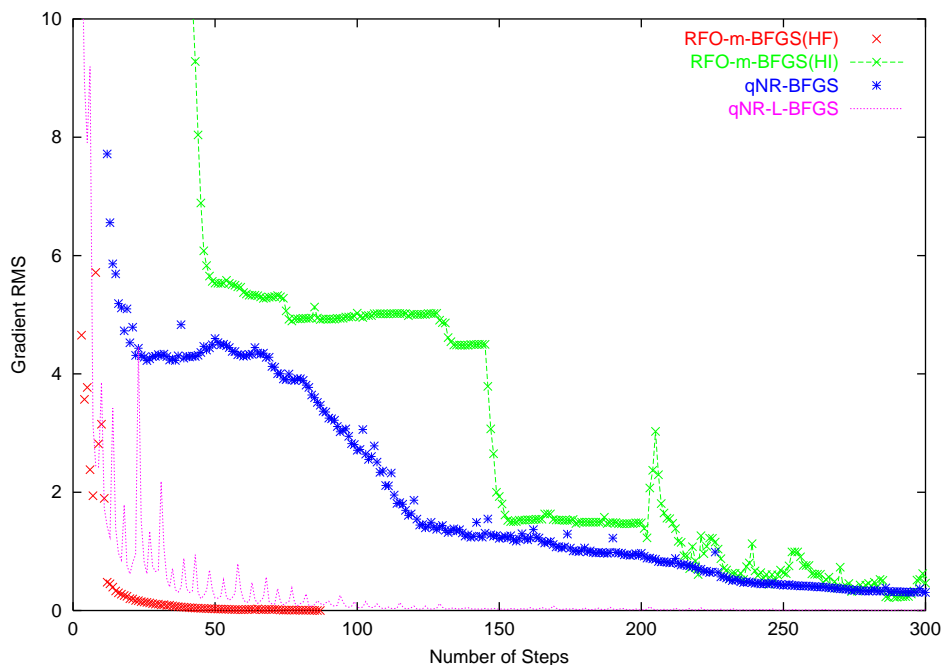


Figure 3.4: Gradient RMS ($\text{kcal mol}^{-1} \cdot \text{\AA}^{-1}$) for the QM/MM ANTA system as a function of the number of steps of minimization with four algorithms

seen that RFO-m-BFGS(HF) behaves reasonably well in comparison with qNR-BFGS. In addition, bearing in mind that the number of energy and gradient evaluations compared to the number of steps indicates the efficiency of the step, it is shown that the efficiency of an RFO-m-BFGS(HF) step is greater than that of a qNR-BFGS step because the ratio between those two numbers for qNR-BFGS is always more than 2, whereas the ratio for the RFO-m-BFGS(HF) algorithm is close to 1.

We also studied how the RMS of the gradient behaves during the minimization process. Although we present here only the QM/MM ANTA system as an illustrative example (figure 3.4) the comparative results are similar in all the systems studied. It can be seen that the RFO-m-BFGS(HF) algorithm reaches a low-gradient region faster, and it is in this quasi-converged region where it spends most of the steps. This is true even for the cases for which qNR-BFGS needs fewer steps to reach the minimum. The reason why RFO-m-BFGS(HF) reaches a low-gradient zone faster is obviously due to the information that provides the initially calculated Hessian matrix and prob-

System	QM+MM coordinates	$ g_0 ^a$	ΔE^b	RFO-Powell(HF) ^c
DHAP	30+15	6.86	22.82	56/59
PHTAL	27+30	1.18	4.03	39/54
TIM	21+36	11.17	3.70	65/88
LDH	165+0	6.78	4.19	614/636
LDH	36+129	3.24	0.87	316/392
ANTA	0+270	8.04	7.78	291/315
ANTA	48+249	2.10	1.25	421/598

^a Initial gradient norm in $kcal/(mol \cdot \text{\AA})$

^b Energy difference between the initial geometry and the minimum in $kcal/mol$

^c Number of steps/number of gradient and energy evaluations

Table 3.2: Results corresponding to the location of transition states for the different systems studied

ably the higher RFO efficiency. The reason why once in a quasi-converged RMS gradient region, RFO-m-BFGS(HF) can sometimes require many steps could be due to a recognized behavior of RFO and consequently it does not give the correct shift [245].

Location of transition states:

We report the test of the RFO-Powell algorithm with the same reactive systems as for the minima in table 3.2. We recall that for a transition-state search the BFGS formula cannot be used because, in the TS case, the \mathbf{M} matrix involved in equation 3.5 is not positive-definite.

The initial structure for the transition-state search is usually the most energetic point in a few points scan along the approximated reaction path. During the search we have to ensure that the algorithm is following the correct direction. This direction will be given by the eigenvector with the negative eigenvalue of the current Hessian matrix (Augmented Hessian in our case). In order to follow during the search the same direction, we choose the eigenvector with the maximum overlap with the followed eigenvector of the previous step.

Once the transition state is reached we have characterized the structure found by a numerical calculation of the Hessian matrix.

Overall, the RFO-Powell algorithm performs well in locating transition states. The ratio between the number of steps and the number of gradient and energy evaluations is still close to 1 as previously found during the minimization process. Our implementation allows the location of transition-

state structures, even if they are quite far away from the starting structures as depicted in table 3.2 for the DHAP system (*i.e.*, $\Delta E = 22.82 \text{ kcal/mol}$). Therefore, we can deduce from the previous results that the RFO-Powell algorithm is a solid algorithm to locate transition-state structures of systems from small to medium size, involving different ratios of QM and MM atoms, described in Cartesian coordinates, including link atoms and representing several types of chemical reactions.

3.1.3 Conclusions

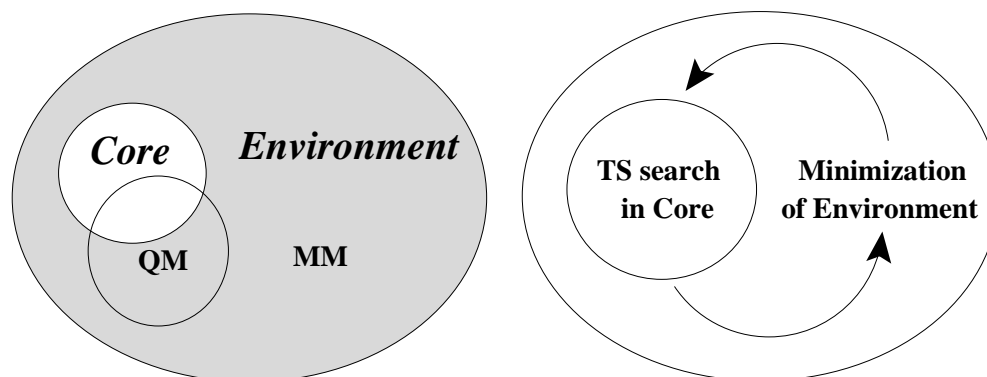
We have presented a robust algorithm able to locate minima and TS structures on QM, MM and QM/MM potential-energy surfaces. It is based on a suitable approximation to an initial full Hessian matrix, a modified BFGS formula or a Powell update formula for the location of a minimum or a transition state, respectively and the RFO.

RFO method avoids the Hessian matrix inversion required by a quasi-Newton-Raphson method. It also introduces in an automatic way a shift that preserves the current behavior of the optimization process. This algorithm has been successfully tested for a variety of chemical and biochemical systems from small to medium size. The good behavior of the algorithm presented here has encouraged us to extend it to locate minima and transition states on QM/MM potential-energy surfaces corresponding to real reactive biochemical systems including thousands of atoms.

In the next sections we will modify this algorithm in order to take into account the special problems derived from the large size of those systems, but still handling the information contained in a full Hessian matrix. In section 3.2 the method is implemented in the micro-iterative scheme. In section 3.4 the problem of the storage and diagonalization of big matrices is studied.

3.2 Micro-iterative method

In the present section we study some of the possible strategies that can be adopted in the micro-iterative method.



A compromise between the need of using a second derivative method and the drawbacks when moving many degrees of freedom is the so-called micro-iterative method[108, 160, 65, 84, 81, 161, 95, 120, 162] described in section 1.3.7.4. In this case a Newton-like method[12] is used to locate a stationary point in a small core zone where the number of atoms to be moved is low enough to avoid computational problems. The rest of atoms of the system do belong to the environment zone which is minimized with an inexpensive method. Both processes alternate one each other until consistency. In our particular case, we make usage of Rational function optimization method implemented in the last section 3.1 as a second order method for the search in the core, while for the environment the L-BFGS method is applied. Although this scheme can be applied to locate any kind of stationary points, from here on we will refer only to saddle point structures.

In the micro-iterative scheme the core and environment zones do not have to match with the QM and MM regions respectively. The quantum region is selected in a QM/MM system as the set of atoms whose interactions need to be described by a QM potential, usually when bond breaking or charge transfer is involved. This selection will define a certain PES. Conversely, the selection of the core zone is just a strategy to locate stationary points and it can change when looking for different stationary points in a reaction mechanism.

The criterion to select this core zone is not based on the interaction energy but on geometrical criteria, that is, the core zone must include the atoms whose movement may play an important role when looking for the stationary point at the current chemical step. We will see that depending on the reaction type a big core (even bigger than the QM zone) must be chosen, while in some other cases a core with few atoms is enough to reach the stationary point easily.

This section is presented in two main parts. In section 3.2.1 the micro-iterative method as a possible solution to our problem is presented, and we discuss the different options that this method can offer. In section 3.2.2 we present the results as they have been obtained from Mandelate Racemase reaction. A discussion is done on the results obtained with the different options of the method. Finally, the conclusions are presented.

3.2.1 Strategies and its implementation

3.2.1.1 Possible options in the micro-iterative method

Different possible options can be chosen in the general micro-iterative scheme, some of them, as we will see below, may be crucial in order to obtain an accurate convergence as fast as possible.

Frequency of the iterated processes:

Although some authors^[108, 160] perform a full minimization of the environment at every TS search step, it is not evident that this is the best way to proceed. We could run a full TS search until convergence before minimizing the environment again. As a matter of fact it is still an open question to know which is the most suitable alternating frequency between the two processes.

The core size:

Another important aspect is how big the core zone must be selected in order to reach convergence as fast as possible. When the core is big the coupling between the two search processes is minimized. This would tend to reduce the number of steps required to converge. On the other hand, an optimization of a larger number of atoms requires intrinsically a higher number of steps to converge, in addition a big core region implies a bigger Hessian matrix and, as a consequence, a more expensive initial Hessian calculation (see section 3.4 for a discussion about other problems related to big Hessian matrices).

The interaction between QM and MM zones:

Another option in the micro-iterative method that deserves more discussion is how to handle the interaction between the core and the environment.

The environment zone is usually bigger than the core zone and it needs to be optimized with a cheap minimizer such as conjugate gradient, ABNR or L-BFGS. Although these methods imply low memory requirements their poor efficiency has as a consequence that they need many of steps to reach convergence. Depending on the QM level the environment minimization could demand a high computational effort if a full QM/MM calculation is performed at every step.

Different strategies have been proposed to perform faster QM/MM energy evaluations [249, 101, 250, 84, 95, 251]. These strategies are based on the idea that an exact SCF evaluation of the QM zone might not be needed at each minimization step in order to calculate the energy of the system. These methods have not only been applied to the location of stationary points [84, 95], but also some work had already been done in QM/MM Monte Carlo to explore the environment configurational space avoiding a full SCF evaluation at every step of the simulation[249, 101, 250, 251].

In order to summarize the different approximations adopted to accelerate the QM/MM calculation we have again to keep in mind the general QM/MM electronic embedding scheme already explained in section 1.2.3.1

The full Hamiltonian can be expressed as:

$$\hat{H} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}} \quad (3.9)$$

where \hat{H}_{QM} is the Hamiltonian describing the set of atoms whose interactions are computed using quantum mechanics, and \hat{H}_{MM} is the molecular mechanics Hamiltonian. The crucial aspect is the calculation $\hat{H}_{\text{QM/MM}}$ that describes the interaction between the two regions treated at the two different levels

$$\hat{H}_{\text{QM/MM}} = V_{\text{QM/MM}}^{\text{van der Waals}} - \sum_i^{\text{electrons classical}} \sum_C \frac{Q_C}{r_{iC}} + \sum_K^{\text{nuclei classical}} \sum_C \frac{Z_K Q_C}{R_{KC}} \quad (3.10)$$

where Q_C are the MM charges, Z_K are the effective nuclear charges of quantum atoms, r_{iC} and R_{KC} are the distances from the MM charges to the electrons and quantum nuclei, respectively. The second term of the right-

hand of equation 3.10 describes the interaction between the MM charges Q_C and the electrons. The Q_C charges will polarize the wavefunction. So when molecular mechanics atoms are moved all the terms in equation 3.9 have to be recomputed in order to consider the effect of MM atoms on the quantum mechanics zone⁴. Indeed when the QM level is highly accurate the energy computation becomes not affordable.

ESP:

A possible attempt to speed up the full QM/MM calculation is to consider a quantum atom as a classical atom during the MM atoms movement. This can be carried out associating point charges to the quantum atoms, for example, computing the electrostatic potential fitted charges (ESP)[74] from an electron density obtained by means of a full QM/MM calculation. So, in equation 3.10 we can join the second and third term giving the expression shown in equation 3.11

$$\hat{H}_{\text{QM/MM}} = V_{\text{ESP/MM}}^{\text{van der Waals}} + \sum_K^{\text{quantum}} \sum_C^{\text{classical}} \frac{Q_{\text{ESP}K} Q_C}{R_{KC}} \quad (3.11)$$

The ESP charges for the quantum atoms will be constant during the environment minimization and they will be recomputed at the next core step or at the next full QM/MM evaluation. In this case it is evident that the core zone must always include all the quantum atoms.

In this work the ESP charges are taken from a PM3 wavefunction instead of using an *ab initio* one[84] but this will not include any systematic error in the obtained results.

It is important to note that the ESP/MM strategy will reach a geometry that is not a stationary point on the real QM/MM surface. It is not evident that the interaction between classical charges and ESP charges is equivalent to the second and the third term in equation 3.10. A better approximation is a method which is based on the original QM/MM expression that will be called 1SCF method from here on in this paper.

1SCF:

Note that in the QM/MM equation 3.10 the QM/MM interaction Hamiltonian only contributes to the total one-electron core quantum Hamiltonian but not to the two-electron integrals. This means that for a fixed geometry

⁴ This is an important difference with the IMOMM-ONIOM scheme [108, 162] where no explicit polarization of the MM part is included; and probably this is why a full and yet cheap minimization was done at every core step

of the quantum atoms if we move the MM atoms the two-electron integrals have not to be recomputed and the one-electron integrals can be easily updated at the new Q_C configuration

If we save the last converged wave function (Ψ_{frozen}) and perform only one SCF cycle the Fock matrix will not be diagonal but we will obtain a good and cheap approximation to the exact QM/MM energy (see equation 3.12)

$$E_{1\text{SCF}} = \langle \Psi_{\text{frozen}} | \hat{H}_{\text{QM}} - \sum_i^{\text{electrons}} \sum_C^{\text{classical}} \frac{Q_C}{r_{iC}} | \Psi_{\text{frozen}} \rangle + \hat{H}_{\text{MM}} + V_{\text{QM/MM}}^{\text{van der Waals}} + \sum_K^{\text{nuclei}} \sum_C^{\text{classical}} \frac{Z_K Q_C}{R_{KC}} \quad (3.12)$$

Our proposal is to use this method to save computing time during the minimization of the environment in the micro-iterative scheme. The efficiency of this strategy has been evaluated by Evans and Truong [252] in a Monte Carlo sampling of the environment. Keeping the wavefunction frozen will be a good approximation as long as the perturbation of the classical charges Q_C is small. That is, as long as the distribution of charges Q_C during the minimization do not change too much or they are not too close to the QM part to polarize the real $|\Psi\rangle$ in a very different manner, the 1SCF method will work well.

3.2.1.2 Implementation

The implementation of the micro-iterative method described here must be flexible enough to permit the different possibilities on the sizes between the two regions, its interaction and on the number of micro-iterations at every alternating process.

Schematically the flux diagram can be depicted as follows:

- Firstly the geometry is read from the input. The set of atoms that belong to the core and environment zones must be specified (the hydrogen link atom will always be at zone where its QM-host belongs). When any of the described approximated interaction between the core and the environment is chosen the core zone must include at least all the QM zone. In addition any geometrical restriction can be imposed.

- A first full energy QM/MM is performed in order to evaluate the gradient norms of the different zones.
- Decide what process to do first (TS search in core / minimization of environment) and how many iterations can be carried out.
- Here starts a conditional loop that will not stop until both optimizations fulfill the convergence criteria or the maximum number of iterations is reached.
- When TS search in core starts the procedure is the same as described in section 3.1. There will be only one initial Hessian calculation and this matrix will be updated during the whole process. It will be a full Hessian rather than the one (square+vector) depicted in figure 3.1 in page 117.
- For the minimization the L-BFGS optimization is activated. When ESP/MM or 1SCF/MM approximations are set a first full QM/MM energy must be carried out in order to calculate the ESP charges from the electron density or to save the one-electron and two-electron integrals.

For a more detailed description of the implementation and subroutines see the appendix section B in page 201.

3.2.2 Tests on Mandelate Racemase

Two model reactions:

We have performed several series of tests on the same model of Mandelate Racemase enzyme studied in section 2.3. In particular, we took as starting points the mechanism found with the propargylglycolate substrate. Between the different steps of the three mechanisms we have selected two of them as representative examples of opposite cases. The first one is step 1 of mechanism II, which consists of a proton transfer from Glu317 to a carboxylate oxygen of propargylglycolate. The second one is step 4 of mechanism I, which essentially involves the configuration inversion of the stereogenic carbon of propargylglycolate.

The analysis of the components of the transition vectors associated with the respective transition state structures shows the difference between both steps, which, for the sake of clarity, will be called from here on the proton

transfer step and the carbon configuration inversion step. Few atoms are expected to rearrange during the proton transfer step, whose transition vector mainly includes the motion of the transferring proton and a small contribution of the proton donor oxygen atom of Glu317. That is, the changes along this step are restricted to a small local zone involving a very reduced number of atoms. Conversely, many atoms are involved in the components of the transition vector corresponding to the carbon configuration inversion step: the stereogenic carbon atom and the atoms of propargylglycolate directly attached to it, along with the proton of His297 that will be transferred to the substrate. In addition, some participation of Lys166 at the first stages of the step is also expected. Then, this step involves a rather global motion of several groups and residues including an important number of atoms.

Procedure:

In this test series we have looked for the transition state structures of the two above mentioned steps starting from the highest energy point of the profile built up along the adequate reaction coordinate as described in section 2.3.

We have preferred to test our algorithm on the location of transition states rather than for minima. The reason is because although the algorithm can be applied to both cases, the transition state search is always more problematic and this will help us to discriminate and discuss between the several options considered here.

In what follows we compare the results obtained with the micro-iterative method using the different options mentioned above. Prior to this we have to stress that some uncertainty accompanies the quantitative value of most of the results obtained in this work.⁵ The number of iterations required to reach the transition state structures is an example. The convergence of each minimization of the environment or each core search is fulfilled when the suitable convergence criterion of the gradient norm is reached. However, sometimes the low-gradient region can be attained fast, but the algorithm can spend some time in this quasi-converged region conferring some aleatory character to the crossing of the gradient norm threshold (RMS=0.005 kcal/(mol·Å)). Indeed this fact also affects the total CPU time spent in each transition state structure location. In spite of that, we think that several important qualitative trends emerge from the analysis of our

⁵ This relative uncertainty will be exemplified in the appendix section testing different compilers and several optimization levels in compilation

core / env	(iter.core/ iter.env)	energy ^a	totalCPU ^b	HessCPU ^b	%Hess
3 / 1295	(186/1859)	-7453.26	17772	150	1
12 / 1286	(219/2056)	-7453.26	20317	604	3
23 / 1275	(172/725)	-7453.27	8994	1158	13
80 / 1218	(174/479)	-7453.28	9791	4017	41
138 / 1160	(206/765)	-7453.27	15551	6989	45
194 / 1104	(187/596)	-7453.28	17084	10002	59
399 / 899	(205/530)	-7453.28	27926	20263	73

^a In kcal/mol

^b In seconds

Table 3.3: Results testing the core/environment size for the proton transfer step (full TS search)

numerical results. The tests were run in a 2.0 GHz Pentium IV computer.

3.2.2.1 Results on the core size

Proton transfer reaction:

First we have carried out the series of tests corresponding to the proton transfer step when a full TS search in the core until convergence is performed before minimizing again the environment. Each test corresponds to a micro-iterative search of the transition state structure using a particular core/environment distribution of the 1298 moving atoms.

The results are presented in Table 3.3. From left to right the different columns indicate, respectively, the number of atoms included in the core/environment regions, the number of iterations made in each zone, the final energy of the located transition state structure, the total CPU time of the location, the CPU time devoted to calculate the initial Hessian and its percentage over the total CPU time.

The first test just includes 3 atoms in the core: the shifting proton and both the proton donor and acceptor oxygen atoms. The successive tests progressively increase the number of atoms around the first three that are incorporated in the core region. The third column shows that the final energy has always the same value, what indicates that, in this case, whatever core / environment partition which includes the three atoms directly involved in the proton transfer leads to the right transition state structure. However, each test behaves in a different way as for the number of iterations

and the CPU time. When the core is small a lot of environment iterations are required due to both the huge coupling between the two regions and the great size of the environment region. Conversely, the coupling is low when the core is big (what in addition implies a small environment region), what reduces significantly the number of environment iterations.

It seems that it exists a range of intermediate core sizes which involves a reduced number of core iterations. This probably comes from a compromise between a lower core/environment coupling and the efficiency of the RFO method in handling a progressively increasing number of core atoms as the size of the core grows.

These last results show clearly that when the core/environment partition is selected in an adequate way, the part of the Hessian matrix of the entire system corresponding to the core-core and environment-environment diagonal blocks are numerically much more important than the core-environment non-diagonal ones. A Hessian of this type is the most suitable to be used for any type of optimization[253]. And this fact justifies the efficiency of the micro-iterative method.

As for the CPU time, there is a good correlation between the total number of iterations and the difference between the total CPU time and the time devoted to the calculation of the initial Hessian. So, for instance, the core/environment partition 80/1218 converges with the smallest number of total iterations (653) lasting the most reduced CPU time (5774 s), the Hessian calculation excluded. On the other hand, as expected, the CPU time corresponding to the calculation of the initial Hessian increases monotonically very fast with the core size, requiring the 1% of the total CPU time when just 3 atoms are included in the core, but as much as the 73% for a core with 399 atoms.

Joining all the factors described above, it can be easily understood why it exists an interval of medium core sizes that minimizes the total CPU time, as seen in the fourth column, the optimal partition being 23/1275 (within the discrete, limited series studied here).

Carbon configuration inversion reaction:

The results of the series of tests corresponding to the carbon configuration inversion step when a full TS search in the core until convergence is performed before minimizing again the environment are exhibited in Table 3.4. The first test has 7 atoms in the core, including the stereogenic carbon atom

core / env	(iter.core/ iter.env)	energy ^a	totalCPU ^b	HessCPU ^b	%Hess
7 / 1291	(145/2061)	-7443.25	19041	346	2
17 / 1281	(150/1965)	-7442.31	18677	840	4
23 / 1275	(140/1862)	-7442.31	18172	1136	6
34 / 1264	(102/1524)	-7442.32	15453	1686	11
80 / 1218	(113/1827)	-7442.34	20401	3978	19
138 / 1160	(62/731)	-7440.50	17809	9025	51
194 / 1104	(102/891)	-7440.49	23960	12826	54
643 / 655	(291/508)	-7440.53	58303	42504	73

^a In kcal/mol

^b In seconds

Table 3.4: Results testing the core/environment size for the carbon configuration inversion step (full TS search)

and several atoms of His297 and Lys166. As for the number of iterations and the CPU time these results follow the same trends as the proton transfer step (Table 3.3). The difference lies on the final energies. It can be seen that three distinct sets of energies are obtained, with a difference of roughly 3 kcal/mol between the test with 7 core atoms and the tests of 138 or more core atoms. This indicates that three different progressive approximations to the transition state structure are located, a number of 138 or more core atoms (in this discrete, limited series) being required to converge to the right transition state structure as the core size increases. (We consider that the right transition state structure of the corresponding potential energy valley is the one that would be obtained including all the 1298 moving atoms in the core).

This fact highly contrasts with the proton transfer step where a core of three atoms is enough. This is a consequence of the global character of the changes involved in the carbon configuration step (in front of the local character of the proton transfer step), whose TS search requires a core including all the atoms that participate significantly in the reaction coordinate to be carried out successfully. Then, at first glance, it would seem that the best core/environment partition could be 34/1264 which requires 15453 s of total CPU time. However, this partition does not lead to a good enough approximation to the right transition state structure yet. Then, the optimal partition for this step is rather 138/1160 which lasts more total

CPU time (17809 s) due to the bigger CPU time required to calculate the initial Hessian, although it involves clearly fewer iterations and spends less time in the location of the transition state structure once the initial Hessian has been calculated. Indeed this partition leads to the right TS. That is, as the different atoms which define the reaction coordinate are incorporated in the core, the description of the TS of the corresponding potential energy valley is progressively improved, until a good enough approximation to the right TS is reached.

Finally, it has to be mentioned that the high total CPU time of the test corresponding to the partition 643/655 is due not only to the CPU time required to calculate the initial Hessian, but also to the time employed to partially diagonalize this big matrix Hessian at each iteration⁶.

3.2.2.2 Results on the frequency of the iterated processes

The results of the series of tests corresponding to the proton transfer and the carbon configuration inversion steps, when a full minimization of the environment is carried out at every TS search step in the core, are given in Tables 3.5 and 3.6, respectively. Comparing with Tables 3.3 and 3.4, it can be observed that the minimization of the environment at every TS search step in the core clearly reduces the number of core iterations, although in turn tends to augment the number of environment iterations. As a result, the difference between both strategies does not seem to be very significant, although the minimization of the environment at every TS search step rather tends to increase the total number of iterations in the two reaction steps studied here. As expected, the energy column shows that the final transition state structure for each core/environment partition is independent on the strategy used.

Our experience shows that since the location of TS points is complicated, the strategy of a full TS search before a next minimization is recommendable to unsure at a primary stage of the search that the final convergence will be reached. On the contrary, when the minimization is performed at every TS search step we have no information about the desired success of the search until the end of the whole process.

⁶ The partial diagonalization process extract the 5 lowest eigenpairs to calculate the displacement in RFO and to check the curvature in the current region of the PES.

core / env	(iter.core/iter.env)	energy ^a
3 / 1295	(96/1896)	-7453.26
12 / 1286	(105/2075)	-7453.25
23 / 1275	(42/1211)	-7453.27
80 / 1218	(46/976)	-7453.29
138 / 1160	(105/1088)	-7453.27
194 / 1104	(62/752)	-7453.27
399 / 899	(80/793)	-7453.28

^a In kcal/mol

Table 3.5: Results testing the core TS search / environment minimization switch for the proton transfer step. At every TS search step a full minimization of the environment is performed.

core / env	(iter.core/iter.env)	energy ^a
7 / 1291	(88/2092)	-7443.25
17 / 1281	(143/1923)	-7442.31
23 / 1275	(115/1908)	-7442.31
34 / 1264	(66/1631)	-7442.32
80 / 1218	(69/1861)	-7442.33
138 / 1160	(52/758)	-7440.50
194 / 1104	(73/920)	-7440.50
643 / 655	(117/710)	-7440.54

^a In kcal/mol

Table 3.6: Results testing the core TS search / environment minimization switch for the carbon configuration inversion step. At every TS search step a full minimization of the environment is performed.

core / env	(iter.core/iter.env)	energy ^a	totalCPUtime ^b
80 / 1218	(184/557)	-7453.28	8677
140 / 1158	(164/658)	-7453.28	12005
194 / 1104	(183/616)	-7453.27	14776
399 / 899	(190/525)	-7453.28	26250
643 / 655	(229/509)	-7453.29	44321

^a In kcal/mol

^b In seconds

Table 3.7: Results testing the QM(1SCF)/MM approach for the proton transfer step (full TS search)

3.2.2.3 Results on the interaction between QM and MM zones

ESP/MM results:

We have tested the ESP/MM approach during the environment minimization in the proton transfer step using different core/environment partitions. A full TS search in the core until convergence has been performed before minimizing again the environment. Unfortunately most of the calculations are unable to reach convergence. The few ones that converge require a huge amount of iterations and total CPU time. For instance, for the partition 138/1160 the ESP/MM approach needs 947/4942 core/environment iterations and 52690 s of total CPU time, in front of 206/765 iterations and 15551 s when the QM/MM calculations are carried out. Probably the problems in the convergence of the micro-iterative method are due to the fact that the TS search in the core and the environment minimization follow different potential energy surfaces.

QM(1SCF)/MM results:

We present now the results of the series of tests in which only one SCF cycle has been performed to approximate the QM/MM energy during the minimization of the environment. A full TS search in the core until convergence has been performed before minimizing again the environment. The cases corresponding to the proton transfer and carbon configuration inversion steps are given in Tables 3.7 and 3.8, respectively. Although this 1SCF option has no definite effect on the total number of iterations, it is clear that the total CPU time is noticeably smaller. Indeed this is due to the fact that each 1SCF energy evaluation during the environment minimization is faster than a complete QM/MM energy calculation. On the other hand,

core / env	(iter.core/iter.env)	energy ^a	totalCPUtime ^b
80 / 1218	(426/2426)	-7442.34	19842
140 / 1158	(84/839)	-7440.50	12407
194 / 1104	(105/841)	-7440.49	15525
643 / 655	(327/545)	-7440.54	47773

^a In kcal/mol

^b In seconds

Table 3.8: Results testing the QM(1SCF)/MM approach for the carbon configuration inversion step (full TS search)

the energy column shows that the 1SCF option leads to the same transition state structures as the complete QM/MM calculations. These results confirm that this 1SCF approach is reliable and provides a cheaper alternative to be used in the micro-iterative method.

3.2.3 Conclusions

In this section we have studied how the efficiency of the micro-iterative method for locating transition state structures in QM/MM potential energy surfaces of very high dimensionality can be optimized. Several series of calculations testing different options have been run on the potential energy surfaces corresponding to two of the reaction steps of the mechanisms by which Mandelate Racemase enzyme catalyzes the reversible interconversion of the (S)- and (R)-enantiomers of the substrate propargylglycolate. A total of 3962 atoms constitute the whole QM/MM model, 1298 of which are moved during the location of the transition state structures.

The micro-iterative method divides the whole system in two parts, a core zone where an accurate second order search (a Rational function optimization method in our case) of the transition state structure is done, and an environment that is kept minimized with a cheap first order method (an L-BFGS method in this case). Our results show that the core has to include at least all the atoms that participate significantly in the reaction coordinate of the corresponding reaction step. Otherwise, the right transition state structures are not reached. Indeed, this threshold size of the core depends a lot on each particular reaction step. Beyond this minimum core size, there is an interval of medium core sizes that minimizes the total CPU time. This arises from a compromise among a lower core/environment coupling, the efficiency

of the Rational function optimization method in handling a progressively increasing number of core atoms, and the monotone augment of the CPU time required to calculate the initial Hessian matrix, as the core size grows. As a consequence, the use of a core as great as possible is not advised. In other words, above that threshold size of the core, which depends of the relevant motions taking place during the corresponding chemical step, it is not true that the larger the core size the more efficient the TS search.

A considerable amount of CPU time can be saved if only one SCF cycle is performed to evaluate the potential energy during the environment minimization. This option is clearly faster than a complete QM/MM energy calculation and leads to the right transition state structures. Conversely, the use of the ESP charges to simplify the calculation of the interaction energy between the QM and the MM regions seems to be less efficient.

Finally, we have to remark that the location of the transition state structures in enzyme catalysis is needed to define reliable reaction pathways along which a set of generalized free energies can be calculated. In this sense, an extreme accuracy in the geometrical or energetic parameters of the transition state structures is not required. We just need that the right pathway can be built up starting from the transition state structure, so avoiding the use of reaction paths that run on wrong potential energy valleys. However, when a QM(*ab initio*)/MM scheme is used, a configurational sampling becomes prohibitive. Therefore, an accurate location of stationary points will be the only strategy to provide meaningful results.

3.3 How important is an accurate optimization

At this point of the chapter we have already designed, implemented and tested a micro-iterative method capable to find accurate stationary points. In section 2.3 we have given an energetic profile for the different mechanisms of two substrates of Mandelate Racemase. Now we must check the difference between the profile sketched with the "coordinate scan" method and the results coming from the more accurate micro-iterative method.

In this section the need to use the second derivatives direct algorithm to refine the location of transition state structures obtained in enzymatic systems will be analyzed. The 25 approximate QM/MM transition state structures previously found by means of a reaction coordinate approach for the three mechanisms of racemization of the mandelate and propargylglycolate by Mandelate Racemase enzyme have been refined using the micro-iterative method.

We want to show that the refinement of transition state structures is especially useful to assure that a structure, found as the highest potential energy point on a profile depicted by a particular reaction coordinate, lies in the correct quadratic region. This is more important in those steps of the enzymatic process where the selected reaction coordinate may not reflect quite accurately the geometrical changes taking place in the active site.

The point now is how precise the location of the transition state structures has to be to produce a reliable reaction path. Is the location of the transition state structure as the maximum energy point along an energy profile built up as a function of a conveniently chosen reaction coordinate sufficient or after that the transition state structures have to be refined? The purpose of this section is to use the different reaction channels we have previously found[1] for the racemization of mandelate and propargylglycolate by Mandelate Racemase enzyme to shed some light to that question.

Furthermore, we advance that an accurate exploration of the PES is a useful preliminary task for free energy computations. As we will see, the choice of a geometrical reaction coordinate for a PMF calculation cannot be evident without the previous knowledge of the PES.

3.3.1 Procedure

The common procedure has been to take the structures obtained with the reaction coordinate method as the input for the micro-iterative method. After the successful search a second derivatives analysis has been done to check the correct number of negative eigenvalues. Even though, there were always very small frequencies impossible to eliminate and probably caused by the typical *coarse-grained* surface inherent of enzymatic systems.

Obviously the method for the potential energy used here is the same used in section 2.3.

The options on the micro-iterative method described in the last section 3.2 are the following:

Core size The core size has been selected in particular for every reaction. For most of proton transfer reactions where few atoms are expected to rearrange the core size is smaller than the quantum part. A good strategy is that when the search becomes difficult we can select a very small core (*e.g.* donor, acceptor and transferring proton) and perform, keeping the rest frozen, a fast TS search on this small system. Then this will be a very good initial structure for the micro-iterative refinement.

For the carbon configuration inversion the core includes all the residues that have QM atoms (as we already said, this reaction needs a bigger core size).

Frequency of both process We always recommend a full TS search in the core before a new minimization of environment is restarted.

Interaction between the two zones Since these PM3/AMBER calculations are not very expensive we do not select any of the approximations described in section 3.2. If the QM level became more expensive the QM(1SCF)/MM would be the most suitable strategy. Moreover, a full QM/MM evaluation at every minimization step permits the inclusion of some quantum atoms in the environment zone, in other words, a more flexible core size selection.

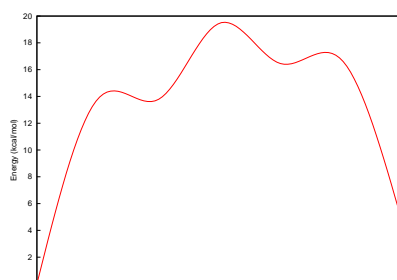


Figure 3.5: New energy profile for mechanism III for the racemization of mandelate substrate after the refinement.

3.3.2 Comparison between Mandelate Racemase mechanism structures

All the structures found by the micro-iterative method correspond to the same chemical step found with the coordinate scan method described in section 2.3. However, we have been able to locate another intermediate for the mechanism III in mandelate substrate. The energetic profile would be as figure 3.5. So in this new case the mechanism III turns out to be a symmetric mechanism step. The central TS is the configuration inversion $S \rightarrow R$, and the two intermediates are due to a weak stabilization of the carbanion with the conjugate acid. These shallow minima have few chemical meaning and they probably due to a recognized failure of PM3 semiempirical Hamiltonian. In conclusion, they will not be included in the following discussion and then the mechanism III for mandelate will be considered practically as a one-step mechanism.

3.3.2.1 Energetic comparison

The potential energy barriers obtained by using the reaction coordinate and the micro-iterative methods for the racemization of mandelate and propargylglycolate are shown in Tables 3.9 and 3.10, respectively. No significant differences are observed among the two sets of energy barriers in the case of mandelate, no other than those that could be practically removed by performing a scanning with more intermediate points along the reaction coordinate. On the contrary, in the case of propargylglycolate important differences exist among some energy barriers obtained with the two methods. These differences are mainly centered in step 4, when stereogenic center

Structures	Mech. I	Mech. II	Mech. III
S	0.00	0.00	0.00
TS1	17.77 (18.24)	17.69 (17.78)	
TS2	19.52 (19.65)	14.77 (14.46)	
TS3	20.04 (20.06)	14.55 (14.95)	
TS4	22.54 (22.56)		20.19 (19.50)
TS5	25.15 (25.75)	23.57 (23.83)	
TS6	27.22 (27.28)	28.14 (28.18)	
R	6.74	6.74	4.63

Table 3.9: Potential energies (kcal/mol) for the racemization of mandelate using the reaction coordinate method. In brackets, the corresponding values obtained from the micro-iterative method.

Structures	Mech. I	Mech. II	Mech. III
S	0.00	0.00	0.00
TS1	15.90 (19.66)	11.83 (11.24)	
TS2	19.88 (19.92)	15.33 (15.32)	
TS3	22.18 (22.20)	16.96 (16.99)	
TS4	30.12 (22.20)	28.11 (20.68)	22.05 (21.98)
TS5	23.16 (22.60)	19.23 (20.08)	
TS6	23.89 (24.35)	26.45 (24.59)	
R	3.34	3.34	3.34

Table 3.10: Potential energies (kcal/mol) for the racemization of propargylglycolate using the reaction coordinate method. In brackets, the corresponding values obtained from the micro-iterative method.

C_α changes its configuration. Another significant discrepancy can be seen at the TS1 of the mechanism I. A geometric analysis of the corresponding transition state structures will shed some light to this point.

3.3.2.2 Geometric comparison

The more relevant distances for the transition state structures corresponding to the racemization of mandelate and propargylglycolate according to mechanisms I, II and III (located using the reaction coordinate method and the micro-iterative method) are presented in Tables 3.11, 3.12 and 3.13, respectively.

Mechanism I	TS1	TS2	TS3	TS4	TS5	TS6
mandelate						
Lys166N-H ₁₆₆	1.81 (1.81)	1.21 (1.27)	1.09 (1.09)	1.01 (1.01)	1.00 (1.00)	1.00 (1.01)
H ₁₆₆ -C _α	1.16 (1.16)	1.48 (1.46)	1.67 (1.67)	2.45 (2.47)	2.76 (2.76)	2.85 (2.85)
C _α -H ₂₉₇	3.28 (3.29)	3.45 (3.44)	3.12 (3.13)	1.99 (1.95)	1.49 (1.45)	1.18 (1.18)
H ₂₉₇ -NHis297	1.02 (1.02)	1.02 (1.02)	1.00 (1.00)	1.02 (1.03)	1.18 (1.31)	1.76 (1.76)
propargylglycolate						
Lys166N-H ₁₆₆	1.84 (1.85)	1.19 (1.23)	1.13 (1.13)	1.01 (1.00)	1.01 (1.01)	1.01 (1.01)
H ₁₆₆ -C _α	1.16 (1.16)	1.50 (1.47)	1.58 (1.58)	2.69 (2.38)	2.66 (2.70)	2.76 (2.73)
C _α -H ₂₉₇	3.81 (3.82)	3.72 (3.73)	3.31 (3.33)	1.61 (2.02)	1.51 (1.45)	1.19 (1.18)
H ₂₉₇ -NHis297	1.03 (1.03)	1.03 (1.03)	1.00 (1.00)	1.10 (1.02)	1.16 (1.31)	1.74 (1.74)

Table 3.11: More relevant distances (Å) for the transition state structures corresponding to the racemization of mandelate and propargylglycolate according to mechanism I using the reaction coordinate method. The same distances obtained using the micro-iterative method are given in brackets.

Mechanism II	TS1	TS2	TS3	TS4	TS5	TS6
mandelate						
Lys166N-H ₁₆₆	1.78 (1.78)	1.25 (1.28)	1.09 (1.09)		1.00 (1.01)	1.01 (1.00)
H ₁₆₆ -C _α	1.16 (1.16)	1.47 (1.46)	1.66 (1.66)		2.68 (2.68)	2.70 (2.77)
C _α -H ₂₉₇	3.19 (3.20)	3.35 (3.20)	3.15 (3.05)		1.40 (1.42)	1.18 (1.19)
H ₂₉₇ -NHis297	1.02 (1.02)	1.02 (1.01)	1.01 (1.00)		1.49 (1.37)	1.74 (1.71)
propargylglycolate						
Lys166N-H ₁₆₆	1.84 (1.85)	1.19 (1.19)	1.15 (1.15)	1.00 (1.01)	1.00 (1.01)	1.01 (1.01)
H ₁₆₆ -C _α	1.16 (1.16)	1.51 (1.51)	1.55 (1.55)	2.81 (2.47)	2.72 (2.71)	2.80 (2.80)
C _α -H ₂₉₇	3.82 (3.80)	3.67 (3.68)	3.41 (3.42)	1.60 (1.99)	1.40 (1.43)	1.20 (1.19)
H ₂₉₇ -NHis297	1.02 (1.02)	1.02 (1.02)	1.00 (1.00)	1.09 (1.02)	1.52 (1.34)	1.71 (1.71)

Table 3.12: More relevant distances (Å) for the transition state structures corresponding to the racemization of mandelate and propargylglycolate according to mechanism II using the reaction coordinate method. The same distances obtained using the micro-iterative method are given in brackets.

Mechanism III	TS
mandelate	
Lys166N-H ₁₆₆	1.03 (1.01)
H ₁₆₆ -C _α	2.10 (2.26)
C _α -H ₂₉₇	1.78 (2.09)
H ₂₉₇ -NHis297	1.06 (1.01)
propargylglycolate	
Lys166N-H ₁₆₆	1.07 (1.02)
H ₁₆₆ -C _α	1.76 (2.15)
C _α -H ₂₉₇	2.06 (2.12)
H ₂₉₇ -NHis297	1.02 (1.01)

Table 3.13: More relevant distances (\AA) for the transition state structures corresponding to the racemization of mandelate and propargylglycolate according to mechanism III using the reaction coordinate method. The same distances obtained using the micro-iterative method are given in brackets.

Since many atoms move in each step, we have also compared the positions of the main residues of the active center at the transition state structures located using the two methods. To this aim we have calculated the root mean square (RMS) of the difference between the coordinates of the atoms at the transition state structures obtained employing the reaction coordinate method and the structures located by means of the micro-iterative method. These RMS values are shown in tables 3.14, 3.15 and 3.16 for mechanisms I, II and III, respectively. The rest of the residues of the active center give lower values of RMS and are not included in the tables.

Mechanism I	TS1	TS2	TS3	TS4	TS5	TS6
mandelate						
Substrate	3.38(-03)	5.79(-03)	9.43(-04)	1.45(-02)	9.88(-03)	2.52(-03)
Lys166	1.88(-03)	7.00(-03)	6.42(-04)	2.43(-03)	7.73(-03)	6.25(-04)
His297	9.81(-04)	9.08(-04)	5.85(-03)	3.98(-03)	1.96(-02)	1.01(-03)
Lys164	2.49(-02)	1.38(-03)	6.73(-04)	2.89(-03)	3.02(-03)	3.65(-03)
Glu317	6.08(-03)	1.82(-03)	4.68(-04)	1.06(-03)	3.33(-03)	2.63(-03)
Asp195	2.90(-03)	1.43(-03)	6.34(-04)	3.51(-03)	1.62(-03)	7.51(-04)
Glu221	4.74(-03)	2.03(-03)	1.36(-03)	2.33(-03)	2.33(-03)	2.67(-03)
Glu247	3.81(-03)	4.22(-03)	1.74(-03)	7.58(-03)	7.24(-03)	1.04(-03)
propargylglycolate						
Substrate	1.96(-02)	4.79(-03)	7.87(-04)	2.79(-01)	4.79(-02)	9.47(-03)
Lys166	6.72(-03)	4.23(-03)	6.30(-04)	1.99(-01)	1.56(-02)	3.21(-03)
His297	1.46(-02)	1.32(-03)	8.16(-03)	5.82(-02)	2.65(-02)	6.95(-03)
Lys164	1.81(-02)	1.42(-03)	2.63(-04)	1.43(-02)	1.73(-02)	3.53(-02)
Glu317	1.31(-02)	9.10(-04)	6.13(-04)	2.76(-02)	1.76(-02)	5.23(-03)
Asp195	1.84(-02)	1.16(-03)	8.97(-04)	4.10(-02)	3.25(-03)	9.72(-03)
Glu221	8.65(-03)	1.94(-03)	8.86(-04)	6.74(-02)	3.27(-02)	1.54(-02)
Glu247	1.00(-02)	3.72(-03)	3.72(-03)	1.26(-01)	2.19(-01)	1.83(-02)

Table 3.14: RMS (\AA) for the transition state structures corresponding to the racemization of mandelate and propargylglycolate according to mechanism I. See text. Powers of ten in parenthesis.

Mechanism II	TS1	TS2	TS3	TS4	TS5	TS6
mandelate						
Substrate	4.47(-03)	1.43(-01)	5.99(-03)		7.76(-03)	5.95(-02)
Lys166	2.04(-03)	1.58(-01)	4.73(-03)		5.39(-03)	1.21(-02)
His297	2.06(-03)	3.00(-01)	1.44(-02)		2.10(-02)	3.22(-02)
Lys164	1.85(-03)	1.23(-02)	4.58(-03)		1.07(-03)	3.09(-02)
Glu317	5.32(-03)	8.88(-02)	8.38(-03)		3.91(-03)	6.46(-02)
Asp195	1.77(-03)	1.52(-02)	2.01(-03)		2.51(-03)	1.15(-02)
Glu221	3.73(-03)	3.07(-02)	3.38(-03)		1.79(-03)	2.50(-02)
Glu247	2.86(-03)	3.87(-01)	1.39(-02)		1.29(-02)	8.42(-03)
propargylglycolate						
Substrate	4.78(-02)	1.37(-03)	1.03(-03)	1.92(-01)	2.36(-02)	2.61(-02)
Lys166	3.05(-02)	2.93(-03)	8.90(-04)	5.46(-02)	8.24(-03)	5.74(-03)
His297	1.71(-02)	1.25(-03)	9.23(-03)	1.76(-02)	2.99(-02)	1.86(-02)
Lys164	1.99(-02)	2.28(-03)	7.47(-04)	8.68(-03)	3.50(-03)	1.81(-02)
Glu317	4.37(-02)	8.64(-04)	1.00(-03)	2.15(-02)	1.34(-02)	4.57(-02)
Asp195	1.70(-02)	1.20(-03)	8.24(-04)	1.69(-02)	3.59(-03)	9.98(-03)
Glu221	1.97(-02)	2.23(-03)	8.68(-04)	3.65(-02)	2.26(-03)	1.98(-02)
Glu247	8.22(-03)	2.81(-03)	1.62(-03)	1.39(-01)	6.44(-03)	2.15(-02)

Table 3.15: RMS (\AA) for the transition state structures corresponding to the racemization of mandelate and propargylglycolate according to mechanism II. See text. Powers of ten in parenthesis.

It can be seen that there are not significant divergences between both sets of transition state structures in most of the steps of the mechanisms. The discrepancies in general have no important consequence on the potential energy barriers: See, for instance, that a deviation of even 0.15 Å (Table 3.12) between the TS2 corresponding to mechanism II for mandelate, along with some RMS values of the order of 0.1 Å (table 3.15) of the residues in the active center, produces a change of only -0.31 kcal/mol in the corresponding energy barrier (table 3.9).

The step 4 of mechanism I and II of propargylglycolate is a especial case in which the transition state structures obtained with the two methods differ significantly (Tables 3.11, 3.12, 3.14 and 3.15). This explains the energy difference already seen in this step as well. The differences appear in residue Lys166 and in the substrate. In Figure 3.6 we can do a visual inspection of what is happening. That is, the hydroxyl group of propargylglycolate points to different directions, the conformation of the amino group in Lys166 is slightly different and the C_α atom presents a different degree of configurational change. The Asp195 residue is depicted to realize the different conformation of Lys166. Then, we can see in this case that the differences are not only in the distances associated to the transferring hydrogens (Tables 3.11 and 3.12) but in the immediate surrounding. Note that this step 4 consists basically of the configuration change of the C_α atom, the reaction coordinate chosen in this case (the distance between the transferring hydrogen and the acceptor heavy atom) not being perhaps the most adequate.

It is important to remark that in all cases but in the step 4 (mechanisms I and II) for propargylglycolate we have found a negative eigenvalue from the beginning and a corresponding eigenvector that describes the step. In step 4 the initial structure (that is, the transition state structure obtained from the reaction coordinate method) had no transition vector. This means that the proposed structure according to the reaction coordinate method was away from the quadratic region with the suitable curvature corresponding to the actual transition state structure of this step. So in this case we are not just dealing with the convenience of the refinement of the coordinates and the potential energy of the located structure but with the problem of a transition state structure provided by the reaction coordinate method that may not fulfill the adequate mathematical conditions to be considered at least, as

Mechanism III	TS
mandelate	
Substrate	6.36(-02)
Lys166	4.15(-02)
His297	4.14(-02)
Lys164	8.09(-03)
Glu317	5.28(-03)
Asp195	9.78(-03)
Glu221	1.01(-02)
Glu247	3.68(-02)
propargylglycolate	
Substrate	9.18(-02)
Lys166	5.41(-02)
His297	3.60(-02)
Lys164	1.23(-02)
Glu317	4.82(-03)
Asp195	5.69(-03)
Glu221	1.05(-02)
Glu247	7.30(-03)

Table 3.16: RMS (\AA) for the transition state structure corresponding to the racemization of mandelate and propargylglycolate according to mechanism III. See text. Powers of ten in parenthesis.

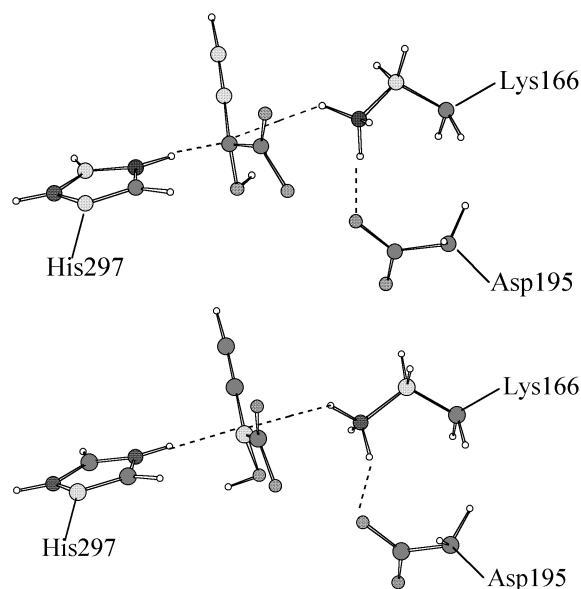


Figure 3.6: Transition state structures located with the reaction coordinate method (upper structure) and the micro-iterative method (lower structure) for the step 4 in mechanism I of the substrate propargylglycolate.

an approximation to the real transition state structure of the corresponding step.

As for the TS1 of the mechanism I of propargylglycolate, the main divergence between both transition state structures comes from the two distances associated with the transferring hydrogen from Lys164 to the appropriate carboxylate oxygen atom of the substrate: 1.451 Å and 1.056 Å for the Lys164-N \cdots H and the H \cdots OOC distances, respectively, according to the reaction coordinate method, but 1.298 Å and 1.209 Å, respectively, arising from the micro-iterative method. These differences, which could be avoided using a denser grid along the reaction coordinate method, leads to the corresponding potential energy barrier disagreement in Table 3.10.

3.3.3 Conclusions

In this section we have performed a practical test of the micro-iterative method to locate transition state structures in enzymatic reactions. We have used this method to refine the approximate QM/MM transition state structures obtained using a reaction coordinate method for the racemization of mandelate and propargylglycolate by Mandelate Racemase enzyme found

in section 2.3.

The results show that the transition state structures located by means of the reaction coordinate method are, in general, good enough to define a reliable reaction path of the reaction, taking into account both the potential energy barriers and the geometrical structures, provided that a suitable reaction coordinate has been defined for each step of the reaction. However, if the geometrical parameters chosen to define the reaction coordinate in a concrete step do not involve the main geometrical changes that take place in that step, the transition state structure supplied by the reaction coordinate method can turn out to be quite different from the actual one. In this sense the application of a second derivatives direct method, like the micro-iterative, is always recommended, rather than to warrant the real nature of transition state of the located structure (in other words, that the located structure lies on the quadratic region corresponding to the actual transition state structure), than to refine the concrete values of the potential energy barriers and the geometrical structure.

Indeed this work concerns just to a particular enzymatic reaction. However, we have tested the racemization of two substrates, that takes place through three different mechanisms, involving many distinct steps. In all, 25 transition states have been located, which provide a critical mass of information, probably enough to think that the conclusions derived here can be quite general for the enzymatic reactions involving thousands of atoms.

3.4 Avoiding the memory problem

In this section we recover the RFO equations of section 3.1 and design a method to be used for thousands of atoms. Three main problems arise when working with a big Hessian, that is, the initial Hessian calculation, its diagonalization at every step of the process and its storage. The initial Hessian computational effort is related to the method at which the energy is calculated while the two other issues are only related to the size of the matrix.

The reason why standard Newton methods are not used in systems with a great number of atoms is that computational difficulties are encountered in a big Hessian manipulation. In this chapter we have seen how micro-iterative methods overcome this problem but we must keep in mind that an expensive QM method or an increasing of the core size will provoke these computational problems to reappear.

As we said in the introductory section 1.3.7, memory problems arise in matrices with dimension of $\sim 10\,000$. An example that shows how the computational requirements increase with the size is given in the table below. In this table it is shown the time required to extract the lowest five eigenpairs along with the time needed to calculate the Numerical Hessian at different dimensions of a matrix.

Dimension ¹	Partial diagonalization ²	Hessian calculation ²
99	0.01	2291
285	0.16	6719
420	1.1	9878
978	17.3	23071
3894	1195.0	92196

¹ The given matrix dimension n should properly be indicated as $n \times n$

² Time given in seconds on a Pentium IV 2.0 GHz

The above table is only to exemplify that although matrix diagonalization scales at higher rate, the absolute computational effort devoted to the Hessian calculation in this range of dimensions is greater.

We do not need bigger Hessian matrices in enzymatic reactions:

One of the conclusions in the previous sections is that a core size higher than a thousand of atoms is not useful. The storage of a matrix of few hundreds of atoms is not a real problem in nowadays computers. In addition

to the storage, the full diagonalization can be avoided with a standard partial diagonalization which provides the two lowest eigenpairs needed to calculate the displacement in the RFO scheme.

Depending on the QM level a compromise must be found between the quality of the initial Hessian and the steps required to converge. That is, a very cheap second derivatives matrix could be used but this would imply a bigger number of steps. Since the number of steps required to converge with a bad initial Hessian can increase very rapidly, our recommendation would be that the higher the quality of the Hessian the better. Of course in a transition state search this last statement is crucial. It has also been shown[151, 161] that an adequate initial Hessian eliminates the known problems of coupling between Cartesian coordinates.

We still need a general algorithm:

The above discussion is valid for transition state search of a chemical step in enzymatic reactions. However, we must take into account that trying to find a general second-order algorithm capable to move a huge number of atoms is still a very active area. This ideal algorithm would be needed to minimize a very big number of atoms or to locate saddle points when many atoms are involved in the transition (*e.g.* phase transition in solid state[142]). In this case, the problem of storage of a big Hessian matrix emerges as the most important bottleneck in the algorithm.

Keeping this in mind we have designed a second order method that avoids the storage of the Hessian matrix. It is based on a limited-memory update combined by a Lanczos-type iterative diagonalization that permits to extract the displacement eigenpair at very low memory cost.

This procedure has already been proposed by Bofill and co-workers[155] but in that case only small systems were tested and the usage of internal coordinates may have notable implications that we will comment immediately. Besides, here we analyze the shape of different initial Hessians and we will alert to the problems related to the convergence in the iterative diagonalization.

3.4.1 Initial Hessians, sparsity and storage

Before we explain the general scheme for the optimization procedure, a short discussion on the amount of memory needed for a successful optimization will be made, mainly, the initial Hessian characteristics.

The initial Hessian matrix needs of a special consideration. The matrix shape employed in section 3.1 displayed in page 117 with a small squared Hessian for a core and a diagonal vector for the environment was suitable for a very localized reaction. But when many atoms have to rearrange during the optimization a bigger square Hessian would be needed.

The sparsity of the Hessian matrix:

It is interesting to see how a Hessian looks like. In figure 3.7 there is a representation of the Hessian elements bigger than a threshold. The Hessian matrix corresponds to 148 atoms in the active site of Mandelate Racemase enzyme. The Hessian is calculated numerically in Cartesian coordinates representation. We can see that the biggest elements are in the diagonal, but there are also elements bigger than $10 \text{ kcal}/(\text{mol}\cdot\text{\AA}^2)$ even quite far away from the diagonal. This fact is probably due to the Cartesian coordinates coupling and the long-range interaction in the enzymatic system. In the next section we will have to recall this fact, in other words, this kind of Hessian is not like a MC-SCF Hamiltonian which is rather sparse, concentrating the significative elements in the diagonal or near-diagonal region[243].

When there are many significative elements out of the diagonal the initial Hessian matrix represented as a small matrix plus a vector is not the most suitable. A sparse Hessian when only some elements are stored has been suggested for the Newton-like optimizations[157].

Optimization with a sparse Hessian:

We have tested how important is the exclusion of some elements in the Hessian matrix. The antamanide polypeptide studied in section 3.1 is used here to carry out several TS optimizations using RFO method. The tests consist in taking as initial Hessian matrix a sparse Hessian where only the elements bigger than a threshold are stored.

Procedure: In order to test only the influence of a sparse Hessian we work with the same LAPACK diagonalization subroutine [246] that needs the whole matrix in memory. In the next section we implement a real save memory strategy with the iterative Lanczos diagonalization.⁷

We calculate the Hessian numerically. When the calculated element matrix is smaller than the given threshold this element is set to zero. This matrix element will be zero all along the optimization process and the Powell

⁷ Obviously, since the Hessian is a symmetric matrix we are always working with the triangular part

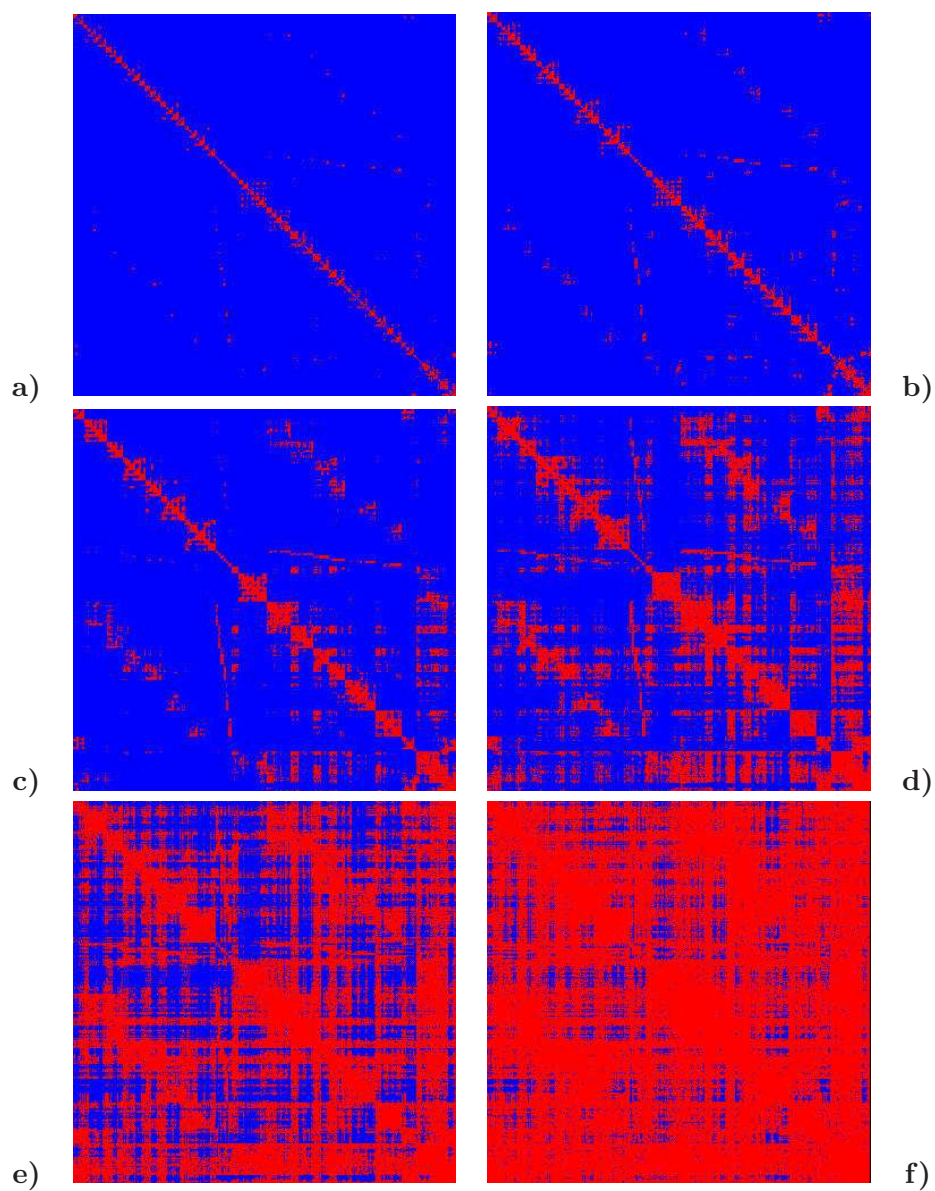


Figure 3.7: Elements of a (444×444) Hessian matrix bigger than a threshold. **a)** 10^2 , **b)** 10^1 , **c)** 10^0 , **d)** 10^{-1} , **e)** 10^{-2} , **f)** 10^{-3} in kcal/(mol·Å²) This Hessian corresponds to the numerical calculation in Cartesian coordinates of the active site of Mandelate Racemase

Threshold ^a	$ g_0 = 8.1$	$ g_0 = 0.1$	approx % stored ^b
10	<i>not converged</i>	<i>not converged</i>	6.2
1	142	25	10.5
10^{-1}	165	24	22.1
10^{-2}	193	6	41.1
10^{-3}	195	9	65.3
10^{-4}	147	9	85.7
10^{-5}	158	9	96.3
10^{-6}	149	9	99.4

^a In (kcal/(mol·Å²))

^b The percentage depends on the initial structure, but in both cases are very similar

Table 3.17: Steps required to converge two different structures taking as initial Hessian a sparse matrix where only elements bigger than a threshold (first column) are stored

Hessian update will not be permitted to change it. RFO equations are solved and the optimization process is performed as explained in section 3.1. The convergence criteria is 10^{-4} kcal/(mol·Å) on the gradient norm.

Results: In table 3.17 we show the results of two different structure optimizations. We represent the steps required to converge depending on the initial sparse Hessian matrix.

From the table 3.17 we can extract some conclusions. When the initial structure is too far from the stationary point ($|g_0| = 8.1$) the initial Hessian has few influence on the number of iterations. After one hundred of steps the Hessian has been updated one hundred of times and there is no trace of the initial matrix shape. On the contrary, when the initial gradient is closer to the stationary point the optimization obviously takes less steps to converge. It can be seen that we obtain the same results storing the elements bigger than 10.0^{-2} kcal/(mol·Å²), that is, storing only the 41.1 % of a Hessian, than storing all the Hessian, In addition, we obtain reasonable results storing only the 22.1 % and even the 10.5 %. The optimization of both structures fail when the threshold is too big (10 kcal/(mol·Å²)), this means that the essential matrix elements are not stored.

From the above results a rapid conclusion is that a considerable number of Hessian elements can be discarded in the storage and consequently a method to save memory is a worthy task.

In the next section we will test these kind of Hessians with the general

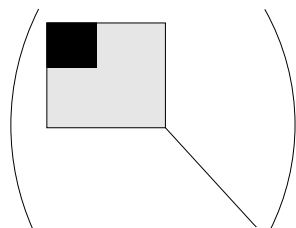


Figure 3.8: General scheme for an initial Hessian. In black a core where all elements are computed, in grey sparse Hessian where only elements bigger than a threshold are stored. The rest is represented by a diagonal vector

shape displayed in figure 3.8. From the input we will decide which zone is represented with a full Hessian (in black), sparse with a threshold (grey) and with a diagonal vector.

3.4.2 General scheme for the strategy

A solution to the storage: limited-memory

We have seen in section 1.3.7.1, where L-BFGS is described, that the limited memory strategy avoids the construction of the Hessian. We store the position and the gradient of a limited number of previous steps and by means of the limited-memory update formulae (equation 1.93 in page 48) we could re-build the whole Hessian. Actually we never build the big matrix, but perform the matrix vector product needed to obtain the Newton-Raphson displacement.

Two aspects differentiate the L-BFGS case with ours. i) We want a more accurate control of the curvature of the PES, so it means that we have to deal with the Hessian instead of its inverse. ii) We want an initial Hessian better than a diagonal matrix.

So, if the Hessian we want does not have a trivial inverse it means that we cannot work with original Newton-Raphson equation and its expensive matrix inversion. It turns out that the RFO formulation is a more suitable strategy for this kind of problem⁸.

The limited-memory-RFO at iteration k would have the known equation

⁸See truncated Newton Raphson as another alternative [156, 157]

already explained in section 3.1

$$\begin{pmatrix} 0 & \mathbf{g}_k^T \\ \mathbf{g}_k & \mathbf{B}_k \end{pmatrix} \mathbf{v}_\theta^{(k)} = \lambda_\theta^{(k)} \mathbf{v}_\theta^{(k)} \quad \forall \quad \theta = 1, \dots, N+1 \quad (3.13)$$

Where in this case the Hessian update is only extended to the last m previous steps.

$$\mathbf{B}_{k+1} = \mathbf{B}_0 + \sum_{i=k-m}^k [\mathbf{j}_i \mathbf{u}_i^T + \mathbf{u}_i (\mathbf{j}_i^T - (\mathbf{j}_i^T \Delta \mathbf{q}_i) \mathbf{u}_i^T)] \quad (3.14)$$

where \mathbf{j} and \mathbf{u} have been explained in page 116.

The implementation of equations 3.13 and 3.14 would still be problematic because to solve the eigenvalue equation 3.13 we must keep in memory all the matrix. So, we need a diagonalization process able to extract the two lowest eigenpair needed to calculate the displacement without the storage of any big matrix.

**A solution to the diagonalization:
iterative diagonalization without storage**

A specially suitable solution to this requirement is the family of Lanczos-like diagonalization methods which only require as input a matrix-vector product [254].

In the Normal Mode Analysis the task of diagonalization of a big Hessian is the central computational step [14, 255]. There are available on the web free of charge some packages already optimized to solve such big eigenvalue problems. An example of this is the ARPACK and PARPACK. We think that these strategies will give similar results to those we have obtained here.

In the direct multiconfiguration SCF electronic problems big Hamiltonians have to be diagonalized. Iterative diagonalizations have been widely used. Davidson diagonalization is one of the most popular methods. In these kind of problems, only the eigenvalues of the lowest molecular orbitals are needed and the huge Hamiltonian matrix is usually stored in disk rather than in the rapid access memory, so that only matrix-vector product can be calculated.

In the case of geometry optimizations the matrix-vector products can be performed *on-the-fly* by the update Hessian formula. It is actually the

equation 3.14 multiplied by a vector \mathbf{v} .

$$\mathbf{B}_{k+1}\mathbf{v} = \mathbf{B}_0\mathbf{v} + \sum_{i=k-m}^k [\mathbf{j}_i\mathbf{u}_i^T\mathbf{v} + \mathbf{u}_i(\mathbf{j}_i^T\mathbf{v} - (\mathbf{j}_i^T\Delta\mathbf{q}_i)\mathbf{u}_i^T\mathbf{v})] \quad (3.15)$$

Note that although we display the expression for the Hessian \mathbf{B}_{k+1} actually we have to deal with the diagonalization of the Augmented Hessian of equation 3.13. This is an easy step where the gradient has to be added in the matrix-vector multiplication.

Bofill and co-workers developed an iterative diagonalization method that improves the Davidson method in many aspects [243, 256]. It is able to extract not only the lowest or the highest but also the degenerate roots and any eigenpair that fits to a guess given as input. Like all the iterative methods, it is based on the Ritz-Galerkin algorithm. It means that a subspace is constructed where the eigenvalue equation 3.16 is projected and solved. In the Lanczos-like methods this subspace is increased at every iteration. The efficiency of the method will rely on the efficiency of the iterative construction of the subspace basis vectors and the procedure to solve the equation 3.16 in this subspace.

The method developed by Bofill *et. al.* will not be covered in detail here. Check the references [257, 258, 243, 256] for an accurate description of the equations and the possible alternatives. As a general perspective, the main characteristics of the method are that the vectors to build the iterative subspace are generated first using an approximated or inexact Lagrange-Newton-Raphson (LNR) formula and later, after an orthonormalization procedure, the exact Lagrange-Newton-Raphson equation for the desired root is solved in the iterative subspace.

In more detail, we can transform a generic diagonalization problem

$$\mathbf{B}\mathbf{V} = \mathbf{V}\mathbf{\Lambda} \quad (3.16)$$

into a minimization process through a Lagrangian function ⁹:

$$L(\mathbf{v}, \lambda) = \mathbf{v}^T\mathbf{B}\mathbf{v} - \lambda(\mathbf{v}^T\mathbf{v} - 1) \quad (3.17)$$

Expanding equation 3.17 up to second order and applying the stationary

⁹This technique is in the inverse direction of the Hartree-Fock-Roothan equations where a minimization of the energy leads to a diagonalization of the Fock matrix

conditions leads to the Lagrange-Newton-Raphson equations. This permits, after some manipulations, to obtain an improvement for the eigenvector \mathbf{v} and eigenvalue λ .

$$\delta\mathbf{v} = -(\mathbf{B} - \lambda\mathbf{I})^{-1}(\mathbf{B}\mathbf{v} - (\lambda + \delta\lambda)\mathbf{v}) \quad (3.18)$$

$$\delta\lambda = \frac{\mathbf{v}^T(\mathbf{B} - \lambda\mathbf{I})^{-1}(\mathbf{B}\mathbf{v} - \lambda\mathbf{v})}{\mathbf{v}^T(\mathbf{B} - \lambda\mathbf{I})^{-1}\mathbf{v}} \quad (3.19)$$

A critical point is also the inversion of the full matrix in the denominator of equation 3.19 that can be approximated by a diagonal(Davidson) or by a square+vector (figure 3.1 in page 117).

This iterative diagonalization task involves a sort of internal optimization process to obtain every new vector of the subspace. Such internal optimization of the subspace basis vector is carried out by Newton-Raphson method or by a Rational Function Optimization. Note that in equation 3.18 $\mathbf{B} - \lambda\mathbf{I}$ would be the Hessian and $\mathbf{B}\mathbf{v} - (\lambda + \delta\lambda)\mathbf{v}$ the gradient. This is important because if the iterative diagonalization process takes many steps to converge the subspace will increase its dimension and the internal Hessian diagonalizations could become problematic.

As a resume the procedure of diagonalization of a matrix \mathbf{B} can be outlined as follows:

1. Start with a normalized guess eigenvector \mathbf{v} . Compute the corresponding eigenvalue $\lambda = \mathbf{v}^T\mathbf{B}\mathbf{v}$
2. Using the inexact LNR equations 3.18 and 3.19 compute the improvements to \mathbf{v} and λ . (they are inexact because involve the inversion of the big matrix that must be approximated by its diagonal or the square+vector form)
3. Orthonormalize the new improved vector with the rest of the already constructed subspace, through a Gram-Schmidt procedure [259] for example.
4. Project the LNR equations to the subspace and solve it by a Newton-Raphson or RFO procedure.
5. The solution of this internal optimization will be the improvement $\delta\mathbf{v}$ to the initial eigenvector \mathbf{v} .

6. If this improvement is bigger than a convergence criteria, return to point 1. If not, the whole process is ended with the eigenpair (\mathbf{v}, λ) as a result.

If the convergence is reached after k iterations the number of matrix-vector products $\mathbf{B}\mathbf{v}$ will be $k + 1$. The algorithm only needs to store three vectors in the high speed memory at every iteration.

3.4.3 Results and discussion

We coupled the Bofill's LNR diagonalization method to the RFO geometry optimization source code commented in section 3.1. In addition we tested several shapes for the initial Hessian as displayed in figure 3.8.

After many different implementations the results were not satisfactory. The iterative diagonalization did not converge after 100 steps. Recall that 100 steps means that the internal subspace has a dimension 100 and at every iteration an internal optimization manipulating a Hessian as big as the number of external iterations has to be performed. This fact increases the computational cost. In conclusion, it seems that a method that is very powerful to diagonalize any kind of Hamiltonian matrix [256] was unable to converge for the Augmented Hessian diagonalization.

In small test cases we observed that in the inexact LNR equations, when the matrix in the denominator of equation 3.19 which must have an affordable inversion is less approximated, the inexact LNR equations are not that inexact and the diagonalization process converges along with the geometry optimization. This is the case when this approximated matrix is represented by a square+vector where the square part has about the half of the dimension of the whole space. This case in big matrices is not affordable and the square part must be smaller.

We compared the LNR iterative diagonalization with the original Davidson method and the latter failed also. We tried to increase the convergence criteria for the iterative diagonalization, and proceed computing the corresponding geometry displacement and tried to optimize the geometry. The displacement vectors were not good enough to reach the stationary point. The search was not successful even combining the method with an improvement of the displacement vector by DIIS strategy (introduction section 1.3.4.3).

In conclusion, after these failed attempts we conclude that our Augmented Hessian was very difficult to diagonalize by these kind of methods. The origin of the problem could be due to the non-sparsity of a Hessian in Cartesian coordinates representation or due to an intrinsic redundancy of the Augmented Hessian matrix (it can be shown that the gradient can be obtained by linear combination of the eigenvectors of the Hessian). This last argument is less important if we take into account that Bofill and co-workers obtained successful results in small systems where internal coordinates are used [155].

Chapter 4

Molecular Dynamics and Free Energy on Mandelate Racemase

In the current chapter we carry out molecular simulations to calculate the free energy profile corresponding to the mandelate racemization in Mandelate Racemase enzyme.

During the last years, there has been much effort towards the efficient calculation of free energy differences in condensed phase systems. Among the different possibilities, the so-called Potential of Mean Force (PMF) is the magnitude applied to enzymatic reactivity in order to elucidate the reaction mechanism and for predicting the rates of chemical processes. The PMF is obtained using molecular dynamics or Monte Carlo sampling methods, but many aspects in the simulation technique still require an improvement to avoid sources of error and to go beyond the qualitative results. For instance, an adequate potential energy is always required, the length of the sampling must be large enough to ensure an adequate exploration. In addition, the PMF must be performed along a distinguished coordinate (reaction coordinate) from reactant to product region, that must be known *a priori* and capable to adequately describe the reaction. The choice of the reaction coordinate will play a central role in the discussion of the present chapter.

We will calculate the PMF using the information about the racemization reaction acquired in the preceeding chapters by means of optimization methods. In particular, the reaction path obtained previously will be a helpful

guide for obtaining the PMF, mainly in the crucial choice of an adequate distinguished reaction coordinate to be used in the PMF calculation.

From the chemical point of view, we will see the influence of the inclusion of temperature effects, and whether it makes the free energy profile very different from the potential energy profile or not. The PMF will be also obtained for the N197A mutant in order to explain the lower enzymatic activity observed in the mutagenesis experiment [223].

Many chemical steps in enzymatic reactions are proton transfers or nucleophilic substitutions. The geometrical reaction coordinate of these type of reactions can be well described by the difference between the bond breaking and bond forming distances $r_c = r_{form} - r_{break}$.

We saw that Mandelate Racemase enzyme catalyzes the mandelate substrate racemization through three different mechanisms. We found that the most favorable path is mechanism III which potential energy profile is depicted in figure 4.1. Unlike many other reactions, mechanism III cannot be described by a unique bond breaking / bond forming process. As we commented in the preceding chapters, the central configuration inversion step is accompanied by two asynchronous proton transfers through the acid-base catalytic residues Lys166 and His297. We will see that the shallow minima, represented in figure 4.1 by Is and Ir, are not minima in the free energy profile computed in the present chapter. Therefore mechanism III might be seen as a concerted mechanism, although rather asynchronous, where there exist two proton transfers and a configuration inversion of the stereogenic center. In conclusion, in order to compute the PMF of this mechanism the geometrical reaction coordinate must describe the concerted process contemplating the different chemical changes already observed by optimization techniques.

4.1 Mandelate Racemase: model and setup for simulation

The procedure that we follow for the setup and sampling of Mandelate Racemase system is not the same that the one we used in the preceding chapters. A different QM/MM level with a different QM/MM partition, along with different solvation model and MD approach will be employed.

The same 2.0 Å resolution structure of the complex of *Pseudomonas*

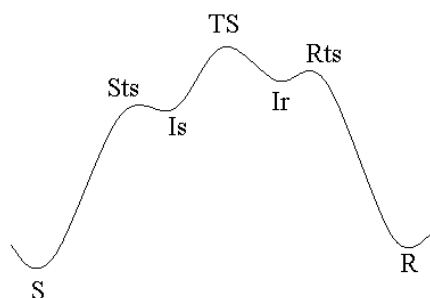


Figure 4.1: Potential energy profile found for Mechanism III in the racemization of mandelate substrate by means of optimization methods

putida Mandelate Racemase enzyme (PDB code 1MNS [233]) has been used again as the starting point for the calculations. In this case we have used a modified version of the c28b2 CHARMM package of programs[53]. The waters from the X-ray structure have been kept, and the protonation of the non-titrable hydrogens have been added with the HBUILD facility of CHARMM.

With the exception of the active site the protonation state of titrable residues have been set at pH 7. The histidine residues are neutral with the hydrogen at N_ϵ or N_δ depending on the possibility of hydrogen bond formation. As for the active site the protonation of the relevant residues corresponds to the structure of reactant S. The mandelate substrate has been added matching the maximum number of atoms with the inhibitor (R)- α -phenylglycidate found in the PDB structure. This process that may be too rough in some cases, in mandelate substrate is quite appropriate since the experimental results show that the binding of (R)- α -phenylglycidate is very similar to the natural substrate.

4.1.1 Potential Energy Surface

The QM/MM method used in this chapter is slightly different than the one used in the previous chapter with ROAR. The set of atoms selected as QM zone are depicted in figure 4.2. The QM part will be treated with the PM3 semiempirical Hamiltonian, and it contains the mandelate substrate, the general acid-base catalyst Lys166 and His297 along with the charge stabilization residues Lys164 and Glu317. This selection of atoms has been labeled as model 2 in the gas phase calculation in page 103. The magnesium

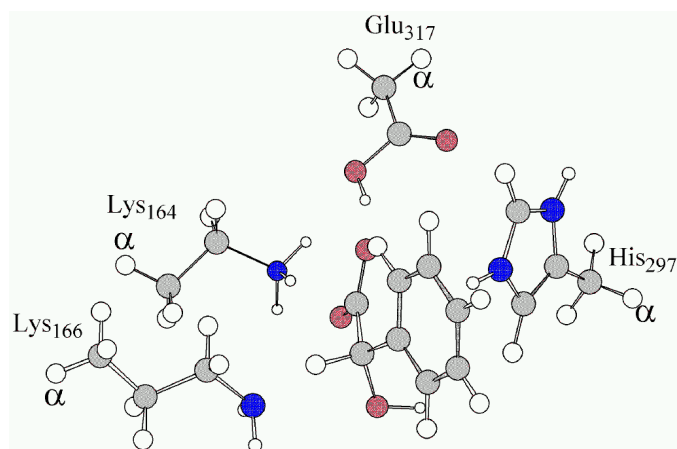


Figure 4.2: Atoms included in the QM part. The α mark indicates the boundary atom treated with GHO method.

cation and its four other ligands are not included in the QM part anymore. The QM part has been shortened for several computational reasons. We encountered some problems to converge the SCF process with the bigger model. But mainly we decided to work with a QM model of 63 atoms because a 91 atoms model was too expensive for an appropriate sampling in the PMF calculation.

Boundary Atoms:

The partition of covalent bonds in the QM/MM frontier has been treated with the Generalized Hybrid Orbital framework[91], in particular using the PM3(GHO) method published recently[93]. The boundary between the QM and MM zones is a specially delicate issue. The GHO method is a more refined strategy than the link atom used in previous sections. As we commented in the introductory section (section 1.2.3.1), in GHO no additional atoms are included, the boundary atom must be a sp^3 carbon whose hybrid orbitals are locally optimized. The boundary atoms used in this model are displayed in figure 4.2 with the α mark. The rest of the enzyme will be represented with the all-atom CHARMM22 force field[54], while the solvent(not yet commented) with the three point charge TIP3P model[237].

Non-bonded interactions:

The van der Waals parameters for the quantum atoms have not been optimized with this new QM/MM selection. The optimization of parameters is recommended when the quantum part interacts directly with the MM

part[77]. A particular example of this fact is the QM/MM study[6] of a S_N2 reaction in haloalkane dehalogenase where the enzymatic efficiency depends strongly on how a quantum chloride is stabilized by two MM triptophanes. In this case, the reproduction of an accurate non-bond interaction is crucial.

In our QM/MM partition the only problematic non-bonding interaction is between mandelate substrate and magnesium cation (not displayed in figure 4.2, see preceding chapters, for example figure 2.12 in page 104). We already commented that the magnesium cation stabilizes the transition state removing charge density from the substrate. However, after some tests, we have seen that the QM/MM interaction mandelate(PM3)-magnesium(CHARMM) using the standard parameters from CHARMM22 force field is very similar in energy and structure to the full PM3-SRP results.

The rest of classical ligands bound to the magnesium cation, including the water, is stable enough to keep its coordination over all our 2 nanosecond simulation.

The non-bonded interactions have been calculated with the following characteristics. The pairlist is built on the basis of a group based cutoff using a distance of 13.5 Å (both QM and MM regions). The pairlist is updated every 35 steps during the dynamics simulation. The van der Waals are calculated using a shifting function with a cutoff at 13 Å. The electrostatics have been calculated using a switching function (see introductory section 1.2.2.2 in page 23) activated at 12 Å at which the smoothing function begins to reduce the potential and eliminated completely at 13 Å.

4.1.2 Molecular Dynamics

To mimic the aqueous environment the Stochastic Boundary molecular dynamics method (SBMD) has been used. A particular method of solvation will be used due to this fact. We have used a sphere of pre-equilibrated waters with a radius of 24 Å to solvate the system. This sphere was centered on the stereogenic carbon of the substrate (C_α). All the crystallographic waters beyond the sphere were removed along with any water which oxygen is closer than 2.5 Å near any heavy atom of the protein. A soft and deformable boundary potential is applied at the edge of the sphere of waters to mimic the effect of the inexistent bulk solvent.

The classical dynamics is carried out partitioning the system into three

zones (see section 1.4.4 for more details in SBMD method). The dynamics region which consists of atoms within a distance of 20 Å from the center; the buffer region which contains the atoms surrounding the dynamics region from 20 Å up to 24 Å; and the reservoir region which includes the remainder of the system and is excluded in the explicit dynamics simulation. The sizes of the dynamics and buffer region are large enough to ensure that the active site and all the possible rearrangements in its surroundings during the reaction are adequately modeled.

The partition of the dynamical system is not done atom-wise, that is, all atoms of an aminoacidic residue are included in the dynamics region if any atom of the residue is within 20 Å from the reference point. If none of the atoms of a residue are in the 24 Å sphere the aminoacid will belong to the reservoir region. While the rest of enzyme atoms that may be found between 20 and 24 Å will be contained in the buffer region. The labels for the protein are assigned at the beginning of the simulation and kept all along the simulation. This will not be the case for the water molecules which will be permitted to diffuse between the dynamics and buffer region. The label list containing the waters of every region is updated every five steps during the simulation.

The parameters needed in the SBMD framework are taken from the original publications [186] which have been tested thoroughly in more recent works [187, 189]. The trajectory in the dynamics region is propagated using the Newton's equation of motion, while in the buffer region the Langevin equations with a stochastic term are used. The constants for boundary forces are $K_C = 1.30$, $K_O = 1.22$, $K_N = 1.30$ kcal/(mol·Å²) for the backbone atoms and 0.73 kcal/(mol·Å²) for the lateral chain heavy atoms. The friction coefficients are 200 ps⁻¹ for all protein atoms and 62 ps⁻¹ for oxygen atom in water molecules (this last value corresponds to the self-diffusion constant of bulk water at 300 K [183]). Both friction coefficients and boundary forces applied to protein atoms are scaled by a screening function that depends on the distance from the center (see section 1.4.4).

The final dimension of the different regions is:

ATOMS	protein + substrate	waters	Total
QM atoms	63	0	63
MM atoms	5406	2730	8136
Dynamics zone	3276	1242	4518
Buffer zone	1208	1488	2686
Reservoir zone	985	0	985
All	5469	2730	8199

The leapfrog algorithm to integrate the MD equations is used in all cases with a timestep of 1 fs. The hydrogen atoms are constrained using the SHAKE algorithm (page 56). The process of heating and equilibrating the system can be outlined as follows:

1. The 24 Å sphere of waters is added on the protein three times at three different orientations. The corresponding remove criterion is applied every time.
2. A short steepest descent minimization is performed into the dynamics and buffer region to avoid bad contacts
3. A 5 ps NVT Langevin dynamics of the solvent is performed in order to equilibrate the waters.
4. The three previous steps are repeated to obtain a good solvation of the system.
5. 10 ps at 100, 200 and 298 K SBMD are performed to heat the system.
6. The final system is equilibrated during 50 ps at 298 K.

4.1.3 Potential of Mean Force

The molecular dynamics simulation described above has as a principal goal the calculation of the PMF $\Delta G(Rc)$ of the reaction. In our case the PMF calculation has been performed using the umbrella sampling technique[196, 15].

In the umbrella sampling method, outlined in section 1.5.2, at every simulation window the reaction coordinate is restrained within a limited range by imposing a harmonic umbrella potential.

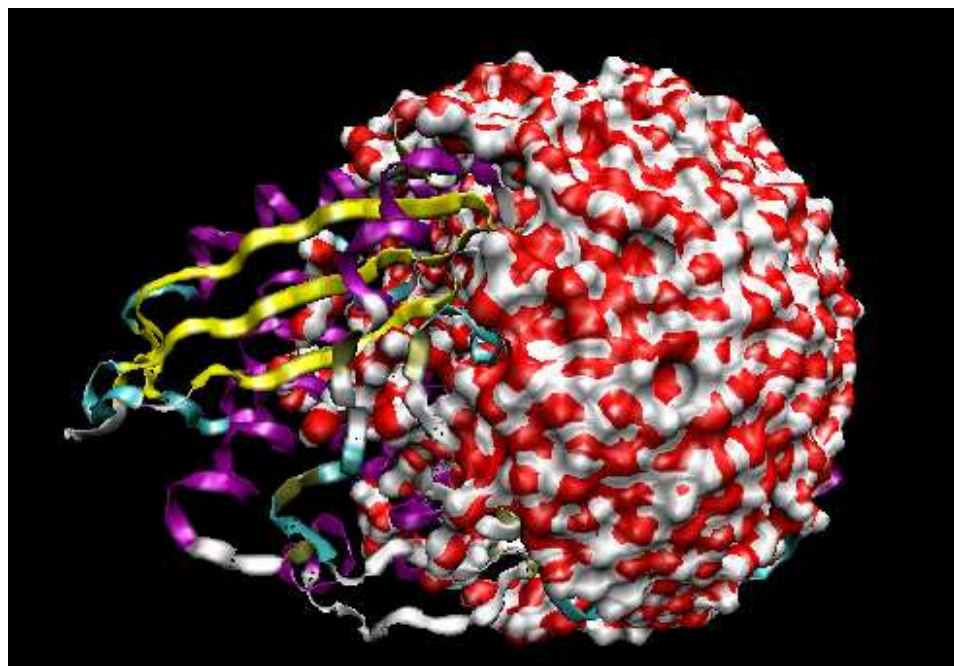


Figure 4.3: Schematic representation of the solvated Mandelate Racemase model

The simulations are carried out by overlapping regions (windows) until covering the entire reaction coordinate. For every window we run 15 ps for equilibration and 50 ps for sampling. The reaction coordinate is divided in finite bins of length 0.01 \AA . During the sampling the probability distribution $P(Rc)$ is built counting the configurations that fall in the range $Rc \pm 0.005 \text{ \AA}$ of every bin.

The WHAM technique has been used to join the different windows of simulations. The WHAM results have been compared with those obtained by directly matching every two overlapping windows and few differences have been encountered¹. For a detailed description of umbrella sampling and WHAM technique see the introductory section 1.5.2. Unless indicated explicitly the free energy profile using WHAM will be shown.

In general, the simulation techniques need many heuristic parameters. Magnitudes such as the length of a dynamics for equilibration or sampling, the number of waters to solvate, the cutoff for non-bonded interactions ... *etc.* Most of these parameters and procedures are not tested here, they

¹See the appendix for the source code developed in order to obtain the PMF profile

are taken from the literature or by personal communication from other experienced researchers.

Multi-dimensional PMF:

There has been several publications paving the methodology needed for multi-dimensional PMF calculations [201, 199, 200]. A multi-dimensional PMF would be useful for many reasons, mainly for the usage of several independent parameters to fully represent the reaction path. A two-dimensional free energy surface would permit to observe how two reaction coordinates participate in the reaction mechanism. The decision in our study of preferring a one-dimensional PMF was mainly due to the computational cost.

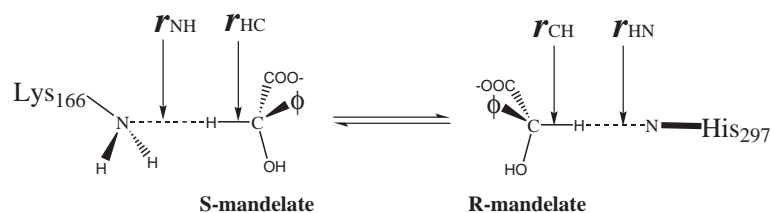
An approximated prediction of the computational requirements for a 2D PMF applied to our system is the following. It took about 30 hours of CPU for every 65 ps window (15ps equilibration + 50ps sampling). In order to scan from reactants to products the geometrical coordinate (see next section) R_{NHC} goes from -0.7 to 1.9 Å, R_{CHN} from -1.9 to 0.6 Å. If the coordinate increment for two adjacent windows is 0.2 Å, we need 169 windows, or 5070 hours of computation. Unless an automated and parallelized process is designed it means 211 days of uninterrupted computation time.

4.2 Potential of Mean Force on different Reaction coordinates

4.2.1 Selection of a reaction coordinate

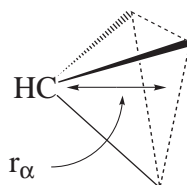
From the stationary points found along the reaction path for mechanism III we can select an appropriate reaction coordinate. We will use a geometrical coordinate which is a magnitude easy to define. However, there exist some works using an energetic or solvent coordinate[114]. A recent comparative study by Gao *et. al.* [260] demonstrates that calculating a PMF with the adequate sampling the results using an energetic or a geometric coordinate are equivalent.

The reaction coordinate must be a combination of the main geometrical parameters that participate in the reaction, namely, the distance between heavy atoms and transferred hydrogens r_{NH} , r_{HC} , r_{CH} , r_{HN} as indicated in the following graphic.



In addition we can define r_{NC} and r_{CN} as the distances between the three heavy atoms that participate in the concerted step, that is, $N_{Lys166} - C$ and $C - N_{His297}$

These distances mainly describe the double proton transfer. An additional parameter for the configuration inversion could be the solid angle that indicates the sp^3 hybridization, or as an alternative, the distance between the stereogenic carbon and the plane defined by its three bound substituents.



In table 4.1 the evolution of geometrical parameters during the reaction from S reactant to the R product is shown.

	r_{NH}	r_{HC}	r_{CH}	r_{HN}	r_{NC}	r_{CN}	r_{α}	ΔE
S	1.804	1.158	2.927	.995	2.884	3.844	0.4217	0.0
Sts	1.194	1.522	2.865	.996	2.645	3.765	0.4127	13.81
Is	1.151	1.570	2.845	.995	2.644	3.750	0.3984	13.79
TS	1.011	2.258	2.093	1.013	3.085	3.077	-0.0576	19.50
Ir	1.005	2.612	1.607	1.117	3.435	2.716	-0.3064	16.45
Rts	1.004	2.736	1.509	1.200	3.566	2.706	-0.3116	16.75
R	1.004	2.868	1.165	1.772	3.708	2.926	-0.3501	4.63

Table 4.1: Relevant geometrical parameters corresponding to the stationary points that described mechanism III found in the previous chapters

def.	R_{NHC} $r_{HC}-r_{NH}$	R_{CHN} $r_{HN}-r_{CH}$	$R_{NHC-CHN}$ $r_{HC}-r_{NH} + r_{HN}-r_{CH}$	R_{HCH} $r_{HC}-r_{CH}$	R_{NCN} $r_{NC}-r_{CN}$
S	-0.646	-1.932	-2.578	-1.769	-0.960
Sts	0.328	-1.869	-1.541	-1.343	-1.120
Is	0.419	-1.850	-1.431	-1.275	-1.106
TS	1.247	-1.080	0.167	0.165	0.008
Ir	1.607	-0.490	1.117	1.005	0.719
Rts	1.732	-0.309	1.423	1.227	0.860
R	1.864	0.607	2.471	1.703	0.782

Table 4.2: Possible combinations of geometrical parameters to be used as distinguished reaction coordinate in the PMF calculation

There is not a unique combination of parameters capable to describe the whole process. In table 4.2 we show the most intuitive combinations R_{xx} for an appropriate reaction coordinate.

While R_{NHC} and R_{CHN} would be the most appropriate reaction coordinate for the proton transfer step between substrate and Lys166 and His297 respectively, we already see that they only describe partially the evolution from S to R structures.

Finally, a special computational care must be adopted when reaction coordinate R_c includes a distance from hydrogen of Lys166. Because then, the three hydrogens of amino group may act as the proton to be transferred. In consequence, although CHARMM package builds the reaction coordinate from the input atom labels, Lys166 during a MD may rotate and the labile hydrogen may change. To prevent from scanning a bad coordinate, at every step in MD we must check for the most appropriate hydrogen to be

transferred, in this case, the closest hydrogen to the alpha carbon.

In what follows several reaction coordinates are employed in the mono-dimensional PMF calculation.

4.2.2 Combining two bond distances

NHC and CHN:

Starting from the equilibrated S structure the reaction coordinate R_{NHC} , as defined in table 4.2, is scanned by the different windows in the S \rightarrow R direction. After this simulation the final equilibrated R structure is taken to scan the R_{CHN} reaction coordinate in the R \rightarrow S direction. Unless a major sampling is needed the reference reaction coordinate Rc_0 used in the umbrella potential is increased by units of 0.2 Å.

A harmonic restraining potential is often used as biasing function. Sometimes, in order to obtain an efficient sampling an additional biasing potential is used[261]. A cubic spline function fitted to a number of points along the reaction coordinate is added to the harmonic restraining potential.

$$U_{tot} = U_{QM/MM} + k(Rc - Rc_0)^2 + U_{spline} \quad (4.1)$$

This strategy would obtain an enhanced exploration of the low probability zones without increasing the harmonic constant. The drawback is that this additional biasing function should be ideally equal to the negative of the unknown PMF ($-\Delta G(Rc)$) that is actually what we want to calculate. Then the points to which the spline is fitted must be determined by trial and error until obtaining the expected sampling.

We have used this additional biasing in the PMF for R_{NHC} and R_{CHN} reaction coordinates. In this case the spline function for an appropriate sampling was found easily thanks to the energy profile obtained in the preceding chapters by optimization of the structures. In the rest of PMF calculations such trial error effort was not worthwhile and the sampling is only improved by increasing the restraining harmonic constant. In the present case the harmonic constant is 20.0 kcal/(molÅ²) while for the rest of reaction coordinates is increased to 30.0 kcal/(molÅ²).

In figure 4.4 we show the histograms for the probability distribution $P(Rc)$ for the R_{NHC} and R_{CHN} reaction coordinates. $P(Rc)$ is converted to the probability density $\rho(Rc)$ by removing the biasing effect in the posterior

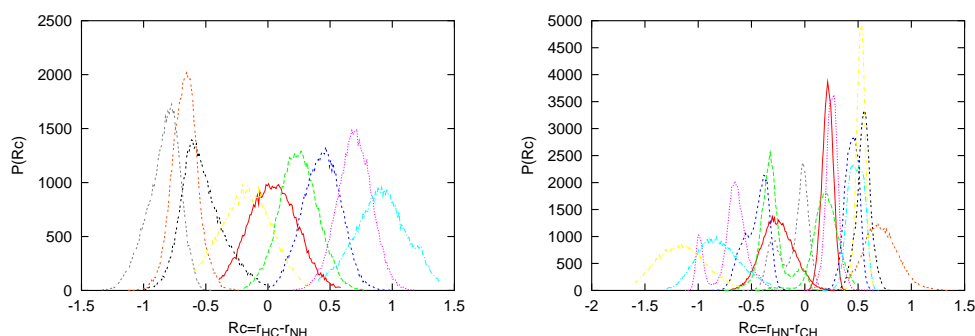


Figure 4.4: Representation of the several probability distributions $P(Rc)$ collected for the different windows. Units are given in number of configurations

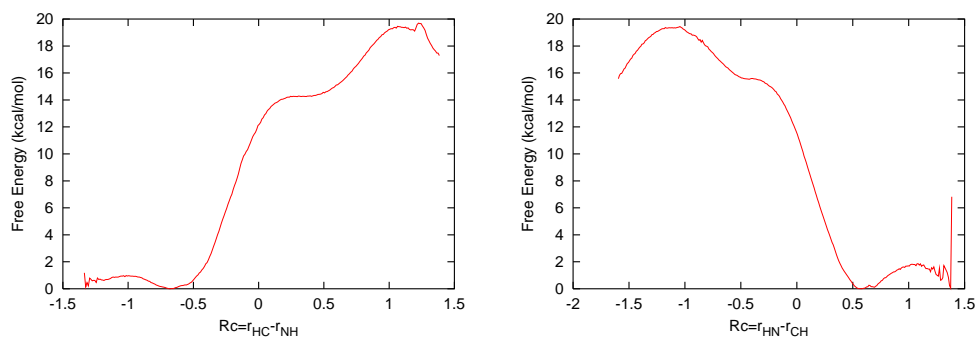


Figure 4.5: PMF profile for R_{NHC} (left) and R_{CHN} (right)

statistical analysis (see introductory section 1.5.2). It is considered that the sampling has converged when the bin is visited more than approximately 500 times².

In figure 4.5 the potential of mean force is shown for the R_{NHC} ($S \rightarrow R$) and R_{CHN} ($R \rightarrow S$).

The scanning of the reaction coordinate R_{NHC} (on the left of figure 4.5) reaches a zone after the proton transfer from substrate to Lys166 ($R_{NHC} \sim 1.0$) where this coordinate is unable to describe the carbon configuration inversion. After this region the simulation explores a zone that does not belong to the reaction path anymore. At this point Lys166 moves away from the substrate, leaving the stereogenic carbon in S configuration.

The PMF calculation using R_{CHN} (the profile on the right of figure

²Personal communication from Prof. Jiali Gao

Reaction coordinate	R_{NHC}	R_{CHN}
Minimum	S: 0 (-0.675)	R: 0 (0.575)
Inflection barrier	14.29 (0.305)	15.59 (-0.405)
Inflection minimum	14.25 (0.365)	15.53 (-0.425)
Highest point	19.46 (1.075)	19.44 (-1.045)

Table 4.3: Energetics in kcal/mol corresponding to the PMF using R_{NHC} and R_{CHN} . In brackets the corresponding reaction coordinate bin (± 0.005 Å)

4.5) has different behavior but similar consequences. R_{CHN} is not able to describe the whole reaction either. In the R→S direction, after His297 abstracts the hydrogen from the substrate ($R_{CHN} \sim -1$) and after some progress in the configuration inversion, there is a simulation window where the molecular dynamics falls into the steep valley directly to the reactant S region without describing the other half of the reaction. This different behavior with the R_{NHC} case can be attributed to the fact that the central TS (the saddle point located in chapter 3) is rather product-like (R-like). Consequently the reaction coordinate R_{CHN} is able to reach some configurations that belong to the TS and falling down to the bottom at the S configuration.

The two flat zones encountered after the two corresponding proton transfers (in figure 4.5 are around $R_{NHC} \sim 0.4$ and $R_{CHN} \sim -0.5$) and before the configuration inversions step are not thermally stable at 298 K. Therefore, although these two structures were minima with very low barrier saddle points, they cannot be considered intermediates in terms of free energy. As a consequence, the mechanism cannot be considered stepwise and a unique reaction coordinate should be able to describe the whole step.

The energetics for the free energy profiles are collected in the table 4.3. Despite the highest energy point is meaningless, the energy corresponding to the inflection zones is well described and they will be valuable in the following sections.

We tried to include the coordinate r_α (see table 4.1) to describe the central step where the configuration inversion takes place. However, any movement of the substrate substituents bounded to the alpha carbon (hydroxyl, phenyl and carboxyl groups) provokes a remarkable variation in the value of r_α and this makes the scanning of this coordinate a very difficult task.

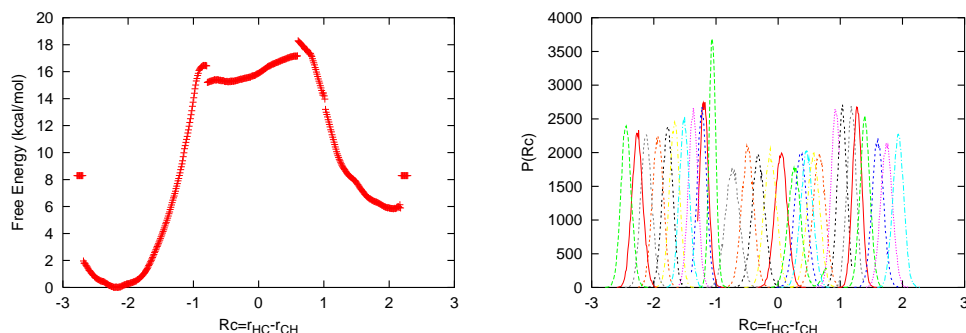


Figure 4.6: **Left:** The PMF profile using R_{HCH} as reaction coordinate. **Right:** The histogram showing $P(R_c)$ for the different simulations. In this particular case some windows had a bigger restriction umbrella potential to improve the sampling

HCH:

Another option for R_c still using a combination of two distances consists in employing the distance between the two transferred hydrogens with respect to the alpha carbon $R_{HCH}=r_{HC}-r_{CH}$.

In figure 4.6 we show the PMF and the $P(R_c)$ for every simulation window using R_{HCH} . In some regions the restraining harmonic constant was increased to obtain a better sampling and an adequate overlap between consecutive windows. The PMF profile shown in the left of figure 4.6 has been built by WHAM technique. The PMF profile by matching the different windows (not shown) had some sharp points in the regions where WHAM displays a discontinuity.

The discontinuity in the free energy profile shows clearly the inadequacy of this reaction coordinate. Although we have been able to connect the reactants and products through different overlapping simulations, this reaction coordinate describes the racemization process but lacks of the adequate exploration of some regions along the reaction process.

This fact means that there are relevant movements that belong to the reaction path not contained in R_{HCH} when scanning the distances $H_{166} \cdots C$ and $C \cdots H_{297}$. A confirmation of this assessment is shown in figure 4.7 where the average and the fluctuations of the two bond distances r_{HC} and r_{CH} for each simulation window are shown. The evolution of the distances shows an abrupt jump that explains the discontinuity in the free energy profile. In particular, from S→R direction, the initial approximation of Lys166 to

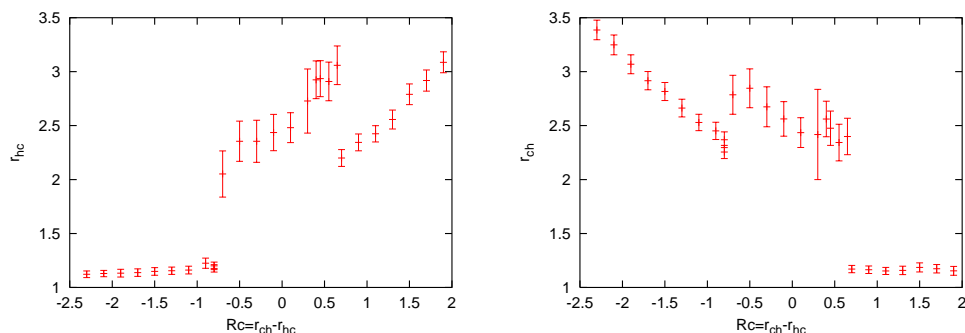


Figure 4.7: Evolution of the $H_{166} \cdots C$ (left) and $C \cdots H_{297}$ (right) distances. The average value for every window is plotted along with the standard deviation.

the substrate needed for the first proton transfer is not reproduced and the proton transfer takes place at $Rc \sim -0.8$ too brusquely. This sudden change is represented in the PMF profile by the discontinuity. The situation is the same in the *R-side* for the proton transfer between His297 and the substrate. Although the configuration inversion is monitored adequately, the posterior proton transfer from His297 takes place again with a brusque jump between two adjacent and overlapping windows around $Rc \sim 0.6$.

4.2.3 Combining four bond distances

Since the precedent two-distances reaction coordinates failed in the PMF computation, we included the four bond distances that describe the two proton transfers, namely:

$$R_4 \equiv R_{NHC-CHN} = R_{NHC} + R_{CHN} = r_{HC} - r_{NH} + r_{HN} - r_{CH} \quad (4.2)$$

Intuitively, we thought that if R_{NHC} reproduced adequately the *S-side* of the reaction and R_{CHN} the *R-side*, it indicates that joining the two of them could be an adequate way to describe the whole process.

In addition, this coordinate gives freedom to the variation of the four distances since the biasing potential only penalizes the whole combination R_4 . Therefore this would permit the observation of an hypothetical stepwise mechanism. R_4 is scanned from negative values (see table 4.2) by increments of 0.2 \AA until $\sim 2.5 \text{ \AA}$.

In figure 4.8 the histogram of $P(Rc)$ and the PMF profiles are shown. The

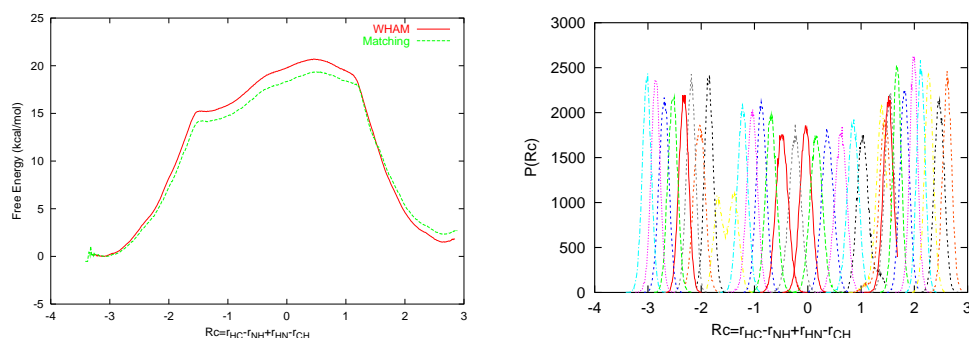


Figure 4.8: **Left:** the PMF profile using R_4 reaction coordinate. The two profiles correspond to the PMF built by WHAM or by a direct match of the overlapping windows. **Right:** Histogram diagram displaying the probability distribution.

two free energy profile correspond to the one obtained by WHAM technique and the other by matching the different windows.

Despite of some difficulties that we comment below referring to the differences in the statistical analysis of $P(R_c)$, the PMF computed using this reaction coordinate is able to draw the free energy profile all along the process. The shape of the profile is not the same as the one we encountered using R_{NHC} in the *S-side* and R_{CHN} in the *R-side* (figure 4.5), the main difference is that the inflection zone in the *R-side* is now less detailed. However, the energy for the early small barriers representing the proton transfer is approximately the same.

In addition, and what makes R_4 the most adequate coordinate, is that now we can describe the central step, where the configuration inversion takes places and which is the highest point in the PMF, along with the approximation of Lys166 and His297 for the two proton transfers.

PMF: WHAM vs matching:

There are some differences between the results coming from WHAM analysis and by adjusting automatically the several adjacent windows by a simple match criterion. The free energy profile is in some points up to ~ 1 kcal/mol different. As we already said in section 1.5.2, WHAM technique is a more sophisticated iterative procedure that takes into account all the data without discarding the overlapping regions in the simulation and that avoids the uncertainty involved in the matching process. However, we cannot ensure for certain that WHAM method gives the correct energy without a deeper

PMF Technique	Matching	WHAM
Minimum S	0 (-3.095)	0 (-0.395)
Inflection Barrier (Sts)	14.202 (-1.395)	15.25 (-1.475)
Inflection Minimum (Is)	14.116 (-1.355)	15.20 (-1.415)
TS	19.38 (0.575)	20.70 (0.455)
Inflection (Rts)	18.13 (1.135)	19.00 (1.125)
R	2.32 (2.675)	1.49 (2.675)

Table 4.4: Free energy in kcal/mol for the PMF using R_4 as a reaction coordinate. In brackets the corresponding reaction coordinate bin ($\pm 0.005 \text{ \AA}$)

analysis. In order to examine which profile is the right one, a comparison between the two free energy profiles obtained for R_4 with those obtained previously will be useful. In table 4.4 the energetics of the different points is shown. Since R_4 is unable to describe the *R-side* as accurately as R_{CHN} we only give an inflection point as the point where the two slopes cross. A comparison between table 4.4 and table 4.3 can be made at the two inflection points. If we accept that R_{NHC} and R_{CHN} can reproduce adequately the *S-side* and *R-side* respectively we must conclude that the better free energy profile computed with R_4 will come from the matching technique. The energy corresponding to the inflection points in table 4.3 are very similar to the points Sts, Is and Rts in table 4.4 matching the windows.

In figure 4.9 a detail of the matching windows process is shown. The different free energy profiles are matched at the maximum overlap point. What is shown in figure 4.9 is the unbiased and matched free energy profiles for the different windows. In this case it is included the overlap zones that will be discarded to obtain a single free energy profile (left of figure 4.8). We can see that there exist a simulation at the left of the graphic 4.9 in which the free energy increases abnormally with respect to the others. This simulation window is exploring a bad region in the $R \rightarrow S$ direction before the hydrogen from His297 is transferred. This fact explains the difference between the two PMF profiles computed by WHAM and by direct matching of the windows. While WHAM takes into account all the given data, including this abnormal simulation, the matching process only employs the most populated bins and discard the overlapping regions. In a posterior examination we discarded the abnormal simulations in the WHAM analysis and then the free energy profiles from the two techniques coincided and both gave the same free

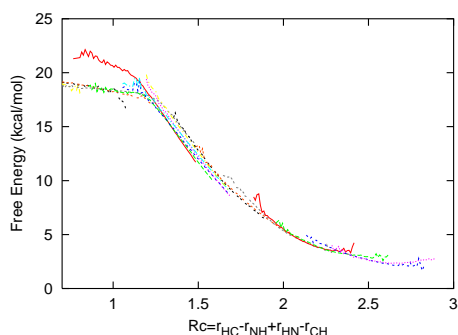


Figure 4.9: A detail of the matching windows process. Free energy profiles for every window without removing the overlapping region.

energy profile.

In the histogram of figure 4.8 we can see how the probability distribution of the reaction coordinate between two consecutive windows have a good overlap. However, the difference between the WHAM analysis and direct matching of windows forces us to think that there must be an abrupt change in some of the distances involved in the reaction coordinate.

A further analysis comes from monitoring the evolution of the four bond distances during the reaction. The plots displayed in figure 4.10 show the average of the four distances that combine in R_4 at every simulation window.

The distances r_{HC} and r_{CH} do not change brusquely, but the approximation of Lys166 and His297 represented by the distances r_{NH} and r_{HN} , respectively, is in some zones not well reproduced. An analysis of the geometries obtained during the reaction shows that when the approximation of one residue is not energetically favorable it is compensated in the reaction coordinate combination by the approximation of the conjugated residue whose movement is more labile. For example, when the first proton abstraction by Lys166 to the substrate must take place, His297 which is not coordinated to any residue and whose movement is rather free, tends to approximate to the (S)-substrate in order to compensate in the four distances combination R_4 the more energetic proton abstraction by Lys166. After the proton transfer has taken place His297 goes back to its position.

This situation is repeated more remarkably in the *R-side* with the Lys166 free movement in order to compensate the proton transfer between the (R)-mandelate substrate and His297. This is why residue Lys166 and His297 go forwards and backwards during the corresponding proton transfer in which

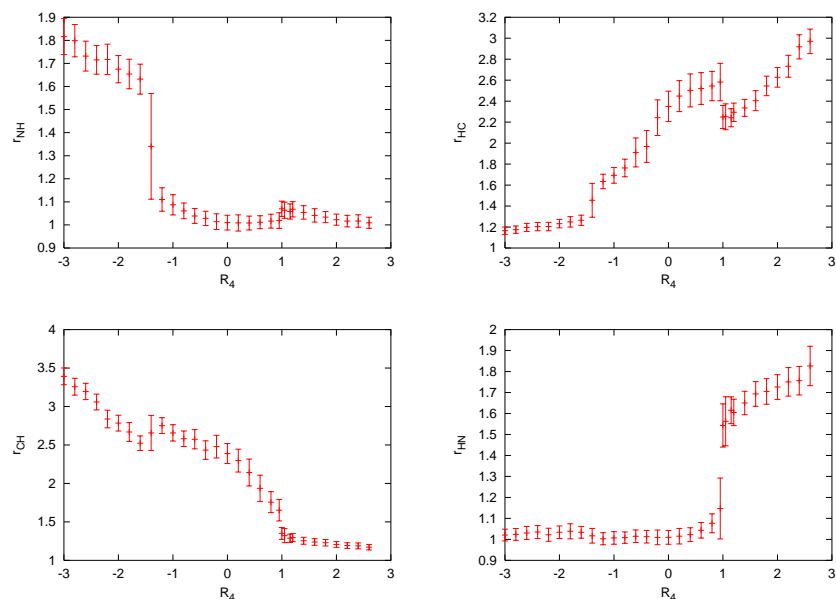


Figure 4.10: Evolution of the four bond distances in R_4 . The average value for every window is plotted along with its standard deviation.

they do not participate actively.

A solution to this problem could be weighting the distance combination in order to give preference to the important movements. For example, during the *S-side* of the reaction where Lys166 must abstract the proton, we could penalize the movements of His297 and do the same for the *R-side* with Lys166. We tried different weighting schemes mainly based on the evolution of the four distances obtained by optimization techniques (table 4.1) but the results obtained did not improve the sampling already obtained without the weighting.

4.2.4 Reactivity for the mutant N197A

Several mutagenesis experiments on Mandelate Racemase reactivity have been reported (see page 71). Some mutations on the active site show a decrease on the enzymatic reactivity rather easy to explain on the basis of the reaction mechanism that we already know (*e.g.* K166R, H297N). On the other hand, other mutation experiments such as the substitution of Asparagine 197 to an alanine, N197A, have provided a slight change in the racemization rate whose experimental explanation [223] is not clear from

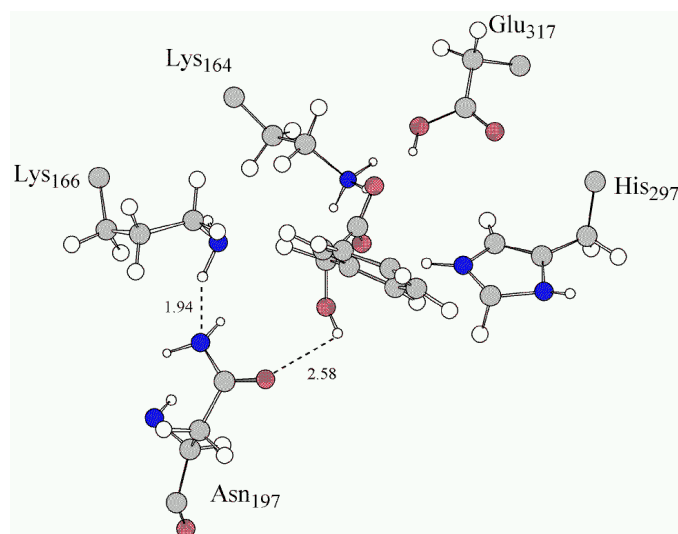


Figure 4.11: Representation of the wild type active site in a snapshot of the S reactant. Asn197 is frequently interacting with Lys166 and the substrate

the mechanistic point of view. In figure 4.11 we show the coordination of Asn197 in the active site taken from a MD snapshot of MR wild type with mandelate substrate.

The experimental work on the N197A mutation [223] points to the important role of Asn197 for the binding of intermediates analogues (α -hydroxybenzylphosphonate and benzohydroxamate derivatives) through an interaction with the α -OH group of the ligand. The enzyme binds the TS analogue with affinities 100-fold greater than that observed for the substrate, while the mutant N197A binds the analogues with less affinity. In addition to the binding energy there is also a reduction of the k_{cat} of 30-fold for (R)-mandelate and 179-fold for (S)-mandelate relative to wild-type MR.

Then it is concluded that Asn197 must stabilize the TS when the racemization with the natural substrate takes place. From the chemical point of view it is not clear the stabilization interaction between the substrate and Asn197. The amide group has low capacity to withdraw electron density to stabilize the transition state, and a possible hydrogen bond between the amide group and the hydroxyl group in the substrate should not involve any important energetic change.

On the basis of the above discussion we have performed a PMF calculation on the N197A mutant for the same mechanism III step studied in the

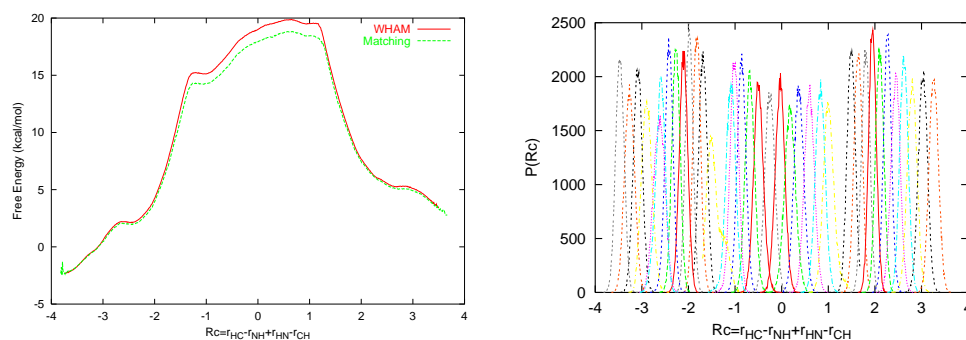


Figure 4.12: **Left:** the PMF profile using R_4 reaction coordinate. The two profiles correspond to the PMF built by WHAM or by a direct match of the overlapping windows. **Right:** Histogram diagram displaying the probability distribution.

previous sections. Although the interaction substrate(QM) and Asn197(MM) should be adequately represented we assume that the non-bonded interaction already gives the tendency that should explain the mutagenesis experiment.

By circular dichroism spectroscopy it has been seen that the mutation of Asn197 to alanine does not cause any gross structural perturbation to the secondary structure[223]. Consequently, we used the same PDB coordinates that we used for the simulation in wild type enzyme. The mutation has been done manually and the setup followed the same procedure specified above for the wild type enzyme.

We have run a PMF calculation using the same reaction coordinate R_4 combination of the four relevant bond distances. In figure 4.12 the histogram of $P(R_c)$ and the free energy profile are shown.

Some differences appear in the PMF for the mutant N197A racemization reaction with respect to the wild type MR. As we can see in the free energy profile in figure 4.12 the S and R minimum zones have not a flat profile and the system evolves towards a more stable zone. This metastability of S and R structures is due to a new configuration of Lys166 that in wild type enzyme was not found. The flat zone representing the new free energy minimum cannot be reached by the simulation because in the new configuration Lys166 is farther from the substrate and this movement cannot be scanned by the reaction coordinate R_4 .

There could be the possibility that the observed change is provoked by

an inadequate setup instead of the mutation N197A. In order to discard this possibility a longer unconstrained molecular dynamics should be propagated. However the new configuration of Lys166 was stable during 100 ps of unconstrained molecular dynamics.

This fact would explain the lowering of enzyme activity for the mutant N197A. The change of Asn197 to Ala197 removes the possibility of the hydrogen bond between Asn197 and Lys166 as displayed in figure 4.11. This interaction was stable enough to adequately orient Lys166 in order to abstract the hydrogen in (S) configuration of the substrate. When Asn197 is removed this interaction is not present and this fact has as a consequence that Lys166 is not pointing to the substrate anymore.

We cannot energetically quantify this change because, as we said, reaction coordinate R_4 does not take into account the displacement of Lys166 to the new configuration. The only information that we can extract from the free energy profile in figure 4.12 is that the system just before and after the racemization reaction is found in a configuration where Lys166 is not able to abstract the proton from the substrate. A geometrical change must occur to proceed with the reaction, and the additional energy required to promote a configurational change may explain the lower rate of racemization reaction observed in the N197A mutagenesis experiment.

4.3 Discussion and conclusions

In this final chapter we have revisited the Mandelate Racemase reactivity with its natural substrate. Molecular Dynamics umbrella-sampling simulations have been used to calculate the Potential of Mean Force. The activation free energy is computed for the mechanism III previously described. In addition, a computational simulation of the mutagenesis experiment N197A has shed some light on the molecular basis that could qualitatively explain the lowering in the rate of racemization of the substrate by the mutant enzyme.

The inclusion of temperature effects is not only a clue step to compute energy barriers. Molecular Dynamics simulations include the different parallel pathways found in condensed phase systems with optimization methods. Statistical methods collect in only one magnitude the multiplicity of pathways and eliminates the uncertainty problem in optimization methods when they fall down into a certain valley that may determine in a different way

the potential energy profile.

However, the location of saddle points on the PES has been very valuable for choosing an adequate reaction coordinate used for PMF calculations. The reaction coordinate labeled as R_4 is a combination of the four bond distances, and it is the most adequate reaction coordinate to describe the whole process. We have tried different reaction coordinates although we knew from the beginning that the whole process would not be well described by a too simple reaction coordinate because the mechanism implies two proton transfer and an inversion of configuration.

Note that the problem encountered with some reaction coordinates is not the lack of sampling in some regions, because the overlap between two adjacent windows is quite adequate. The problems reside in the inadequacy of the reaction coordinate to explore some regions whereas relevant chemical changes are taking place.

This particular case of Mandelate Racemase would probably be solved by a two-dimensional PMF calculation, which despite being very expensive is still feasible. However, this discussion could be extrapolated to enzymatic reactions where more than two or even four coordinates significantly participate in the chemical step. Even in Mandelate Racemase enzyme it has also been proposed the participation of Glu317 as a candidate residue which could withdraw negative charge during the racemization process. Glu270 coordinated to His297 has also been proposed as a catalytic dyad in order to decrease the pK_a of His297. These residues altogether would imply asynchronously the participation of four proton transfers and eight bond distances. A PMF of this complexity is very difficult to calculate whereas the location of the saddle points would give a very useful information in order to decide the reaction coordinate.

Chapter 5

Final Conclusions

1. The QM/MM model of Mandelate Racemase studied in chapter 2 is able to explain the experimental reactivity trends for the three different substrates.
2. However from this QM/MM study some problems need to be clarified: We need an accurate and efficient optimizer to locate saddle points. An enzymatic study is not complete until we include the temperature effects to collect the ensemble of parallel reaction pathways in one single thermodynamic magnitude
3. The QM gas phase study in a Mandelate Racemase model is problematic due to the inherent flexibility of the active site.
4. To solve the problem of the location of stationary points in Mandelate Racemase enzyme we designed, implemented and tested the second order RFO method which optimizes minima and TS structures on QM/MM surfaces. RFO does not need the full expensive inversion of a Hessian and it has an implicit step length determination.
5. The combination of RFO with a cheap minimizer such as L-BFGS in the micro-iterative scheme is a good method to locate stationary points in big systems.
6. Some features in the micro-iterative method such as the size of the core and environment, the frequency of alternating the optimization processes in the core and in the environment and finally the interaction

between the two regions has to be carefully calibrated in order to improve the micro-iterative method.

7. In the micro-iterative method there exist a core size which is a compromise between computational cost and efficiency in the search. This optimal core size will depend in particular on the number of atoms involved in the reaction step under study.
8. The 1SCF/MM method is a very good strategy to obtain the same results as full-SCF-QM/MM but with a cheaper computational cost. The 1SCF/MM method will be very useful when the QM level is expensive or many micro-iterations are required.
9. We show that an accurate TS location is always recommended. It prevents from wrong conclusions about the mechanism when the TS structure is located with a rough method such as the coordinate scan.
10. We include a section where a method to avoid the storage of very big Hessians is tested. Unfortunately the method fails in the convergence of the iterative diagonalization of the Augmented Hessian matrix. The problems may be due to the non-sparsity of a Cartesian coordinate Hessian and the possible linear dependency of the AH matrix.
11. The calculations of the potential of mean force for the Mandelate Racemase reaction permits to obtain the free energy profile. From the previous location of TS we concluded that the central step of the reaction is a double proton transfer coupled to a configuration inversion of the α carbon. This makes that the geometrical reaction coordinate for the PMF calculation is not obvious.
12. A combination of four interatomic distances is the only reaction coordinate that permits to sample rather smoothly all the regions of the racemization reaction path.
13. As a conclusion of the conclusions the main thesis of this thesis is that the previous exploration of the PES of an enzymatic system is a very recommended task before a free energy computation is performed, for its low computational cost and its valuable insight into the reaction pathways.

14. However we cannot conclude without looking to the future. An improvement on the quality of the PES for enzymatic systems would permit us to move from qualitative to predictive results. Moreover, the PMF calculations still have an extreme dependence on the too short phase space exploration carried out by MD simulations and on the uncertainty involved in the *a priori* choice of a reaction coordinate.

Appendix A

Computer resources and informatics

Computers play today such an important role in theoretical chemistry that it deserves at least few words. Actually, we owe a very important part of the emergence of theoretical chemistry to the high power and the low cost of nowadays computers.

Van Gunsteren and Mark [262] stated the quality of the simulated properties of a molecular system will depend on

- i) the quality of the theory or model,
- ii) the accuracy of the interatomic interaction function or force field,
- iii) the degree of sampling, statistics and convergence reached in the simulation,
- iv) the quality of the simulation software, and
- v) how competently the simulation software is used.

The two last points depend on how capable is the theoretical chemist with the computer and the computer tools used.

The computational tools developed in this thesis need software of good quality that is nowadays available at no cost thanks to the GNU¹ initiative and many other initiatives that distribute their free software on the web. Most of this software can be installed under the GNU/Linux operative system which is also free and and it can be installed in a cheap and powerful personal computer. It means that we can have useful tools such as numerical and graphic libraries, different compilers, databases and many different

¹ GNU: <http://www.gnu.org/>

kinds of software at very low cost. Obviously the production of results needs very powerful computational resources and sometimes the availability of efficient computer facilities requires a constant maintenance work which is also expensive.

In any case, it is out of discussion that a deep knowledge of the tools in informatics and computational science improves the efficiency and accuracy of the theoretical chemist work, not only to compute the numerical results but to manipulate, validate and analyze the data. I will give as example, two cases in which there are factors that can influence on the numerical accuracy of the computed results.

A.1 An example of numerical computation

Most of results of this thesis are based on numerical computations carried out in a computer. In some sections we have discussed the efficiency of a strategy depending on the number of steps to converge or on the CPU time that the process requires. But we must know that these results have a certain level of uncertainty. The performance of a numerical computation goes through several steps that have its own numerical error. The writing of the source code can be critic. Moreover, we need some external libraries such as diagonalization subroutines as well as the compiler and the compiling options that must have to be chosen with care.

In order to exemplify this fact we have performed two type of systematic benchmarks. The first type is a comparison of diagonalization subroutines and the second a minimization of a QM/MM model of Mandelate Racemase.

Diagonalization of an Augmented Hessian:

We have diagonalized an Augmented Hessian matrix with a dimension of 445×445 . In table A.1 we show the RMS between the first six eigenvectors obtained by three standard diagonalizers. The jacobi, MHQR II (the standard diagonalizer from mopac 4.0 source code) and DSPEV from LAPACK[246] have been compiled at four different levels of optimization with the GNU-g77 Fortran compiler 2.95.

The only conclusions that we want to extract from table A.1 is that although there is small difference between the results obtained at different compiler options and by different subroutines, it is not always zero. Therefore, a process such as the optimization of a structure that needs to diago-

Diagonalizer	Optimization Level in g77			
	-O	-O2	-O3	-O6
Jacobi vs MHQRII	0.00449	0.00449	0.00449	0.00449
	0.00449	0.00449	0.00449	0.00449
	0.00449	0.00449	0.00449	0.00449
	0.00449	0.00449	0.00449	0.00449
	3.89E-13	3.89E-13	3.89E-13	3.89E-13
	0.00449	0.00449	0.00449	0.00449
Jacobi vs DSPEV	0.00449	0.	0.00449	0.00449
	0.	0.	0.	0.
	5.02E-13	5.02E-13	5.02E-13	5.02E-13
	3.17E-13	3.17E-13	3.17E-13	3.17E-13
	3.89E-13	3.17E-13	3.17E-13	3.17E-13
	0.00449	0.00449	0.00449	0.00449
MHQRII vs DSPEV	0.	0.00449	0.	0.
	0.00449	0.00449	0.00449	0.00449
	0.00449	0.00449	0.00449	0.00449
	0.00449	0.00449	0.00449	0.00449
	0.	2.242E-13	2.247E-13	2.247E-13
	0.	0.	0.	0.

Table A.1: Benchmarks testing the compiler optimization level and the subroutine of a matrix diagonalization . The RMS between the two eigenvectors obtained with two different subroutines and compiled with different options is given.

nalize many matrices will accumulate a numerical error that will influence the final results.

Minimization of a QM/MM system:

In table A.2 we show a minimization of Mandelate Racemase model studied in section 2.3 by LBFGS procedure. We can see that different compilers take a different number of iterations to converge and different CPU time. In any case, all processes find the same stationary point with the same energy. In a systematic study on the efficiency of a method the compiler and compiler options must be kept fixed in order to obtain reliable results.

Optimization level	Compiler		
	g77 ¹	ifc ²	pgf77 ³
-O	496(4315.70)	475(1434.23)	493(1761.34)
-O2	484(4453.83)	475(1435.54)	493(1755.41)
-O3	525(4577.82)	465(1424.35)	493(1761.15)

¹ GNU fortran compiler 2.95

² Intel fortran compiler 5.0.1

³ Portland Group Fortran compiler 4.0-2

Table A.2: Number of steps required to minimize the Mandelate Racemase QM/MM system shown in section 2.3. In brackets the CPU time spent. The convergence criteria is $5 \cdot 10^{-3}$ kcal/(mol·Å).

Appendix B

Source Code

The source code developed in this thesis to carry out the molecular optimizations (TSSEARCH) has been implemented in ROAR 2.0 [235] and CHARMM [53] c28b2. The version implemented in ROAR is called from the main program (sander subroutine) and it can be downloaded from the web

<http://www.qf.uab.es/prat/prog/ts>

The version interfaced with CHARMM package is called from the minimization module (minimize.src) and it will be also available along with a short documentation file.

The general scheme of TSSEARCH code is outlined in the scheme B.1 and the name of the subroutines are intended to be autoexplicative. It requires as input the geometry in Cartesian coordinates and the labels that indicate the partition of the system in core and environment zones. As the program proceeds it requires the energy and the gradient from the external program. The subroutine energy_qmmm is the interface between TSSEARCH and external energy and gradient code and it must be modified depending on the requirements of the external program.

In some cases the original energy subroutines had to be modified, for instance in the ESP/MM or QM(1SCF)/MM approximations. Since either the original or modified subroutines cannot be distributed please contact the author of this thesis for any question.

Finally you may find also on the web the source code that calculates the potential of mean force from the distribution function output by CHARMM used in chapter 4.

<http://www.qf.uab.es/prat/prog/pmf>

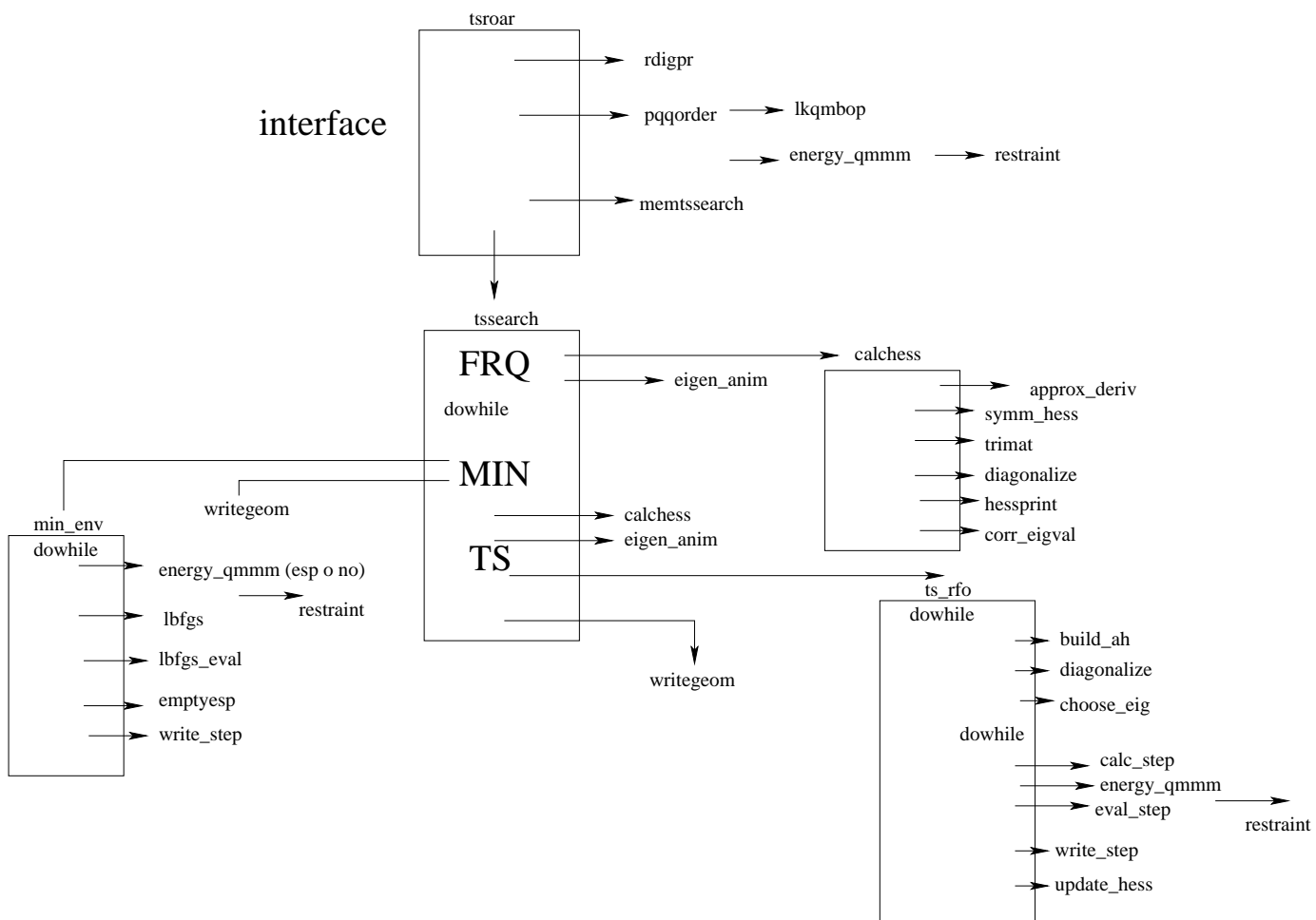


Figure B.1: General scheme for TSSEARCH program

Acronyms

- AH** Augmented Hessian in RFO scheme (section [3.1](#))
- AM1** Austin Model 1. Semiempirical Hamiltonian by Dewar and co-workers [[31](#)]
- CPU** Central Processing Unit
- DIIS** Direct Inversion of Iterative Space optimization method [[133](#)]
- DFT** Density Functional Theory (section [1.2.1.6](#))
- GHO** Generalized Hybrid Orbital method for the QM/MM frontier [[91](#)]
- IRC** Intrinsic Reaction Coordinate (section [1.3.5.2](#))
- HF** Hartree-Fock method (section [1.2.1.4](#))
- L-BFGS** Limited memory Broyden Fletcher Goldfarg Shanno method [[153](#)]
- LNR** Lagrange-Newton-Raphson iterative diagonalization technique (page [163](#))
- MC** Monte Carlo (section [1.4](#))
- MD** Molecular Dynamics (section [1.4](#))
- MEP** Minimum Energy Path (section [1.3.5](#))
- MM** Molecular Mechanics (section [1.2.2.1](#))
- MR** Mandelate Racemase Enzyme (section [2.1](#))
- PBC** Periodic Boundary Conditions (section [1.2.2.2](#))
- PES** Potential Energy Surface (section [1.2.1.2](#))

- PGA** phenylglycidate, inhibitor of Mandelate Racemase (section 2.1)
- PM3** Parametrized Model 3. Semiempirical Hamiltonian [32]
- qNR** Quasi Newton-Raphson. Newton Raphson optimization using an approximated second derivative matrix. Quasi Newton methods (section 1.3.4.1)
- qNR-BFGS** Quasi Newton-Raphson with BFGS update (see page 124)
- qNR-LBFGS** Quasi Newton-Raphson with limited memory BFGS update (see page 124)
- QM/MM** quantum mechanics /molecular mechanics strategy [66]
- RFO** Rational Function Optimization [130]
- RFO-mBFGS** Rational Function Optimization with modified BFGS update (see page 124)
- RMS** Root Mean Square. In this thesis the term is used in particular for the gradient norm.
- SBMD** Stochastic Boundary Molecular Dynamics method [185]
- SD** Steepest Descent (section 1.3.3)
- SRP** Specific Reaction Parameters. In semiempirical context some empirical parameters are modified for an specific reaction
- TS** Transition state. Depending on the context this can be the term for a first-order saddle point on a PES.
- ZPE** Zero Point Energy

Bibliography

Bibliography

- [1] Prat-Resina, X., Garcia-Viloca, M., González-Lafont, A., Lluch, J. M. On the modulation of the substrate activity for the racemization catalyzed by mandelate racemase enzyme. a qm/mm study. *Phys. Chem. Chem. Phys.* 4:5365–5371, 2002.
- [2] Prat-Resina, X., Garcia-Viloca, M., Monard, G., González-Lafont, A., Lluch, J. M., Anglada, J. M., Bofill, J. M. The search for stationary points on a quantum mechanical/molecular mechanical potential-energy surface. *Theor. Chem. Acc.* 107:147–153, 2002.
- [3] Prat-Resina, X., Bofill, J. M., González-Lafont, A., Lluch, J. M. Geometry optimization and transition state search in enzymes: Different options in the micro-iterative method. *Int. J. Quant. Chem.* 98(4):367–377, 2004.
- [4] Prat-Resina, X., González-Lafont, A., Lluch, J. M. How important is the refinement of transition state structures in enzymatic reactions? *J. Mol. Struct. (Theochem)* 632:297–307, 2003.
- [5] Monard, G., Prat-Resina, X., González-Lafont, A., Lluch, J. M. Determination of enzymatic reaction pathways using qm/mm methods. *Int. J. Quant. Chem.* 93:229–244, 2003.
- [6] Nam, K., Prat-Resina, X., Garcia-Viloca, M., Devi-Kesavan, L. S., Gao, J. Dynamics of an enzymatic substitution reaction in haloalkane dehalogenase. *J. Am. Chem. Soc.* 126:1369–1376, 2004.
- [7] Prat-Resina, X., González-Lafont, A., Lluch, J. M. Free energy calculations on different reaction coordinates of mandelate racemase. in preparation.
- [8] Pauling, L., Wilson, E. B. *Introduction to Quantum Mechanics. With Applications to Chemistry.* New York: Dover. 1985.
- [9] Pilar, F. L. *Elementary Quantum Chemistry: McGraw-Hill Inc.* 1968.
- [10] Daudel, R., Leroy, G., Peters, D., Sana, M. *Quantum Chemistry.* New York: John Wiley & Sons. 1983.
- [11] Jensen, F. *Introduction to Computational Chemistry.* West Sussex, England: John Wiley & Sons. 1999.
- [12] Fletcher, R. *Practical methods of optimization.* 2nd Ed. Tiptree, Essex, United Kingdom: John Wiley & Sons. 1987.
- [13] Leach, A. R. *Molecular Modelling. Principles and Applications.* 2nd edition Ed. Essex, England: Pearson Education. 2001.
- [14] Schlick, T. *Molecular modeling and simulation. An Interdisciplinary Guide.* New York: Springer. 2002.
- [15] Allen, M. P., Tildesley, D. J. *Computer Simulation of Liquids.* Oxford: Oxford University Press. 1987.
- [16] McQuarrie, D. A. *Statistical Mechanics.* Sausalito, California: University Science Books. 2000.
- [17] Nye, M. J. *From Chemical Philosophy to Theoretical Chemistry: Dynamics of Matter and Dynamics of Disciplines, 1800-1950: University of California Press.* 1994.
- [18] Levine, I. N. *Química Cuántica. Spanish version from the original "quantum chemistry"* Ed. Madrid: AC. 1977.
- [19] Cramer, C. J. *Essentials of Computational Chemistry : Theories and Models: John Wiley & sons.* 2002.
- [20] Nakamura, H. *Nonadiabatic Transition: Concepts, Basic Theories and Applications.* Singapore: World Scientific. 2002.
- [21] Baer, M. Introduction to the theory of electronic non-adiabatic coupling terms in molecular systems. *Phys. Rev.* 358:75–142, 2002.
- [22] Handy, N. C., Yamaguchi, Y., Schaefer III, H. F. The diagonal correction to the born-oppenheimer approximation: Its effect on the singlet-triplet splitting of ch_2 and other molecular effects. *J. Chem. Phys.* 84(8):4481–4484, 1986.
- [23] Wilson, E., Decius, J. C., Cross, P. *Molecular Vibrations.* New York: Dover. 1980.
- [24] Kosloff, R. Time-dependent quantum-mechanical methods for molecular dynamics. *J. Phys. Chem.* 92:2087–2100, 1988.

- [25] Miller, W. H. The semiclassical initial value representation: A potentially practical way for adding quantum effects to classical molecular dynamics simulations. *J. Phys. Chem. A* 105:2942–2955, 2001.
- [26] Beck, M., JaKckle, A., Worth, G., Meyer, H.-D. The multiconfiguration time-dependent hartree (mctdh) method: a highly efficient algorithm for propagating wavepackets. *Phys. Rep.* 324:1–105, 2000.
- [27] Issue dedicated to time-dependent quantum molecular dynamics. *J. Phys. Chem. A* 103(47).
- [28] Szabo, A., Ostlund, N. S. *Modern Quantum Chemistry*. New York: Dover. 1983.
- [29] Goedecker, S. Linear scaling electronic structure methods. *Rev. Mod. Phys.* 71(4):1085–1123, 1999.
- [30] Dewar, M. J. S., Thiel, W. Ground states of molecules, 38. the mndo method. approximations and parameters. *J. Am. Chem. Soc.* 99:4899–4907, 1977.
- [31] Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., Stewart, J. J. P. Am1: A new general purpose quantum mechanical model. *J. Am. Chem. Soc.* 107:3902–3909, 1985.
- [32] Stewart, J. J. P. Optimization of parameters for semiempirical methods. i. method. *J. Comp. Chem.* 10:209–220, 1989.
- [33] Stewart, J. J. P. Optimization of parameters for semiempirical methods. ii. applications. *J. Comp. Chem.* 10:221–264, 1989.
- [34] Clark, T. Quo vadis semiempirical mo-theory? *J. Mol. Struct. (Theochem)* 530:1–10, 2000.
- [35] Pople, J. A., Beveridge, D. L. *Approximate Molecular Orbital Theory*. New York: McGraw-Hill. 1970.
- [36] Bernal-Uruchurtu, M. I., Ruiz-Lopez, M. F. Basic ideas for the correction of semiempirical methods describing h-bonded systems. *Chem. Phys. Lett* 330:118–124, 2000.
- [37] Hohenberg, P., Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* 136:864, 1964.
- [38] Ziegler, T. Approximate density functional theory as a practical tool in molecular energetics and dynamics. *Chem. Rev.* 91:651–667, 1991.
- [39] Kohn, W. Nobel lecture: Electronic structure of matter wave functions and density functionals. *Rev. Mod. Phys.* 71(5):1253–1266, 1999.
- [40] Kohn, W., Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev. A* 140:1133, 1965.
- [41] Parr, R. G., Yang, W. *Density Functional Theory*: Oxford University Press. 1989.
- [42] Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* 38:3098–3100, 1988.
- [43] Lee, C., Yang, W., Parr, R. G. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* 37:785–789, 1988.
- [44] Becke, A. D. Density-functional thermochemistry. iii. the role of exact exchange. *J. Chem. Phys.* 98(7):5648–5652, 1993.
- [45] Koch, W., Holthausen, M. C. *Chemist’s Guide to Density Functional Theory*. Weinheim: Wiley-VCH. 2000.
- [46] Luchow, A., Anderson, J. B. Monte Carlo methods in electronic structures for large systems. *Annu. Rev. Phys. Chem.* 51:501–526, 2000.
- [47] Head-Gordon, M. Quantum chemistry and molecular processes. *J. Phys. Chem.* 100(31):13213–13225, 1996.
- [48] Steinbach, P. J., Brooks, B. R. New spherical-cut-off methods for long-range forces in macromolecular simulation. *J. Comput. Chem* 15:667–683, 1994.
- [49] Feller, S. E., Pastor, R. W., Rojnuckarin, A., Bogusz, S., Brooks, B. R. Effect of electrostatic force truncation on interfacial and transport properties of water. *J. Phys. Chem.* 100:17011–17020, 1996.
- [50] Frenkel, D., Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*. San Diego, CA: Academic Press. 1996.
- [51] Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G. S., Profeta, J., Weiner, P. -. *J. Am. Chem. Soc.* 106:765, 1984.
- [52] Pearlman, D., Case, D., Caldwell, J., Ross, W., Cheatham, T., Debolt, S., Ferguson, D., Seibel, G., Kollman, P. Amber, a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Comm.* 91(1-3):1–41, 1995.
- [53] Brooks, B. R., Brucoleri, 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.* 4(2):187–217, 1983.
- [54] 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., Wiórkiewicz-Kuczera, J., Yin, D., Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102:3586–3616, 1998.

- [55] Jorgensen, W. L., Maxwell, D. S., Tirado-Rives, J. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118(45):11225–11236, 1996.
- [56] Scott, W. R. P., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Krüger, P., van Gunsteren, W. F. The gromos biomolecular simulation program package. *J. Phys. Chem. A* 103(19):3596–3607, 1999.
- [57] Price, D. J., Brooks III, C. L. Modern protein force fields behave comparably in molecular dynamics simulations. *J. Comput. Chem.* 23(23):1045–1057, 2002.
- [58] Cramer, C. J., Truhlar, D. G. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.* 99:2161–2200, 1999.
- [59] Roux, B. *Computational biochemistry & biophysics*. New York: Marcel Dekker Inc. 2001.
- [60] Cui, Q. Combining implicit solvation models with hybrid quantum mechanical/molecular mechanical methods: A critical test with glycine. *J. Chem. Phys.* 117(10):4720–4728, 2002.
- [61] Orozco, M., Luque, F. J. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.* 100:4187–4225, 2000.
- [62] Garcia-Viloca, M., Gao, J., Karplus, M., G.Truhlar, D. How enzymes work: analysis by modern rate theory and computer simulations. *Science* 303:186–195, 2004.
- [63] Matsubara, T., Maseras, F., Koga, N., Morokuma, K. Application of the new "integrated mo + mm" (imomm) method to the organometallic reaction $\text{pt}(\text{pr}_3)_2 + \text{h}_2$ ($r = \text{h, me, t-bu, and ph}$). *J. Phys. Chem.* 100(7):2573–2580, 1996.
- [64] Maseras, F., Lledós, A. *Computational modeling of homogeneous catalysis*. Dordrecht (Holland): Kluwer. 2002: 1–21.
- [65] Sierka, M., Sauer, J. Finding transition structures in extended systems: A strategy based on a combined quantum mechanics empirical valence bond approach. *J. Chem. Phys.* 112(16):6983–6996, 2000.
- [66] Warshel, A., Levitt, M. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* 103:227–249, 1976.
- [67] Singh, U. C., Kollman, P. A. A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: applications to the $\text{ch}_3\text{cl} + \text{cl}^-$ exchange reaction and gas phase protonation of polyethers. *J. Comput. Chem.* 7(6):718–730, 1986.
- [68] Field, M. J., Bash, P. A., Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* 11(6):700–733, 1990.
- [69] Warshel, A., Karplus, M. Calculation of ground and excited state potential surfaces of conjugated molecules. i. formulation and parametrization. *J. Am. Chem. Soc.* 94(16):5612–5625, 1972.
- [70] Gao, J., Thompson, M. A., eds. *Combined Quantum Mechanical and Molecular Mechanical Methods*. ACS Symposium Series 712 Washington D.C.: American Chemical Society. 1998.
- [71] Monard, G., Merz, K. M. Combined quantum mechanical/molecular mechanical methodologies applied to biomolecular systems. *Acc. Chem. Res.* 32:904–911, 1999.
- [72] Field, M. J. Simulating enzyme reactions: challenges and perspectives. *J. Comput. Chem.* 23(1):48–58, 2002.
- [73] Bakowies, D., Thiel, W. Hybrid models for combined quantum mechanical and molecular mechanical approaches. *J. Phys. Chem.* 100:10580–10594, 1996.
- [74] Besler, B. H., Merz Jr., K. M., Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* 11:431–439, 1990.
- [75] Luque, F. J., Reuter, N., Cartier, A., Ruiz-Lopez, M. F. Calibration of the quantum/classical hamiltonian in semiempirical qm/mm am1 and pm3 methods. *J. Phys. Chem. A* 104:10923–10931, 2000.
- [76] Cummins, P. L., Gready, J. L. Coupled semiempirical quantum mechanics and molecular mechanics (qm/mm) calculations on the aqueous solvation free energies of ionized molecules. *J. Comput. Chem.* 20:1028, 1999.
- [77] Freindorf, M., Gao, J. Optimization of the lennard-jones parameters for a combined ab-initio quantum-mechanical and molecular mechanical potential using the 3-21g basis-set. *J. Comput. Chem.* 17(4):386–395, 1996.
- [78] Martin, M. E., Aguilar, M. A., Chalmet, S., Ruiz-Lopez, M. F. An iterative procedure to determine lennard-jones parameters for their use in quantum mechanics/molecular mechanics liquid state simulations. *Chem. Phys.* 284:607–614, 2002.
- [79] Ranganathan, S., Gready, J. E. Hybrid quantum and molecular mechanical (qm/mm) studies on the pyruvate to l-lactate interconversion in l-lactate dehydrogenase. *J. Phys. Chem. B* 101(28):5614–5618, 1997.
- [80] Reuter, N., Dejaegere, A., Maigret, B., Karplus, M. Frontier bonds in qm/mm methods: a comparison of different approaches. *J. Phys. Chem. A* 104(8):1720–1735, 1999.

- [81] Hall, R. J., Hindle, S. A., Burton, N. A., Hillier, I. H. Aspects of hybrid qm/mm calculations: the treatment of the qm/mm interface region and geometry optimization with an application to chorismate mutase. *J. Comput. Chem.* 21(16):1433–1441, 2000.
- [82] Antes, I., Thiel, W. Adjusted connection atoms for combined quantum mechanical and molecular mechanical methods. *J. Phys. Chem. B* 103:9290–9295, 1999.
- [83] Cummins, P. L., Gready, J. E. Combined quantum and molecular mechanics (qm/mm) study of the ionization state of 8-methylpterin substrate bound to dihydrofolate reductase. *J. Phys. Chem. B* 104(18):4503–4510, 2000.
- [84] Zhang, Y., Liu, H., Yang, W. Free energy calculations on enzyme reactions with an efficient iterative procedure to determine minimum energy paths on a combined ab initio qm/mm potential energy surface. *J. Chem. Phys.* 112(8):3483–3492, 2000.
- [85] Das, D., Eurenus, K. P., Billings, E. M., Sherwood, P., Chatfield, D. C., Hodoscek, M., Brooks, B. R. Optimization of quantum mechanical molecular mechanical partitioning schemes: Gaussian delocalization of molecular mechanical charges and the double link atom method. *J. Chem. Phys.* 117(23):10534–10547, 2002.
- [86] DiLabio, G. A., Hurley, M. M., Christiansen, P. A. Simple one-electron quantum capping potentials for use in hybrid qm/mm studies of biological molecules. *J. Chem. Phys.* 116(22):9578–9584, 2002.
- [87] Swart, M. Addremove: A new link model for use in qm/mm studies. *Int. J. Quant. Chem.* 91:177–183, 2003.
- [88] Théry, V., Rinaldi, D., Rivail, J.-L., Maigret, B., Frency, B. Quantum mechanical computations on very large molecular systems: the local self-consistent field method. *J. Comp. Chem.* 15:269–282.
- [89] Monard, G., Loos, M., Théry, V., Baka, K., Rivail, J.-L. Hybrid classical quantum force field for modeling very large molecules. *Int. J. Quantum Chem.* 58:153–159, 1996.
- [90] Nicolas Ferré, J.-L. R. Xavier Assfeld. Specific force field parameters determination for the hybrid ab initio qm/mm lscf method. *J. Comp. Chem.* 23:610–624, 2002.
- [91] Gao, J., Amara, P., Alhambra, C., Field, M. J. A generalized hybrid orbital (gho) method for the treatment of boundary atoms in combined qm/mm calculations. *J. Phys. Chem. A* 12(24):4714–4721, 1998.
- [92] Amara, P., Field, M. J., Alhambra, C., Gao, J. The generalized hybrid orbital (gho) method for combined qm/mm calculations: Formulation and tests of the analytical derivatives. *Theor. Chem. Acc.* 104:336–343, 2000.
- [93] Garcia-Viloca, M., Gao, J. Generalized hybrid orbital for the treatment of boundary atoms in combined quantum mechanical and molecular mechanical calculations using the semiempirical parameterized model 3 method. *Theor. Chem. Acc. ASAP.*
- [94] Pu, J., Gao, J., Truhlar, D. G. Generalized hybrid orbital (gho) method for combining ab initio hartree-fock wave functions with molecular mechanics. *J. Phys. Chem. A* 108(4):632–650, 2004.
- [95] Murphy, R., Philipp, D., Friesner, R. Frozen orbital qm/mm methods for density functional theory. *Chem. Phys. Lett.* 321:113–120, 2000.
- [96] Hyperchem. HyperChem Users Manual. 1998.
- [97] Cui, Q., Elstner, M., Kaxiras, E., Frauenheim, T., Karplus, M. A qm/mm implementation of the self-consistent charge density functional tight binding (scc-dftb) method. *J. Phys. Chem. B* 105:569–585, 2001.
- [98] Formanek, M. S., Li, G., Zhang, X., Cui, Q. Modeling zinc in biomolecules with the self consistent charge-density functional tight binding (scc-dftb) method: applications to structural and energetic analysis. *J. Comp. Chem.* 24:565–581, 2003.
- [99] Lee, Y. S., Worthington, S. E., Krauss, M., Brooks, B. R. Reaction mechanism of chorismate mutase studied by the combined potentials of quantum mechanics and molecular mechanics. *J. Phys. Chem. B* 106:12059–12065, 2002.
- [100] Lyne, P. D., Hodoscek, M., Karplus, M. A hybrid qm-mm potential employing hartree-fock or density functional methods in the quantum region. *J. Phys. Chem. A* 103:3462–3471, 1999.
- [101] Tuñón, I., Martins-Costa, M., Millot, C., Ruiz-López, M., Rivail, J. A coupled density functional-molecular mechanics monte carlo simulation method: The water molecule in liquid water. *J. Comput. Chem.* 17(1):19–29, 1996.
- [102] Tuñón, I., Martins-Costa, M., Millot, C., Ruiz-López, M. Molecular dynamics simulations of elementary chemical processes in liquid water using combined density functional and molecular mechanics potentials. i. proton transfer in strongly h-bonded complexes. *J. Chem. Phys.* 106(9):3633–3642, 1997.
- [103] Friesner, R. A., Dunietz, B. D. Large-scale ab initio quantum chemical calculations on biological systems. *Acc. Chem. Res.* 34:351–358, 2001.
- [104] 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., c, A. J. T., Billeter, S., Terstegen, F., Thiel, S., Kendrick, J., Rogers, S. C., Casci, J., Watson, M., King, F., Karlsen, E., Sjøvoll, M., Fahmi, A., Schafer, A., Lennartz, C. Quasi: A general purpose implementation of the qm/mm approach and its application to problems in catalysis. *J. Mol. Struct. (Theochem)* 632:1–28, 2003.

- [105] Kongsted, J., Osted, A., Mikkelsen, K. V., Christiansen, O. Coupled cluster/molecular mechanics method: Implementation and application to liquid water. *J. Phys. Chem. B* 107:2578–2588, 2003.
- [106] Car, R., Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* 55(22):2471–2474, 1985.
- [107] Carloni, P., Rothlisberger, U., Parrinello, M. The role and perspective of ab initio molecular dynamics in the study of biological systems. *Acc. Chem. Res.* 35:455–464, 2002.
- [108] Maseras, F., Morokuma, K. Imomm: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J. Comput. Chem.* 16(9):1170–1179, 1995.
- [109] Dapprich, S., Komáromi, I., Byun, K. S., Morokuma, K., Frisch, M. J. A new oniom implementation in gaussian98. part i. the calculation of energies, gradients, vibrational frequencies and electric field derivatives. *J. Mol. Struct. (Theochem)* 461:1–21, 1999.
- [110] Hurley, M. M., Wright, J. B., Lushington, G. H., White, W. E. Quantum mechanics and mixed quantum mechanics/molecular mechanics simulations of model nerve agents with acetylcholinesterase. *Theor. Chem. Acc.* 109:160–168, 2003.
- [111] Vreven, T., Morokuma, K. Investigation of the $s_0 \rightarrow s_1$ excitation in bacteriorhodopsin with the oniom (mo:mm) hybrid method. *Theor. Chem. Acc.* 109(3):125–132, 2003.
- [112] Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*: New York, 1992.
- [113] Warshel, A. Molecular dynamics simulations of enzymatic reactions. *Acc. Chem. Res.* 35:385–395, 2002.
- [114] Warshel, A., Parson, W. W. Dynamics of biochemical and biophysical reactions : insight from computer simulations. *Quarterly Review of Biophysics* 4:563–679, 2001.
- [115] Warshel, A. Computer simulations of enzyme catalysis: Methods, progress, and insights. *Ann. Rev. Biophys. Biomol. Struct.* 32:425–443, 2003.
- [116] Mo, Y., Gao, J. An ab initio molecular orbital-valence bond (movb) method for simulating chemical reactions in solution. *J. Phys. Chem. A* 104:3012–3020, 2000.
- [117] Devi-Kesavan, L. S., Garcia-Viloca, M., Gao, J. Semiempirical qm/mm potential with simple valence bond (svb) for enzyme reactions. application to the nucleophilic addition reaction in haloalkane dehalogenase. *Theor. Chem. Acc.* 109(3):133–139, 2003.
- [118] Hong, G., Strajbl, M., Wesolowski, T. A., Warshel, A. Constraining the electron densities in dft method as an effective way for ab initio studies of metal-catalyzed reactions. *J. Comp. Chem.* 21(16):1554–1561, 2000.
- [119] Cummins, P. L., Gready, J. E. Computational methods for the study of enzymic reaction mechanisms. ii. an overlapping mechanically embedded method for hybrid semiempirical-qm/mm calculations. *J. Mol. Struct. (Theochem)* 632:247–257, 2003.
- [120] Hayashi, S., Ohmine, I. Proton transfer in bacteriorhodopsin: Structure, excitation, ir spectra, and potential energy surface analyses by an ab initio qm/mm method. *J. Phys. Chem. B* 104:10678–10691, 2000.
- [121] Poteau, R., Ortega, I., Alary, F., Solis, A. R., Barthelat, J.-C., Daudey, J.-P. Effective group potentials. 1. method. *J. Phys. Chem. A* 105:198–205, 2001.
- [122] Kairys, V., Jensen, J. H. Qm/mm boundaries across covalent bonds: A frozen localized molecular orbital-based approach for the effective fragment potential method. *J. Phys. Chem. A* 104:6656–6665, 2000.
- [123] Gogonea, V., Westerhoff, L. M., Merz Jr., K. M. Quantum mechanical/quantum mechanical methods. i. a divide and conquer strategy for solving the schrodinger equation for large molecular systems using a composite density functional semiempirical hamiltonian. *J. Chem. Phys.* 113(14):5604–5613, 2000.
- [124] Cui, Q., Guo, H., Karplus, M. Combining ab initio and density functional theories with semiempirical methods. *J. Chem. Phys.* 117(12):5617–5631, 2002.
- [125] Komeiji, Y., Nakano, T., Fukuzawa, K., Ueno, Y., Inadomi, Y., Nemoto, T., Uebayasi, M., G.Fedorov, D., a, K. K. Fragment molecular orbital method:i application to molecular dynamics simulation, ab initio fmo-md. *Chem. Phys. Lett.* 372:342–347, 2003.
- [126] Zhang, D. W., Zhang, J. Z. H. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein molecule interaction energy. *J. Chem. Phys.* 119(7):3599–3605, 2003.
- [127] Cui, Q., Karplus, M. Molecular properties from combined qm/mm methods. i. analytical second derivative and vibrational calculations. *J. Chem. Phys.* 112(3):1133–1149, 2000.
- [128] Wales, D. J. A microscopic basis for the global appearance of energy landscapes. *Science* 293:2067–2070, 2001.
- [129] Onuchic, J. N., Luthey-Schulten, Z., Wolynes, P. G. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600, 1997.

- [130] Simons, J., Jorgensen, P., Taylor, H., Ozment, J. *J. Phys. Chem.* 87:2745, 1983.
- [131] Banerjee, A., Adams, N., Simons, J., Shepard, R. Search for stationary points on surfaces. *J. Phys. Chem.* 89:52–57, 1985.
- [132] Pulay, P. Improved scf convergence acceleration. *J. Comput. Chem.* 3:556, 1982.
- [133] Császár, P., Pulay, P. Geometry optimization by direct inversion in the iterative subspace. *J. Mol. Struct.* 114:31–34, 1984.
- [134] Wittbrodt, J. M., Schlegel, H. B. Estimating stretching force constants for geometry optimization. *J. Mol. Struct. (Theochem)* 398:55–61, 1997.
- [135] Bofill, J. M. Updated hessian matrix and the restricted step method for locating transition structures. *J. Comput. Chem.* 15(1):1–11, 1994.
- [136] Anglada, J. M., Bofill, J. M. How good is a broyden-fletcher-goldfarb-shanno-like update hessian formula to locate transition structures? specific reformulation of broyden-fletcher-goldfarb-shanno for optimizing saddle points. *J. Comput. Chem.* 19(3):349–362, 1998.
- [137] Bolhuis, P. G., Chandler, D., Dellago, C., Geissler, P. L. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* 53:291–318, 2002.
- [138] Schlegel, B. Exploring potential energy surfaces for chemical reactions: An overview of some practical methods. *J. Comput. Chem.* 24(12):1515–1527, 2003.
- [139] Heidrich, D., ed. *The Reaction Path in Chemistry: Current Approaches and Perspectives*. Dordrecht: Kluwer Academic, 1995.
- [140] Eurenium, K. P., Chatfield, D. C., Brooks, B. R., Hodoscek, M. Enzyme mechanisms with hybrid quantum mechanical and molecular mechanical potentials. i. theoretical considerations. *Int. J. Quantum Chem.* 60(6):1189–1200, 1996.
- [141] Fukui, K. The path of chemical reactions - the irc approach. *Acc. Chem. Res.* 14(12):363–368, 1981.
- [142] Henkelman, G., Jóhannesson, G., Jónsson, H. Methods for finding saddle points and minimum energy paths. In: *Progress on Theoretical Chemistry and Physics*. Schwartz, S. D. ed. . Kluwer Academic Publishers 2000 269–300.
- [143] Henkelman, G., Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* 113(22):9978–9985, 2001.
- [144] Chu, J.-W., Trout, B. L., Brooks, B. R. A super-linear minimization scheme for the nudged elastic band method. *J. Chem. Phys.* 119(24):12708–12717, 2003.
- [145] Crehuet, R., Field, M. J. A temperature-dependent nudged-elastic-band algorithm. *J. Chem. Phys.* 118(21):9563–9571, 2003.
- [146] Fischer, S., Karplus, M. Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chem. Phys. Lett.* 194(3):252–261, 1992.
- [147] Dutzler, R., Schirmer, T., Karplus, M., Fischer, S. Translocation mechanism of long sugar chains across the maltoporin membrane channel. *Structure* 10(9):1273–1284, 2002.
- [148] Woodcock, H. L., Hodoscek, M., Sherwood, P., Lee, Y. S., Schaefer III, H. F., R. Brooks, B. Exploring the quantum mechanical/molecular mechanical replica path method: a pathway optimization of the chorismate to prephenate claisen rearrangement catalyzed by chorismate mutase. *Theor. Chem. Acc.* 109(3):140–148, 2003.
- [149] Pulay, P., Fogarasi, G. Geometry optimization in redundant internal coordinates. *J. Chem. Phys.* 96(4):2856–2860, 1992.
- [150] Kudin, K. N., Scuseria, G. E., Schlegel, H. B. A redundant internal coordinate algorithm for optimization of periodic systems. *J. Chem. Phys.* 114(7):2919–2923, 2001.
- [151] Eckert, F., Pulay, P., Werner, H.-J. Ab initio geometry optimization for large molecules. *J. Comput. Chem.* 18(12):1473–1483, 1997.
- [152] Nocedal, J. Updating quasi-newton matrices with limited storage. *Mathematics of Computation* 35(151):773–782, 1980.
- [153] Liu, D. C., Nocedal, J. On the limited memory bfgs method for large scale optimization. *Math. Programming* 45:503–528, 1989.
- [154] Moreales, J. L., Nocedal, J. Enriched methods for large-scale unconstrained optimization. *Comp. Opt. Appl.* 21(2):143–154, 2002.
- [155] Anglada, J. M., Besalú, E., Bofill, J. M., Rubio, J. Another way to implement the powell formula for updating hessian matrices related to transition structures. *J. Math. Chem.* 25:85–92, 1999.
- [156] Schlick, T., Overton, M. *J. Comput. Chem.* 8:1025, 1987.
- [157] Derremaux, P., Zhang, G., Schlick, T., Brooks, B. A truncated newton minimizer adapted for charmm and biomolecular applications. *J. Comput. Chem.* 15(5):532–552, 1994.
- [158] Thomas, A., Field, M. J. Reaction mechanism of the hgxpptase from *plasmodium falciparum*: A hybrid potential quantum mechanical/ molecular mechanical study. *J. Am. Chem. Soc.* 124:32–12438, 2002.

- [159] Moliner, V., Turner, A. J., Williams, I. H. Transition-state structural refinement with grace and charmm: realistic modelling of lactate dehydrogenase using a combined quantum/classical method. *J. Chem. Soc. Chem. Comm.* 14:1271–1272, 1997.
- [160] Turner, A. J., Moliner, V., Williams, I. H. Transition-state structural refinement with grace and charmm: Flexible qm/mm modelling for lactate dehydrogenase. *Phys. Chem. Chem. Phys.* 1:1323–1331, 1999.
- [161] 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.* 2:2177–2186, 2000.
- [162] Vreven, T., Morokuma, K., Farkas, Ö., Schlegel, H. B., Frisch, M. J. Geometry optimization with combined methods. i. micro-iterations and constraints. *J. Comput. Chem.* 24(6):760–769, 2003.
- [163] Németh, K., Coulaud, O., Monard, G., Ángyán, J. G. Linear scaling algorithm for the coordinate transformation problem of molecular geometry optimization. *J. Chem. Phys.* 113(6):5598–5603, 2000.
- [164] Németh, K., Coulaud, O., Monard, G., Ángyán, J. G. An efficient method for the coordinate transformation problem of massively three-dimensional networks. *J. Chem. Phys.* 114(22):9747–9753, 2001.
- [165] Paizs, B., Baker, J., Suhai, S., Pulay, P. Geometry optimization of large biomolecules in redundant internal coordinates. *J. Chem. Phys.* 113(16):6566–6572, 2000.
- [166] Tuckerman, M. E., Martyna, G. J. Understanding modern molecular dynamics: Techniques and applications. *J. Phys. Chem. B* 104:159–178, 2000.
- [167] Goldstein, H. *Mecánica Clásica*. Barcelona: Reverté. 1996. Spanish translation of the second english edition.
- [168] Truhlar, D. G., Gao, J., Alhambra, C., García-Viloca, M., Corchado, J., Sánchez, M. L., Villá, J. The incorporation of quantum effects in enzyme kinetics modeling. *Acc. Chem. Res.* 35(6):341–349, 2002.
- [169] Elber, R., Ghosh, A., Cárdenas, A. Long time dynamics of complex systems. *Acc. Chem. Res.* 35:396–403, 2002.
- [170] Issue dedicated to molecular dynamics simulations of biomolecules. *Acc. Chem. Res.* 35(6).
- [171] Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* 72(4):2384–2393, 1980.
- [172] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81(8):3684–3690, 1984.
- [173] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* 81(1):511–519, 1984.
- [174] Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* 31(3):1695–1697, 1985.
- [175] Martyna, G. J., and, M. L. K. Nosé-hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* 97(4):2635–2643, 1992.
- [176] Cheng, A., Merz Jr., K. M. Application of the nosé-hoover chain algorithm to the study of protein dynamics. *J. Phys. Chem.* 100(5):1927–1937, 1996.
- [177] Zhang, Y., Feller, S. E., Brooks, B. R., Pastor, R. W. Computer simulation of liquid/liquid interfaces. i. theory and application to octane/water. *J. Chem. Phys.* 103(23):10252–10266, 1995.
- [178] Ryckaert, J.-P., Ciccotti, G., Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.* 23:327–341, 1977.
- [179] Kräutler, V., Gunsteren, W. F. V., Hünenberger, P. H. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* 22(5):501–508, 2001.
- [180] Tobias, D. J., Brooks III, C. L. Molecular dynamics with internal coordinate constraints. *J. Chem. Phys.* 89(9):5115–5127, 1988.
- [181] Coluzza, I., Sprik, M., Ciccotti, G., Fom, A. Constrained reaction coordinate dynamics for systems with constraints. *Mol. Phys.* 101(18):2885–2894, 2003.
- [182] Issue dedicated to computational molecular biophysics. *J. Comput. Phys.* 1(151).
- [183] Brooks III, C. L., Karplus, M. Deformable stochastic boundaries in molecular dynamics. *J. Chem. Phys.* 79(12):6312–6325, 1983.
- [184] Brunger, A., Brooks III, C. L., Karplus, M. Stochastic boundary conditions for molecular dynamics simulations of st2 water. *Chem. Phys. Lett.* 105(5):495–500, 1984.
- [185] Brooks III, C. L., Brunger, A., Karplus, M. Active site dynamics in protein molecules: A stochastic boundary molecular-dynamics approach. *Biopolymers* 24:843–865, 1985.
- [186] Brooks III, C. L., Karplus, M. Solvent effects on protein motion and protein effects on solvent motion. dynamics of the active site region of lysozyme. *J. Mol. Biol.* 208:159–181, 1989.
- [187] Alhambra, C., Gao, J. Hydrogen bonding interactions in the active site of a low molecular weight protein tyrosine phosphatase. *J. Comput. Chem.* 21:1192–1203, 2000.

- [188] Li, G., Zhang, X., Cui, Q. Free energy perturbation calculations with combined qm/mm potentials complications, simplifications, and applications to redox potential calculations. *J. Phys. Chem. B* 107:8643–8653, 2003.
- [189] Garcia-Viloca, M., Alhambra, C., G. Truhlar, D., Gao, J. Hydride transfer catalyzed by xylose isomerase: Mechanism and quantum effects. *J. Comp. Chem.* 24:177–190, 2003.
- [190] Tolman, R. C. *The Principles of Statistical Mechanics*. New York: Dover Publications Inc. 1979.
- [191] Chandler, D. *Introduction to Modern Statistical Mechanics*. New York: Oxford University Press. 1987.
- [192] Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* 93:2395–2417, 1993.
- [193] Masgrau, L., Àngels González-Lafont, Lluch, J. M. Dependence of the rate constants on the treatment of internal rotation modes: The reaction $\text{oh} + \text{ch}_3\text{sh} \rightarrow \text{ch}_3\text{s} + \text{h}_2\text{o}$ as an example. *J. Comput. Chem.* 24(6):701–706, 2003.
- [194] Garcia-Viloca, M., Alhambra, C., Truhlar, D. G., Gao, J. Inclusion of quantum-mechanical vibrational energy in reactive potentials of mean force. *J. Chem. Phys.* 114(22):9953–9958, 2001.
- [195] Stanton, R. V., Dixon, S. L., Merz, Jr., K. M. Free energy perturbation calculations within quantum mechanical methodologies. In: *Computational Approaches to Biochemical reactivity*. Naray-Szabo, G., Warshel, A. eds. . Kluwer Academic Publishers Dordrecht 1997 103–123.
- [196] Torrie, G. M., Valleau, J. P. Nonphysical sampling distributions in monte carlo free energy estimation: Umbrella sampling. *J. Comput. Phys.* 23:187–199, 1977.
- [197] González-Lafont, A., Lluch, J. M., Bertrán, J. Monte carlo simulations of chemical reactions in solution. In: *Solvent Effects and Chemical Reactivity*. Tapia, O., Bertrán, J. eds. Understanding Chemical Reactivity. Kluwer Academic Publishers Dordrecht 1996 125–177.
- [198] Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A., Rosenberg, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.* 13(8):1992, 1011–1021.
- [199] Boczeko, E. M., Brooks III, C. L. Constant-temperature free energy surfaces for physical and chemical processes. *J. Phys. Chem.* 97:4509–4513, 1993.
- [200] Rajamani, R., Naidoo, K. J., Gao, J. Implementation of an adaptive umbrella sampling method for the calculation of multidimensional potential of mean force of chemical reactions in solution. *J. Comput. Chem.* 24:1775–1781, 2003.
- [201] Roux, B. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* 91:275–282, 1995.
- [202] Eyring, H. The activated complex and the absolute rate of chemical reactions. *Chem. Rev.* 17(1):65–77, 1935.
- [203] Truhlar, D., Garrett, B., Klippenstein, S. Current status of transition-state theory. *J. Phys. Chem.* 100(31):12771–12800, 1996.
- [204] Geissler, P. L., Dellago, C., Chandler, D., Hutter, J., Parrinello, M. Autoionization in liquid water. *Science* 291:2121–2124, 2001.
- [205] Gao, J., Truhlar, D. Quantum mechanical methods for enzyme kinetics. *Ann. Rev. Phys. Chem.* 53:467–505, 2002.
- [206] Himo, F., Siegbahn, P. E. M. Quantum chemical studies of radical-containing enzymes. *Chem. Rev.* 103:2421–2456, 2003.
- [207] Noodleman, L., Lovell, T., Han, W.-G., Li, J., Himo, F. Quantum chemical studies of intermediates and reaction pathways in selected enzymes and catalytic synthetic systems. *Chem. Rev.* 104(2):459–508, 2004.
- [208] York, D. M., Lee, T.-S., Yang, W. Quantum mechanical treatment of biological macromolecules in solution using linear-scaling electronic structure methods. *Phys. Rev. Lett.* 80(22):5011–5014, 1998.
- [209] Boero, M., Terakura, K., Tateno, M. Catalytic role of metal ion in the selection of competing reaction paths: A first principles molecular dynamics study of the enzymatic reaction in ribozyme. *J. Am. Chem. Soc.* 124:8949–8957, 2002.
- [210] Benkovic, S. J., Hammes-Schiffer, S. A perspective on enzyme catalysis. *Science* 301:1196–1202, 2003.
- [211] Villà, J., Warshel, A. Energetics and dynamics of enzymatic reactions. *J. Phys. Chem. B* 105:7887–7907, 2001.
- [212] Garcia-Viloca, M., González-Lafont, A., Lluch, J. M. A qm/mm study of the racemization of vinylglycolate catalyzed by mandelate racemase enzyme. *J. Am. Chem. Soc.* 123:709–721, 2001.
- [213] Gerlt, J. A., Kozarich, J. W., Kenyon, G. L., Gassman, P. G. Electrophilic catalysis can explain the unexpected acidity of carbon acids in enzyme-catalyzed reactions. *J. Am. Chem. Soc.* 113(25):9667–9669, 1991.
- [214] Mitra, B., Kallarakal, A. T., Kozarich, J. W., Gerlt, J. A., Clifton, J. R., Petsko, G. A., Kenyon, G. L. Mechanism of the reaction catalyzed by mandelate racemase: Importance of electrophilic catalysis by glutamic acid 317. *Biochemistry* 34(9):2777–2787, 1995.

- [215] Gerlt, J. A., Gassman, P. G. An explanation for rapid enzyme-catalyzed proton abstraction from carbon acids: importance of late transition states in concerted mechanisms. *J. Am. Chem. Soc.* 115(24):11552–11568, 1993.
- [216] Bearne, S. L., Wolfenden, R. Mandelate racemase in pieces: Effective concentrations of enzyme functional groups in the transition state. *Biochemistry* 36(7):1646–1656, 1997.
- [217] Kenyon, G. L., Gerlt, J. A., Petsko, G. A., Kozarich, J. W. Mandelate racemase: Structure-function studies of a pseudosymmetric enzyme. *Acc. Chem. Res.* 28:178–186, 1995.
- [218] Humphrey, W., Dalke, A., Schulten, K. Vmd: Visual molecular dynamics. *J. Mol. Graph.* 14(1):33–38, 1996.
- [219] Li, R., Powers, V. M., Kozarich, J. W., Kenyon, G. L. Racemization of vinylglycolate catalyzed by mandelate racemase. *J. Org. Chem.* 60(11):3347–3351, 1995.
- [220] Landro, J. A., Kenyon, G. L., Kozarich, J. W. Mechanism-based inactivation of mandelate racemase by propargylglycolate. *Bioorg. Med. Chem. Lett.* 2(11):1411–1418, 1992.
- [221] Goriup, M., Strauss, U. T., Felfer, U., Kroutil, W., Faber, K. Substrate spectrum of mandelate racemase part 1: Variation of the α -hydroxy acid moiety. *J. Mol. Catal. B* 15:207–212, 2001.
- [222] Whitman, C. P., Hegeman, G. D., Cleland, W. W., Kenyon, G. L. Symmetry and asymmetry in mandelate racemase catalysis. *Biochemistry* 24(15):3936–3942, 1985.
- [223] Maurice, M. S., Bearne, S. L. Reaction intermediate analogues for mandelate racemase: Interaction between asn 197 and the r -hydroxyl of the substrate promotes catalysis. *Biochemistry* 39(44):13324–13335, 2000.
- [224] Goriup, M., Strauss, U. T., Felfer, U., Kroutil, W., Faber, K. Substrate spectrum of mandelate racemase part 2. (hetero)-aryl-substituted mandelate derivatives and modulation of activity. *J. Mol. Catal. B* 15:213–222, 2001.
- [225] Maurice, M. S., Bearne, S. L. The low barrier hydrogen bond in enzymatic catalysis. *J. Biol. Chem.* 39(44):13324–13335, 2000.
- [226] Schnell, B., Faber, K., Kroutil, W. Enzymatic racemisation and its application to synthetic biotransformations. *Adv. Synth. Catal.* 345:653–666, 2003.
- [227] Kallarakal, A. T., Mitra, B., Kozarich, J. W., Gerlt, J. A., Clifton, J. R., Petsko, G. A., Kenyon, G. L. Mechanism of the reaction catalyzed by mandelate racemase: Structure and mechanistic properties of the k166r mutant. *Biochemistry* 34(9):2788–2797, 1995.
- [228] Neidhart, D. J., Howell, P. L., Petsko, G. A., Powers, V. M., Li, R., Kenyon, G. L., Gerlt, J. A. Mechanism of the reaction catalyzed by mandelate racemase. 2. crystal structure of mandelate racemase at 2.5 Å resolution: Identification of the active site and possible catalytic residues. *Biochemistry* 30(38):9264–9273, 1991.
- [229] Schafer, S. L., Barrett, W. C., Kallarakal, A. T., Mitra, B., Kozarich, J. W., Gerlt, J. A. Mechanism of the reaction catalyzed by mandelate racemase: Structure and mechanistic properties of the d270n mutant. *Biochemistry* 35(18):5662–5669, 1996.
- [230] Northrop, D. B. Follow the protons: A low-barrier hydrogen bond unifies the mechanism of the aspartic proteases. *Acc. Chem. Res.* 34(10):790–797, 2001.
- [231] Guthrie, J. P., Kluger, R. Electrostatic stabilization can explain the unexpected acidity of carbon acids in enzyme-catalyzed reactions. *J. Am. Chem. Soc.* 115(24):11569–11572, 1993.
- [232] Alagona, G., Ghio, C., Kollman, P. A. Ab initio explorative survey of the mechanism catalyzed by mandelate racemase. *J. Mol. Struct. (Theochem)* 390:217–223, 1997.
- [233] Landro, J. A., Gerlt, J. A., Kozarich, J. W., Koo, C. W., Shah, V. J., Kenyon, G. L., Neidhart, D. J., Fujita, S., Petsko, G. A. The role of lysine 166 in the mechanism of mandelate racemase from *Pseudomonas putida*: Mechanistic and crystallographic evidence for stereospecific alkylation by (r)- α -phenylglycidate. *Biochemistry* 33:635–643, 1994.
- [234] Hutter, M. C., Hughes, J. M., Reimers, J. R., Hus, N. S. Modeling the bacterial photosynthetic reaction center. 2. a combined quantum mechanical/molecular mechanical study of the structure of the cofactors in the reaction centers of purple bacteria. *J. Phys. Chem. B* 103(23):4906–4915, 1999.
- [235] Cheng, A., Stanton, R. S., Vincent, J. J., van der Vaart, A., Damodaran, K. V., Dixon, S. L., Hartsough, D. S., Mori, M., Best, S. A., Monard, G., Garcia-Viloca, M., Zant, L. C. V., Merz, Jr., K. M. ROAR 2.0. The Pennsylvania State University. 1999.
- [236] Merz Jr., K. M., Banci, L. Binding of azide to human carbonic anhydrase ii: The role electrostatic complementarity plays in selecting the preferred resonance structure of azide. *J. Phys. Chem.* 100(43):17414–17420, 1996.
- [237] Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R. W., Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935, 1983.
- [238] Gerlt, J. A., Gassman, P. G. Understanding the rates of certain enzyme-catalyzed reactions: Proton abstraction from carbon acids, acyl-transfer reactions, and displacement reactions of phosphodiester. *Biochemistry* 32(45):11943–11952, 1993.

- [239] Frisch, M. J., Trucks, G. W. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Zakrzewski, V. G., Montgomery Jr., J. A., Stratmann, R. E., Burant, J. C., Dapprich, S., Millam, J. M., Daniels, A. D., Kudin, K. N., Strain, M. C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G. A., Ayala, P. Y., Cui, Q., Morokuma, K., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Cioslowski, J., Ortiz, J. V., Baboul, A. G., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Andres, J. L., Gonzalez, C., Head-Gordon, M., Replogle, E. S., , A., P. J. Gaussian 98, Revision A.9, Gaussian, Inc., Pittsburgh, PA, USA. 1998.
- [240] Schlick, T., Skeel, R., Brunger, A. T., Kalé, L. V., Board, J. A., Hermans, J., Schulten, K. Algorithmic challenges in computational molecular biophysics. *J. Comput. Phys.* 151:9–48, 1999.
- [241] Elber, R., Shalloway, D. Temperature dependent reaction coordinates. *J. Chem. Phys.* 112(13):5539–5545, 2000.
- [242] Das, B., Meirovitch, H., Navon, I. M. Performance of hybrid methods for large-scale unconstrained optimization as applied to models of proteins. *J. Comput. Chem.* 24:1222–1231, 2003.
- [243] Anglada, J. M., Besalú, E., Bofill, J. M. Remarks on large-scale matrix diagonalization using a lagrange-newton-raphson minimization in a subspace. *Theor. Chem. Acc.* 103:163–165, 1999.
- [244] Leininger, M. L., Sherrill, C. D., Allen, W. D., Schaefer III, H. F. Systematic study of selected diagonalization methods for configuration interaction matrices. *J. Comp. Chem.* 22(13):1574–1589, 2001.
- [245] Besalú, E., Bofill, J. M. On the automatic restricted-step rational-function-optimization. *Theor. Chem. Acc.* 100:265–274, 1998.
- [246] Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Croz, J. D., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D. LAPACK User's Guide. 3rd Ed. Philadelphia: SIAM. 1999.
- [247] Alagona, G., Ghio, C., Kollman, P. A. Do enzymes stabilize transition states by electrostatic interactions or pka balance: The case of triose phosphate isomerase (tim)? *J. Am. Chem. Soc.* 117(39):9855–9862, 1995.
- [248] Andrés, J., Moliner, V., Krechl, J., Silla, E. *J. Chem. Soc. Perkin Trans.* 2:1551, 1995.
- [249] Gao, J. Absolute free energy of solvation from monte carlo simulation using combined quantum and molecular mechanical potentials. *J. Phys. Chem.* 96:537–540, 1992.
- [250] Truong, T. N., Stefanovich, E. V. Development of a perturbative approach for monte carlo simulations using a hybrid ab initio qm/mm method. *Chem. Phys. Lett.* 256:348–352, 1996.
- [251] Cubero, E., Luque, F. J., Orozco, M., Gao, J. Perturbation approach to combined qm/mm simulation of solute-solvent interactions in solution. *J. Phys. Chem. B* 107:1664–1671, 2003.
- [252] Evans, T. J., Truong, T. N. Optimizing efficiency of perturbative monte carlo method. *J. Comput. Chem.* 19(14):1632–1638, 1998.
- [253] Bell, S., Crighton, J. S., Fletcher, R. A new efficient method for locating saddle points. *Chem. Phys. Lett.* 82:122–126, 1981.
- [254] Golub, G. H., Loan, C. F. V. *Matrix Computations*. 3rd Ed. USA: The John Hopkins University Press. 1996.
- [255] Li, G., Cui, Q. Analysis of functional motions in brownian molecular machines with an efficient block normal mode approach: Myosin-ii and ca^{+2} -atpase. *Bioph. J.* 86(2):743–763, 2004.
- [256] Bofill, J. M., de Pinho Ribeiro Moreira, I., Anglada, J. M., Illas, F. Accurate and efficient determination of higher roots in diagonalization of larges matrices based in function restricted optimization algorithms. *J. Comput. Chem.* 21(15):1375–1386, 2000.
- [257] Bofill, J. M., Anglada, J. M. Some remarks on the use of the three-term recurrence method in the configuration interaction eigenvalue problem. *Chem. Phys.* 183:19–26, 1994.
- [258] Besalú, E., Bofill, J. M. Calculation of clustered eigenvalues of large matrices using variance minimization method. *J. Comput. Chem.* 19(15):1777–1785, 1998.
- [259] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*. 2nd Ed.: Cambridge University Press. copyright 1986-1992.
- [260] Gao, J., Garcia-Viloca, M., Poulsen, T. D., Mo, Y. Solvent effects, reaction coordinates, and reorganization energies on nucleophilic substitution reactions in aqueous solution. In: *Advances in Physical Organic Chemistry*. Vol. 38. Vol. 38. Elsevier 2003 161–181.
- [261] Gao, J. A priori computation of a solvent-enhanced $sn2$ reaction profile in water: the menshutkin reaction. *J. Am. Chem. Soc.* 113(20):7796–7797, 1991.
- [262] van Gunsteren, W. F., Mark, A. E. Validation of molecular dynamics simulation. *J. Chem. Phys.* 108(15):6109–6116, 1998.