# Local Correlation Methods in Classical and Quantum Mechanics Hybrid Schemes

Von der Fakultät Chemie der Universität Stuttgart
zur Erlangung der Würde eines Doktors der
Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von
**Ricardo André Fernandes da Mata**
aus Lissabon

| | |
|---|---|
| Hauptberichter: | Prof. Dr. H.-J. Werner, Universität Stuttgart |
| Mitberichter: | Prof. Dr. G. Rauhut, Universität Stuttgart |
| Tag der mündlichen Prüfung: | 8 November 2007 |

Institut für Theoretische Chemie der Universität Stuttgart

2007

# Contents

# Acknowledgments

To my parents for their support ...

... to all (former and present) coworkers from the Stuttgart group, especially Andreas Nicklass, Robert Polly, Alexander Mitrushchenkov and Klaus Pflüger for their patience in answering all my nagging questions, and also to those who helped me by proof-reading this manuscript: Thomas Adler (who had it worst), Erich Goll and Christoph Köppl ...

... to Prof. Stoll and Prof. Rauhut, for sharing their research and knowledge ...

... to my friends in Stuttgart, for four years I will never forget ...

... to my brother Tiago and Christine, for always being there for me ...

... and to Prof. Werner, who so wholeheartedly accepted me in his group and guided me through this work (and let us not forget, taught me how to put the back of an envelope to good use) ...

... I owe my thanks, and am forever in your debt.

# Abstract

The computation of reaction barriers in molecular systems has been since its birth one of the major challenges in Theoretical Chemistry. It is of vital importance in understanding and predicting catalysis phenomena, and in rationalizing our knowledge of chemical reactivity in general using Transition State Theory (TST).

The Hartree-Fock (HF) approximation offers a "mean field" approach to the solution of the Schrödinger Equation, and is the starting point for most of quantum chemical methods. It does however not account for electron correlation effects, i.e., the instantaneous Coulomb repulsion between electrons. This effect is of prime importance in describing chemical reactivity, due to the changes in electron interaction during bond breaking/formation. The HF method normally has errors in the range of 100-500% for reaction barriers. The approximate treatment of electron correlation through Density Functional Theory (DFT) is an inexpensive way to include some of these effects in the energy estimate. However, its results depend strongly on the parametrization made and a functional which consistently delivers good results for all chemical systems has still not been found. The post-HF *ab initio* family of methods offers a systematic way to approach a converged result. However, the high scaling of computational cost with molecular size only allows quantitative calculations for small-sized systems (up to 15 atoms).

Local correlation methods avoid the steep scaling of conventional canonical methods by using local spaces to describe occupied and virtual orbitals. The excitations are limited by distance criteria, and the correlation of electron pairs is approximated in an hierarchical manner, with higher levels used for neighboring orbitals, and neglecting very distant pair contributions.

In this PhD work several advances have been made in the application of local correlation methods to the computation of reaction paths and barriers. A new procedure was implemented in the `Molpro` program package to compensate for the geometry dependence of excitation domains. This dependence can lead to noncontinuous potential energy surfaces and used to be a drawback in the use of local methods for tracing reaction paths. The main focus however, was in the implementation and use of Quantum Mechanical/Molecular Mechanics (QM/MM), as well as a combined Quantum Mechanics/Quantum Mechanics (QM/QM) approaches for the computation of reaction barriers. Although the local methods approach asymptotically the linear scaling regime, they can only be routinely applied to systems of up to 100-150 atoms. Enzymatic systems include well above 1000 atoms and further approximations are needed. In the QM/MM case, the environment is treated by MM force fields, and the active site by regular quantum mechanical methods. The use of

this approach together with local methods provided reaction enthalpies of high accuracy for two enzymatic systems (Chorismate Mutase and *p*-Hydroxybenzoate Hydroxylase). The second coupling (QM/QM) is made by classifying orbital pairs according to regions of different chemical interest (normally separating the active site from the environment) and applying different correlation schemes. This has shown promising results for medium to large sized systems.

# Citations to Published Work

Most of Chapter 3 has been published as

> "Computation of smooth Potential Surfaces using Local Correlation Methods",
>
> R. A. Mata, H.-J. Werner, J. Chem. Phys. **125**, 184110 (2006)

Chapter 4 is also partly featured in

> "Local Correlation Methods with a Natural Localized Molecular Orbital Basis",
>
> R. A. Mata, H.-J. Werner, *in press*

Chapter 5 is featured in

> "Correlation regions within a localized molecular orbital approach",
>
> R. A. Mata, M. Schütz, and H.-J. Werner, *to be submitted*

Large portions of Chapter 6 have been or will be published in the following papers:

> "High-Accuracy Computation of Reaction Barriers in Enzymes",
>
> F. Clayessens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel, and H.-J. Werner, Angew. Chemie **118**, 7010-7013 (2006)
>
> "Towards accurate barriers for enzymatic reactions: QM/MM case study on *p*-hydroxybenzoate hydroxylase",
>
> R. A. Mata, S. Thiel, W. Thiel, and H.-J. Werner, *to be submitted*

# Abbreviations

| | |
|---|---|
| 3,4-DOHB | 3,4-dihydroxybenzoate |
| AM1 | Austin Model 1 |
| BP | Boughton-Pulay |
| BSSE | Basis Set Superposition Error |
| CBS | Complete Basis Set |
| CC | Coupled Cluster |
| CCSD | Singles and Doubles Coupled Cluster |
| CCSD(T) | Singles and Doubles Coupled Cluster with perturbative triples |
| CM | Chorismate Mutase |
| CP | Counterpoise |
| DF | Density Fitting |
| DFT | Density Functional Theory |
| FAD | Flavin adenine |
| FCI | Full Configuration Interaction |
| HF | Hartree-Fock |
| HOMO | Highest Occupied Molecular Orbital |
| IMOMO | Integrated MO/MO |
| KS | Kohn-Sham |
| LCCSD | Local CCSD |
| LCCSD(T) | Local CCSD(T) |
| LCCSD(T0) | Local CCSD with non-iterative triples |
| LMO | Localized Molecular Orbital |
| LMP2 | Local MP2 |
| LUMO | Lowest Unoccupied Molecular Orbital |
| MD | Molecular Dynamics |
| MM | Molecular Mechanics |

| | |
|---|---|
| MNDO | Modified Neglect of Diatomic Overlap |
| MO | Molecular Orbital |
| MP | Møller-Plesset |
| MP2 | Second Order Møller-Plesset Perturbation Theory |
| NAC | Near Attack Conformation |
| NAO | Natural Atomic Orbital |
| NBO | Natural Bond Orbital |
| NDDO | Neglect of Diatomic Differential Overlap |
| NHO | Natural Hybrid Orbital |
| NLMO | Natural Localized Molecular Orbital |
| NMB | Natural Minimal Basis |
| NPA | Natural Population Analysis |
| NRB | Natural Rydberg Basis |
| PAO | Projected Atomic Orbital |
| PES | Potential Energy Surface |
| PHBH | *para*-Hydroxybenzoate Hydroxylase |
| PM | Pipek-Mezey |
| PM3 | Parametric Method Number 3 |
| pOHB | *para*-Hydroxybenzoate |
| QM | Quantum Mechanics |
| QM/MM | Quantum Mechanics/Molecular Mechanics |
| SCF | Self-Consistent Field |
| SCS | Spin Component Scaled |
| TS | Transition State |
| TST | Transition State Theory |
| WT | Wild Type |
| ZDO | Zero Differential Overlap |
| ZPVE | Zero Point Vibrational Energy |

# Chapter 1

# Introduction

*We are perhaps not far removed from the time when we shall be able to submit the bulk of chemical phenomena to calculation.*

Joseph Louis Gay-Lussac,
*Memoires de la Société D'Arcueil*, **2**, 207 (1808)

Theoretical Chemistry is a vibrant and expanding research area. It covers virtually all aspects of Chemistry, from calorimetry to spectroscopy, biochemistry and solid state, organic and inorganic materials. Fulfilling the wishes of Gay-Lussac, even if with some delay, one is currently able to predict and understand phenomena in ever larger time or size scales, almost exclusively on the basis of calculations. But even considering all the advances made in the last decades, size still remains a central concern for the theoretical chemist.

If we consider the different methods and applications of Theoretical Chemistry as a function of the system size, we will find a broad spectrum, changing not only in subject, but also in character. On one end, we would find the study of very small systems in the gas phase, typically up to 10 atoms. The theoretical chemist on this side of the spectrum will be concentrated on questions of "quantity". His calculations can reach an accuracy below 1 kcal $mol^{-1}$ (or even sub kJ $mol^{-1}$), and therefore rival with experiments. These studies give a valuable support to the lab-chemist, often identifying errors in tabulated experimental data or providing assistance in the interpretation of various spectra. At the other end, stands the study of condensed/biological systems. These involve the study of thousands to hundreds of thousands of atoms. In this case, one mostly hopes for a qualitative description, like information on conformational stability or preferential reactivity.

The reasons behind these differences are manifold. The questions posed on an enzymatic system are of course not the same when dealing with a two-atom molecule. But there are still many common interests. The ability to make quantitative predictions is always desirable, whether one is dealing with a large or small amount of atoms. There is however a barrier in the way. The computational cost of accurate quantum chemical methods scales exponentially with system size. This means that beyond a given number of atoms, one will face an *exponential wall*. The cost increases too steeply and it is impossible to add any element to the calculation without exceeding the available resources. Since computer technology evolves at best linearly, little progress is to be expected just by waiting for the new computers to come along. Therefore, improving the scaling of quantum chemical methods should be one of our top priorities.

If the exponential wall is to be overcome, the scaling of the computational cost with molecular size must be made at least linear. In this case, when one doubles the system

size, the CPU memory and/or disk space requirements will only double. If the task can be parallelized the resources can be distributed over several machines, the quality of the calculation can be preserved and the requirements per CPU kept constant. Several progresses have been made in this direction over the last few years. Many of them even took place here in Stuttgart. Based on the ideas first put forward by Peter Pulay, linear scaling correlated methods have been successfully programmed and tested. They are referred to as local correlation methods, due to the use of approximations based on the locality of electron correlation. This family of methods critically decreases the cost of conventional methods, and are today's top reference in the field of linear scaling quantum chemical algorithms. But we still may have a long wait ahead before we will be able to perform a full quantum mechanical calculation on a protein with the same accuracy as in a diatomic. The prefactors involved are still too large, and one would be extremely limited in performing these applications at such a large scale (need for supercomputers, small number of calculations, limiting the search of the conformational space). On the other hand, one could ask whether this is actually necessary. Does one need a full calculation to extract information from the system?

Fortunately, when discussing chemical phenomena the answer is in many cases "no". Most of the effects of interest in reactivity are local in nature. In a bond breaking/formation, only the near-lying groups will have a significant influence on the process. The surrounding environment can be included approximately or even neglected. The ideas detailed above denote the fundaments for my work. It should be possible to treat large molecular systems with unprecedented accuracy by coupling local correlation methods with lower level approaches. Such approaches are already in use today. The innovation lies in the use of local methods. They allow for larger active sites and higher accuracies than their conventional counterparts. Also, since one avoids the exponential scaling, it is possible to increase the active site in the calculation without great increase in the computational cost. This can be used to test the approximations involved or simply to improve the overall accuracy of the calculation. Hybrid schemes involving both coupling of quantum mechanical and molecular mechanical methods (QM/MM) as well as quantum mechanics with quantum mechanics (QM/QM) have been implemented and/or used in this work, building up the centerpiece of this Thesis.

The structure of this Thesis is as follows. In the first two sections of Chapter 2 a short review of the various methods used in this work is given. The large amount of information led me to cut some parts short, but I believe the connection between the different methodologies has been evidenced. The level of detail and the wide spectrum of techniques featured make the text an advisable pre-graduate reading material. The informed reader may how-

ever skip most of it without great loss. The main emphasis lies in the description of local correlation methods and the quantum mechanics/molecular mechanics coupling schemes. This is required for the following Chapters 5 and 6. In the last section, the general theory on rate constants is presented, in support of the discussion featured in Chapter 6.

In Chapter 3, the computation of potential energy surfaces using local correlation methods is discussed. These methods are known to generate non-continuous potential energy surfaces in cases of bond breaking or significant geometric displacements. This is due to the use of geometry-dependent excitation spaces. This problem has been investigated and a simple procedure is presented in order to generate smooth surfaces in such cases. The procedure is also found to improve the description of some reaction energetics.

Chapter 4 discusses the use of Natural Localized Molecular Orbitals as a new occupied space for local methods in general. A new single-parameter domain criterion is also presented. The combination of this proposed occupied space with the new selection proves to be remarkably more stable than previous implementations, and is a promising development for establishing local correlation methods as well defined computational models. First results for absolute energies, together with a thorough comparison between the new procedure and the previous methods are shown.

Chapter 5 describes the hybrid QM/QM implementation at the heart of this work. It presents a novel approach to the problem of separating the system into constituent parts of different accuracy. It is also the second method to date which allows a combined use of quantum mechanical methods. Test calculations are presented for biological systems, and comparison is made with other proposed QM/QM coupling schemes. The method is shown to deliver similar or better accuracy, but with significant advantages relative to other models.

The use of QM/MM coupling schemes for local correlation methods did not involve any new theoretical developments. Therefore, only the applications will be discussed. The general theory is given in Chapter 2 and in Appendix D. The applications are discussed in Chapter 6. These involved extensive collaboration with other groups in Bristol and at the Max-Planck Institute in Mülheim. References to their work will accordingly be made throughout the Chapter.

# Chapter 2

# Theoretical Background

## 2.1 Quantum Mechanics

In *quantum mechanics*, the *state* of a system is defined by its *wave function* $\Psi$. In this work only time-independent functions will be considered, so that $\Psi = \Psi(\mathbf{x})$, where $\mathbf{x}$ is a vector representing the systems generalized coordinates (spatial and spin). Since a state is fully described by its wave function, both terms will be used interchangeably.

The Schrödinger equation[1]

$$\hat{H}\Psi = E\Psi, \tag{2.1}$$

determines which states $\Psi$ are allowed in a system described by the Hamiltonian operator $\hat{H}$. The wave function must be an *eigenvector* of the operator, with the energy $E$ as the corresponding *eigenvalue*. For a system composed of $N$ electron and $M$ nuclei, the Hamiltonian is written (in atomic units) as

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \frac{1}{2}\sum_{m=1}^{M}\frac{1}{M_m}\nabla_m^2 - \sum_{m=1}^{M}\sum_{i=1}^{N}\frac{Z_m}{r_{im}} + \sum_{i<j}\frac{1}{r_{ij}} + \sum_{m<n}\frac{Z_m Z_n}{r_{mn}}, \tag{2.2}$$

where the indices $i$ and $j$ refer to electron indices, $m$ and $n$ to nuclei indices. The masses and charges of the nuclei are represented by $M_m$ and $Z_m$, respectively, and the distances between two particles as $r_{xy}$. The first two sums in Eq. (2.2) account for the electron and nuclear kinetic energy, respectively, while the remaining terms describe Coulombic interactions between the particles.

Due to the high-dimensionality of the problem, this equation is exactly soluble just for very simple cases. Especially troublesome are the Coulombic terms which couple the movement of all particles in the system. Therefore, instead of analytically solving the equation, one is forced to use approximate representations of the wave function[2] and/or the Hamiltonian. A fundamental assumption made in all of the methods to be later discussed is the Born-Oppenheimer approximation. Since the electrons travel much faster than the nuclei, the movement of both can be decoupled. The problem is then divided into two parts. The solution of an electronic Hamiltonian for fixed nuclear coordinates, and of a nuclear Hamiltonian for an effective electron potential. The concept of a Potential Energy Surface (PES), so often discussed in Chemistry, is thereof derived. The nuclei move subject to forces derived from the solution of the electronic Hamiltonian and the Coulomb nuclear repulsion. This approximation is generally valid, except when discussing phenomena such as conic intersections, or if one is interested in high accuracy calculations. The energy of a

[1] The Schrödinger equation featured is also time-independent.

[2] As long as they comply to the Pauli principle for fermions, which states that the exchange of two electrons leads to a change of the sign in the wave function - antisymmetry.

given electronic state is obtained by solving Eq. (2.1), with $\hat{H}$ substituted by an electronic Hamiltonian of the form

$$\hat{H}_{el} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{m=1}^{M}\sum_{i=1}^{N}\frac{Z_m}{r_{im}} + \sum_{i<j}\frac{1}{r_{ij}} = \sum_{i=1}^{N}\hat{h}(i) + \sum_{i<j}\frac{1}{r_{ij}} \tag{2.3}$$

The nuclear repulsion potential is added *a posteriori*. In the following sections, allusions to the Hamiltonian operator will be implicit references to Eq. (2.3).

### 2.1.1 Hartree Fock

In the Hartree-Fock (HF) approximation, the wave function is represented by a Slater determinant

$$\Psi^{\mathrm{HF}} = \frac{1}{\sqrt{N!}}\begin{vmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & \dots & \psi_N(\mathbf{x}_1) \\ \psi_1(\mathbf{x}_2) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \psi_1(\mathbf{x}_N) & \dots & \dots & \psi_N(\mathbf{x}_N) \end{vmatrix} = |\psi_1(\mathbf{x}_1)\psi_2(\mathbf{x}_2)\dots\psi_N(\mathbf{x}_N)> . \tag{2.4}$$

The electron coordinates are given as a vector $\mathbf{x}_i = \{\mathbf{r}_i, \mathbf{s}_i\}$, for spatial and spin coordinates respectively. The total wave function is an anti-symmetrized product of molecular spin-orbitals $\{\psi_i\}$. In cases where the number of $\alpha$ and $\beta$ spin electrons are the same, one may use the same spatial orbitals $\{\phi_i\}$ for both sets. This is referred to as a *closed shell* representation. The closed-shell energy expectation value is of the form (according to the Slater Condon rules[14])

$$< \Psi^{\mathrm{HF}}|\hat{H}|\Psi^{\mathrm{HF}} >= \sum_{i}^{N/2} 2 < i|\hat{h}|i > + \sum_{ij}^{N/2}\left[2(ii|jj) - (ij|ji)\right], \tag{2.5}$$

where use has been made of the following notation for one and two-electron integrals

$$< i|\hat{h}|j > = \int \phi_i^*(\mathbf{r}_1)\hat{h}(\mathbf{r}_1)\phi_j(\mathbf{r}_1)d\mathbf{r}_1 \tag{2.6}$$

$$(ij|kl) = \int \phi_i^*(\mathbf{r}_1)\phi_j(\mathbf{r}_1)r_{12}^{-1}\phi_k^*(\mathbf{r}_2)\phi_l(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2. \tag{2.7}$$

Integration over the spin coordinates has already been carried out, so that the theory is from now on spin-free.

The Hartree-Fock method is by construction a variational method, which means that

the energy obtained with any test wave function will always be an upper bound to the exact energy of the system[3]. This can be used as a criterion to optimize the form of our wave function. The lower the energy expectation value, the closer one should be to the best possible description of the system. The condition for an optimized HF wave function is that the derivative of Eq. (2.5) with respect to orbital changes is equal to zero (minimum point), under the orthonormality restriction. For this reason, one does not minimize the energy expression, but instead the *Lagrangian function*

$$\mathscr{L} = < \Psi^{\mathrm{HF}} | \hat{H} | \Psi^{\mathrm{HF}} > -2 \sum_{ij}^{N/2} \varepsilon_{ji} \left[ < i | j > -\delta_{ij} \right], \tag{2.8}$$

where $\varepsilon_{ji}$ are the lagrangian multipliers and $< i|j >$ is an overlap integral. Setting the derivative of the Lagrangian with respect to each orbital to zero, one obtains the *HF equations*

$$\hat{f}|i> = \sum_{j}^{N/2} \varepsilon_{ji} |j> \tag{2.9}$$

with the *Fock operator* $\hat{f}$ defined as

$$\hat{f}(i) = \hat{h}(i) + \sum_{j} \left[ 2\hat{J}_j(i) - \hat{K}_j(i) \right] = \hat{h}(i) + \hat{g}(i). \tag{2.10}$$

The two operators $\hat{J}_j(i)$ and $\hat{K}_j(i)$, referred to as *Coulomb* and *exchange* operators, respectively, are defined by their effect when operating on a spatial orbital

$$\hat{J}_j(1)\phi_i(\mathbf{r}_1) = \int \phi_j^*(\mathbf{r}_2) \frac{1}{r_{12}} \phi_j(\mathbf{r}_2)\phi_i(\mathbf{r}_1) d\mathbf{r}_2, \tag{2.11}$$

$$\hat{K}_j(1)\phi_i(\mathbf{r}_1) = \int \phi_j^*(\mathbf{r}_2) \frac{1}{r_{12}} \phi_j(\mathbf{r}_1)\phi_i(\mathbf{r}_2) d\mathbf{r}_2. \tag{2.12}$$

Comparing Eq. (2.3) and Eq. (2.10), it is easy to identify the approximation made in HF theory. The electron-electron interaction operator $r_{ij}^{-1}$ is replaced by a mean-field electron repulsion in the form of $\hat{g}(i)$. Each electron only "feels" an averaged interaction with the remaining electrons.

Since the energy is invariant with respect to unitary transformations among the occupied orbitals, a simpler form for Eq. (2.9) is possible by using a basis where the fock operator

[3]The proof for the variational theorem is relatively trivial, and can be found in almost any Quantum Chemistry book. I would however advise looking into Ref. [15]

is diagonal, the *canonical HF equations*

$$\hat{f}|i>=\varepsilon_i|i>. \tag{2.13}$$

Such a basis is referred to as a canonical orbital basis. The diagonal elements of the Fock matrix are the orbital energies $\varepsilon_i$. Notice that the HF total energy is not equal to the sum of the occupied orbital energies. This would lead to double-counting of the electron-electron interaction.

Up to this point, the theory has been derived without any consideration over the form of the spatial orbitals $\{\phi_i\}$ which are used. If one would use an infinite basis - also referred to as complete basis set (CBS) - to expand the orbital space, the variational method would find the "right" solution under the HF mean-field approximation. This is of course not possible, since we always have to restrict ourselves to a finite-sized expansion. In most of the molecular structure programs in use today, these orbitals are built as a linear combination of atom-like gaussian functions[4] - the Atomic Orbital (AO) set $\{\chi_\mu\}$. Much effort has been put in to achieve values close to the CBS limit using the smallest number of functions possible. The spatial orbitals are defined in the Linear Combination of Atomic Orbitals (LCAO) approximation as

$$\phi_i(\mathbf{r}) = \sum_\mu C_{\mu i}\chi_\mu(\mathbf{r}). \tag{2.14}$$

Introduction of Eq. (2.14) into (2.5) gives the Hartree-Fock energy in dependence of the atomic orbital integrals

$$\begin{aligned} E_{\mathrm{HF}} &= \sum_{\mu\nu} D_{\mu\nu}\left\{h_{\mu\nu} + \frac{1}{2}\sum_{\rho\sigma} D_{\rho\sigma}\left[(\mu\nu|\rho\sigma) - \frac{1}{2}(\mu\sigma|\rho\nu)\right]\right\} \\ &= \frac{1}{2}\sum_{\mu\nu} D_{\mu\nu}\left(h_{\mu\nu} + f_{\mu\nu}\right). \end{aligned} \tag{2.15}$$

Both matrices $h_{\mu\nu}$ and $f_{\mu\nu}$ are integrals over the operators $\hat{h}$ and $\hat{f}$ in the AO basis. The matrix **D** is the *one-electron density matrix*, with

$$D_{\mu\nu} = 2\sum_i^{\mathrm{N}/2} C_{\mu i}C^*_{\nu i} \tag{2.16}$$

The sum runs only over the occupied orbital indices. Therefore, the HF energy is only dependent of the AO basis and the first $N/2$ molecular orbitals (MOs) coefficients. The remaining $N_{\mathrm{AO}} - N/2$ orbitals ($N_{\mathrm{AO}}$ is the number of AOs used in the expansion of Eq. (2.14))

[4]The basis functions, however, do not have to be necessarily gaussian functions.

are referred to as *virtual* orbitals. They have no special significance in HF theory, except for the Koopman's Theorem[5], but are of vital importance for post-HF treatments. Although the HF energy commonly gives about 99% of the total energy, the electron-electron interaction is still approximated through the use of a mean-field. The remaining 1% describing the instantaneous correlation between the electrons as they move is often important for describing chemical phenomena. This is called the *correlation energy*, and methods which include this contribution are referred to as *correlated methods*. To obtain the full energy of the system for a given AO basis, the wavefunction would have to be built out of a linear combination of all possible Slater determinants, each with different occupations for the $N_{\mathrm{AO}}$ molecular orbitals. This is normally referred to as a Full Configuration Interaction (FCI) method. Such an approach is however too costly, and only feasible for very small systems and AO expansions. Other methods have been developed which scale significantly better while providing good estimates for the correlation contribution. These methods are the subject for the next few Sections.

### 2.1.2 Møller-Plesset Perturbation Theory

One of the simplest approaches to the correlation problem is to consider the HF solution a sufficiently good approximation to the total energy of the system, and to obtain the missing contributions through a perturbation expansion. The Hamiltonian is split into a reference $\hat{H}^{(0)}$ and a perturbation $\hat{H}^{(1)}$

$$(\hat{H}^{(0)} + \lambda \hat{H}^{(1)})|\Psi >= E|\Psi >, \tag{2.17}$$

with

$$\hat{H}^{(0)} = \sum_{i}^{N} \hat{f}(i) = \sum_{i}^{N} \left[\hat{h}(i) + \hat{g}(i)\right] \tag{2.18}$$

$$\hat{H}^{(1)} = \hat{H} - \hat{H}^{(0)} \tag{2.19}$$

[5]According to Koopman's Theorem, the electron affinity of a molecular system will correspond to the orbital energy of the lowest lying virtual orbital. This definition, however, rarely provides reliable results, as the virtual orbital energies do not converge to defined values upon increasing the basis sets, and the physical meaning of them is in fact dubious.

This choice of Hamiltonian is referred to as Møller-Plesset (MP) Perturbation Theory. By expanding the energy and the wave function in a Taylor series

$$E = \sum_{k=0} \lambda^k E^{(k)} \qquad , \qquad |\Psi >= \sum_{k=0} \lambda^k |\Psi^{(k)} > \tag{2.20}$$

and inserting them into Eq. (2.17), one obtains

$$(\hat{H}^{(0)} - E^{(0)})|\Psi^{(0)} > +\lambda \left[ (\hat{H}^{(0)} - E^{(0)})|\Psi^{(1)} > +(\hat{H}^{(1)} - E^{(1)})|\Psi^{(0)} > \right] + \cdots = 0, \tag{2.21}$$

Since Eq. (2.21) must hold for any value of $\lambda$, there are in fact $n+1$ equations to be solved, where $n$ is the expansion limit. For each power of $\lambda$, the associated terms must equal zero. Such an expansion contains energy corrections up to order $n+1$, although the wavefunction only has to be known up to order $n$.

The reference wave function $\Psi^{(0)}$ will be the HF wave function, which is an eigenfunction of $\hat{H}^{(0)}$. It is easy to show that the energies of order $n = 0, 1$

$$E^{(0)} =< \Psi^{\mathrm{HF}}|\hat{H}^{(0)}|\Psi^{\mathrm{HF}} >=< \Psi^{\mathrm{HF}}|\sum_i^N \hat{f}_i|\Psi^{\mathrm{HF}} >= \sum_i^N \varepsilon_i \tag{2.22}$$

$$E^{(1)} =< \Psi^{\mathrm{HF}}|\hat{H}^{(1)}|\Psi^{\mathrm{HF}} >= -\frac{1}{2}\sum_{ij}[2(ii|jj) - (ij|ji)] \tag{2.23}$$

summed together will give the HF energy (compare the above result to Eq.(2.5)). The first correction is therefore contained in $E^{(2)}$. This energy term already involves the first order wave function $\Psi^{(1)}$, which must be given. According to the Brillouins Theorem, singly excited configurations do not interact with the reference and, therefore, do not contribute in first order to the energy. The first order wave function is built as a combination of the doubly excited configurations

$$|\Psi^{(1)} >= \frac{1}{2}\sum_{ij}\sum_{ab} T_{ab}^{ij} |\Phi_{ij}^{ab} > . \tag{2.24}$$

The $|\Phi_{ij}^{ab} >$ functions are defined as

$$|\Phi_{ij}^{ab} >= \hat{E}_{ai}\hat{E}_{bj}|\Psi^{\mathrm{HF}} >, \tag{2.25}$$

where $\hat{E}_{ai}$ is a spin-adapted operator which excites an electron from an occupied orbital $i$ to a virtual orbital $a$. However, since the $\Phi_{ij}^{ab}$ configurations are not orthogonal nor normal-

ized, it is convenient to make use of *contravariant* configurations and amplitudes

$$\tilde{\Phi}_{ij}^{ab} = \frac{1}{6}(2\Phi_{ij}^{ab} + \Phi_{ji}^{ab}) \qquad , \qquad \tilde{T}_{ab}^{ij} = 2T_{ab}^{ij} - T_{ab}^{ji}, \tag{2.26}$$

which have the following properties:

$$\begin{aligned} < \tilde{\Phi}_{ij}^{ab} | \Phi_{kl}^{cd} > &= \delta_{ac}\delta_{bd}\delta_{ik}\delta_{jl} + \delta_{ad}\delta_{bc}\delta_{il}\delta_{jk}, & (2.27)\\ < \tilde{\Phi}_{ij}^{ab} | \Psi^{(1)} > &= T_{ab}^{ij}, & (2.28)\\ < \tilde{\Phi}_{ij}^{ab} | \hat{H} | \Psi^{\mathrm{HF}} > &= K_{ab}^{ij}. & (2.29) \end{aligned}$$

The use of contravariant configurations and amplitudes greatly simplify the end formulae of matrix elements involving excited configurations. The MP2 correlation energy gives the first correction to the HF value, and is easily computed as

$$\begin{aligned} \Delta E_{\mathrm{MP2}} = E^{(2)} =& < \Psi^{\mathrm{HF}} | \hat{H} | \Psi^{(1)} > = \sum_{ij}\sum_{ab} < \Psi^{\mathrm{HF}} | \hat{H} | \tilde{\Phi}_{ij}^{ab} > \tilde{T}_{ab}^{ij} \\ &= \sum_{ij}\sum_{ab} K_{ab}^{ij} \tilde{T}_{ab}^{ij}. \end{aligned} \tag{2.30}$$

The new term which has been introduced is an exchange integral $K_{ab}^{ij} = (ia|jb)$.

The amplitudes are calculated by taking the second term of Eq. (2.21), and multiplying from the left by a contravariant configuration. As stated before, the expression should equal zero for the converged solution

$$R_{ab}^{ij} = < \tilde{\Phi}_{ij}^{ab} | \hat{H}^{(0)} - E^{(0)} | \Psi^{(1)} > + < \tilde{\Phi}_{ij}^{ab} | \hat{H} | \Psi^{(0)} > = 0, \tag{2.31}$$

where $R_{ab}^{ij}$ is referred to as the *doubles residual* for the given electron pair excitation. This expression can be evaluated with help of second quantization to

$$R_{ab}^{ij} = K_{ab}^{ij} + \sum_{c}(f_{ac}T_{cb}^{ij} + T_{ac}^{ij}f_{cb}) - \sum_{k}(f_{ik}T_{ab}^{kj} + T_{ab}^{ik}f_{kj}), \tag{2.32}$$

where $f_{rs}$ are elements of the Fock matrix. In the case of canonical orbitals, the matrix is diagonal and the expression reduces to

$$R_{ab}^{ij} = K_{ab}^{ij} + (\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j)T_{ab}^{ij}. \tag{2.33}$$

The amplitudes can be calculated directly as long as the $K_{ab}^{ij}$ integrals have been computed.

Substitution into Eq. (2.30) gives the canonical MP2 energy

$$\Delta E_{\mathrm{MP2}} = \sum_{ij}\sum_{ab} \frac{K_{ab}^{ij}(2K_{ab}^{ij} - K_{ab}^{ji})}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}. \tag{2.34}$$

Higher order perturbations can also be used, leading to the MP$n$ series. There is however no given proof that the series necessarily converges, and it is often the case that the correlation energy oscillates or even diverges when including higher excitations. It is also difficult to judge the quality of the MP2 estimate. In molecular systems with a small HOMO-LUMO gap, the denominator in Eq. (2.34) will be small and might lead to large errors in the energy. Also, in cases where the HF reference gives a bad energy estimate, MP2 is well known to overestimate the correlation contribution (some empirical corrections have however been proposed with some success[16]).

Nonetheless, it is one of the most commonly used post-HF methods. It is size-consistent (although not variational) and has a very low cost compared to other correlation methods. Formally, the computational cost of canonical MP2 scales with $\mathscr{O}(\mathscr{N}^5)$, where $\mathscr{N}$ stands for the size of the system. This is due to the transformation of the two-electron AO integrals to build the matrices $\mathbf{K}^{ij}$. The scaling can be reduced by integral screening or by reducing the number and/or size of these matrices. Some of these approximations will be later discussed in the text.

### 2.1.3 Coupled Cluster Theory

Coupled Cluster (CC) Theory is (as in the MP2 case) a non-variational size consistent method. It has gained great popularity in the last years, mostly due to the latter property, and to the fact that it converges towards the FCI limit in going to higher order excitations. The CC wave function has the form

$$|\Psi^{\mathrm{CC}}> = e^{\hat{T}}|\Psi^{\mathrm{HF}}> . \tag{2.35}$$

The cluster operator $\hat{T}$ includes excitation operators up to a given order as $\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \ldots$, where the excitation operators are defined as

$$\hat{T}_1 = \sum_i \sum_a \hat{E}_{ai} t_a^i \tag{2.36}$$

$$\hat{T}_2 = \frac{1}{2} \sum_{ij} \sum_{ab} \hat{E}_{ai} \hat{E}_{bj} T_{ab}^{ij}. \tag{2.37}$$

$$\vdots$$

The exponential function shown in Eq. (2.35) may be expanded into a Taylor series

$$e^{\hat{T}} = 1 + \hat{T} + \frac{\hat{T}^2}{2!} + \frac{\hat{T}^3}{3!} + \ldots \tag{2.38}$$

Including excitation operators up to $N$th order (with $N$ the number of electrons in the system) would give the FCI result. However, the $\hat{T}$ operators can be truncated to include only the lowest excitations, since the higher orders should have a smaller contribution. This leads to the hierarchy of coupled cluster models available today. One of the most common truncations is up to double excitations, and the method is termed as Coupled Cluster Singles and Doubles (CCSD). The expansion in (2.38) can in this case be rewritten as

$$e^{\hat{T}_1 + \hat{T}_2} = 1 + \hat{T}_1 + \left( \hat{T}_2 + \frac{\hat{T}_1^2}{2} \right) + \left( \hat{T}_2 \hat{T}_1 + \frac{\hat{T}_1^3}{6} \right) + \ldots \tag{2.39}$$

As seen above, this includes higher order excitations by products of single and double operators. This is the key for the size-extensivity of the method.

The CCSD correlation energy is obtained by applying the Hamiltonian operator to the CCSD wave function, and projecting from the left with the reference function

$$\begin{aligned}
\Delta E^{\mathrm{CCSD}} &= < \Psi^{\mathrm{HF}} | \hat{H} | \Psi^{\mathrm{CCSD}} > \\
&= < \Psi^{\mathrm{HF}} | \hat{H} (\hat{T}_1 + \hat{T}_2 + \frac{1}{2} \hat{T}_1^2) | \Psi^{\mathrm{HF}} > \\
&= < \Psi^{\mathrm{HF}} | \hat{H} \hat{T}_1 | \Psi^{\mathrm{HF}} > + < \Psi^{\mathrm{HF}} | \hat{H} (\hat{T}_2 + \frac{1}{2} \hat{T}_1^2) | \Psi^{\mathrm{HF}} > \\
&= \sum_{ai} < \Psi^{\mathrm{HF}} | \hat{H} | \Phi_i^a > t_a^i + \sum_{ij} \sum_{ab} < \Psi^{\mathrm{HF}} | \hat{H} | \Phi_{ij}^{ab} > \left( T_{ab}^{ij} + \frac{1}{2} t_a^i t_b^j \right)
\end{aligned} \tag{2.40}$$

Again, with help of second quantization, it is possible to transform the equation into a more

practical form

$$\Delta E^{\mathrm{CCSD}} = \sum_{ai} 2 f_{ai} t_a^i + \sum_{abij} \left[2\mathbf{K}^{ij} - \mathbf{K}^{ji}\right]_{ab} \left(T_{ab}^{ij} + \frac{1}{2} t_a^i t_b^j\right). \tag{2.41}$$

The singles and doubles residuals are obtained in a similar way, by projecting from the left with the corresponding contravariant configurations

$$r_a^i = < \tilde{\Phi}_i^a | \hat{H} - E^{\mathrm{CCSD}} | \Psi^{\mathrm{CCSD}} > \tag{2.42}$$

$$R_{ab}^{ij} = < \tilde{\Phi}_{ij}^{ab} | \hat{H} - E^{\mathrm{CCSD}} | \Psi^{\mathrm{CCSD}} > \tag{2.43}$$

The form of the residuals is somewhat more involved than the MP2 case and, for simplicity, I will only give the doubles residual equation leaving out the singles terms. This corresponds to the CCD residual, and is given by

$$R_{ab}^{ij} = K_{ab}^{ij} + \left[\mathbf{K}(\mathbf{T}^{ij})\right]_{ab} + \sum_{kl} \left[K_{ij}^{kl} + \mathrm{tr}(\mathbf{T}^{ij}\mathbf{K}^{lk})\right] T_{ab}^{kl} + G_{ab}^{ij} + G_{ab}^{ji}. \tag{2.44}$$

The matrices hold the same meaning as before, the only new elements introduced are the matrices $\mathbf{K}(\mathbf{T}^{ij})$ and $\mathbf{G}^{ij}$. The former is an *external exchange operator*

$$\left[\mathbf{K}(\mathbf{T}^{ij})\right]_{ab} = \sum_{cd} T_{cd}^{ij} (ac|db), \tag{2.45}$$

the transformed 4-external integrals are contracted with the amplitudes. The **G** matrix accounts for the contribution of other two-electron integrals, and will only be discussed later in the context of local correlation approximations. It is however given in Refs. [17, 18]. The full residuals for CCSD are featured also in Ref. [19].

Although the coupled cluster family of methods converges relatively quickly to the FCI result with inclusion of higher-order excitations[6], it is regularly the case that triple excitations still give an important contribution to the correlation estimate. However, including the full triples is computationally demanding, scaling with $\mathscr{N}^8$. An alternative is the use of perturbation theory to include the triples effect, the CCSD(T) method.[20, 21, 22] This model has become quite popular over the last years, becoming a standard method for high accuracy single reference calculations.

In the CCSD(T) method, the correlation energy is given by the CCSD contribution of

[6] CC methods have a faster convergence pattern in comparison to configuration interaction methods.

Eq. (2.41), and a perturbative correction of the form

$$\Delta E^{(\mathrm{T})} = < \Psi^{\mathrm{HF}} | (\hat{T}_1 + \hat{T}_2)^{\dagger} \hat{V} \hat{T}_3 | \Psi^{\mathrm{HF}} > \tag{2.46}$$

where $\hat{V}$ denotes the perturbation operator. In the canonical case, there is no coupling between individual amplitudes and the correction can be obtained non-iteratively. The computational cost still scales with $\mathscr{N}^7$, but the CCSD iterations are done independently, which leads to a much more cost effective approach than using the full triples. It is also found that the CCSD(T) model is generally more accurate than CCSDT.

### 2.1.4 Local Correlation Methods

The problem of correlation has been connected for a long time to local representations of orbitals. As Kutzelnigg once remarked, "the clearest pictorial description of correlation is (...) the one based on localized orbitals".[23] It is known that dynamic electron correlation is a short range effect, decaying with $\approx r^{-6}$ as the dispersion energy. Conventional correlation methods, however, can make no use of the electron locality since they employ delocalized canonical orbitals. This leads to a quadratic increase of the number of amplitudes needed to correlate each electron pair, and a quartic increase in the number of parameters to be computed.

In order to avoid a steep scaling of computational cost with the molecular size, several methods using a local description of electron correlation have been presented over the years.[24, 25, 26, 27, 28, 29] One of the most successful to date has been the one first proposed by Pulay.[1] His suggestion was to transform the occupied space into a local orbital basis

$$\begin{aligned} |\phi_i^{\mathrm{loc}} > &= \sum_{\mu} |\chi_{\mu} > L_{\mu i} \\ &= \sum_{k} |\phi_i^{\mathrm{can}} > U_{ki} \end{aligned} \tag{2.47}$$

by means of an unitary transformation $\mathbf{U}$, with $\mathbf{L} = \mathbf{CU}$. The transformation can be chosen from any of the several localization algorithms proposed.[11, 30, 31] The Pipek and Mezey scheme[11] is often preferred, since it keeps the separation between $\sigma$ and $\pi$ orbitals.

The virtual orbitals are obtained by projecting out the occupied space from the AO basis

$$|\tilde{\chi}_r > = \left( 1 - \sum_i |\phi_i^{\mathrm{loc}} >< \phi_i^{\mathrm{loc}}| \right) |\chi_r > = \sum_{\mu} |\chi_{\mu} > P_{\mu r}, \tag{2.48}$$

and are commonly called Projected Atomic Orbitals (PAOs). The projection matrix **P** is computed as

$$\mathbf{P} = \mathbf{1} - \mathbf{L}\mathbf{L}^{\dagger}\mathbf{S}. \tag{2.49}$$

This particular selection of occupied and virtual spaces has the following properties:

(1) the occupied orbitals are kept orthogonal among themselves and the virtual space.

(2) the virtual orbitals are however no longer orthogonal, with overlap

$$< \tilde{\chi}_r | \tilde{\chi}_s >= (\mathbf{P}^{\dagger}\mathbf{S}\mathbf{P})_{rs} = \tilde{\mathbf{S}}_{rs} \tag{2.50}$$

(3) there are linear dependencies in the virtual space. The number of PAOs is equal to the size of the AO basis ($N_{\mathrm{AO}}$), although it spans an $N_{\mathrm{AO}} - n_{\mathrm{occ}}$ dimensional space. This is further discussed later in the text.

(4) both occupied and virtual space are inherently local.

The non-orthogonality of the PAOs leads to somewhat more complicated working equations, but the advantages of using a local orbital space greatly compensate this disadvantage. The methods based on these approximations are referred to as local methods, and named after the canonical counterpart with an "L" prefix added. For second order Møller-Plesset theory this will be LMP2, for CCSD the corresponding local method is denoted as LCCSD, and so on. The approximations involved are now detailed.

### Orbital domains

The fundamental approximation in local correlation methods is to restrict excitations from an occupied orbital $|\phi_i^{\mathrm{loc}}>$ to virtual orbitals in its vicinity. For this purpose, the PAOs are grouped together according to the centers of the original AOs. A group of atoms can then be selected for each LMO based on a locality criterion. The respective group of PAOs build up the *orbital domain* $[i]$. Which criterion should be used is, however, not straightforward. Small domains increase the errors in the correlation energy estimate, large domains will slow down the calculation.

The selection is usually done as first suggested by Boughton and Pulay.[9] The current `Molpro` version [8] contains only minor modifications to this procedure. One starts by ordering the atoms according to decreasing Löwdin charges

$$l_A^i = 2 \sum_{\mu \in A} [\mathbf{S}^{1/2}\mathbf{L}]_{\mu i} \tag{2.51}$$

All atoms with charges above a given threshold are automatically added to the domain list. Further centers may be added according to the overlap criterion of Boughton and Pulay. An approximate LMO $|\hat{\phi}_i^{\rm loc}>$ is built using the AO basis from the selected centers

$$|\hat{\phi}_i^{\rm loc}>=\sum_{A\in[i]}\sum_{\mu\in A}|\chi_\mu>\hat{L}_{\mu i}. \tag{2.52}$$

The coefficients $\hat{L}_{\mu i}$ are determined by maximizing the overlap with the original LMO. The Boughton-Pulay criterion is of the form

$$B^i=1-\int|(\phi_i-\hat{\phi}_i)^2|d\tau>T_{\rm BP}. \tag{2.53}$$

Atoms are added to the domain list until $B^i$ is above the threshold $T_{\rm BP}$. The value is normally varied as a function of the basis set, with $T_{\rm BP}=0.980$ and $T_{\rm BP}=0.985$ recommended for double and triple-zeta basis sets, respectively.

Only excitations $i\rightarrow r,r\in[i]$ will be allowed. For double excitations, *pair domains* are built as the union of two single domains $[ij]=[i]\cup[j]$, with similar restrictions imposed $ij\rightarrow rs,rs\in[ij]$. The error introduced by truncation of the virtual space has been proven to be small, more than 98% of the correlation energy is usually recovered. This fraction even increases with the basis size, with part of the loss being linked to a reduction in the basis set superposition errors (BSSE).[32, 33, 34, 35, 36, 37] This is however not the only reason for the difference. As detailed in previous studies,[34] ionic excitations are also left out, and they should account for the error near the CBS limit.

The linear dependencies in the domains, which have already been mentioned, are in the current `Molpro` implementation by default removed individually in each pair domain. The PAO overlap matrix $\tilde{\mathbf{S}}^{[ij]}$ is built for each domain $[ij]$ separately and diagonalized. The eigenvectors which correspond to the smallest eigenvalues are then deleted, or the individual basis functions with the largest coefficients in these eigenvectors.

### The weak pair approximation

The domain approximation alone does not lead to linear scaling with respect to the molecular size. The number of orbital pairs rises quadratically, and without any further approximation this would be the minimal scaling regime. However, since correlation is a short-range effect, orbital pairs located far apart from each other should give small contributions.

It is therefore reasonable to neglect contributions from pairs with large separations. For any molecular system, it is trivial to prove that within a given distance (smaller than the system diameter) the number of neighbors is a constant number. Only outside this sphere does

Table 2.1: Pair types in local coupled cluster calculations. The default values for $R_c$, $R_w$, $R_d$ and $R_{vd}$ are 1, 3, 8 and 15 Bohr respectively. The distance between the orbital centers is given by $r_p$.

| | | |
|---|---|---|
| strong | $R_c > r_p$ | treated at the CCSD level and in the triples (see text). |
| close | $R_c \leq r_p < R_w$ | treated at the MP2 level, and included in the triples (see text). |
| weak | $R_w \leq r_p < R_d$ | treated at the MP2 level. |
| distant | $R_d \leq r_p < R_{vd}$ | treated at the MP2 level through an approximate multipole method (not used in this work). |
| very distant | $r_p \geq R_{vd}$ | neglected. |

the number of neighbors scale quadratically. By defining a cutoff distance, one can keep the number of pairs in a linear scaling regime. The neglected pairs are referred to as *very distant*. Further approximations can be implemented by defining other distance parameters. At medium distances, where the contributions are relatively small but not negligible, lower level correlation methods may be used. In a LCCSD calculation, these pair energies can be estimated by LMP2 theory. Table 2.1 gives a summary of the distance parameters and different pair classifications used in the `Molpro` LCCSD(T) implementation. However, it should be noted that the only condition for asymptotically linear scaling is the neglect of very distant pairs. All other approximations only affect prefactors and its onset.

The effect of pair approximations has been thoroughly tested. It has been found that correlating orbital pairs which share one center (*strong pairs*) at the LCCSD level, and the remaining ones with LMP2 (*weak pairs*), reaction energies and general molecular properties are well reproduced. The triples are somewhat more sensible to this cutoff and only including strong pairs the percentage of correlation energy recovered was found to be below 70% . An extra class - *close pairs* - was introduced to correct for this problem. The triples will be computed for orbitals $(ijk)$ under the condition that one of the pairs $(ij)$, $(ik)$ or $(jk)$ is strong; the two other pairs can either be strong or close. This has been found to bring the percentage up to 90% .[6]

The default values should be seen as a compromise between computational cost and accuracy, and the effect of these approximations should be carefully monitored[7].

[7]Many benchmarking studies have been made in the last years. K. Pflüger has done extensive work in reaction energies and polarizabilities,[38] G. Rauhut and T. Hrenar on frequencies,[39] so that for these type of calculations the approximations have been tested. Exceptions can however still be found when calculating, for example, activation barriers. This will be further discussed in Chapter 6.

### The local equations and the linear scaling regime

To calculate the LMP2 energy correction, the residual in Eq. (2.32) has to be rewritten in the local basis. This involves transforming from the canonical virtual basis to the (non-orthogonal) PAO basis. The steps are detailed elsewhere,[18] involving relatively simple transformations of each element, and resulting in the expression

$$\begin{aligned} R_{rs}^{ij} =& K_{rs}^{ij} + \sum_{tu\in[ij]} f_{rt} T_{tu}^{ij} \tilde{S}_{us} + \sum_{tu\in[ij]} \tilde{S}_{rt} T_{tu}^{ij} f_{us} \\ & - \sum_{k} \left[ \sum_{tu\in[kj]} \tilde{S}_{rt} f_{ik} T_{tu}^{kj} \tilde{S}_{us} + \sum_{tu\in[ki]} \tilde{S}_{rt} f_{kj} T_{tu}^{ik} \tilde{S}_{us} \right]. \end{aligned} \tag{2.54}$$

Although a bit more intricate, it still bears a great resemblance to Eq. (2.32). The major difference lies in the extra matrix multiplications with the PAO overlap matrix from Eq. (2.50). Since this is done with matrix blocks instead of a full matrix, the extra effort is relatively small. The size of these blocks are determined by the size of the orbital domains, which should be more or less independent of the molecular size.

The LMP2 energy can be written as

$$E^{(2)} = \sum_{ij} \sum_{rs\in[ij]} K_{rs}^{ij} (2T_{rs}^{ij} - T_{sr}^{ij}) \tag{2.55}$$

with the amplitudes in the PAO basis being optimized by solving Eq. (2.54).[18] In local MP2 calculations, due to the choice of the occupied and virtual spaces, the equations have to be solved iteratively. This is actually a minor disadvantage in comparison to the canonical counterpart. In about 7-8 iterations the amplitudes are converged. The advantages on the other hand are manifold. As pointed out above, all summations are carried only over the domains. The matrices are of reduced dimensions and can be kept in memory (avoiding the slowdown caused by I/O operations on disk). Also the occupied indices are constrained. Only energies (and residuals) for $ij$ pairs in the pair list have to be computed. As such, the number of terms will scale linearly with the molecular size if very distant pairs are neglected. The sum over $k$ is also restricted according to the pair list.

Local Coupled Cluster theory can be derived essentially in the same way as in the MP2 case. The residuals have to be transformed from the virtual MO to the PAO basis, and the final equations will resemble the canonical result, except for the extra multiplications with the PAO overlap matrix. In Chapter 5, some comments will be made about some of the residual terms and the restrictions made to orbital pairs and domains. The linear scaling properties are however guaranteed in a similar fashion as in the LMP2 case.

### 2.1.5 Density Functional Theory

In all quantum mechanical methods discussed up to this point, the wave function has been the main quantity of interest. Reaching a converged solution for the orbital indices or amplitudes is just another way of saying that one has been able to determine a wave function representation of the system. For $N$-electrons, this corresponds to dealing with $3N$ coordinates, a high-dimensional problem. It would be desirable to find an alternative function from which the energy could be retrieved with a reduced number of variables.

In the seminal work by Hohenberg and Kohn,[40] proof was given of an univocal relation between the energy and the electron density. This made it possible to establish an entirely new approach to the Schrödinger equation problem. Instead of looking for an eigenvector of the Hamiltonian, one could theoretically use a functional connecting a density to an energy value. This leads to a drastic reduction in the dimensionality of the problem, since the electron density only depends on three coordinates. The only problem remaining is to find the right functional for the system we wish to describe.

The general form for a (spin independent) density functional can be given as

$$E[\rho(\mathbf{r})] = -\frac{1}{2}\sum_i < \phi_i|\nabla^2|\phi_i > + \int v(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + \frac{1}{2}\int\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' + E_{xc}[\rho(\mathbf{r})]. \tag{2.56}$$

In a conventional molecular system, the external potential $v(\mathbf{r})$ is given by the potential of the nuclei, so that the second term is the electron-nuclei Coulombic interaction energy. The third term is the electron-electron classical interaction (Coulombic repulsion of two electron clouds). The factor $1/2$ avoids double-counting. The last term is an exchange-correlation functional, which distinguishes the various density functionals among each other.

A further comment should be made about the first term. It is clearly the kinetic energy contribution, but instead of using the density as a function, molecular orbitals are explicitly used to represent the density. In the beginning, density functionals were also used for the kinetic energy. This would correspond to an *orbital-free* theory, since one could always build the system density with any type of functions. However, this leads to large errors in the kinetic energy and to the nonbinding problem[8]. Kohn and Sham were the first to

[8]The Nonbinding Theorem presented by Lieb(1973) and Simon(1977) proves that under the Thomas-Fermi-Dirac model - which uses a density functional form for the kinetic energy - no molecular system would be stable relative to dissociation into constituent fragments. In short: "Goodbye World!".

propose the use of orbitals, by solving self-consistently the equation

$$\left[-\frac{1}{2}\nabla^2 + v(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})}\right]|\phi_i> = \varepsilon_i|\phi_i> \tag{2.57}$$

for each orbital, and defining the electron density as

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{N}\phi_i^*\phi_i. \tag{2.58}$$

In short, although the main function of interest is the density, one needs to define orbitals and to solve the problem self-consistently. The meaning of these Kohn-Sham (KS) orbitals and respective energies is however not clear. There is no proven one-to-one correspondence between a density and the orbitals, and the Koopman's Theorem also does not apply[9].

As stated above, the form of the exchange-correlation term defines the density functional, and several groups have proposed different solutions to the energy correspondence problem. There is an endless list of functionals in the literature, and sometimes little support for making a decision on the method that can better describe the system under study. The most commonly used density functional today is the B3LYP functional,[41] which has widespread applications in almost all fields of chemistry, from organic compounds to metals. The exchange correlation term is given as

$$E_{xc}^{\mathrm{B3LYP}} = 0.2E_x^{\mathrm{HF}} + 0.72E_x^{\mathrm{B88}} + 0.08E_x^{\mathrm{S}} + 0.81E_c^{\mathrm{LYP}} + 0.19E_c^{\mathrm{VWN80}}. \tag{2.59}$$

It includes the exact Hartree-Fock exchange $E_x^{\mathrm{HF}}$, obtained as in Eq. (2.12), mixed with exchange from both the gradient corrected B88 functional[42] and the Slater-Dirac exchange[43]. The correlation is also built from both gradient corrected and local approximations, the LYP[44] and the VWN80[45] functionals, respectively. The numerical parameters were fitted to reproduce atomization energies, proton affinities and ionization potentials of the G1 molecule test set.

Several criticisms have been directed at Density Functional Theory over the years. Although the theory promises an universally valid functional, this has still to be found, and most density functionals will only work sensibly in systems for which they were parametrized. Secondly, there is no systematic way to improve on a DFT estimate. Contrary to wave function-based theory, there is no hierarchy of methods which approaches a converged result in the n-particle space. Another common problem with DFT functionals is

[9]Nevertheless, the KS orbital energies have been used over the years for comparison with photoelectron spectra. These values are, however, often contested, even if in good agreement with experiment.

the difficulty in describing van der Waals (vdW) interactions. These interactions result from non-overlapping electron densities, and therefore cannot be captured by a simple density based functional form.

### 2.1.6 Semiempirical Methods

The main bottleneck in wave function methods is the calculation and transformation of two-electron integrals. Conventional implementations of these methods scale formally with the fourth power of the number of basis functions. *Semiempirical methods* offer a way to overcome this problem by reducing the number of integrals. The valence electrons are represented with a minimal basis set and the core electrons replaced by functions to represent the combined nuclei and the inner shells. The one-electron operator for a valence electron $i$ can then be rewritten as

$$h(i) = -\frac{1}{2}\nabla_i^2 - \sum_{m=1}^{M} \frac{Z'_m}{r_{im}} \tag{2.60}$$

where $Z'_m$ is the reduced nuclear charge due to the core electrons. This leads to a drastic reduction in the number of integrals, since the number of explicit electrons is kept at a minimum.

It is also common in the semiempirical family of methods to assume that the product of two functions lying on different centers will be zero. This is referred to as Zero Differential Overlap (ZDO) approximation, although sometimes also under the name of Neglect of Diatomic Differential Overlap (NDDO). Since one uses a minimal basis set, the quantum numbers will always be different for same center functions and the overlap matrix is unitary.

The most popular semiempirical methods to date are modified NDDO models, parametrized with atomistic or molecular data (such as enthalpies of formation, atomic spectra...). They are commonly referred to under the term Modified Neglect of Diatomic Overlap (MNDO).[46] The minimal basis set is built up of Slater type orbitals up to $p$-type functions. Extension to $d$-orbitals have also been made[47], allowing for the treatment of heavier atoms and polarization effects in outer-valence electrons. However, the most popular semiempirical methods include only lower angular momentum functions, and the discussion will be restricted to this case.

In order to simplify the notation, a different orbital labeling scheme will be used in this Section. Since there is only one type of orbitals in semiempirical methods (in the other methods, one discusses AO and MO orbitals), and the basis set is minimal (a set of $s$ and $p$ orbitals on each atom) an orbital will be generally be named $\mu_A$, where $\mu$ represents the angular momentum of the function (which can only vary between $s$ and $p$-type), and $A$

stands for the atom at which the function is centered. This is of relative importance due to the ZDO approximation.

Adjustable parameters are included in the one and two-electron integrals, and in the core-core repulsion (the Coulombic interaction of the reduced nuclei charges). The one-electron integrals are given in the form

$$h_{\mu\nu} = < \mu_A | \hat{h} | \nu_A > = \delta_{\mu\nu} U_\mu - \sum_{m \neq A}^{M} Z'_A < \mu_A s_m | \nu_A s_m >, \tag{2.61}$$

where $U_\mu$ corresponds to the energy of a single electron experiencing the full nuclear charge. The values for this quantity are parametrized for each atom. The second term is the potential due to all the other nuclei in the system and is parametrized in terms of reduced nuclear charges and a two-electron integral. This involves the valence orbitals as well as extra nuclei-centered $s$-type functions which model the removed core electrons.

The two-center one-electron integrals are approximated as

$$< \mu_A | h | \nu_B > = S_{\mu\nu} \frac{1}{2} (\beta_\mu + \beta_\nu). \tag{2.62}$$

The $\beta$ values are referred to as "resonance" parameters, and the $\mu$ and $\nu$ labels, as before, are either $s$ or $p$-type functions. Contrary to the NDDO approximation, these modified models calculate $S_{\mu\nu}$ explicitly (this is actually the reason why these methods are called "modified" NDDO models).

The two-electron integrals are modelled as interactions between multipoles, being separated into Coulomb terms or exchange. There are relatively few of them, since they are combinations of only $s$ and $p$-type orbitals.

The core-core repulsion is the main difference between the various semiempirical models. In the MNDO methods, one uses the form

$$V^{\mathrm{MNDO}}(A,B) = Z'_A Z'_B < s_A s_B | s_A s_B > \left(1 + e^{-\alpha_A r_{AB}} + e^{-\alpha_B r_{AB}}\right) \tag{2.63}$$

for any general two atoms $A$ and $B$, except for O-H and N-H bonds, where the expression is slightly modified. The fitting parameters are the $\alpha$ exponents. The reason why the repulsion is computed in this way and not by a Coulomb interaction expression, is that a simple Coulomb force, due to the MNDO approximations, is not canceled by the long distance electron interactions. These expressions guarantee the correct limiting behavior.

The Austin Model 1 (AM1) method [48] uses a diferent set of parameters, with the two-electron integrals fitted to atomic spectra. The core-core repulsion is also somewhat

different

$$V^{\mathrm{AM1}}(A,B) = V^{\mathrm{MNDO}}(A,B) + \frac{Z'_A Z'_B}{r_{AB}} \times \left( \sum_k a_{kA} e^{-b_{kA}(r_{AB}-c_{kA})^2} + \sum_k a_{kB} e^{-b_{kB}(r_{AB}-c_{kB})^2} \right), \tag{2.64}$$

with constants $a$, $b$ and $c$ being introduced as new parameters. These have been fitted to molecular data.

At last, the Parametric Method Number 3 (PM3) [49] is a reparametrization, not by hand, as in the AM1 and MNDO models, but fitting all parameters simultaneously with the help of an error function. The core-core repulsion term is the same as in AM1, with the difference that only two Gaussians were assigned to each atom and allowed to vary during the fitting.

Although at heart the semiempirical family of methods is an approximation to HF, the two approaches are not to be mistaken. Due to the parametrization to experimental values, correlation is partly included in the semiempirical Hamiltonian. It is therefore relatively difficult to judge *a priori* how both approaches will compare to each other. In a few cases semiempirical methods may give better agreement to experiment (or higher-level estimates) than HF. The physics should however not be ignored, and caution should be taken when making use of these methodologies. There are many documented cases where these approaches bluntly fail.

## 2.2 Molecular Mechanics

The methods discussed up to this point offer approximations to the solution of the electronic Schrödinger equation. However, for system sizes ranging above one thousand atoms, such an approach is technically impossible by todays standards. Even if the methods computational cost does scale linearly with the system size, the prefactors or the onset of the scaling regime do not allow for such calculations to be performed routinely on biological systems. An alternative way to solve this problem is to define the energy as a parametric function of the nuclear coordinates. The parameters can be chosen to fit experimental data or accurate quantum chemistry results in small systems. The computational cost is strongly reduced, since one needs only to compute some simple analytic functions. At the same time, the accuracy can remain mostly unchanged, as long as the system is found in similar conditions to the ones used for parametrization, and that the reference data is chosen accordingly.

There are several potential functions available, separated into two major types. The

*all-atom* force fields treat explicitly all atoms contained in the structure. In the *united-atom* approach, some of the non-polar hydrogen atoms are joined together with their bonded atom into a single pseudo-atom. The force field is parametrized in such a way that the atom describes the moiety as a whole. The latter force fields are computationally somewhat simpler, since the number of parameters is reduced, but are in most cases also less accurate. Independent of its type, the force field takes the general form

$$V = E_{\text{str}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{vdW}} + E_{\text{el}} + E_{\text{cross}}. \tag{2.65}$$

In other words, the total energy is a sum of various terms, each describing a different physical contribution to the potential. Bond stretches ($E_{\text{str}}$), bond angles ($E_{\text{bend}}$) and torsions ($E_{\text{tors}}$) are included, as well as van der Waals ($E_{\text{vdW}}$) and electrostatic terms ($E_{\text{el}}$). The last term $E_{\text{cross}}$ is called a cross term and introduces a coupling between other components (e.g., between an angle and a distance).

Discussion will from now on be restrained to the all-atom case, although most of it is extensive to the other approach. One of the most commonly used force fields of this type is CHARMM.[50] It will be used as an example to ilustrate the form of the various potential energy contributions. The total energy for this force field is written as

$$\begin{aligned} V_{\text{CHARMM}} &= \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} k_\phi[1+cos(n\phi-\delta)] \\ &+ \sum_{\text{impropers}} k_\omega(\omega-\omega_0)^2 + \sum_{\text{Urey-Bradley}} k_u(u-u_0)^2 \\ &+ \sum_{\text{nonbonded}} \varepsilon\left[\left(\frac{R_{\min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{\min_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon r_{ij}}. \end{aligned} \tag{2.66}$$

Both bond stretches $E_{\text{str}}$ and angles $E_{\text{bend}}$ (first two sums) are parametrized through a harmonic potential, and each needs two parameters to be fitted - the force constant ($k_b$, $k_\theta$) and the equilibrium value ($b_0$, $\theta_0$). Both parameters will vary depending on the bonded atom types. These functions perform best for values of $b$ or $\theta$ close to the minima. Bond breaking phenomena cannot be described by such an expression, since it fails to describe any dissociative behavior[10] The torsion potential $E_{\text{tors}}$ is described by two sums, one running over regular torsion angles, the other over improper torsions, or out-of-plane bendings. The dihedrals are parametrized through the use of a periodic function, with $k_\phi$ being the dihedral force constant, $n$ the multiplicity of the function, $\phi$ the dihedral angle and $\delta$ the phase shift. Since most dihedral potentials are in fact close to a sine form, the potential will work

[10]The harmonic potential would have to be replaced by a Morse potential or some other function which for $b \to +\infty$ would go to zero.

reasonably well for a wide-range of $\phi$. The out-of-plane bendings are fitted to a harmonic potential to describe the natural plane rigidity. The out-of-plane angle is represented by $\omega$. The cross-term $E_{\text{cross}}$ corresponds to the sum of Urey-Bradley components, which are non-bonding interactions between 1,3 neighbors. The variable $u$ is defined as the distance between the 1,3 atoms in the harmonic potential. This introduces in an approximate way a coupling between the atom distances and the 1,3 angle[11]. The van der Waals term $E_{\text{vdW}}$ corresponds to the sum over nonbonded atoms $i$ and $j$, and is calculated with a standard 12-6 Lennard-Jones potential. The $R_{\min_{ij}}$ term is not the minimum of the potential, but rather where the Lennard-Jones potential is zero. The last term corresponds to $E_{\text{el}}$, a simple Coulombic interaction with permittivity $\varepsilon$ included.

The gradient and Hessian matrices can be computed with little effort (compared to the previously discussed methods) since they also have a very simple analytic form. Molecular Mechanics force fields can be routinely applied to systems spanning thousands of atoms. They are quite reliable for conformational studies, describing well (due to their highly parametrized form) otherwise challenging interactions, such as $\pi$-stacking, general vdW interactions or hydrogen bonding.

## 2.3 Quantum Mechanics/Molecular Mechanics

The study of reactivity in biological systems is a major challenge for computational chemistry methods. Whether discussing a system in solution or an enzymatically catalyzed process, the effects of the environment have to be included in the calculation. Although the reaction normally takes place in a relatively confined space, referred to as *active site*, the surrounding system can influence the reaction rate by many orders of magnitude. However, most systems of interest are too large to be treated by Quantum Mechanics methods, and Molecular Mechanics, although inexpensive, cannot treat bond breaking/formation phenomena. Hybrid methods have therefore been developed which combine both approaches.

The Quantum Mechanics / Molecular Mechanics (QM/MM) method [51] was the first to combine two different levels of theory in a single calculation. It can in general be used with any electronic structure method and force field. The system is partitioned into two sections - the QM and MM parts. The former will normally correspond to the region where the reaction takes place and will be treated at the higher quantum mechanical level. This allows for the study of reactivity without the parametrization problem of Molecular

[11] In the case of water, if only bond and angle potentials would be used, the O-H distance would be independent of the H-O-H angle. By introducing a harmonic potential for the two hydrogens (1,3), all the energy terms will be effectively coupled.

Mechanics. The second part, modeled by a force field, ensures that the environmental effects are incorporated into the reaction.

Let us consider a system separated into an active site and the environment. For the time being, we will neglect the existence of bonds between the two sections. There are $M$ atoms treated at the QM level, $X$ atoms in the MM region, and $N$ electrons in the QM region. The total Hamiltonian for this system will be

$$\hat{H}_{\text{tot}} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}}. \tag{2.67}$$

The first term corresponds to the Hamiltonian of the QM particles (nuclei and electrons) in vacuo, as given by Eq. (2.3). The second term describes the interaction between the MM atoms, and is simply given by the force field energy. The coupling between the two regions

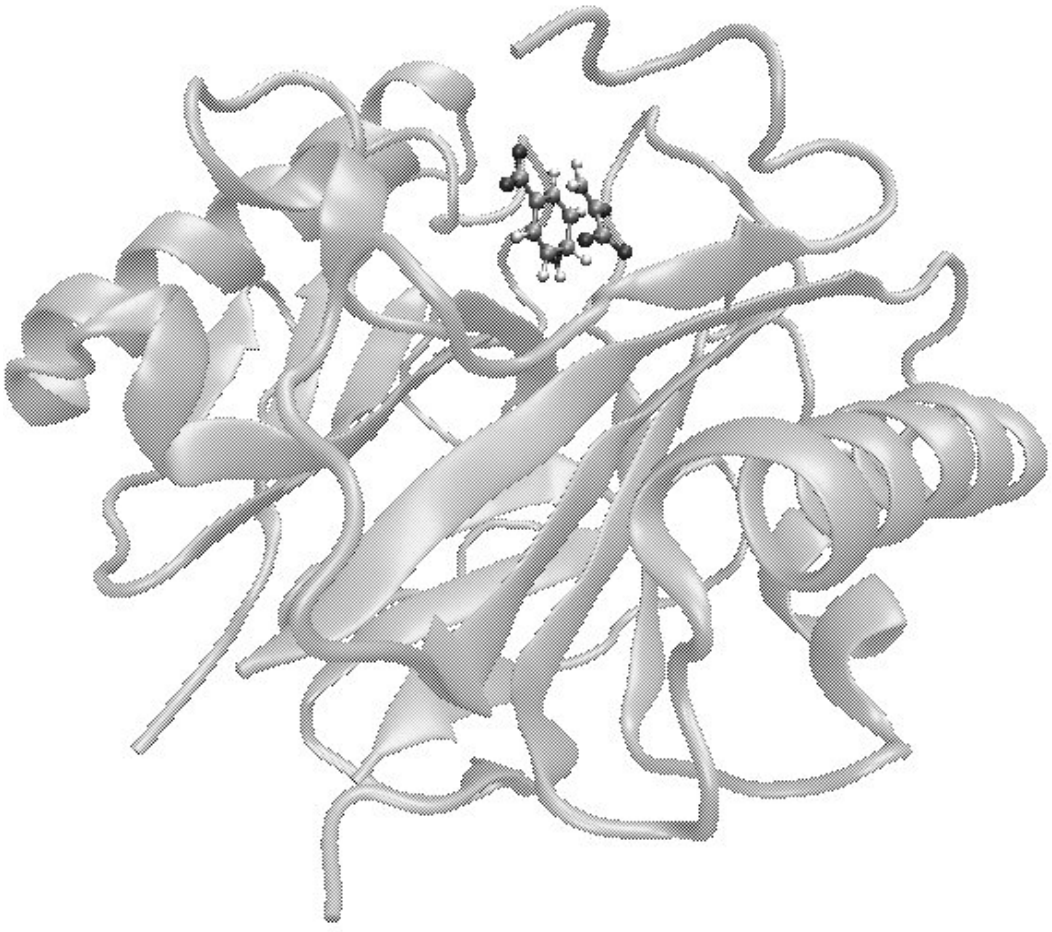


Figure 2.1: Depiction of an enzyme-substrate complex. In a QM/MM calculation the environment is treated with the help of force fields and the active site at higher levels of theory.

is contained in $\hat{H}_{\mathrm{QM/MM}}$, with

$$\hat{H}_{\mathrm{QM/MM}} = -\sum_{x=1}^{X}\sum_{i=1}^{N}\frac{q_x}{r_{ix}} + \sum_{m=1}^{M}\sum_{x=1}^{X}\frac{q_x Z_m}{r_{mx}} + E^{\mathrm{vdW}}_{\mathrm{QM/MM}}, \tag{2.68}$$

where $q_x$ stands for the charge of the $x$th MM atom, $r_{ix}$ the distance between an electron and a MM atom and $r_{mx}$ between two atoms, one in the QM, the other in the MM region. The first two terms represent the electrostatic interaction between the two sections, and are computed by the QM program. The last term stands for the van der Waals interactions. Instead of using an operator, this term is commonly an energy contribution computed by the MM program. It is a simple Lennard-Jones potential as in Eq. (2.66), where standard parameters for the QM atoms are used. The above Hamiltonian accounts for polarization of the QM region due to the MM charges (through the first term of (2.68) and the SCF procedure). This is referred to as *electrostatic embedding*. Backward polarization of the MM system due to the QM part is in this case neglected. One could introduce such terms, but only in conjuction with a polarizable force field, and a self consistent cycle for the QM/MM coupling. Such force fields are, however, rarely used. The polarization effects are already included in an averaged way in the parametrization of the force field so that these effects are generally assumed to be small.

The coupling between the two regions described up till now is referred to as *additive scheme*. The name is due to the addition of the coupling terms present in $\hat{H}_{\mathrm{QM/MM}}$. Only two calculations take place, one for the QM region, another for the total system using MM methods. Other approaches have, however, been presented, which are commonly referred to as *subtractive schemes*. Let us consider the system again divided into two regions, *host* and *cluster*[12], whereby the former denotes the whole system and the latter the cutout to be computed at the QM level. The following formula is used

$$E^{\mathrm{QM/MM}} = E^{\mathrm{MM}}(\mathrm{host}) + \underbrace{E^{\mathrm{QM}}(\mathrm{cluster}) - E^{\mathrm{MM}}(\mathrm{cluster})}_{\Delta E^{\mathrm{sub}}} \tag{2.69}$$

The supercripts indicate the level of theory and the names in parenthesis the geometries used. The total energy for the system is computed at a lower level and a correction is introduced by calculating the difference between the higher and lower levels at the cluster region ($\Delta E^{\mathrm{sub}}$). If both regions would be coincident, this would of course correspond to the exact high level estimate. The size of the cluster should be chosen so that the most relevant energy contributions are contained therein.

[12]The nomenclature is the same as used in the QMPOT program.[52]

Both approaches have advantages and disadvantages. The subtractive scheme is methodologically more general. Observing Eq. (2.69) there is no reason why one should restrict the levels of theory to QM and MM. Since the calculations are decoupled, any level of theory is possible. One may use DFT as a lower level of theory, and combine it with CCSD(T). Also, the number of regions is not restricted to two. One could imagine a sequence of corrections in an onion-like approach, where $\Delta E^{\text{sub}}$ would include the constributions from the inner to the outer region. This is commonly referred to as the ONIOM approach.[53] The same method has also been popularized by Sauer and coworkers [52] in the study of zeolites.

There are however some disadvantages relative to the additive scheme. The correction $\Delta E^{\text{sub}}$ is obtained in the vacuum (i.e., without any information about the environment). The effect of the surrounding molecule(s) is only contained in the $E^{\text{MM}}(\text{host})$ term and is therefore exclusively treated at the lower level. There is no polarization. Recent work has compensated for this fault,[54] but including polarization corrections is not trivial when coupling quantum mechanical methods, since the polarization interaction is in this case not reproducible by including point charges.

## 2.4 Reaction Rate Theory

### 2.4.1 Transition State Theory

Transition State Theory (TST)[13] postulates that a reaction proceeds from one energy minimum to another via a maximum on the potential hypersurface. This maximum is referred to as the transition state (TS). A reaction can therefore be understood in terms of a path connecting reactant and product state, with a hill separating the two minima, as depicted in Fig. 2.2. The TS will correspond to the structure where the potential is maximal.

In classical mechanics, the probability with which the system will progress from reactants to the products should depend on the barrier height, or in other words, on the energy difference between reactant and TS. If one considers a simple Boltzmann distribution[14], then this probability will be proportional to $e^{-\Delta^{\ddagger}G/k_BT}$. $\Delta^{\ddagger}G$ is the free energy difference between the reactant and transition states (see Fig. 2.2)[15]. This however neglects three effects

[13] Sometimes also referred to as Activated Complex Theory.

[14] In a Boltzmann distribution, the probability of finding the system on any given state along the path is given by $e^{-\Delta E/k_BT}$.

[15] The free energy of activation includes not only the energy barrier necessary for a reactant state to be activated (the so called *critical energy*), but also the partition functions of the species involved.

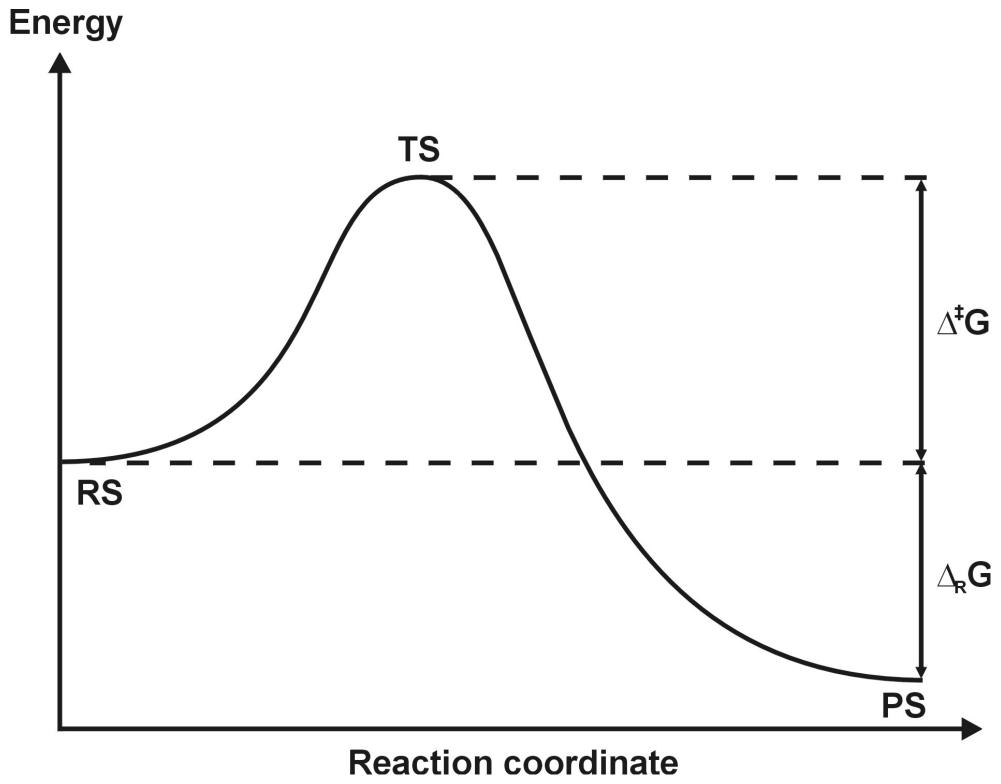



Figure 2.2: Diagram ilustrating the concept of a transition state (TS) intermediate. The potential curve connects the reactant (RS) to the product state (PS), with $\Delta^{\ddagger}G$ representing the energy barrier separating the two states, and $\Delta G_{\mathrm{R}}$ the total reaction energy.

(1) the movement over the TS may be coupled to other movements from the activated complex,

(2) once the TS is reached the system does not necessarily fall into the products and

(3) quantum mechanical tunneling. There should be a small quantum "leaking" to the product side, without necessarily passing over the TS.

The first point would lead to a breakdown of the TST, removing the dependence of the reaction rate on the TS barrier. The other two factors lead to some minor corrections in the formulae. When the energy difference between reactants and products is small, the probability that the system might fall back into the reactants increases, and the speed of the reaction will decrease. On the other hand, quantum tunneling allows the system to travel directly to the product side, increasing the reaction rate. Both factors are taken into consideration by introducing a *transmission coefficient* $\kappa$. For a simple unimolecular

reaction, leading from the reactant A to the product B, the rate of formation is given by

$$\frac{d[B]}{dt} = k(T)[A], \tag{2.70}$$

where [ ] stands for the concentration of each species, and $k(T)$ is called the macroscopic rate constant, which is temperature dependent[16]. This explicit dependence will be from now on dropped. The rate constant is given by

$$k = \kappa \frac{k_B T}{h} e^{-\Delta^{\ddagger} G/RT}. \tag{2.71}$$

Within the limits of the Boltzmann distribution this formulation will hold. It is of general use for liquid or gas state reactions.

The picture of a continuous structural change connecting the reactant to the product states is rather universally accepted. The various PES which have been computed in the last dozens of years, and the high quantitative predictions which have been thereof extracted are undeniable evidence for the existence of these intermediate structures. The most severe criticisms are actually directed at the various assumptions detailed above, and the range of quantitative application which can be given to TST. Particularly in Biochemistry, there is an active debate on the possibility of explaining enzymatic catalysis through the use of such a simplified model.

### 2.4.2 Michaelis-Menten kinetics

Since in this work enzymatically catalized reactions will also be discussed, it is worth at this point to shortly adress Michaelis-Menten kinetics. For a substrate ($S$) which is transformed into a product ($P$) through the mediation of an enzyme ($E$), the following scheme should be valid

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P. \tag{2.72}$$

The first step corresponds to the substrate binding to the enzyme, the enzyme-complex ($ES$) then reacts, forming free enzyme and product. The second step includes product formation and release. Which one of these processes controls the effective kinetic constant, depends on the reaction studied.

The scheme only holds as long as the enzymatic reaction is irreversible, and the product does not rebind to the enzyme. Using Eq. (2.70), and considering the *quasi-steady*

[16] For a more general reaction, with more than one reactant, one just has to multiply the concentrations in the right side, to the power of each order.

*state* approximation[55] that the concentration of enzyme-substrate complex is constant, the relation between the system concentrations is given by

$$[ES] = \frac{[E][S]}{K_m} \tag{2.73}$$

where

$$K_m = \frac{k_{-1} + k_2}{k_1} \tag{2.74}$$

is the *Michaelis-Menten constant*. The rate of product formation can be written as

$$\frac{d[P]}{dt} = k_2([E] + [ES])\frac{[S]}{K_m + [S]} = V_{\max}\frac{[S]}{K_m + [S]}. \tag{2.75}$$

If the concentration of substrate is large compared to the value of $K_m$, the system is said to be saturated, and a maximum velocity of $k_2([E] + [ES])$ is reached.

Enzymes are often characterized by their $K_m$ value. In cases where product formation is rate limiting (large $k_2$), it represents the dissociation constant of the enzyme-substrate complex. Low values indicate a large complex stability and $ES$ will rarely dissociate without the substrate first reacting to form the product.

# Chapter 3

# Computing Potential Energy Surfaces using Local Correlation Methods

## 3.1 The Domain Discontinuity Problem

As detailed before, local correlation methods employ local orbital spaces to restrict the number of excited configurations in the wave function. The use of a truncated virtual space is an essential approximation for achieving linear scaling. By the use of domains, the number of excitations per electron pair becomes independent of the molecular size. The domain lists are also used for defining the orbitals distances, which in turn are necessary for truncating the orbital pair list (neglect of very distant pairs).

These approximations work very well when near equilibrium properties like equilibrium geometries[56, 57, 58], harmonic vibrational frequencies[59, 60, 39], or other properties like dipole moments[38], dipole polarizabilities[38, 61], or NMR chemical shifts[62] are computed. The smoothness of the potentials in geometry optimizations or frequency calculations can be ensured by freezing the domains; in geometry optimizations this is done once the geometry stepsize is smaller than a certain threshold. This is similar to density functional theory, where the definition of the grid must also be fixed at a certain stage.

A more complicated situation arises if activation or reaction energies are considered, and the electronic structure of the system strongly changes along the reaction coordinate. In such cases the results of local correlation calculations can be affected in several ways: when following a reaction path, steps may appear on the potential energy surface (PES) due to discontinuous changes of the localized orbitals or domains. This has been pointed out by Russ and Crawford[63], who used three model systems to investigate this problem. In the homolytic bond breaking of $CH_3$-F, no discontinuity was observed, since in a spin-restricted framework the localized bond orbital stays delocalized over both fragments. However, heterolytic bond dissociation of ketene ($CH_2CO$) and propadienone ($CH_2CCO$) with carbon monoxide as a product revealed in both cases small steps on the PES, caused by changes of the domains. The potential energy surface for ketene, using MP2 and CCSD is depicted in Fig. 3.1, for both local and canonical methods. The difference between both sets $\Delta E_{\mathrm{loc}}(r) = E_{\mathrm{LCCSD}}(r) - E_{\mathrm{CCSD}}(r)$ (or MP2) is given in Fig. 3.2 on a smaller scale. The discontinuities are difficult to recognize in the full path. The steps are only a few milihartree in magnitude, and only by plotting the difference between the two calculations is it possible to clearly observe the effect of domain changes. The computational details will be later discussed in Section 3.3.

Russ and Crawford have pointed out that the discontinuities are of the same order of magnitude as the localization error (due to the truncation of the virtual space) and therefore not negligible. It is desirable to establish a procedure which can produce a continuous PES without affecting the linear scaling behavior of the local methods. Not only due to the

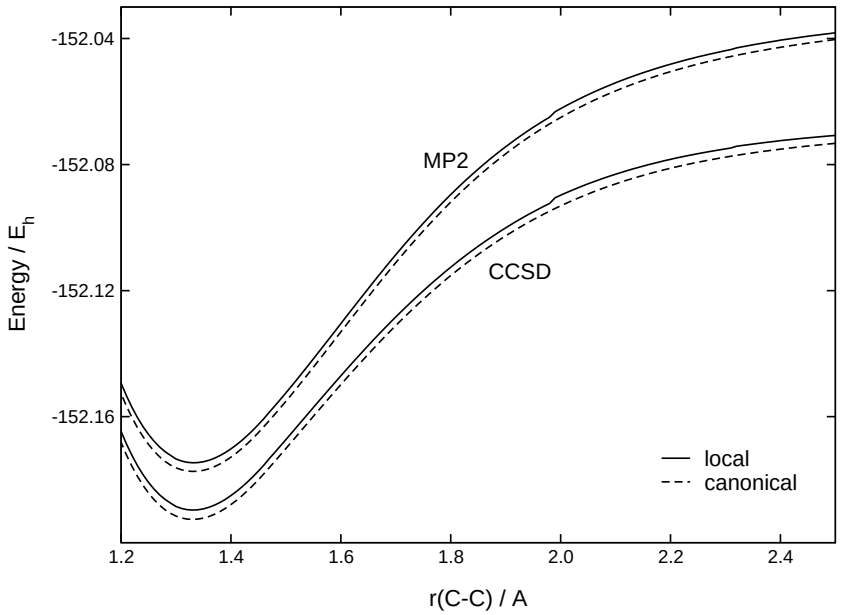



Figure 3.1: (L)MP2 and (L)CCSD potential energy surfaces for the ketene dissociation.

discontinuities, relative energies may also be affected. If the reactants and products have different electronic structures, the domain sizes may change and this can lead to unbalanced results. In particular, it may happen that at the transition state the electronic structure is more delocalized than for reactants and products, and the resulting larger domains at the transition state can then lead to a significant underestimation of computed activation energies.

This subject is of particular concern in the context of this work. Many reactions of biological interest involve complex aromatic structures, and the domain approximation can lead to large errors in relative energies. Due to the system sizes involved, it is difficult to access these effects by comparison to results from canonical methods. An automated procedure is needed to reduce the geometry dependence of the local methods errors, while preserving their low computational cost.

## 3.2 Domain Merging

### 3.2.1 Method

Let us consider the energy $E(\mathbf{x})$ of a molecular system, where $\mathbf{x}$ is a vector representation of the spatial coordinates. If the domain for a given LMO $\phi_i$ at position $\mathbf{x}$ is given by $[i]_{\mathbf{x}}$,

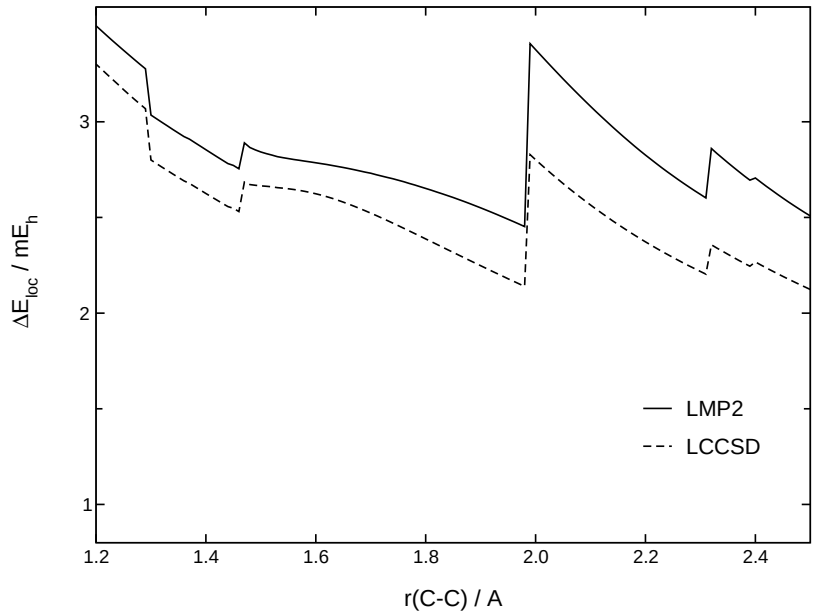



Figure 3.2: Energy difference between local and canonical counterparts, which reveals the discontinuities in the profile.

the domain list is defined as the group of domains

$$\mathbb{D}(\mathbf{x}) = \{\ [i]_{\mathbf{x}}\ \mid\ i \in \text{occupied}\ \}. \tag{3.1}$$

In the case of a reaction, where the system progresses from an initial configuration, defined by $\mathbf{x_i}$, to a final configuration $\mathbf{x_f}$, the energy should remain continuous for any value in the interval, and as a consequence, the domain list constant.

An obvious approach to the problem would be to use a domain definition which would include all of the domains along the reaction path. This would, however, be costly, since it would involve determining domains at all the points along the path. Alternatively, one could also pick the most relevant points on the surface, and only merge the domains for a few selected geometries. This would be enough even in case one uses two points as long as the other domain lists are contained in those of the two selected configurations. The most straightforward solution is to use the initial and final domain lists

$$\mathbb{D}^{\text{merge}} = \{\ [i]_{\mathbf{x_i}} \cup [i]_{\mathbf{x_f}}\ \mid\ i \in \text{occupied}\ \}. \tag{3.2}$$

Building the union of two domain lists is however not straightforward. In order to make use of Eq. (3.2) the orbitals should be in a 1-to-1 correspondence between the two sets $\mathbb{D}(\mathbf{x_i})$ and $\mathbb{D}(\mathbf{x_f})$. This involves finding the orbitals with the largest resemblance at both geome-

tries. A possible criterion for this would be the overlap between the two sets of occupied orbitals. However, for large displacements, the value will be close to 0. Computing the overlap with use of the AO overlap matrix from one of the geometries solves the problem for bond elongations, but might fail in the case of rotations. For example, if a methyl group rotates by about 120°, the overlap matrix will indicate a false ordering for the 3 C-H bond orbitals.

Another possibility to pair the orbitals of one geometry to a second set is to compare the orbital domains and find the pairing for which the orbital domains coincide best. This removes the geometry dependence in the comparison, but can in some cases lead to an arbitrary pairing (in cases where more than one orbital have the same domains). Our procedure uses both criteria, overlap and domain comparison. In this way, the stronger criteria has priority (overlap) and the domain comparison can correct for problems due to rotations.

Once the orbitals in each set are paired, only a small subset of the domain lists should not agree. For each of these orbitals the center lists are merged $[i]_{\mathbf{x_i}} \cup [i]_{\mathbf{x_f}}$. If there are multiple bonds, it can happen that several orbital domains have the same center lists. In this case the merged domains are used for all of these orbitals.

## 3.3 Test Applications

### 3.3.1 Ketene and propadienone bond dissociation

The first tests for the procedure were made for the same systems as studied by Russ and Crawford,[63] namely the ketene and propadienone bond dissociation. The reaction paths were optimized as described in the aforementioned paper: first the equilibrium structures were determined at the CCSD/cc-pVDZ level, and then the C-C bonds were increased while relaxing all other geometry parameters. The resulting structures were used in single point calculations using LMP2 and LCCSD and different basis sets.

For the dissociation reaction of ketene ($CH_2CO$) into singlet methylene and carbon monoxide, the C-C distance was varied in the range from 1.2 Å to 2.5 Å. This region includes the minimum and the four discontinuities observed by Russ and Crawford.[63] At the equilibrium structure ketene has $C_{2v}$-symmetry (C-C distance 1.33 Å), but due to out-of-plane bending the symmetry reduces to $C_s$ at a C-C bond distance of about 1.47 Å.[64, 65] The MP2, LMP2, CCSD and LCCSD energy profiles along the reaction path are shown in Fig. 3.3. The middle panel of Fig. 3.3 shows the difference $\Delta E_{\mathrm{loc}}(r)$ between the local and canonical calculations on a much smaller scale. This plot clearly reveals the four discontinuities at $r = 1.30, 1.47, 1.99, 2.32$ Å. Near the equilibrium structure, the oxygen

out-of-plane lone pair ($b_1$ symmetry) mixes with the C-C $\pi$ bonding orbital in the same symmetry, and this leads to a domain that extends over the 3 heavy atoms. However, the contribution at the methylene C-atom is very small. At shorter ($r < 1.3$ Å) or longer ($r >$ 1.47 Å) distances this atom happens to be not included in the domain when the Boughton-Pulay method is used with a completeness criterion of 0.98. This leads to the lowering of the LMP2 and LCCSD energies in the range 1.3 Å $< r <$ 1.47 Å. At a distance of 1.99 Å one of the C-C bond domains becomes a lone pair ($sp^3$ hybrid) on the $CH_2$ fragment, and at 2.32 Å the second C-C bond domain becomes a lone pair on CO.

The discontinuities disappear when the domain merging procedure is used. In a first test, the first and last points from the path were used to define the domains ($\mathbf{x_i}$=1.2 Å and $\mathbf{x_f}$=2.5 Å). In this case all orbital domains at 2.5 Å are contained in the ones at 1.2 Å (the two C-C bonds are broken and at 2.5 Å one lone pair on each fragment is formed). Thus, the larger domains determined at 1.2 Å are automatically used for all structures. This leaves the energy at short distances unchanged and lowers the energy at long distances. In a second test the domains determined at the equilibrium distance ($\mathbf{x_i}$=1.33 Å) were merged as well. As mentioned above, at this point one of the orbital domains extends over three atoms and consequently these extended domains are used for all three orbitals which involve the CO bond. As expected, the deviations from the canonical CCSD results are somewhat smaller in the latter case. At large distances, the contribution of the basis functions at the methylene C-atom to the description of the CO fragment becomes very small, which explains why the deviations become the same for both test cases.

The flattening of the curve is a positive feature, since it indicates a more systematic deviation from the reference canonical calculation. However, it is observed that the differences between CCSD and LCCSD increase with decreasing C-C distance. This is almost certainly due to the increasing BSSE in the canonical calculation, which is absent (or at least reduced) in the local case[32, 33, 34, 35, 36, 37, 66]. An additional feature is that the increase of this effect is not monotonic, but exhibits a flattening in the region between 1.6 and 1.45 Å. In this region the molecule becomes planar, and this leads to a reduction of the BSSE. In fact, near 1.45 Å there is a valley-ridge inflection point[67, 68], i.e., the second energy derivative with respect to the out-of-plane angle(s) changes from being positive at shorter distances to negative at longer distances. This leads to a very sudden change in the optimized angle as a function of the C-C bond distance (see lower panel of Fig. 3.3).

The same procedure was applied to the lowest singlet state of propadienone. The C-C distance between CO and vinylidene was varied in the same range as ketene. At the equilibrium structure propadienone has a 'kinked' structure ($C_s$ symmetry),[69] but at small C-C distances (1.26 Å) it 'snaps' to $C_{2v}$. Again, as in the case of ketene and shown in the

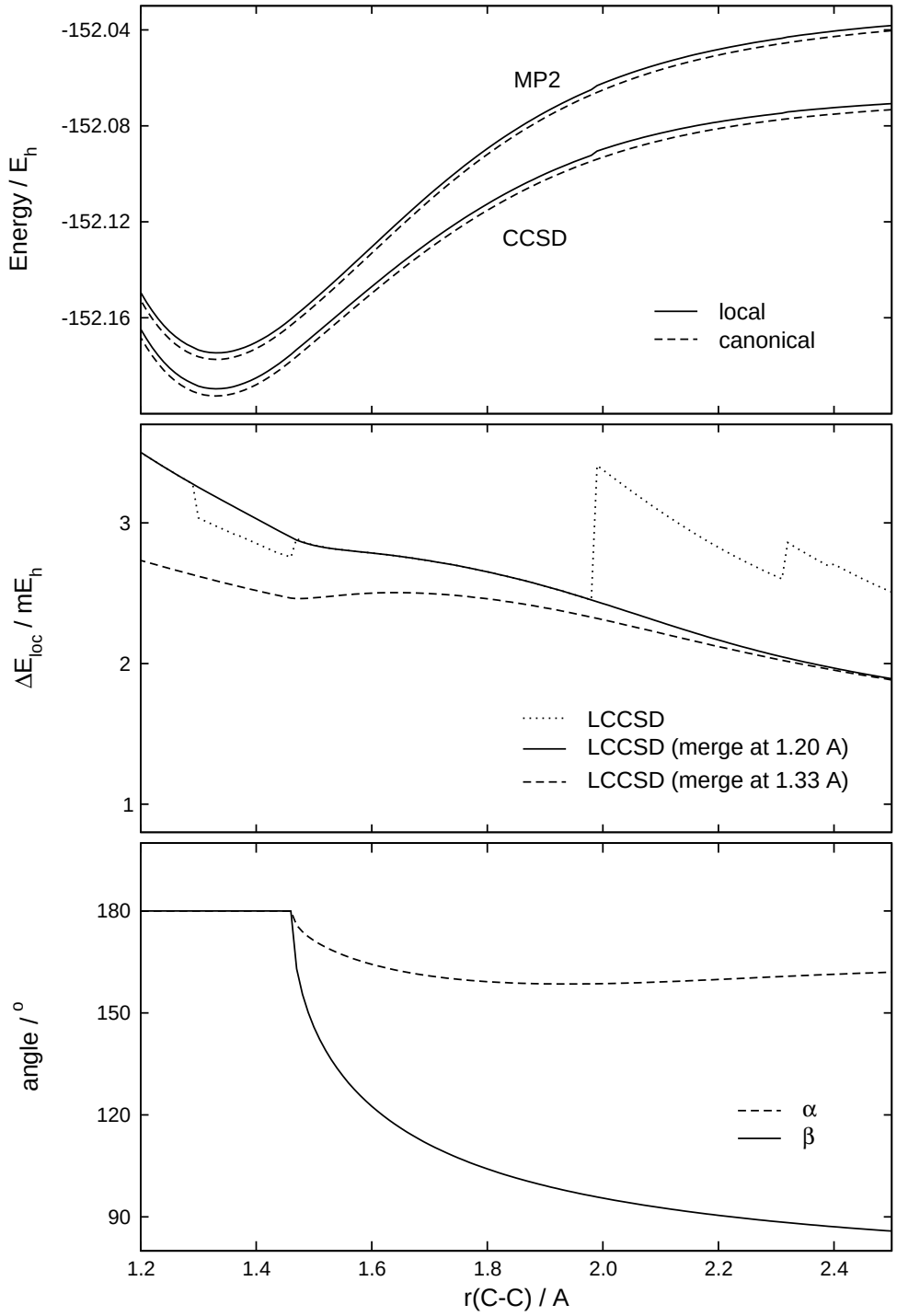



Figure 3.3: Upper panel: (L)MP2 and (L)CCSD potential energy curves for ketene (in a.u.) as a function of the C-C distances. For each C-C distance, all other geometry parameters were optimized at the CCSD/cc-pVDZ level. Middle panel: Difference between the LCCSD and CCSD energies in millihartree. Lower panel: Optimized out-of-plane angles. $\alpha$ is the angle between the $CH_2$ plane and the C-C-bond and $\beta$ is the C-C-O angle.

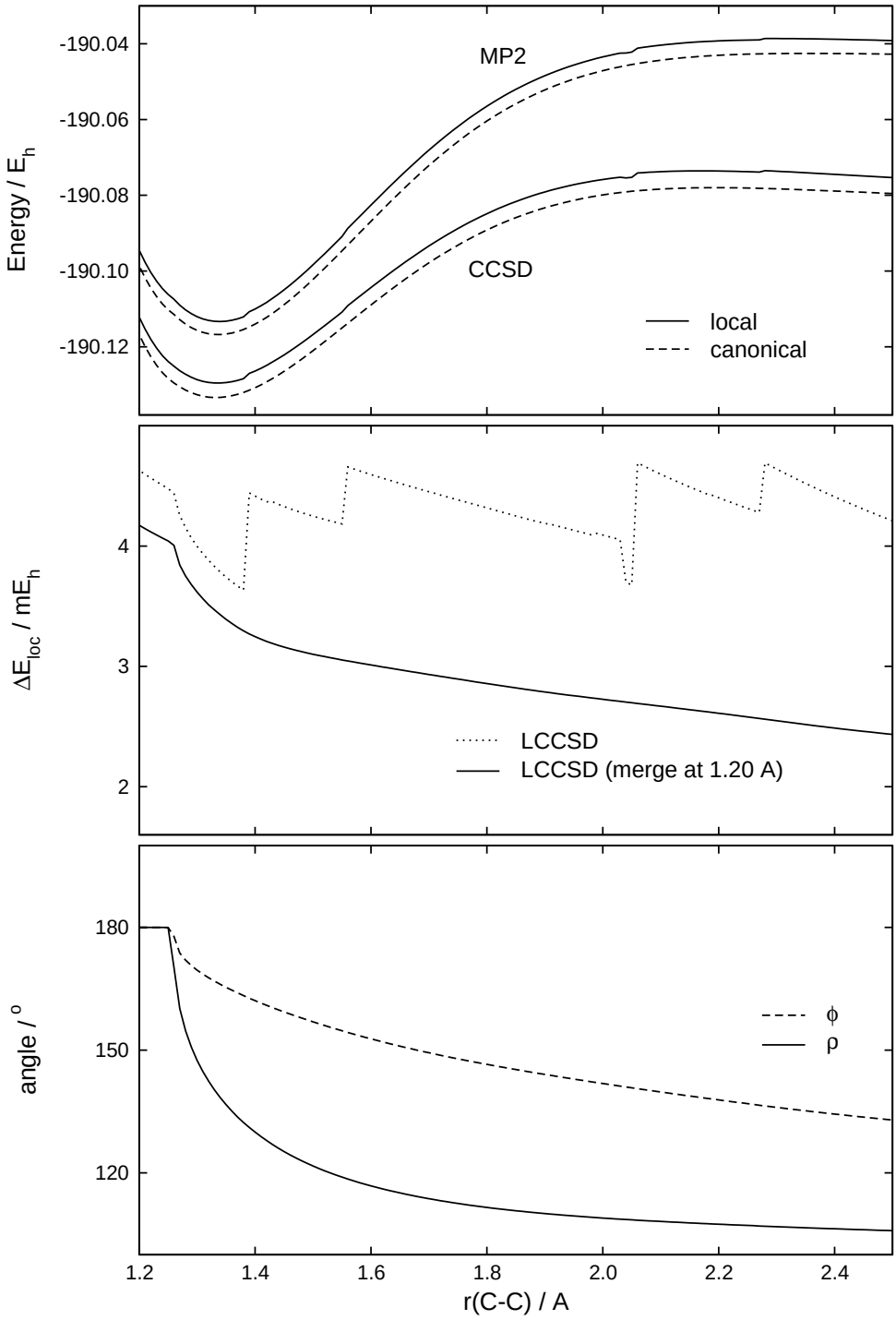



Figure 3.4: Upper panel: (L)MP2 and (L)CCSD potential energy curves for propadienone (in a.u.) as a function of the C-C distance. For each C-C distance, all other geometry parameters were optimized at the CCSD/cc-pVDZ level. Middle panel: Difference between the LCCSD and CCSD energies in millihartree. Lower panel: Optimized out-of-plane angles. $\phi$ is the C-C-O angle and $\rho$ the C-C-C angle.

lower panel of Fig. 3.4, there is a valley-ridge inflection point leading to a very sudden change of the bending angle as a function of the C-C bond distance.

Domain changes occur exclusively for LMOs located at the carbons. At small distances, there are two $\sigma$ C-C bond orbitals and two $\pi$ orbitals delocalized over the three centers. After bond breaking, two lone pairs are left (one on the CO, the other on vinylidene) as well as one $\sigma$ and one $\pi$ orbital on vinylidene. In principle, merging of the two lone pairs would be sufficient to describe the dissociation qualitatively correct. However, at short distances the $\pi$ ($a''$) bond of the vinylidene fragment extends to the third carbon atom. Since the procedure does not consider the symmetry of the orbitals, three merged domains extending over the three C atoms are generated. The profile for the calculation with geometry dependent domains shown in the middle panel of Fig. 3.4 resembles the one obtained by Russ and Crawford at small distances, but lacks some of the discontinuities found by these authors in the bond-breaking region. This may be due to differences in the details of the domain selection procedure.

The use of merged domains produces a smooth curve. Again, the differences between the local and non-local calculations increases with decreasing distance, and at the valley-ridge inflection point the curvature suddenly changes. As already discussed for ketene, the increase of the difference $\Delta E_{\mathrm{loc}}(r)$ between local and canonical energies with decreasing C-C distance is attributed to increasing BSSE in the non-local case.

The effect of the domain approximation for LMP2, LCCSD, and LCCSD(T) was also studied for the propadienone. The effect is seen to be very similar in all cases, which is in agreement with previous studies.[10] Based on this fact, further tests were performed only at the LMP2 level (the computationally cheapest method).

### 3.3.2 SN2 reaction of hydrochlorocarbons with chlorine

The nucleophilic attack of chloride on ethylchloride is a well known textbook example of a SN2 reaction. It is well described by single-reference correlation methods, and various studies at the MP2 level have been published.[70, 71, 72, 73, 74] It is a further example in which lone pairs turn into bonds (and vice versa) during the reaction, and therefore domain changes can be expected. Furthermore, the analogous reactions of 1-propylchloride and 1-butylchloride have been considered in order to investigate the effect of local approximations on computed barrier heights and well depths of the complexes in the entrance channels.

For the $C_2H_5Cl + Cl^-$ reaction, the energy profile was computed along an assumed reaction coordinate $R_2 - R_1$, where $R_1$ and $R_2$ are the two C-Cl bond distances (see Fig. 3.6). The difference was taken since if only one distance was fixed the other one varied very

quickly in some regions, leading to an unexpected shape of the energy profile. The reaction coordinate was varied in steps of 0.04 Å in the region between -2.0 to 2.0 Å. All other internal coordinates were optimized at the MP2 level. The cc-pVTZ(d/p) basis was used for C and H, and the aug-cc-pVTZ(d) basis for Cl ($d$-functions on H and $f$-functions on the other atoms were omitted). In the following, this basis is denoted [aug]-cc-pVTZ(d/p).

Fig. 3.5 shows the MP2 and LMP2 energy profiles along the reaction coordinate. The upper panel shows the absolute energies. The MP2 and LMP2 curves are closely parallel, but the LMP2 energies are about 10 mH higher than the MP2 ones. In agreement with previous studies[70, 74] it is found that the transition state is unsymmetric with respect to $R_1$ and $R_2$, with a $\sigma'$ plane containing the carbons and chlorines. Correspondingly, the reaction coordinate is unsymmetric; in the exit channel, the $CH_3$ group is rotated by 60 degrees about the C-C axis relative to the minimum structure. Thus, in this region the optimized structures correspond to local minima, which are separated from the global minima by rotational barriers.

The middle panel of Fig. 3.5 shows the difference $\Delta E_{\mathrm{loc}} = E_{\mathrm{LMP2}} - E_{\mathrm{MP2}}$ on a smaller scale. The dashed curve corresponds to the calculation with standard domains. Two discontinuities are found in this case on the LMP2 potential, one at positive and one at negative values of $R_2 - R_1$. These correspond to the changes of lone pairs into C-Cl bonds for the entering and exiting $Cl^-$ anions, respectively. Between these discontinuities two domains extend over the whole Cl-C-Cl unit, leading to an energy lowering of about 1 mH. With the automatic domain merging procedure, applied to the reactant, transition state, and product structures, all Cl lone pair and C-Cl domains are united, leading to 8 identical merged domains. In this case a smooth potential function is obtained. Interestingly, $\Delta E_{\mathrm{loc}}$ has a minimum for $R_2 - R_1 \approx 0$ and maxima around $R_2 - R_1 \approx \pm 0.8$ Å. As can be seen in the lower panel of Fig. 3.5, showing $R_1 + R_2$, the average distance between the two chlorine atoms and the ethyl group is minimal in this region; in addition, there is also the closest proximity between the hydrogen atom on the $\sigma'$ plane at the neighboring $CH_3$ group and the incoming $Cl^-$. Therefore, the maxima in $\Delta E_{\mathrm{loc}}$ should be again due to a maximum of the BSSE in the canonical calculation, i.e., the bumps are not caused by the LMP2 but by artificial lowering of the canonical MP2 energy by basis set superposition effects.

We now turn to the question about the effect of the local approximations on the computed well depths and barrier heights for the reaction of $Cl^-$ with ethylchloride, 1-propylchloride and 1-butylchloride. Two stationary points were optimized in each case: the weak complex in the entrance channel, which is stabilized by electrostatic and van der Waals forces, and the transition state, where the electrophilic carbon is partly bonded to the entering ion. These stationary points were optimized with no symmetry constraints

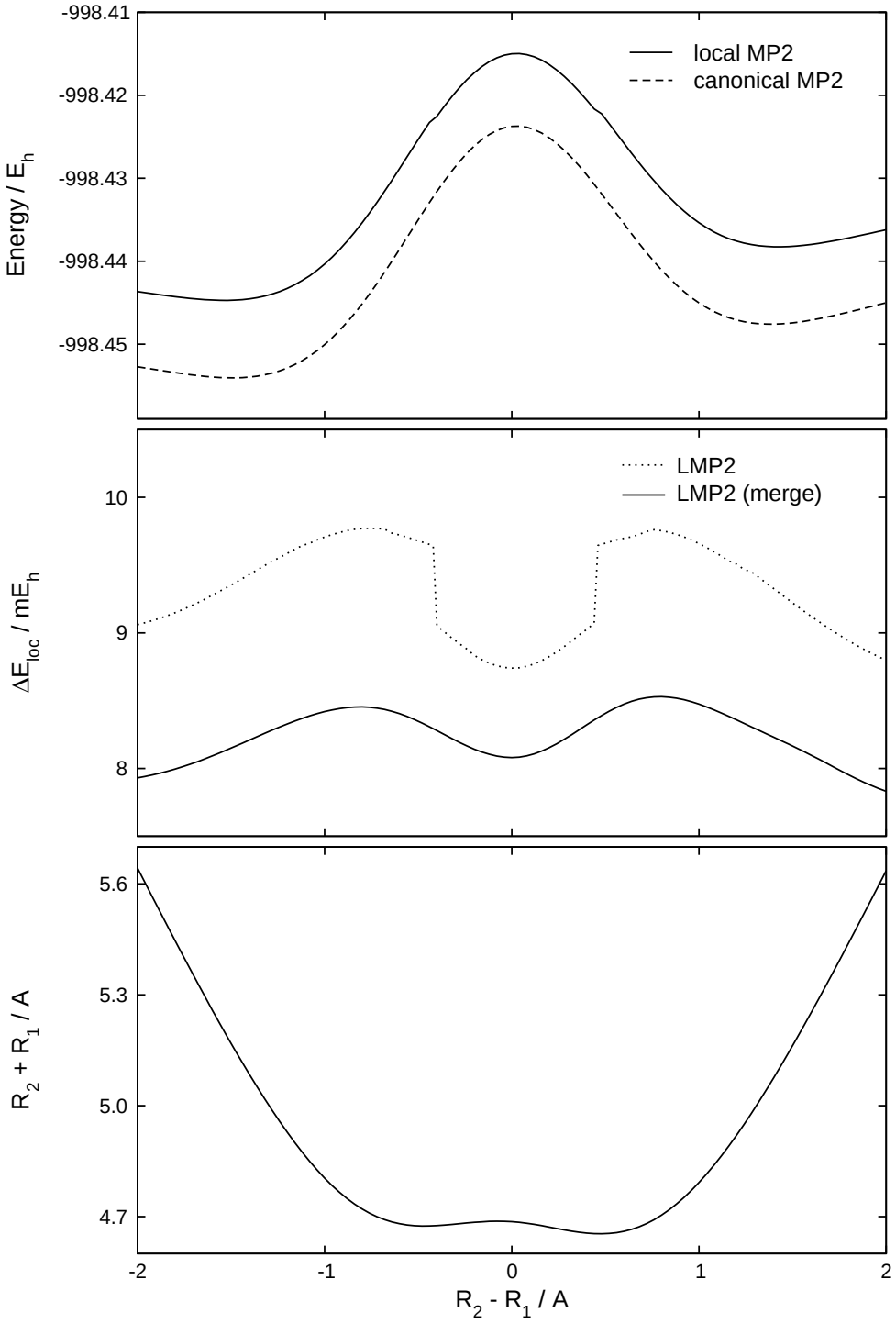



Figure 3.5: MP2 and LMP2 Energy profiles for the $C_2H_5Cl + Cl^-$ reaction. The [aug]-cc-pVTZ(d/p) basis has been used. Upper panel: absolute energies; middle panel: energy difference between MP2 and LMP2 calculations; Lower panel: $R_1 + R_2$ as a function of the reaction coordinate $R_1 - R_2$.

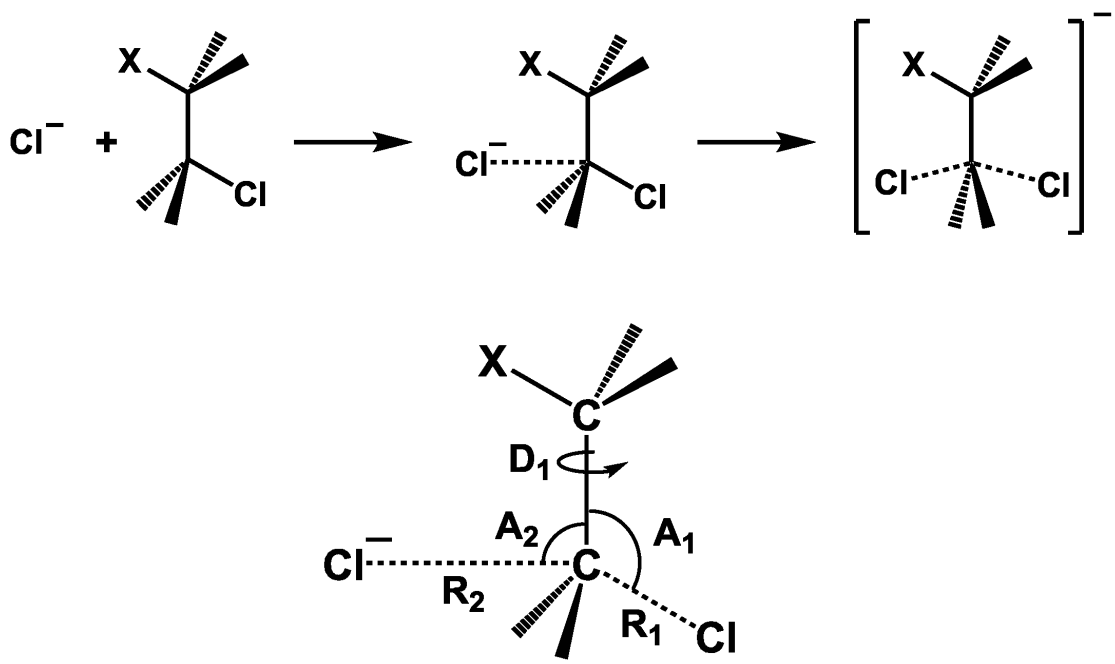



Figure 3.6: Schematic representation of the SN2 reactions of $Cl^-$ with ethylchloride (X=H), 1-propylchloride (X=$CH_3$), and 1-butylchloride (X=$CH_2CH_3$). $D_1$ is the dihedral angle between the entering chlorine, the two carbons and the X fragment.

at the MP2 level. All optimizations were carried out with the [aug]-cc-pVTZ(d/p) basis. The values of the optimized coordinates are summarized in Table 3.1 (see Fig. 3.6 for the definition).

For ethylchloride + $Cl^-$ the structures are in good agreement with those obtained in previous studies[70, 71, 72, 73, 74]; small differences are due to the larger basis set. However, for the propylchloride complex, the structure differs from previous ones[70, 71] in the orientation of the added methyl group. The values for $D_1$ given in the aforementioned references are also presented in Table 3.1. Jensen[70] placed the alkane chain in the direction of the entering chlorine ($D_1 \approx 0^o$), similar to the ethyl reaction. In our calculations, it is instead pointing sideways ($D_1 \approx 90^o$), as in the transition state. No true PES minimum was found for the other structure. Using the same level of theory (MP2/6-31G*) it is, however, possible to locate a saddle point with similar conformational features as the ones published in the previous works.

For butylchloride no previous calculations of the geometries were found. In the optimizations the start geometries were taken from the propylchloride structures, replacing the hydrogen that is farthest from the electrophilic carbon by a methyl group. However, no further extensive search was performed and there is no guarantee that the optimized structures

Table 3.1: Optimized geometry parameters for the stationary points of the SN2 reactions. The structures were optimized at the MP2/[aug]-cc-pVTZ(d/p) level (see text). Distances are in Å and angles in degrees (see Fig. 3.6 for the definition of the coordinates).

| | $R_1$ | $R_2$ | $A_1$ | $A_2$ | $D_1$[a] | $D_1$[b] | $D_1$[c] |
|---|---|---|---|---|---|---|---|
| | | | Complex | | | | |
| ethylchloride | 1.834 | 3.326 | 110.7 | 84.2 | 0.0 | 0.0 | 0.0 |
| 1-propylchloride | 1.833 | 3.369 | 111.5 | 80.6 | -85.7 | 0.0 | 0.0 |
| 1-butylchloride | 1.831 | 3.424 | 111.6 | 76.8 | -81.1 | — | — |
| | | | Transition state | | | | |
| ethylchloride | 2.355 | 2.330 | 94.9 | 100.0 | 2.3 | 0.0 | 2.9 |
| 1-propylchloride | 2.341 | 2.342 | 97.2 | 97.1 | -87.8 | -90.0 | -88.2 |
| 1-butylchloride | 2.341 | 2.340 | 97.1 | 97.0 | -87.9 | — | — |

[a] This work.
[b] Ref. [70]
[c] Ref. [71]

correspond to the global minima.

Single point MP2 and LMP2 calculations were carried out on the optimized structures, using the aug-cc-pVXZ basis sets (X=D,T,Q).[75, 76] The basis will be refered to as AVXZ for short. Due to the large number of basis functions, density fitting (DF) approximations[77] were used throughout. The results are presented in Tables 3.2 and 3.3. In order to study the effect of the local approximations on the relative energies, the LMP2 calculations were performed with three different choices of domains. In the first "standard" case, the domains were determined by the method of Boughton and Pulay[9] at the individual structures. In the second case (denoted merged(A)), the standard domains determined for the reactants and at the transition state were merged. This affects only the lone pairs of the incoming $Cl^-$ ion, yielding four equivalent domains. In the third case (denoted merged(B)) also the products were included in the merge procedure, yielding the same 8 equivalent domains as used in Fig. 3.5. This is relevant to see the effect of using different points for the domain merging.

Table 3.2 shows that in this case the effect of the merging on the computed barrier heights is small for all three systems. The standard and merged results differ by at most 0.2 kcal/mol. For the AVTZ and AVQZ basis sets the agreement with the canonical MP2

Table 3.2: MP2 and LMP2 barrier heights (in kcal mol$^{-1}$) for the SN2 reactions.

| | | DF-LMP2 | | | |
|---|---|---|---|---|---|
| Reactant | Basis[a] | Standard | Merged(A)[b] | Merged(B)[c] | DF-MP2 |
| $C_2H_5Cl$ | AVDZ | 18.2 | 18.1 | 18.3 | 17.6 |
| | AVTZ | 18.5 | 18.5 | 18.5 | 18.5 |
| | AVQZ | 18.8 | 18.8 | 18.9 | 18.9 |
| $C_3H_7Cl$ | AVDZ | 17.3 | 17.2 | 17.3 | 16.6 |
| | AVTZ | 17.9 | 17.8 | 17.9 | 17.8 |
| | AVQZ | 18.2 | 18.1 | 18.2 | 18.2 |
| $C_4H_9Cl$ | AVDZ | 17.6 | 17.4 | 17.4 | 16.9 |
| | AVTZ | 18.1 | 18.1 | 18.1 | 18.1 |
| | AVQZ | 18.4 | 18.4 | 18.5 | 18.6 |

[a] AV$X$Z denotes the aug-cc-pV$X$Z basis sets[76]
[b] Case (A): the domains of reactant and TS are merged.
[c] Case (B): the domains of reactant, TS, and product are merged.

results is excellent. However, in the case with the smaller AVDZ basis set the barrier height is lower in the canonical calculation. Very likely, this effect is due to an artificial lowering of the canonical MP2 barrier by BSSE effects. This is supported by the fact that the convergence with increasing basis set size is faster for LMP2 than for MP2.

The effect of the BSSE can be directly shown for the complex by applying the counterpoise correction (CP).[78] Table 3.3 shows the CP-corrected binding energies. As demonstrated previously for many other cases[32, 33, 34, 35, 36, 37, 66], the CP corrections (in parenthesis) are much smaller for LMP2 than for MP2, and the CP-uncorrected LMP2 values (obtained by subtracting the CP correction) are in excellent agreement with the CP-corrected MP2 values. The binding energies obtained with standard domains are always slightly too small, consistent with previous experience[34, 66]. This is due to the missing class of ionic excitations, which are neglected by construction in the local calculations. When the domains are extended by the merging procedure, the most important contributions of these excitations are included, and the CP corrections slightly increase. This leads to even better agreement between the CP-uncorrected LMP2 and the CP-corrected MP2 binding energies. Due to the remaining BSSE, the CP-uncorrected LMP2 binding energies

Table 3.3: MP2 and LMP2 binding energies (in kcal mol$^{-1}$) of the complexes in the entrance channels of the SN2 reactions. All values are counterpoise corrected, and the BSSE correction is given in parenthesis.

| | | DF-LMP2 | | | |
|---|---|---|---|---|---|
| Reactant | Basis[a] | Standard | Merged(A)[b] | Merged(B)[c] | DF-MP2 |
| $C_2H_5Cl$ | AVDZ | -11.2 (+0.3) | -11.2 (+0.5) | -11.6 (+0.6) | -11.5 (+1.2) |
| | AVTZ | -11.6 (+0.0) | -11.6 (+0.2) | -11.7 (+0.2) | -11.7 (+0.7) |
| | AVQZ | -11.8 (+0.0) | -11.8 (+0.1) | -11.9 (+0.1) | -11.9 (+0.4) |
| $C_3H_7Cl$ | AVDZ | -11.6 (+0.3) | -11.7 (+0.5) | -11.9 (+0.7) | -12.1 (+1.4) |
| | AVTZ | -12.2 (+0.1) | -12.3 (+0.2) | -12.3 (+0.3) | -12.5 (+0.8) |
| | AVQZ | -12.5 (+0.0) | -12.5 (+0.1) | -12.6 (+0.1) | -12.8 (+0.6) |
| $C_4H_9Cl$ | AVDZ | -12.2 (+0.3) | -12.3 (+0.5) | -12.4 (+0.7) | -12.7 (+1.6) |
| | AVTZ | -12.8 (+0.1) | -12.9 (+0.2) | -13.0 (+0.3) | -13.3 (+0.9) |
| | AVQZ | -13.2 (+0.0) | -13.2 (+0.1) | -13.3 (+0.1) | -13.4 (+0.5) |

[a] AV$X$Z denotes the aug-cc-pV$X$Z basis sets[76]
[b] Case (A): the domains of reactant and TS are merged.
[c] Case (B): the domains of reactant, TS, and product are merged.

are in some cases slightly larger than the CP corrected MP2 ones.

### 3.3.3 Hydrogen fluoride addition to double bonds

The addition of hydrogen fluoride (FH)[1] to ethene is a frequently used model system for addition reactions involving unsaturated hydrocarbons. It has been subject of several theoretical works, some of the references are contained in Ref. [79]. The addition follows a concerted mechanism in the gas phase as depicted in Fig. 3.7. The reaction proceeds through the formation of a weak van der Waals complex (RC), leading to a four centered transition state (TS1). Fluoroethane is then formed, but at the eclipsed conformation (TS2) which by rotation leads to the final product (P).

The reaction is a good test candidate for the proposed merge procedure. The system evolves from a double bond to a 4-center bond between FH and ethene, finally leading to a saturated system. Also of interest is the last step, where a methyl rotation takes place. As discussed in Section 3.2.1, the merging procedure depends on a 1-to-1 correspondence

[1] The abbreviation FH will be used throughout the text to avoid possible confusion with Hartree-Fock.

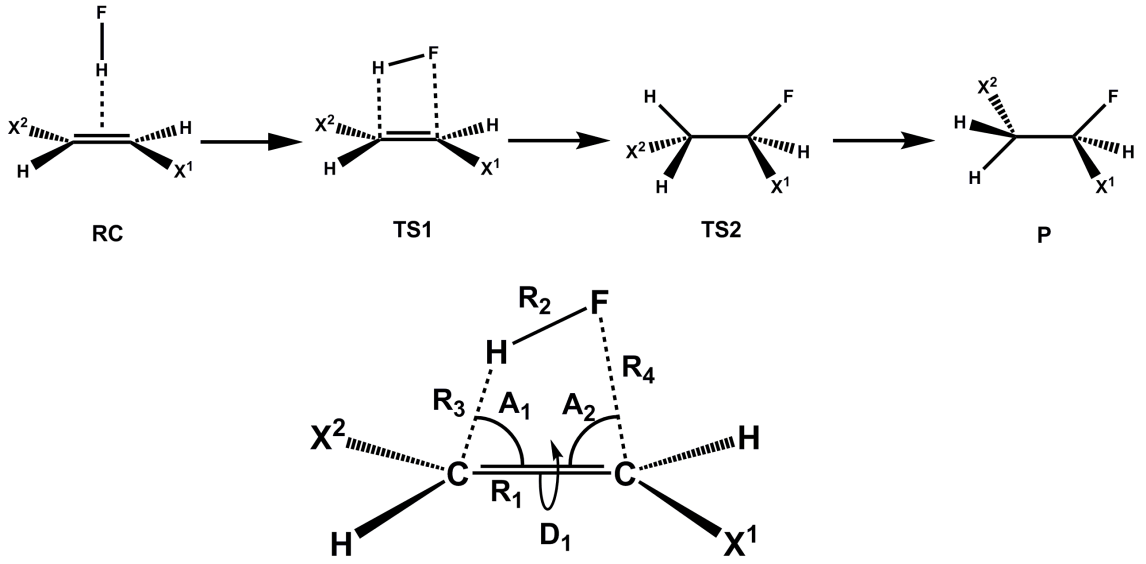



Figure 3.7: Schematic representation of four stationary points of the hydrogen fluoride addition reactions - ethene ($X_1$=$X_2$=H), 1-propene ($X_1$=$CH_3$,$X_2$=H) and 2-butene ($X_1$=$X_2$=$CH_3$). **RC** stands for the van der Waals reactant complex, **TS1** and **TS2** the first and second transition states respectively, and the product state **P**. On the bottom, a representation of **TS1** with a description of the geometric parameters of Table 3.4. $D_1$ stands for the dihedral angle between the entering hydrogen, the carbons and the fluorine.

between orbitals in different geometries. For this purpose, both overlap and domain criteria were used in building the pairing list. The overlap criterion fails in this case and it will be interesting to see how well the pairing algorithm works.

Besides ethene, also the reactions with 1-propene and 2-butene were considered. Just as in the SN2 case, these are systems of increasing size built by substitution of a hydrogen by a methyl group, this time in the vicinity of the double bond. The same mechanism was adopted for the three systems. For ethene and 2-buthene there can only be one final product. For propene there are two possibilities: 1-fluoropropyl and 2-fluoropropyl. The latter product was considered. In all three cases the four stationary points in the PES were found, and the same naming procedure is used as in the work of Cremer *et al.* (see Fig. 3.7 and Table 3.4).[79]

No reaction paths were calculated, only relative energies between the minima/maxima. The reactant domains (computed at large distances) were merged with the TS1 domains. Due to the large domains found for the transition state structures, taking the product domains as reference would be insufficient. The merging procedures leads to relatively small changes. One of the fluorine lone pair domains is augmented to include one of the carbons PAOs, and the entering hydrogen PAOs are added to one of the double bonding carbon domains. Single point calculations were also carried out for MP2 and regular LMP2 calculations, using the cc-pV$X$Z basis sets ($X$=D, T or Q)[75]. Full results are displayed in

Table 3.5.

The results agree with previous findings. The merged values lie between the local and canonical results, or at least quite close. All values converge with increasing basis set size, and the increasing molecular size does not affect this trend. The largest error is found for the $\Delta E_{TS1}$ value, which lies around 50 kcal/mol (and as such, the error is only of the order of 1% in the relative energy). One last notice should be made about the cc-pVQZ value for the butene $\Delta E_{TS1}$, for which the local result still lies about 0.8 kcal/mol above the canonical one. A slow basis set convergence for the correlation energy is observed in this case, probably due to BSSE. Counterpoise corrections carried on the complex structure gave a BSSE of 0.89 kcal/mol, confirming the assumption.

The values for $\Delta E_R$ are all in good agreement, which provides evidence for the robustness of the pairing algorithm. Although the methyl group is strongly rotated, the program still identifies the correct orbital. Ordering the orbitals based on the overlap criterion would lead to an incorrect pairing, and to a large error in the correlation energy.

Table 3.4: Relevant geometric parameters for the stationary points found for the hydrogen fluoride addition to a double bond reactions. Structures were optimized at the MP2/[aug]-cc-pVTZ(d/p) level (diffuse functions added to fluorine). Distances are in Å and angles in degrees (see Fig 3.7 for support).

| | Complex | | | | | | |
|---|---|---|---|---|---|---|---|
| | R1 | R2 | R3 | | | | |
| ethene + HF | 1.337 | 0.932 | 2.228 | | | | |
| propene + HF | 1.340 | 0.934 | 2.164 | | | | |
| butene + HF | 1.342 | 0.936 | 2.145 | | | | |
| | Transition state 1 | | | | | | |
| | R1 | R2 | R3 | R4 | A1 | A2 | D1 |
| ethene + HF | 1.399 | 1.309 | 1.317 | 1.904 | 73.1 | 93.8 | 0.0 |
| propene + HF | 1.403 | 1.329 | 1.292 | 1.958 | 74.7 | 91.7 | 1.4 |
| butene + HF | 1.404 | 1.363 | 1.277 | 1.971 | 73.6 | 92.7 | 3.9 |
| | Transition state 2 | | | | | | |
| | R1 | | R3 | R4 | A1 | A2 | D1 |
| ethene + HF | 1.526 | | 1.086 | 1.404 | 108.6 | 110.0 | 0.0 |
| propene + HF | 1.528 | | 1.086 | 1.413 | 108.9 | 108.6 | 2.9 |
| butene + HF | 1.531 | | 1.088 | 1.413 | 107.3 | 109.0 | 4.8 |
| | Product | | | | | | |
| | R1 | | R3 | R4 | A1 | A2 | D1 |
| ethene + HF | 1.509 | | 1.088 | 1.402 | 109.6 | 109.6 | 180.0 |
| propene + HF | 1.512 | | 1.088 | 1.411 | 110.0 | 107.9 | -177.4 |
| butene + HF | 1.515 | | 1.090 | 1.412 | 107.8 | 108.2 | 177.5 |

Table 3.5: Energies of the FH addition reaction stationary points relative to the reactant complex **RC** (in kcal mol$^{-1}$). These results are obtained from single points on the MP2/[aug]-cc-pVTZ(d/p) (added diffuse functions to the fluorine) optimized structures.

| | | $\Delta E_{TS1}$ | | | $\Delta E_{TS2}$ | | | $\Delta E_P$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | DF-LMP2 | | | DF-LMP2 | | | DF-LMP2 | | |
| Reactant | Basis | Standard | Merged | DF-MP2 | Standard | Merged | DF-MP2 | Standard | Merged | DF-MP2 |
| ethene | cc-pVDZ | 54.3 | 54.4 | 53.5 | -8.7 | -9.4 | -10.0 | -12.3 | -13.1 | -13.8 |
| | cc-pVTZ | 53.3 | 53.4 | 52.9 | -8.0 | -8.5 | -8.5 | -11.3 | -11.9 | -12.0 |
| | cc-pVQZ | 52.0 | 52.1 | 51.9 | -7.8 | -8.0 | -8.0 | -11.1 | -11.4 | -11.4 |
| propene | cc-pVDZ | 51.8 | 51.9 | 50.9 | -9.4 | -10.1 | -10.8 | -12.7 | -13.5 | -14.3 |
| | cc-pVTZ | 50.9 | 51.0 | 50.3 | -8.5 | -9.0 | -9.1 | -11.7 | -12.3 | -12.5 |
| | cc-pVQZ | 49.4 | 49.5 | 49.2 | -8.4 | -8.6 | -8.6 | -11.6 | -11.9 | -11.9 |
| butene | cc-pVDZ | 55.3 | 55.8 | 54.5 | -6.2 | -6.2 | -7.4 | -10.0 | -10.1 | -11.4 |
| | cc-pVTZ | 54.1 | 54.3 | 53.7 | -5.2 | -5.7 | -5.7 | -8.8 | -9.2 | -9.4 |
| | cc-pVQZ | 53.2 | 53.0 | 52.4 | -5.1 | -5.3 | -5.2 | -8.6 | -8.8 | -8.8 |

# Chapter 4

# Natural Localized Molecular Orbitals for Local Correlation Schemes

## 4.1 Critical Assessment of the Boughton-Pulay Criteria

Since the seminal works by Saebø and Pulay,[1, 28] the fundaments of local correlation methods have been kept practically unchanged. There has been significant investment in expanding the range and efficiency of local methods available, and in the pair approximations used, but little on the localization procedure on the basis of the procedure. Also the domain selection criterion, proposed in 1993 by Boughton and Pulay,[9] has remained untouched.

The choice of method for localizing the occupied space is usually of little importance. Pipek-Mezey (PM) localization[11] is preferred in most cases, since it keeps the $\pi$-$\sigma$ separation in planar molecules, but any set of local orbitals could be used. It has been found that the correlation energy is rather insensitive to the localization method. However, the PM procedure is known to be sensible to the use of diffuse basis sets. Since it is based on the AO overlap matrix, near linear dependencies may lead to artificially large LMO coefficients at some atoms. This frequently happens, for instance, in aromatic compounds with basis sets such as aug-cc-pVTZ or aug-cc-pVQZ. Using the Boughton-Pulay (BP) method one then finds unphysical large domains for the $\pi$-orbitals, which include the neighboring H-atoms. One solution to this problem is to remove the most diffuse basis function of each angular momentum type for each atom in the localization criterion. However, this solution is also not straightforward, since again it depends on the basis set used. For larger basis sets or with even more diffuse functions it might be necessary to remove extra functions.

Other problems have been pointed out at the domain selection criterion (detailed in Section 2.1.4). In the BP procedure one computes the overlap of the LMO with a trial function built as linear combination of AOs belonging to the domain centers, adding centers to the list till this value exceeds the BP criterion $T_{\mathrm{BP}}$ (see Eq. (2.53)). The value to which this parameter should be set is, however, unclear, as it also suffers from basis set dependence. Usually, the criterion is more easily fullfilled for larger basis sets, and therefore for a fixed value of the threshold $T_{\mathrm{BP}}$ the domains will become smaller with increasing basis set. To compensate for this, one can use different thresholds for different basis sets, e.g. 0.98 for double zeta, 0.985 for triple zeta, and 0.99 for quadruple zeta. But apart from the fact that this is not a well defined and user-friendly model, the domains often still differ for different basis sets.

Another critical point of the BP criterion is the use of Löwdin populations for ordering the atoms. This population analysis does not converge when going to larger basis sets. In some cases physically unreasonable results can be expected. For example, the partial atomic charge in the methane carbon is predicted to be positive when using the aug-cc-

pVTZ basis set (+0.05 a.u.).[80] The variations can even be quite large when going from a double zeta to a triple zeta basis, changing as much as 0.5 a.u. (see again Ref. [80]). This of course affects the reliability of the whole procedure.

All these problems are an obstacle to the generalized use of local methods, since the user is forced to control too many aspects of the calculation. Besides the BP criterion, many other parameters are used to give stable domains. Limits on the minimum charge population of each atom (depending on whether it is a hydrogen or a heavy atom), as well as parameters for automatically adding atoms with charges above a certain value are today in use. Much of these defaults can be safely used, but exceptions have also been identified. In fact, it is generally advisable to check the orbital domains before carrying out a local correlation calculation. There is an obvious need for a truly automated procedure for localization and domain selection.

## 4.2 Natural Localized Molecular Orbitals

The *Natural Atomic Orbitals* (NAOs) and related population analysis were introduced by Weinholdt and coworkers in 1985.[13] They are defined as the eigenfunctions of the density (see Eq. (2.58)). The matrix representation of the density is defined as

$$\begin{aligned}\Gamma_{\rho\sigma} =& < \chi_\rho|\rho|\chi_\sigma > = \sum_{\mu\nu} < \chi_\rho|\chi_\mu > D_{\mu\nu} < \chi_\nu|\chi_\sigma > \\ &= [\mathbf{SDS}]_{\rho\sigma}. \end{aligned} \tag{4.1}$$

The procedure is started by dividing the matrix $\mathbf{\Gamma}$ into one-center blocks

$$\begin{pmatrix} \mathbf{\Gamma}^{(AA)} & \mathbf{\Gamma}^{(AB)} & \mathbf{\Gamma}^{(AC)} & \dots \\ \mathbf{\Gamma}^{(BA)} & \mathbf{\Gamma}^{(BB)} & \mathbf{\Gamma}^{(BC)} & \dots \\ \mathbf{\Gamma}^{(CA)} & \mathbf{\Gamma}^{(CB)} & \mathbf{\Gamma}^{(CC)} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where for $\mathbf{\Gamma}^{(\mathrm{AB})}_{\mu\nu}$, $\mu \in \{A\}$ and $\nu \in \{B\}$. The AO overlap matrix $\mathbf{S}$ is partitioned in the same way. For each diagonal sub-matrix the generalized eigenvalue problem

$$\mathbf{\Gamma}^{(\mathbf{AA})}\mathbf{X} = \mathbf{S}^{(\mathbf{AA})}\mathbf{XW}, \tag{4.2}$$

where $\mathbf{W}$ is a diagonal matrix holding the eigenvalues, is solved. The eigenvectors $\mathbf{X}$ are the pre-NAOs, and their occupancy is given by the respective eigenvalues $\mathbf{W}$. These orbitals are divided into two sets: the *Natural Minimal Basis* (NMB) and the *Natural Rydberg Basis* (NRB). The division is made by taking the first $N$ orbitals into the NMB set, where $N$ is the number of valence orbitals needed for the representative ground state configuration of the neutral atoms. The orbitals are then orthogonalized among each other, while maintaining the $\mathbf{\Gamma}$ matrix block diagonalization. The steps are as follows:

(1) The NRB orbitals are Schmidt-orthogonalized relative to the NMB set,

(2) the eigenvalue problem of Eq. (4.2) is again solved, this time for the density and overlap matrices in the NRB basis,

(3) the NMB and NRB orbitals of the different centers are orthogonalized using an occupancy-weighted orthogonalization scheme (see Ref. [13]),

(4) the eigenvalue problem is again solved, this time in the basis of the whole orthogonal orbital set.

The diagonalization and orthogonalization matrices multiplied together give the $\mathbf{T}^{\mathrm{NAO}}$ matrix. Transformation of $\mathbf{\Gamma}$ into the NAO basis

$$\tilde{\mathbf{D}} = (\mathbf{T}^{\mathrm{NAO}})^{\dagger}\mathbf{\Gamma}\mathbf{T}^{\mathrm{NAO}}, \tag{4.3}$$

gives a block-diagonalized matrix, and its diagonal elements $\tilde{D}_{rr}$ are the final occupation numbers for the NAO orbital with index $r$. In this way, one can divide the charge among the atoms as

$$P_A = \sum_{r\in\{A\}} \tilde{D}_{rr}. \tag{4.4}$$

This is the so called *Natural Population Analysis* (NPA).[13]

The *Natural Bond Orbitals* (NBOs) are built by diagonalizing one and two-center blocks of the NAO density matrix. The procedure is as follows:

(1) All NAOs with eigenvalues above a given threshold are added to the NBO list as lone-pairs and all lone pair contributions to the density matrix are removed.

(2) The two-center blocks of the NAO density matrix are diagonalized. Again, all orbitals with eigenvalues above the threshold are added to the NBO list. These NBOs are refered to as 2-center bond orbitals.

(3) If the number of NBOs found in this way is equal to the number of electron pairs, the search is stopped and one may proceed to the next step. If not, the threshold is decremented and step 2 is repeated. It would also be possible to expand the search to 3-center bonds, but we found that for the systems included in this study (and in general organic compounds) this was not necessary.

(4) The remaining orbital space (of low occupation) is divided into Rydberg and anti-bonding orbitals. More details can be found in Ref. [81], or in Appendix A.

The NBO orbitals are by construction orthogonal, but should not be used directly in post-SCF calculations since the occupied NBO orbitals do not span the SCF occupied space exactly. Therefore, a final $\mathbf{T}^{\mathrm{NLMO}}$ transformation is performed which rotates the NBO orbitals so that the orbitals with highest occupations (the so called Lewis set) span the SCF valence space. This is done by 2x2 Jacobi rotations which zero the density matrix elements between the Lewis and non-Lewis spaces. For closed-shell SCF wave functions this makes all diagonal elements $D_{ii} = 2$ and all other elements zero. Since the diagonal elements of the NBO density matrix of the Lewis space are already quite close to 2 one only needs a limited number of 2x2 Jacobi rotations and the orbital space stays localized. The procedure is further detailed in Appendix A. In summary, the final *Natural Localized Molecular Orbitals* (NLMO) coefficients are obtained as

$$\mathbf{L} = \mathbf{T}^{\mathrm{NAO}}\mathbf{T}^{\mathrm{NBO}}\mathbf{T}^{\mathrm{NLMO}} = \mathbf{T}^{\mathrm{NAO}}\mathbf{V}. \tag{4.5}$$

which corresponds to a connected series of transformations starting from the AO basis

$$\mathrm{AO} \xrightarrow{\mathbf{T}^{\mathrm{NAO}}} \mathrm{NAO} \xrightarrow{\mathbf{T}^{\mathrm{NBO}}} \mathrm{NBO} \xrightarrow{\mathbf{T}^{\mathrm{NLMO}}} \mathrm{NLMO}. \tag{4.6}$$

All of these sets have been used in the past for analyzing self-consistent wave functions. The NAOs are the basis for the above mentioned NPA analysis, which is free from many of the deficiencies found in Mulliken of Löwdin populations. Of special interest is its stability with respect to the choice of basis set. This can be easily explained by reviewing the way the NAOs are built. After solving Eq. (4.2) for $\mathbf{\Gamma}$, the orbitals are divided into the NMB and NRB sets. All subsequent operations strive to maintain the form of the NMB set (where the majority of the electron population is kept) at the cost of the NRB orbitals. By increasing the basis set, low-populated diffuse functions will be tagged to the latter set and will therefore have little effect on the NPA populations. The NPA derived atomic charges have been shown to converge with respect to basis set size, and also to deliver results which agree well with experimental evidence and/or chemical sense.

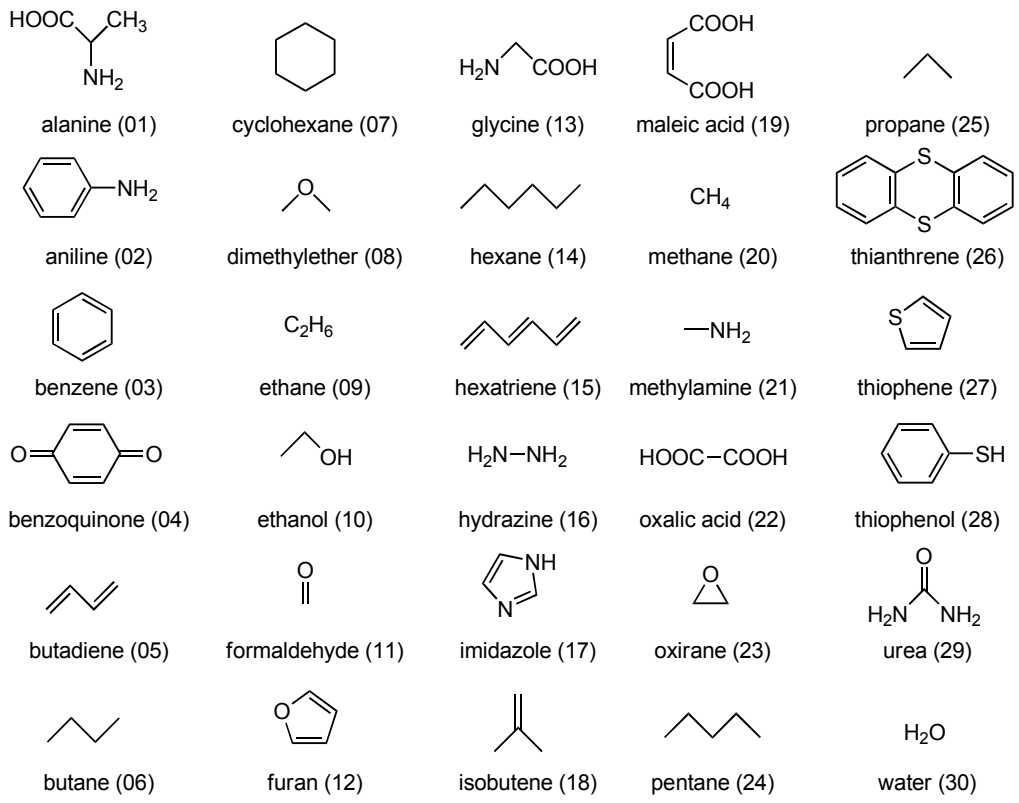



Figure 4.1: Test set of 30 molecules used in this work. All geometries were pre-optimized with B3LYP/cc-pVTZ(d/p) (including up to d functions for the second and third row elements, and up to p functions for the hydrogens). The numbering shown is the same used in some of the diagrams.

The NBO orbitals are an extension to the NAO procedure, whereby also 2-center blocks are diagonalized, giving a Lewis-like description of bonding. Several analysis procedures have been proposed over the years, but this will not be discussed in the text. A review can be found in Ref. [82].

The NLMO orbitals, the last ones in the set, can be used in local correlation treatments. They have in the past compared favorably to Boys[30] and Edmiston-Ruedenberg[31]. Measures of LMO charge centroid distances relative to the atomic nuclei, as well as direct comparison of the orbital coefficients showed similar results for all three localization methods. However, their biggest advantage is the possibility of using the NPA analysis to calculate the charge of each NLMO into charges of centers. If stable enough, a single parameter could be used to control the domain sizes, replacing the somewhat intricate use

of the default PM/BP procedure. For comparing the use of NLMO and PM orbitals in local correlation treatments, a test set comprising 30 molecules was chosen. They are depicted in Fig. 4.1. Included in this list are typical small organic molecules, medium-sized saturated and unsaturated hydrocarbons as well as aromatic systems. Two sets of calculations were run. In both sets, a DF-LMP2/cc-pVTZ calculation was carried out using the BP criterion with a value of $T_{\mathtt{BP}} = 0.985$. In the first set PM localized orbitals were used and in the second NLMOs. The results are shown in Fig. 4.2, where both the percentage of correlation energy recovered (left scale) as well as the difference of the average domain sizes $L_{\mathtt{NLMO}} - L_{\mathtt{PM}}$ (right scale) is plotted. As can be seen in the diagram, this difference is always positive, except for the benzoquinone molecule, meaning that the NLMO domains are slightly larger. There are, however, no major differences between the two. The correlation energy recovered with both orbital sets is very similar, and the differences are mainly due to the different domain sizes. In the cases where the domains are the same, the energies are almost identical, which supports the conclusion that the localization method has very little effect on the energy. These results indicate that replacing Pipek-Mezey by NLMO orbitals should have little effect on the accuracy of local correlation methods, as seen with other localization procedures.[9]

The next Section discusses different ways of partitioning the orbital charge through the centers, making use of the NPA population, and a new domain criterion is proposed.

## 4.3 Natural Population Domain Criterion

### 4.3.1 Orbitals Population

No unique way has been described so far to divide the charge of each NLMO into charges of centers. Löwdin or Mulliken could be used but, as discussed before, they are unreliable. The alternative is the use of NPA charges and/or the coefficients of the only natural orbital set which is uniquely tagged to the atoms - the NAOs. The NBOs contain 2-center orbitals, which do not differentiate between the two centers involved (one could divide the orbital contribution or consider the electronegativity of the atoms involved, but this approach would be somewhat empirical).

In order to determine where the NLMOs are located, one makes use of the $\mathbf{V}$ transformation from NAO to NLMOs and the NAO Density Matrix $\tilde{\mathbf{D}}$. The charge of NLMO $i$ at the center $A$ will be refered to as $P_{Ai}$. There are various possible ways to determine this value:

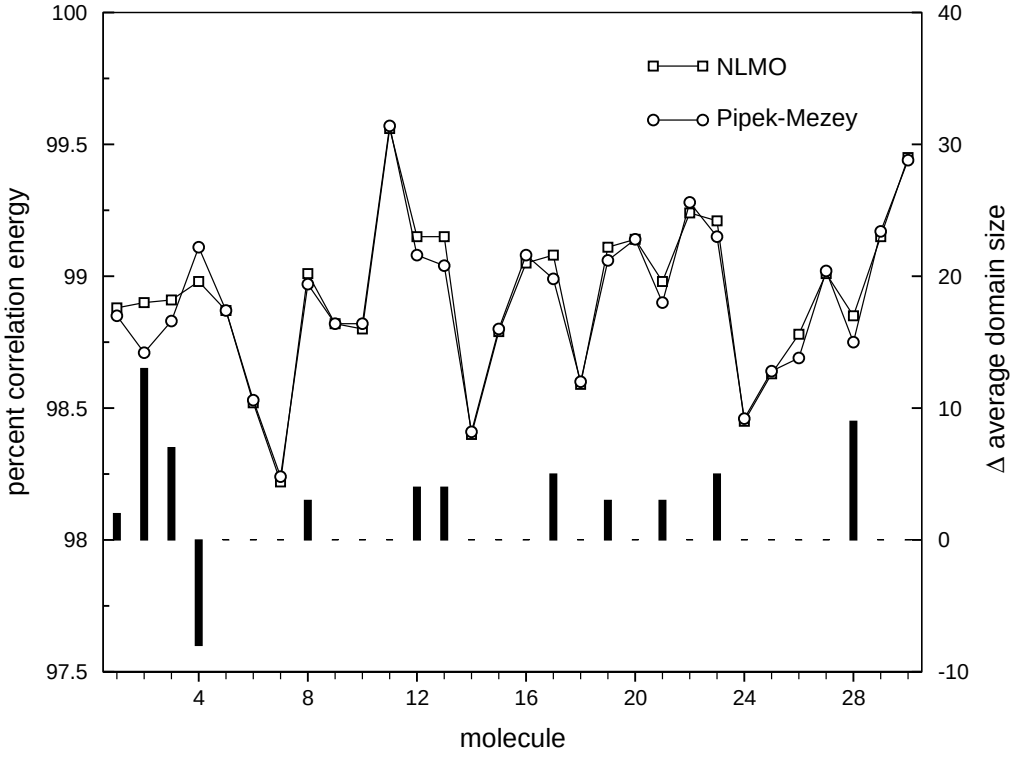



Figure 4.2: Percentage of correlation energy recovered using NLMO and Pipek-Mezey orbitals in LMP2 calculations, and with $T_{\mathrm{BP}} = 0.985$. Also shown (in bars) is the average domain size difference between the two sets. The basis set used was cc-pVTZ. The molecule numbers refer to the ones used in Fig. 4.1.

(1) Use only the transformation coefficients, as in the NBO program[83]

$$P_{Ai} = 2 \sum_{r \in [A]} V_{ri}^2, \tag{4.7}$$

since the NLMOs are normalized, $0 < P_{Ai} < 2$ is valid.

(2) Use the transformation coefficients, with the NAO Density Matrix elements as weighting coefficients

$$P_{Ai} = 2 \frac{\sum_{r \in [A]} V_{ri}^2 \tilde{D}_{rr}}{\sum_r V_{ri}^2 \tilde{D}_{rr}}. \tag{4.8}$$

The sum in the denominator runs through all NAOs.

(3) Use the transformation coefficients to divide the charge through the occupied orbitals

$$P_{Ai} = \sum_{r \in \{A\}} \left( \frac{V_{ri}^2}{\sum_j V_{rj}^2} \right) \tilde{D}_{rr} \tag{4.9}$$

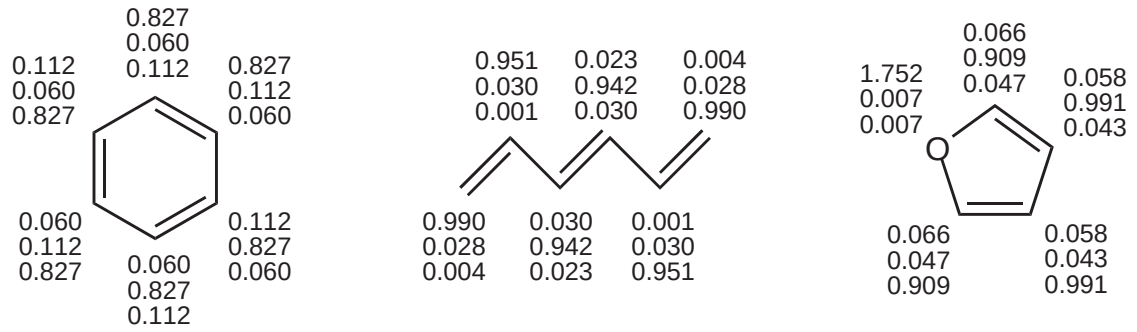



Figure 4.3: NPA calculated $\pi$-orbitals charges for NLMO orbitals. The basis set used was aug-cc-pVQZ. The values are given for each heavy atom, with the orbitals ordered vertically.

The first option is in fact a percentage, giving the weight of each NAO. No use is made of the NPA charges, so that low-occupancy NAOs have the same impact as the ones with higher-occupancy. This choice would lead to an overestimate of the NLMO population on nearby atoms which were included due to the orthogonalization tails, and this effect is undesirable. Option (2) already includes the NPA information, but is not consistent with the individual atomic populations. Consider for example a $CH_4$ molecule. There will be four valence NLMOs, one for each C-H bond. They will be almost perfectly localized, so that each hydrogen atom population could in fact be extracted either by the NPA analysis, or by using the electron charge fraction indicated by the NLMO population. The two values will not coincide, since some of the charge is given to the non-occupied NLMOs. Only by use of Eq. (4.9) are both analyses in agreement. An example for the charge partitioning can be seen in Fig. 4.3. The charges are given for the $\pi$-orbitals of benzene, hexatriene and furan.

### 4.3.2 NPA-based Domain Criterion

It would be desirable to replace the BP procedure by a stable criterion based on the charges obtained using the NPA procedure and Eq. (4.9). There are two approaches for such a criterion. One may define a minimum charge for the LMO domain to be considered "filled". Centers would be added to the domain list in the order of decreasing charges, until $\sum_{A\in[i]} P_{Ai}$ exceeds the threshold. Another possibility is to add all centers to a domain with charge above a given limit. Both approaches should lead to similar results, but the latter alternative was found to be more stable with respect to basis set size, especially in aromatic systems. The new parameter will be refered to as $T_{\mathtt{NPA}}$. For a given LMO $\phi_i$, all atoms for which $P_{Ai} > T_{\mathtt{NPA}}$ are added to the domain list $[i]$.

The value of $T_{\mathtt{NPA}}$ is the only parameter needed for determining the domains. Which value it should take remains, however, an open question. Observing Fig. 4.3, it is clear that it shouldn't be much above 0.10 a.u., otherwise the $\pi$-orbitals of benzene would be double-centric bonds. A too low value is also not advisable, since it would make the selection unstable (for low values the populations start to form a continuum). Values below 0.01 a.u. should be avoided. In order to define the best $T_{\mathtt{NPA}}$ value, further tests were made, namely on reaction energies. In this text, only absolute energies and the qualitative features of the domain selection will be discussed.

I would like to end this Section by adding a further comment on the procedure. The new criterion is extremely reliable in defining domains for $\pi$ orbitals in aromatic systems. By adjusting the value of $T_{\mathtt{NPA}}$, merged $\pi$ domains can be obtained. Values between 0.03-0.01 a.u. are advisable, and could replace the use of merging procedures.[10]

## 4.4 Comparison to Boughton-Pulay

A series of tests were performed to compare the performance of the new domain selection scheme. The combination of NLMO orbitals as ocuppied space and the NPA-based criterion will be from now on denoted as NLMO/NPA. In the tests, the 30 molecule set depicted in Fig. 4.1 was used. All calculations were carried out with the density-fitting variants of HF, MP2 and LMP2.[77, 84] The correlation consistent basis sets of Dunning and co-workers, cc-pV$X$Z [75] and aug-cc-pV$X$Z [76] (with $X$=D, T and Q) were used.

### 4.4.1 Domain Convergence with respect to Basis Set

The variation of the domains with basis set is measured by a parameter $\Delta = \sum_i \Delta_i$, where $\Delta_i$ is the number of non-coinciding atoms in the orbital domain $i$, relative to the domains obtained with the cc-pVDZ basis set. For example, if for two different basis sets the domains of a particular orbital are $C_1$, $C_2$, $H_1$ and $C_1$, $C_2$, $\Delta_i = 1$, while for $C_1$, $C_2$, $H_1$ and $C_1$, $C_2$, $H_2$ $\Delta_i = 2$. Fig. 4.4 shows the variation of $\Delta$ for the largest molecule in the set, thianthrene. Two sets of calculations were carried out with the Pipek-Mezey orbitals. In the first case, fixed parameters were used for the domain selection ($T_{\mathtt{BP}} = 0.985$) and localization for all basis sets. In a second series of calculations, the parameters were changed as a function of the basis set. The BP criterion was set to 0.980 for the double-zeta basis sets, 0.985 for triple and 0.990 for quadruple-zeta. For the augmented basis sets, the contribution of the most diffuse basis function of each angular momentum type for each atom was eliminated in the localization criterion. In the case of aug-cc-pVQZ, the last two were eliminated.

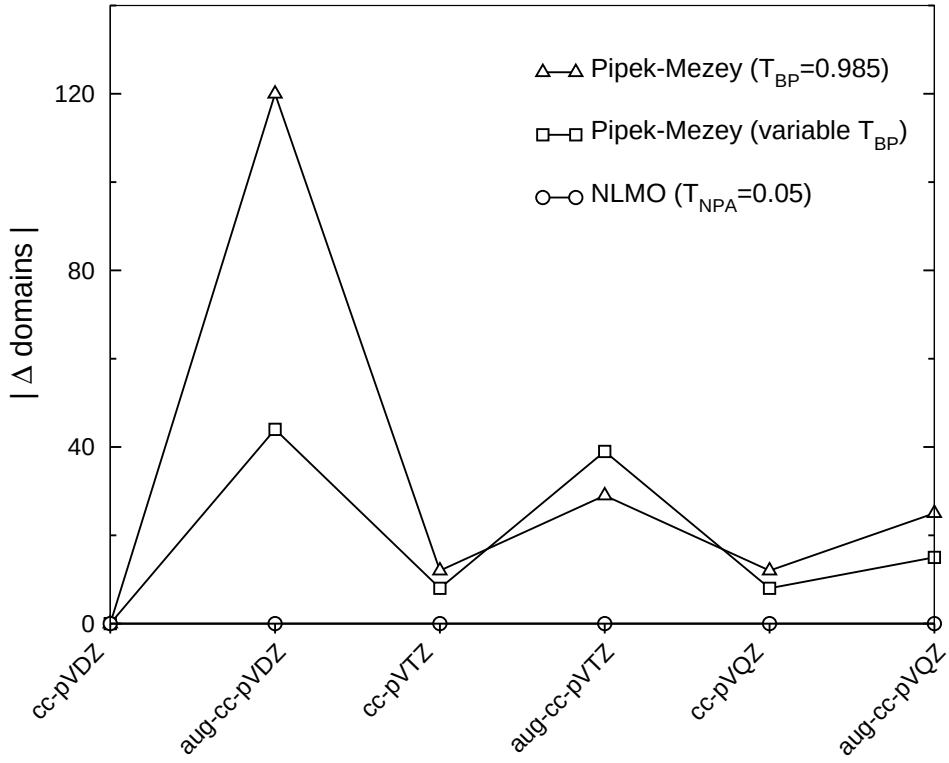



Figure 4.4: Sum of the domain variations (absolute) for the thianthrene molecule with different localization procedures and domain selection test. For the variable $T_{\mathtt{BP}}$ case, different parameters for localization and domain selection were used (see text for more information).

These values are refered to as variable $T_{\mathtt{BP}}$ in Fig. 4.4. This procedure should decrease the basis set dependence. For the NLMO/NPA method, a single value $T_{\mathtt{NPA}} = 0.05$ was used.

The results in Fig. 4.4 show large fluctuations in the domains when using a fixed $T_{\mathtt{BP}}$ value. The use of different parameters for different basis sets helps to decrease these differences, but the changes in the domain lists when using diffuse functions is still measurable. Only in the case of NLMO/NPA is $\Delta = 0$ for all six basis sets, i.e., there is not a single domain change.

Similar tests were done for all 30 molecules using $T_{\mathtt{NPA}}$ values of 0.025, 0.05, and 0.10. The PM/BP domains change significantly as a function of the basis set, not only in the case of aromatic rings, but also in smaller molecules like dimethylether or oxirane. The use of diffuse functions generally leads to a steep increase in the domain sizes. The results are shown in Figs 4.6 and 4.5.

Contrary to the BP defined domains, the NPA-based criterion is extremely robust. For the 30 molecules depicted in Fig. 4.1 and using $T_{\mathtt{NPA}} = 0.05$, all domains were kept. For $T_{\mathtt{NPA}} = 0.025$, there is a difference between the double-zeta domains of the oxalic acid, and

the remaining sets. One of the carboxylic $\pi$-orbitals extend to a neighboring carbon for the larger basis sets. Since there are two carboxylic groups, $\Delta = 2$. The population change in the neighboring carbon is rather small (it changed from 0.023 to 0.026 a.u., cc-pVDZ and cc-pVTZ respectively), but enough to go above the threshold. For $T_{\mathtt{NPA}} = 0.100$, there is also a single change. For the aug-cc-pVTZ basis set, one of the oxygen lone pairs in glycine turns into a double bond, again due to population fluctuations in the order of $\pm 0.001$ a.u. Even with these exceptions, there is an enormous gain in the use of the NPA criterion.

### 4.4.2 Correlation Energies

Fig. 4.7 shows the fraction of correlation energy recovered relative to canonical MP2 using the PM/BP and NLMO/NPA methods. In the latter case, two different thresholds were used for comparison. The results are rather similar in all cases and differ mainly for the aromatic molecules due to the different sizes of the $\pi$-orbital domains. The largest fraction of correlation energy is recovered for the very small molecules water and formaldehyde, the smallest one for alkanes like pentane or cyclohexane. The surprising fact that these most saturated and well localized systems are most strongly affected by the domain approximation has been discussed before[10]. Most likely, this is related to the intramolecular basis set superposition error, which is expected to be largest for molecules in which many atoms have a tetrahedral environment. In the local methods, the BSSE is minimized by construction. Clearly, these variations can have a significant effect on reaction energies. One extreme case, the hydration of benzene to cyclohexane, has been studied in Ref. [10].

### 4.4.3 Local Gradients

An important disadvantage in the use of NLMOs as occupied space is that no single minimization criterion is available. This is relevant for the calculation of analytical gradients. The geometry dependence of the localized orbitals coefficient matrix $\mathbf{L}$ with respect to a nuclear displacement $\lambda$ is given by

$$\mathbf{L}(\lambda) = \mathbf{C}\mathbf{\Delta C}(\lambda)\mathbf{U}\mathbf{\Delta U}(\lambda). \tag{4.10}$$

The matrices $\mathbf{C}$ and $\mathbf{U}$ have already been defined in Eqs. (2.14) and (2.47). In this case, they refer respectively to the canonical orbital coefficients and localization matrices at the reference geometry ($\lambda = 0$). The geometry dependence is given by $\mathbf{\Delta C}(\lambda)$ and $\mathbf{\Delta U}(\lambda)$.

The contribution to the gradient from $\mathbf{\Delta C}(\lambda)$ is computed with help from the Coupled-perturbed Hartree-Fock equations, and bears no relation to the localization method used.

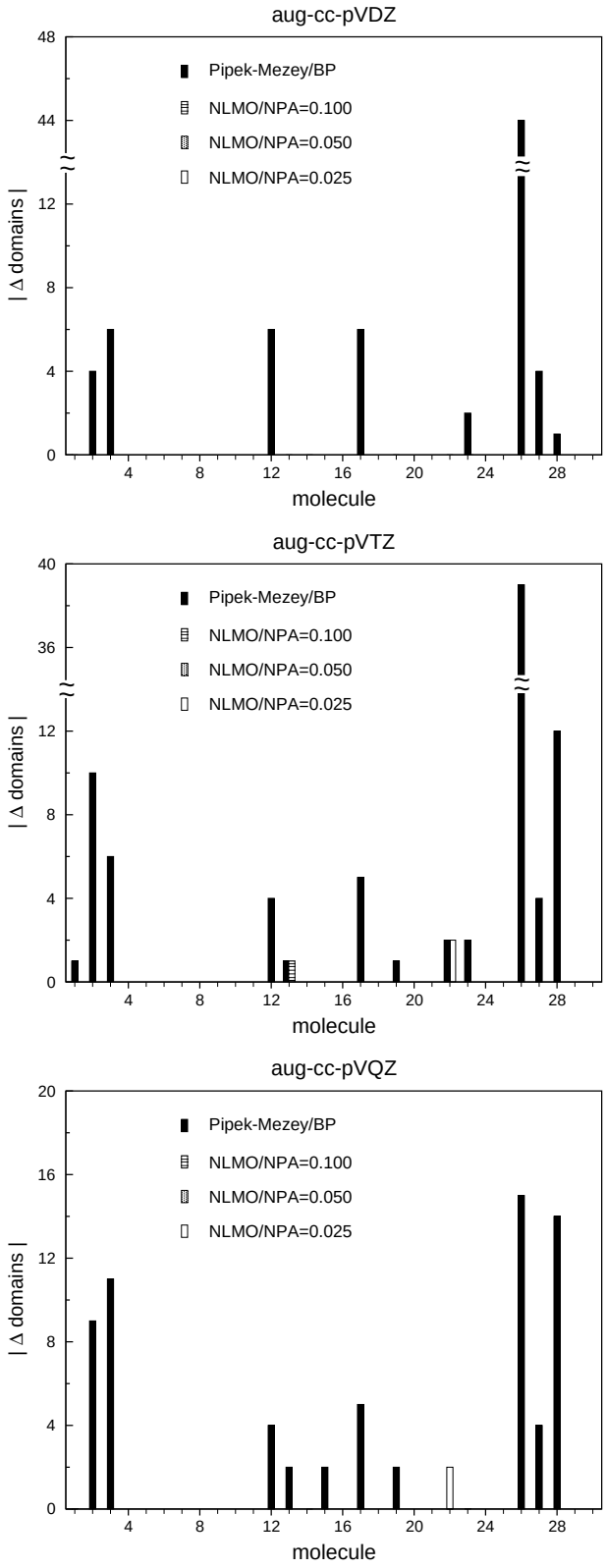



Figure 4.5: Domain changes Δ for the aug-cc-pVDZ, aug-cc-pVTZ and aug-cc-pVQZ basis sets in comparison to the cc-pVDZ basis. See text for a definition of Δ.

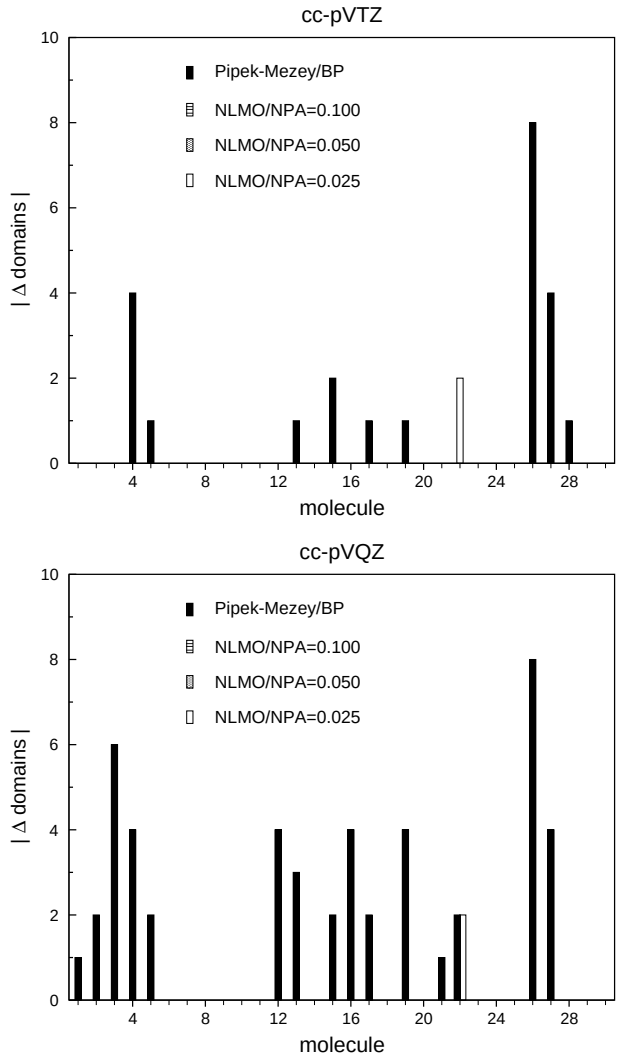



Figure 4.6: Domain changes Δ for the cc-pVTZ and cc-pVQZ basis sets in comparison to the cc-pVDZ basis. See text for a definition of Δ.

The dependence of the localization procedure on the geometry must however be defined. In the Pipek-Mezey localization procedure, one minimizes the number of atoms where the orbital is localized, as measured by the Mulliken population. For orbital $\phi_i$, the Mulliken-charge at atom $A$ is given by

$$\begin{aligned} Q_{Ai} &= \sum_{\mu \in A} \sum_{\nu} L_{\mu i} S_{\mu \nu} L_{\nu i} \\ &= \sum_{jk} U_{ji} U_{ki} \sum_{\mu \in A} \sum_{\nu} C_{\mu j} S_{\mu \nu} C_{\nu k}. \end{aligned} \tag{4.11}$$

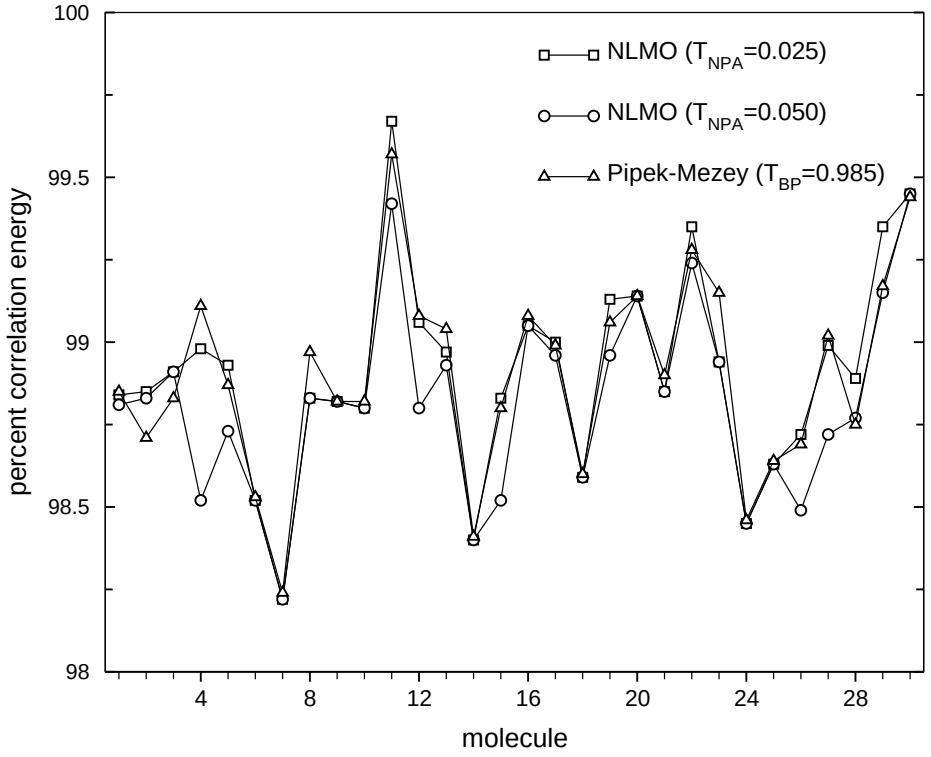



Figure 4.7: Comparison of the fraction of correlation energy relative to canonical MP2 (in percent) using the NLMO/NPA and PM/BP methods for localization and domain selection. The cc-pVTZ basis set was used.

The Pipek-Mezey localization consists in minimizing the value

$$q_i = \left[ \sum_{m=1}^{M} (Q_{mi})^2 \right]^{-1} \tag{4.12}$$

for each occupied orbital, which is equivalent to maximizing the function

$$F = \sum_i \sum_m (Q_{mi})^2 . \tag{4.13}$$

The localization functional must be stationary with respect to infinitesimal changes in the geometry, subject to the orthonormalization constraint. As discussed in Ref. [56], these conditions are fulfilled when

$$\sum_m^M [S_{ll}^m - S_{kk}^m] S_{kl}^m = 0, \qquad \text{for all } \; k > l, \tag{4.14}$$

with

$$S_{kl}^{A} = \sum_{\mu \in A} \sum_{\nu} \left[ L_{\mu k} S_{\mu \nu} L_{\nu l} + L_{\mu l} S_{\mu \nu} L_{\nu k} \right]. \tag{4.15}$$

These correspond to the Coupled-perturbed localization equations. Since NLMO orbitals have no unique minimization criterion, there is no such set of equations in this case. A possible solution to this problem is to use the NPA domain criterion together with Pipek-Mezey orbitals (PM/NPA). The matrix $\mathbf{V}$ in Eq. (4.9) is then substituted by the transformation matrix from NAOs to PM LMOs

$$\mathbf{V}' = \left(\mathbf{T}^{\mathrm{NAO}}\right)^{-1} \mathbf{L}^{\mathrm{PM}}. \tag{4.16}$$

The NLMOs can also be used as a starting guess for the PM localization, in order to keep the PM orbitals as similar as possible to the NLMOs in cases in which PM localization is not unique.

The PM/NPA combination was tested for all 30 molecules and 6 basis sets. Using the standard PM method, the domains were found to still vary considerably, although less than in the PM/BP case. However, significant improvements could be achieved by removing some functions from the localization criterion. For the cc-pV$X$Z basis sets, the most diffuse basis function of each angular momentum type for each atom was removed. For the augmented basis sets the two most diffuse functions were removed (as already mentioned, this can be done by zeroing the corresponding rows and columns of the overlap matrix used in the PM procedure). With these changes the NPA-based center charges are almost as stable as as those of the NLMO/NPA combination. For the recommended value of $T_{\mathtt{NPA}} = 0.05$ only two domain changes were observed for the whole test set. In the benzoquinone and formaldehyde molecules one of the carbonyl oxygen lone pairs changed to a CO bond for some basis sets.

# Chapter 5

# Local Quantum Mechanical Hybrid Scheme

## 5.1 Localized Orbitals as Molecular Subspaces

As discussed in Chapter 2, hybrid QM/MM schemes are nowadays fundamental tools not only in Computational Biochemistry, but also in general solvation and solid state problems. Performing a search in the `Web of Science` with the term "QM/MM", gives a total of $1,051$ results for the period of 2000-2006[1]. This is a clear evidence for their wide range of use and their importance in Chemistry today.

From the two schemes available, the subtractive approach is the most flexible in the combination of methods. It allows not only to couple quantum and molecular mechanics, it also supports coupling of different quantum chemical methods. Applications are not limited to large biological molecules. Even in small to medium-sized systems savings can be made by defining a small group of atoms where a higher level approach should be used. However, this involves cutting out a model system from the whole (what has been previously refered to as cluster in Section 2.3). If this system has a covalent bonding to the rest of the molecule, the dangling bonds must be saturated. In the ONIOM scheme,[71] this is done by adding a hydrogen atom along the broken covalent bond with a predefined distance. By performing a high and a low-level calculation on the model, and calculating the $\Delta E^{\text{sub}}$ correction (see Eq. (2.69)), the errors due to this link atom are mostly cancelled out. Nonetheless, this procedure has several limitations

(1) *The link atom should be able to mimic the properties of the deleted moiety.* The use of a hydrogen atom to cap a dangling C-C bond is usually sufficient, but double or highly polar bonds should be avoided. Cutting through aromatic systems is also not possible.

(2) *Three calculations are needed to obtain a single energy value*, independently of the method used.

(3) *Polarization effects are not included* in the higher level calculation. The higher level correction for the model is computed *in vacuo*, without the effect of the environment. This is true for IMOMO, but electrostatic embedding has been recently implemented in the ONIOM QM/MM coupling (also sometimes referred to as IMOMM).[54]

All of these problems are connected to the use of a different Hamiltonian for the high-level correction. Could it be possible to design another hybrid QM/QM approach which would not suffer from the same faults?

[1] This is actually a modest estimate, since the search is only performed on the title, abstract and keywords. Also, the subtractive schemes usually avoid the QM/MM designation.

The first thing to correct is the way the model system is defined. If atoms are the basis for this definition, it is inevitable to cut through bonds. Although the nuclei are well defined in space and easy to group together, the electrons, on the other hand, are spread out through many atomic centers. It is evident why atoms are not the best choice for defining model systems. In fact, since one is interested in solving an electronic problem, the most straightforward solution would be to split the electrons into different groups. In a converged RHF solution, each electron pair occupies a distinct molecular orbital. These molecular orbitals define subspaces and in theory one could use them to define the model. However, canonical orbitals usually span most of the molecular space, and it would be unreasonable to select a group of orbitals to describe local reaction energetics. There is no criteria to select individual contributions. On the other hand, localized orbitals are perfect candidates since they are well located in a region of the molecule and chemical sense could guide us in the choice of the most significant orbitals for the problem in question.

Another issue is that in correlated post-HF calculations the space spanned by an electron is not only given by the occupied orbital, but also by the virtual orbitals into which it can be excited. In local correlation methods, as discussed in Chapters 2 and 3, the excitation space is also local since domains are used to restrict the virtual space. Each orbital has its own domain and therefore, even in correlated methods, the space spanned by each individual electron is well defined.

The discussion above hints at a new way to couple different quantum chemical methods into a single calculation. Each LMO or group of LMOs, together with their associated domains, can be viewed as a subsection of the system and can be individually treated at a specific level of theory. Similar to IMOMO, a lower level method, e.g., LMP2, can be applied to a large part of the molecule or the whole system, and a higher level method, e.g. LCCSD(T), to a smaller subset of LMOs. As in IMOMO, Eq. (2.69) is effectively used to compute the final energy, but no artificial splitting of the molecule is needed, and only one calculation needs to be performed. Moreover, the same Hartree-Fock orbitals are used in the low-level and high-level calculations, and optionally a coupling of the high-level region to the environment can also be introduced at the correlated level.

## 5.2 Local Regions Approach

### 5.2.1 Method

The new hybrid QM/QM coupling scheme should allow, in a single calculation, for the treatment of molecular regions at different levels of accuracy through the use of HF and local correlation methods. Several local methods have been implemented in the `Molpro` program package over the last years, including Configuration Interaction methods, Møller-Plesset perturbation theory and Coupled Cluster up to perturbative triples (for a list of references, please consult Chapter 2). This discussion will be restricted to the use of LMP2, LCCSD and LCCSD(T), since these are the most commonly used quantum mechanical electronic structure methods to date.

The principle behind the method is to divide the orbitals into *regions*, each with a different correlation level. The same nomenclature is used as in the ONIOM approach. If two regions are defined, with the high-level treatment being LMP2, and the low-level HF, this will be referred to as a LMP2:HF calculation. If another region is added, using LCCSD(T), the name given will be LCCSD(T):LMP2:HF. Due to the similarity to the IMOMO approach, and the use of local methods, the scheme will be referred to as Local Molecular Orbital : Molecular Orbital (or LMOMO for short).

In the LMOMO scheme, one starts by performing a HF calculation for the whole system. After localization, a list of centers is assigned to each LMO as described in Section 2.1.4, and then each LMO is assigned to a region. In the current implementation, a maximum of three regions is allowed: a high-level region to be treated by, e.g., LCCSD(T), a low-level region, to be treated by LMP2, and the remainder, which is not correlated. It would be straightforward to extend this concept further.

The assignment of the LMOs to the regions is done as follows:

(1) A list of atoms and the corresponding correlation method for a region is provided as input.

(2) All LMOs which contain at least one of these centers in their domain lists are assigned to the region.

(3) The second region may be provided, and in this case steps 1 and 2 are repeated. The higher level regions should be assigned at the end (e.g., when coupling LCCSD and LMP2, the LMP2 region should be given first, and the LCCSD one afterwards).

(4) If no further region is provided, all remaining orbitals are treated by a default method; normally, this is HF (i.e. not correlated), but another method, e.g., LMP2, may be

specified as well.

The LMOMO procedure can be implemented as an extension to the pair approximations (discussed in Section 2.1.4). Consider a LCCSD(T):LMP2:HF hybrid calculation. The orbital pairs made up of LCCSD(T) orbitals will be classified as strong pairs, the LMP2 orbitals will build weak pairs, and pairs with HF orbitals are simply neglected (very distant pairs). Mixed pairs will belong to the lowest level regions.

The low-level correlation calculation, usually LMP2, is performed with all orbital pairs that can be formed from the orbitals in the high-level and low-level regions. Thus, the coupling between the different regions is fully included in the LMP2. Optionally, very distant pairs can be removed in order to achieve linear scaling. The final correlation energy is computed as a sum of pair energies; for all strong pairs the LCCSD pair energy is taken, for all other pairs the LMP2 pair energy. Finally, the triples correction is added. This is exactly as in any LCCSD calculation without regions.

The restriction of the pair list automatically leads to a restriction of the number of transformed integrals needed in the calculation. For LMP2, there is a one-to-one correspondence between the amplitudes $T_{rs}^{ij}$ and the required electron repulsion integrals $(ri|sj)$, and therefore the list of integrals is directly determined by the pair list and the associated pair domains. In the LCCSD(T) method further and larger integral classes are needed[4]. However, the occupied integral labels are always related to pair labels, and the virtual ones to domains of correlated pairs. Thus, the number of required integrals is determined by the list of high-level pairs. If the size of the high-level region is fixed, the number of transformed integrals becomes independent of the size of the molecule. This leads to an asymptotic $\mathcal{O}(1)$ scaling.

### 5.2.2 Preliminary Tests

#### Peptide Bond Formation Energies

The first system under study was a condensation reaction between polyglycine residues. Specifically, the formation of $(\mathrm{gly})_8$ from two $(\mathrm{gly})_4$ chains with water as a by-product was considered. Since the chains were assumed to be linear and the structures were not fully optimized, this is just a convenient model to study a number of different approximations at relatively low cost. It is also a system commonly used for evaluating the scaling properties of local correlation methods. The results are shown in Fig. 5.1, where the reaction energies are plotted as a function of the size of the high-level region.

Four sets of calculations were performed: In the first case, the high-level method was LMP2, and the remainder was not correlated (LMP2:HF). In the second case, similar cal-

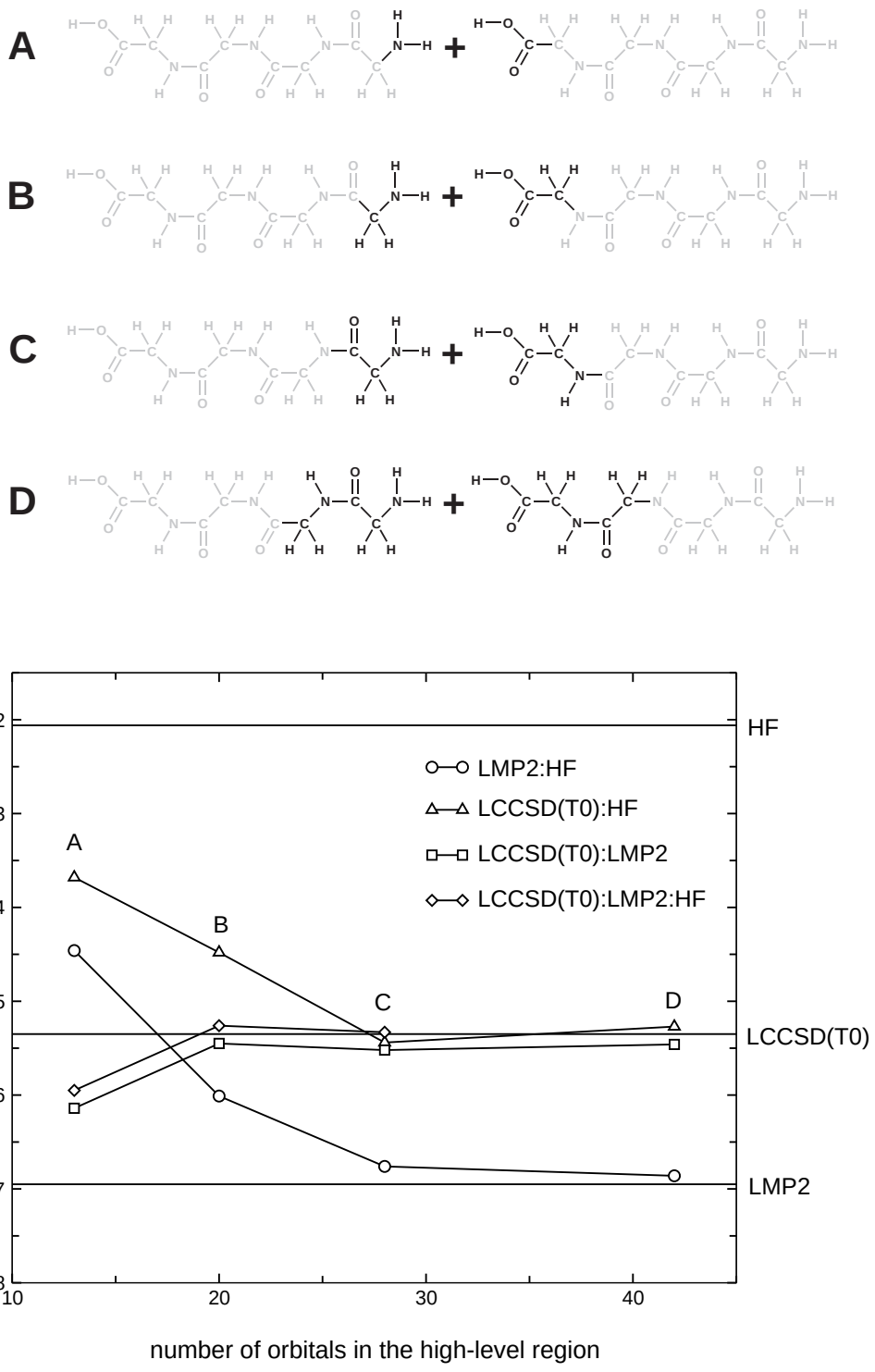



Figure 5.1: Reaction energies $\Delta E_R$ (in kcal mol$^{-1}$) for the peptide bond formation between two $gly_4$ chains computed at various levels of theory with the cc-pVTZ basis set. The horizontal lines represent reference values, according to the legend on the right hand side. The high-level selection is depicted in the Lewis diagrams shown above and further explained in the text. The atoms included in the high-level region are shown in black, the remaining molecule in light gray.

culations were performed with LCCSD(T0) as high-level method, i.e., LCCSD(T0):HF. Third, the high-level method was again LCCSD(T0), but all remaining orbitals were correlated using LMP2 (LCCSD(T0):LMP2). And finally, the low-level LMP2 region was restricted to be region **D**, and the rest of the molecule was left uncorrelated (LCCSD(T0):LMP2:HF). In each of these cases, the size of the high-level region was varied as shown in Fig. 5.1 (regions **A**,**B**,**C**,**D**). In addition, full LMP2 and LCCSD(T0) calculations were performed for comparison. The abscissa of Fig. 5.1 corresponds to the number of orbitals in the high-level region. Zero means the result of just the low-level calculation (HF or LMP2), while on the right hand side of the scale the results of the full LMP2 and LCCSD(T0) calculations are indicated.

The figure shows that HF strongly underestimates the reaction energy, while LMP2 overestimates it. The convergence of the LMP2:HF result with respect to the size of the high-level region is rather slow. Case **A**, which just correlates the terminal $NH_2$ and COOH groups, is clearly insufficient. The error is reduced to about 1 kcal/mol if all atoms up to the second neighboring bonds (relative to the atoms involved in the reaction) are included in the high-level region (case **B**). However, a satisfactory value of the reaction energy is only obtained by including a larger section of the system (cases **C** or **D**). For region **D**, the error amounts to at most 0.1 kcal/mol.

The same trend is visible in the LCCSD(T0):HF case. Only with region **C** is the value close to the full result. The convergence can however be greatly improved by the use of LMP2 in the lower level regions. Both LCCSD(T0):LMP2 and LCCSD(T0):LMP2:HF results are quite similar. The inclusion of the first-neighboring atoms is already sufficient to obtain near LCCSD(T0) values. It should be noted that the cost of LMP2 is small when compared to the coupled cluster calculation, and therefore comparable timings are obtained for region **B** with LCCSD(T0):HF or LCCSD(T0):LMP2 (for further information on timings see Section. 5.2.3).

### Including Environment Effects in the High-Level Region

In Section 2.3, the concept of electronic embedding was introduced in the context of QM/MM coupling schemes. It is defined as the (approximate) inclusion of polarization effects due to the surrounding environment. If the lower level region is represented by a force field, the embedding is performed by including point charges in the QM-Hamiltonian, which represent the environment atoms. If the lower level region is itself treated at the QM level (QM/QM), these effects are left out. A relatively straightforward solution to the problem could be the use of RESP charges, computed from a lower level run of the host, and including the charges on the other two calculations (see Eq. 2.69) just as in the QM/MM

case. However, no work has been done in this direction. In the LMOMO scheme, as previously discussed, electronic embedding is implicitly included. This has an influence not only on the HF value, but also on the correlated calculation, since the reference function is changed.

Another type of effect yet to be considered, is what could be referred to as *correlation embedding*. Let us consider a LCCSD(T):LMP2 calculation. In the LMOMO scheme, the $\mathbf{R}^{ij}$ MP2 residuals will be computed for each orbital pair. On top of this calculation, a group of orbitals will be selected, the doubles and singles CCSD residual equations will be solved and the perturbative triples calculated. For the LCCSD(T) step, only orbital labels belonging to the high-level region will be considered, and the LCCSD(T) contribution will be therefore independent of the LMP2 amplitudes outside the region.

Shown below is the LCCD residual[2] in matrix form

$$\mathbf{R}^{ij} = \mathbf{K}^{ij} + \mathbf{K}(\mathbf{T}^{ij}) + \sum_{kl} \left[ K_{ij}^{kl} + \mathrm{tr}(\mathbf{T}^{ij}\mathbf{K}^{lk}) - \delta_{jl}\beta_{ki} - \delta_{ki}\beta_{lj} \right] \tilde{\mathbf{S}}\mathbf{T}^{kl}\tilde{\mathbf{S}} + \mathbf{G}^{ij} + \mathbf{G}^{ji}, \quad (5.1)$$

with

$$\beta_{ij} = F_{ij} + \sum_{k} \mathrm{tr}\left( \left[ 2\mathbf{K}^{ik} - \mathbf{K}^{ki} \right] \mathbf{T}^{kj} \right). \quad (5.2)$$

The $\mathbf{G}^{ij}$ matrices include many more couplings which need not to be discussed at this time (see Refs. [33, 4]). It is possible to include the LMP2 computed amplitudes of the surrounding environment in the sum running over $kl$ in Eq. (5.1). In the LMOMO scheme, as defined up till now, this sum will only run over pairs of orbitals in the high-level region. By adding these extra terms, one is effectively including correlation effects of the neighboring region into the high-level calculation. As before, this can be implemented through the use of the local pair approximations. Neighboring orbitals in the regions border (one orbital located in the high-level region, the other in the low-level) may be classified as close pairs. The program can then be instructed to include the LMP2 amplitudes into the CC residuals. This requires only a small extra computational cost. The sum will be larger, and the necessary $\mathbf{K}$ operators have to be computed. This is , however, for most purposes insignificant when compared to the total cost of the LCCSD(T) calculation. The most expensive term to compute is the $\mathbf{K}(\mathbf{T}^{ij})$ term, which is diagonal in the pair index.

Calculations have been performed in order to access the effect of this coupling between regions. One should keep in mind that only orbital pairs close to the region border are added (the ones which in a regular LCCSD(T) run would be classified either as strong or as close).

[2]The singles are not discussed to keep the formulae simple and compact. Including the terms arising from the singles would not change the following discussion.

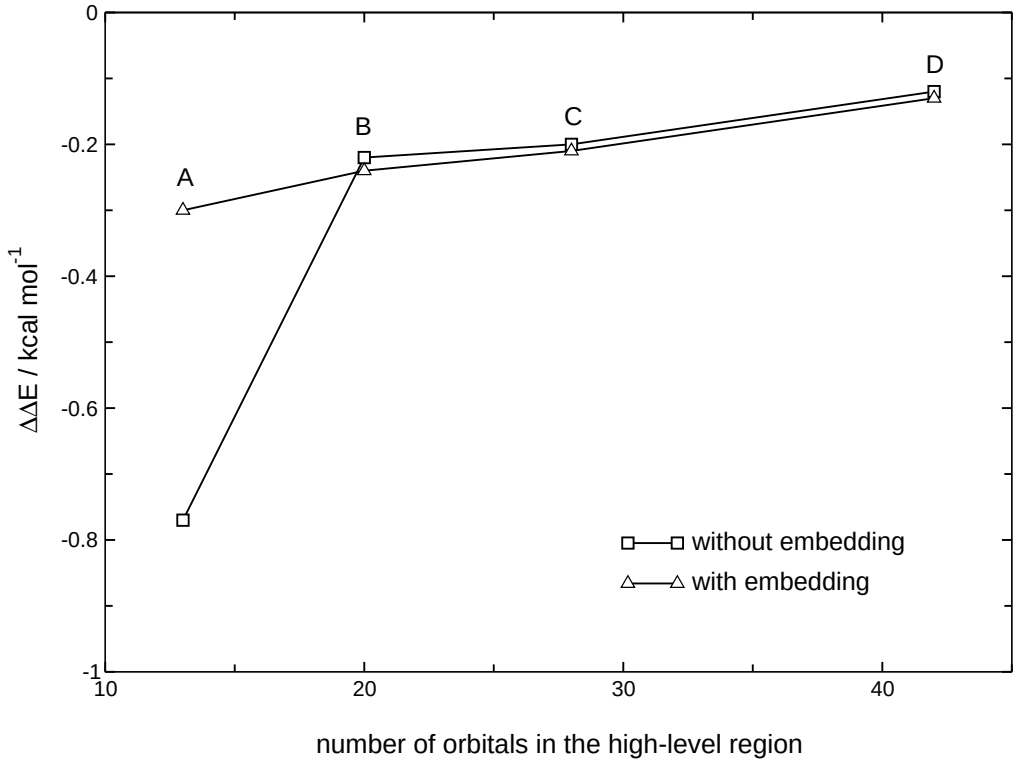



Figure 5.2: Reaction energy error $\Delta\Delta E$ (in kcal mol$^{-1}$) for LCCSD(T0):LMP2 calculations on the condensation reaction of glycine tetrapeptides. The high-level regions are described in Fig. 5.1. In one set of calculations, orbital pairs at the boundary region were classified as close, and their amplitudes were included in the CC residuals (*with embedding*). For the other set, only strong pairs at the high-level region were included (*without embedding*).

The test system was again the condensation reaction of two $gly_4$ residues, and the method in test was LCCSD(T0):LMP2. The results are shown in Fig. 5.2, with the same regions as in Fig. 5.1. Two sets of calculations were carried out, one without including amplitudes from the low-level region, other including close pairs connecting both regions and introducing the LMP2 amplitudes in Eq. (5.1). The figure plots the error of the LMOMO calculation relative to the full LCCSD(T0) result

$$\Delta\Delta E = \Delta E_R(\mathrm{LCCSD(T0):LMP2}) - \Delta E_R(\mathrm{LCCSD(T0)}). \tag{5.3}$$

It is preferable to plot $\Delta\Delta E$ instead of $\Delta E_R$ in comparing both calculations, since the close pairs within the LCCSD(T0) region are also included in the CC residual, and the LCCSD(T0) result converges to a different value. The effect is, however, small. The discussion is also out of the scope of this work.

It is observed in Fig. 5.2 that the effect on the energies is rather small, except for selec-

tion **A**. This effect is obviously due to the small region size. The number of orbitals in the high-level region is so small that the CC amplitudes are not well converged (in comparison to the ones in the full local CC treatment). Comparing the result with embedding for region **A**, and for region **B** without, it is clear to see that the improvement on going from selection **A** to **B** (in the non-embedding case) is mostly due to a better description of correlation in the smaller region, an effect of the added neighboring orbitals in the residuals. It is also interesting to see an almost linear error for the case with embedding.

Calculations were carried out for two other systems, later discussed in Sections 5.3.1 and 5.3.2. The effect was found to be similar, slightly improving the energies for smaller regions, but never more than 1 kcal $\text{mol}^{-1}$. Since the effect of including the close pairs in the CC iterations (for the full calculation) is of the same magnitude, it is questionable whether this option should actually be used.

### 5.2.3 Scaling of the Method

The use of correlation regions should lead asymptotically to the scaling regime of the lower level calculation. As long as the high-level region is fixed, extending the molecular system should not influence its cost. If HF would be the method of choice for the low-level, this would mean an $\mathscr{O}(1)$ scaling[3]. The present implementation, although not optimized for the integral transformation, already shows optimal scaling for the iteration steps and the triples calculation. To evaluate the scaling and computational cost of this approach the peptide bond reaction energy was computed

$$2\ \text{gly}_n \longrightarrow \text{gly}_{2n} + \text{H}_2\text{O}\ ,$$

where $n$ stands for the number of glycine residues in the polypeptide.

#### The LMP2 case

In a LMOMO calculation, the number of pairs only depends on the correlated region size. In the LMP2:HF case, as previously discussed, all quantities are directly connected to the pair list, and the $\mathscr{O}(1)$ scaling should be directly observable.

LMP2:HF/cc-pVDZ calculations were performed for reactions $n = 2,6$, with the same atom selection **C** as in Fig. 5.1. The iteration timings are compared to the full LMP2 counterparts in Fig. 5.3. The LMP2 curve shows the expected linear scaling, while the

[3]This discussion is only aimed at the correlation calculation.

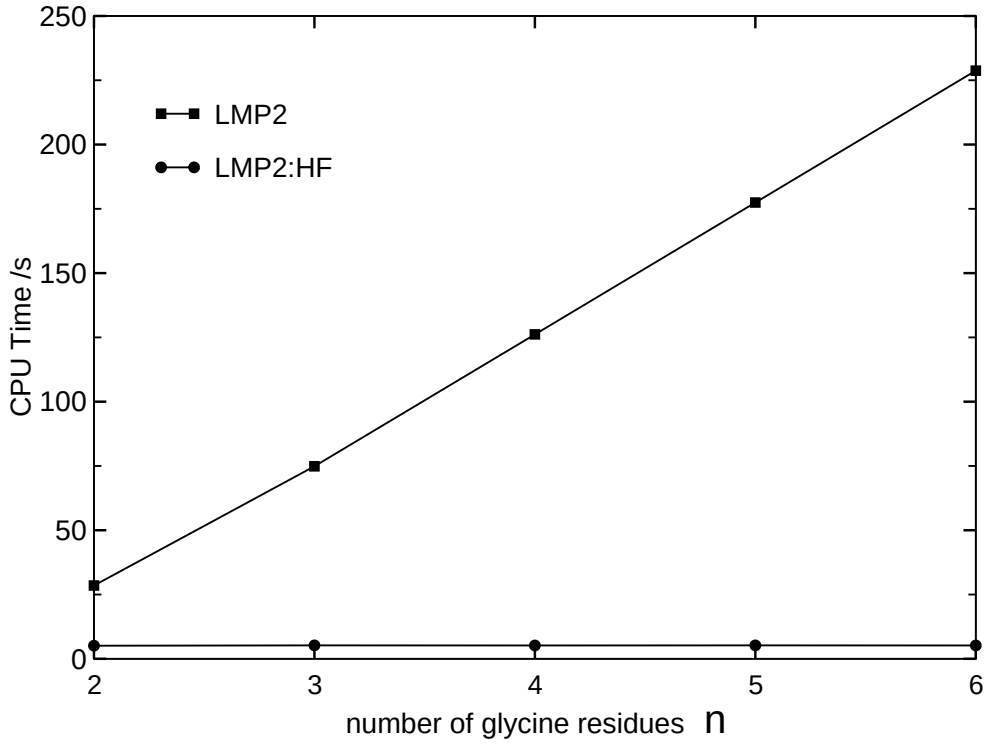



Figure 5.3: Timings (in s) for LMP2/cc-pVDZ and LMP2:HF/cc-pVDZ MP2 iterations on the protein bond reaction. All calculations were performed on an AMD-Opteron MP2400+ dual processor machine. Density fitting approximations were used.

LMP2:HF curve is flat, with $\mathcal{O}(1)$ scaling. The computational cost for the transformation steps is still not optimal (linear scaling) due to a dependence on the basis set size. However, even without the proper asymptotical scaling behavior, significant savings are made.

### The LCCSD case

In the case of LCCSD, obtaining the target $\mathcal{O}(1)$ scaling behavior involves some further approximations, due to the slow decay of some 1- and 3-external integrals. These arise in terms from the interaction between singles and doubles. The first term in question is found in the pair residuals

$$R_{rs}^{ij} = \cdots - \sum_{kl}\sum_{tu} \tilde{S}_{rt} T_{tu}^{lj} \tilde{S}_{us} \sum_{v} t_v^k [2(vk|li) - (vl|ki)]. \tag{5.4}$$

The slow decay is due to 1-external integrals $(vk|ll)$ which decay only with the inverse square of the distance between orbitals $k$ and $l$. The other term is connected to 3-external

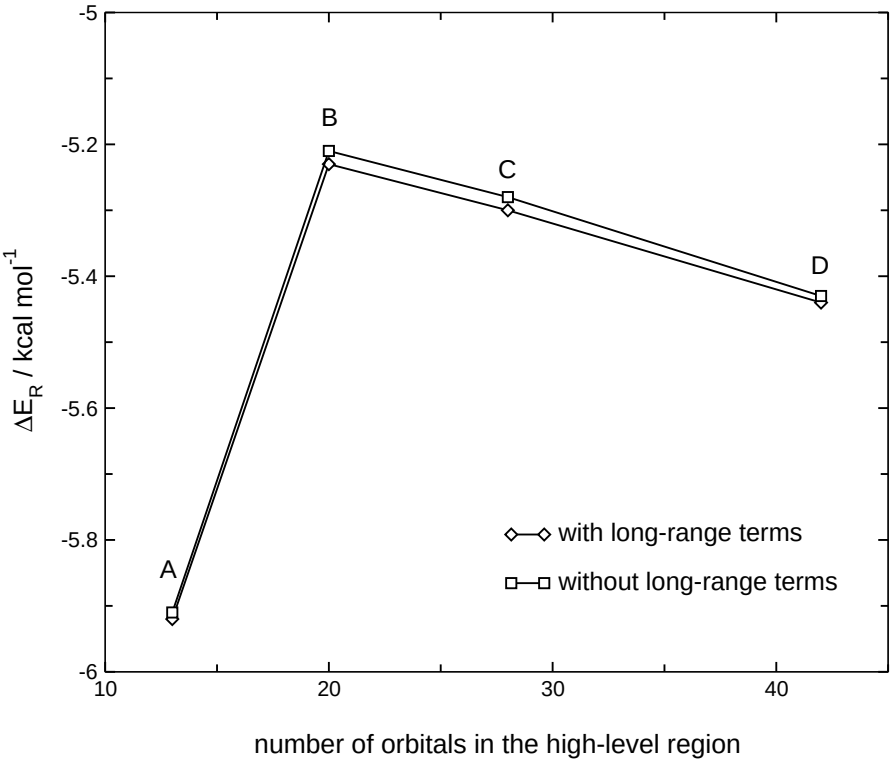



Figure 5.4: Reaction energies $\Delta E_R$ (in kcal mol$^{-1}$) for the peptide bond formation between two $gly_4$ chains computed at the LCCSD(T0):LMP2:HF/cc-pVTZ level of theory. The high-level selection is depicted in the Lewis diagrams shown in Fig. 5.1 and further explained in the text.

integrals

$$G_{rs}^{ij} = \cdots + \sum_{tu}\sum_{k}\sum_{v} \tilde{S}_{rt} T_{tu}^{ij} t_v^k [2(vk|us) - (uk|vs)]. \tag{5.5}$$

The integral $(vk|us)$ has the same dependence as the previously discussed 1-external integral. In the current LCCSD program version, the term in Eq. (5.5) is fully computed, with the cost scaling quadratically. However, the prefactor is small and is only expected to become a bottleneck for very large calculations. The 1-external term is truncated using additional distance criteria, just as discussed in Ref. [5]. In the case of a LMOMO calculation, keeping such terms has a strong effect on the scaling, and leads to a breakdown of the $\mathcal{O}(1)$ asymptotic limit. However, since both terms have opposite signs, and they should cancel out at large distances (to restore the asymptotic $r^{-6}$ distance dependence), one can in fact neglect them, introducing only small errors in the calculation. LMOMO LCCSD(T0):LMP2:HF calculations for the $gly_4$ condensation reaction were repeated, both including the two contributions and removing them. The results are shown in Fig. 5.4.

The maximal error made by neglecting the aforementioned integrals is 0.02 kcal mol$^{-1}$.

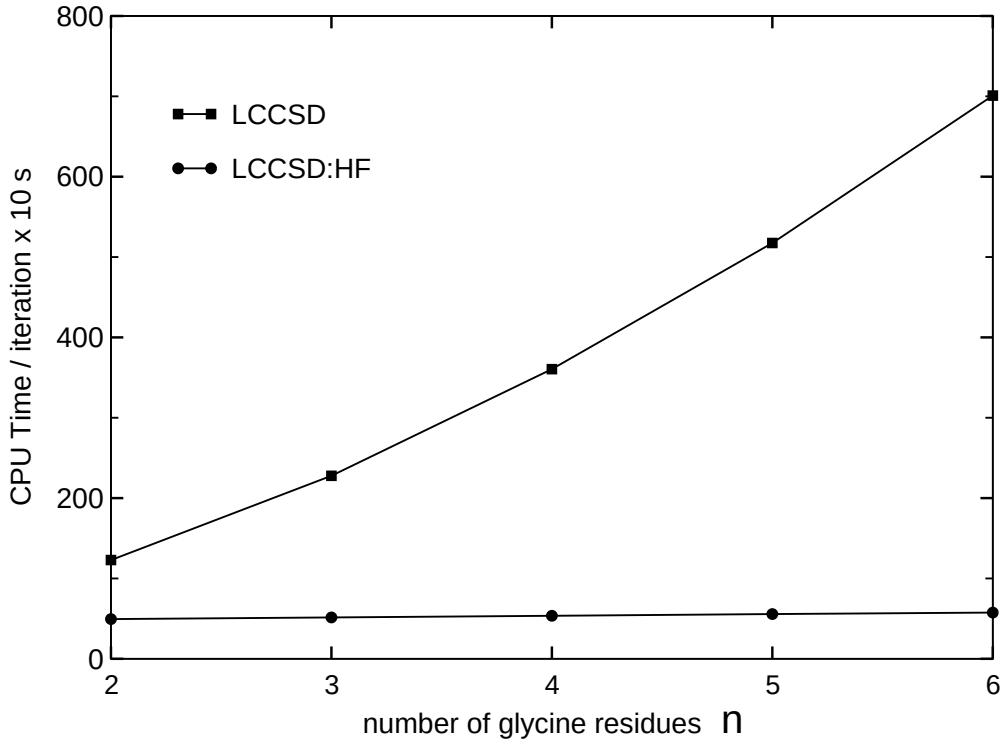



Figure 5.5: Timings (in s) for LCCSD/cc-pVDZ and LCCSD:HF/cc-pVDZ CC iterations on the protein bond reaction. The time is computed as the average of each iteration, and multiplied by a factor 10 (the regular number of cycles needed to converge the CCSD solution). All calculations were performed on an AMD-Opteron MP2400+ dual processor machine. Density fitting approximations were used.

Such an error is much lower than those introduced by the local approximations, and of the same order of magnitude as the density fitting errors. Also of interest is the fact that the error remains constant with increasing CC region size. Therefore, the approximation seems to be well-founded, and can be used to significantly reduce the cost of LMOMO calculations.

At the time of this work, the integral transformation costs still scaled linearly, but the iterations showed the expected $\mathscr{O}(1)$ scaling. The timings for LMOMO LCCSD:HF and LCCSD calculations are shown in Fig. 5.5. The iteration time is defined as the average of all iteration steps multiplied by a factor 10. This is due to the fact that the regular LCCSD calculation takes 11 iterations (or more) to converge, while the LCCSD:HF only takes 10. In order to effectively compare the iteration times, the number of steps should be the same. The reduced number of iterations is due to the reduced number and size of the residuals to compute, which is a further advantage for the hybrid scheme.

The LCCSD timings only show linear scaling behavior beyond $n = 4$, somewhat later

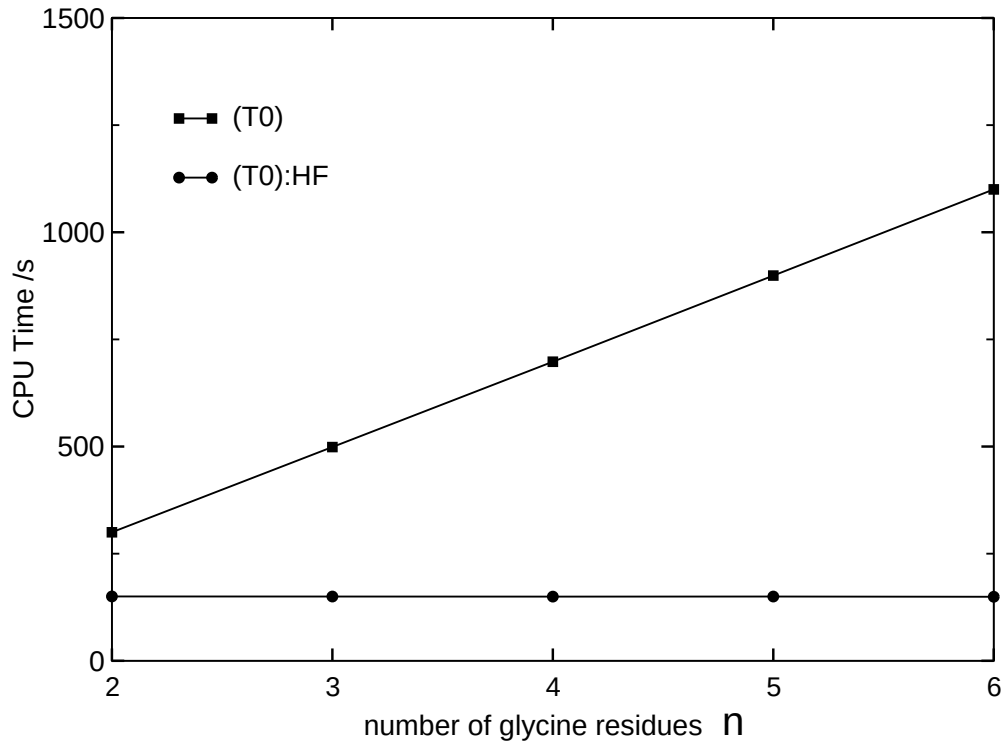



Figure 5.6: Timings (in s) for LCCSD(T0)/cc-pVDZ and LCCSD(T0):HF/cc-pVDZ triples calculations on the protein bond reaction. All calculations were performed on an AMD-Opteron MP2400+ dual processor machine. Density fitting approximations were used.

than in the LMP2 case, certainly due to the larger CC operator lists[4]. The $\mathcal{O}(1)$ scaling behavior of LCCSD:HF is already visible beyond $n = 2$, a rather early onset. However, it is not totally independent of the molecular size, and there are still some steps which slowly grow in weight, although not visible in the graph. These steps may be an outcome from the use of regular sized-matrices or some of the loop structures, but are nevertheless too small in size to become relevant (for any application size where HF can still be computed).

### The (T0) perturbative triples correction

In the triples calculation only excitations from those orbital triples $(ijk)$ are included for which $i$, $j$, and $k$ belong to the high-level region. Again, this list is further reduced by the condition that one of the pairs $(ij)$, $(ik)$, or $(jk)$ is a strong pair; the two other pairs can

[4]In the case of local Coupled Cluster, further distance criteria are used for the operator lists. The $\{K\}$ operator list, for example, will include all pairs which are within 8 Bohr from each other. This means that only for molecules with a diameter significantly above 16 Bohr will the linear scaling regime be visible. The glycine tripeptide measures about 22 Bohr.

either be strong or close[5, 6, 7]. In order to obtain an accurate triples energy, it is necessary to include the close-pair amplitudes in the triples calculation. Only pairs with two orbitals in the high-level region are treated as close pairs; the close-pair amplitudes are determined in the low-level calculation.

Calculations were performed for LCCSD(T0):HF, and the timings for the triples calculation is given in Fig. 5.6. The desired scaling is obtained already for $n = 2$, and considerable savings are made with the use of LMOMO.

## 5.3 Test Applications

### 5.3.1 Proton Transfer

In this section, a first application of the LMOMO scheme is discussed. The reaction is depicted in Fig. 5.7, a proton transfer step for an intermediate involved in the reaction between bis(1,3,4-thiadiazole)-1,3,5-triazinium halides and nitrogen-based nucleophiles.[85] The active site, although well localized (the proton 'jumps' from one nitrogen to a neighbouring one), is difficult to model since the atoms involved are located in a ring structure. For these cases, when using QM/MM or ONIOM-related methods, one would be forced to include the aromatic rings into the model system. This is due to the fact that 'link atoms' cannot accurately replace the aromaticity effect. In the LMOMO scheme, however, cutting through the rings is possible, since most of the aromaticity effect is contained in the SCF, which is always performed for the whole system. The structures have been preoptimized at the B3LYP/cc-pVDZ level.

The reaction energy barrier was computed with different region sizes and methods. The cc-pVTZ basis set[75] was used. The results are shown in Fig. 5.8, plotting the error of the barrier height relative to the pure high-level result as a function of the orbitals included in the region. The first plausible choice for the high-level region are the three atoms involved

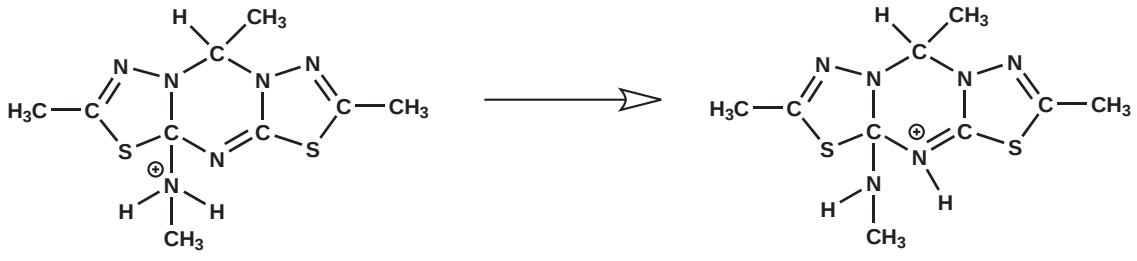



Figure 5.7: Proton transfer reaction.

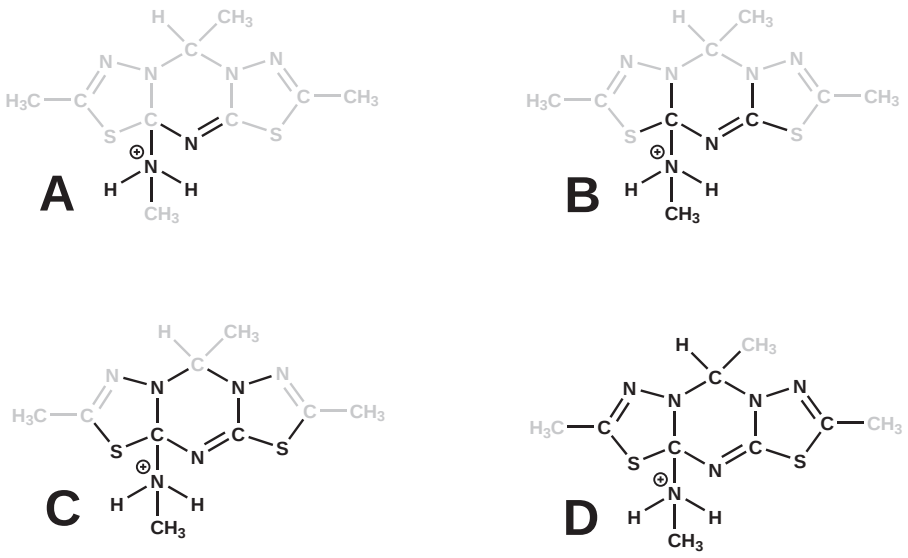
in this reaction, two nitrogens and the migrating proton. In Fig. 5.8 this selection is denoted as region **A**. For the combination of HF with LMP2 or LCCSD(T0), it is seen that already $3/4$ of the correlation correction to the HF value is obtained. The error amounts, however, still to 3-4 kcal mol$^{-1}$. By adding neighboring atoms to the region the results converge only slowly. Only with the selection **C**, which includes most of the atoms in the rings, the target value is approached to within 1 kcal mol$^{-1}$. This shows the physical significance of including all contributions from the aromatic system into the correlation treatment. However, combining LCCSD(T0) with LMP2 for the low-level region (LCCSD(T0):LMP2), one obtains a much more stable result. In this case already region **A** yields a value that is in very close agreement with the full LCCSD(T0) calculation, a very satisfying result. LCCSD(T0):LMP2:HF calculations, with selection **D** as the LMP2 region (leaving only the methyl groups uncorrelated) give the same results as LCCSD(T0):LMP2 to within 0.02 kcal/mol. These values were left out of Fig. 5.8, as they would simply superimpose with the other set.

The effect of removing the contributions shown in Eqs. (5.4) and (5.5) from the CC residuals was tested also for the LCCSD(T0):LMP2:HF case. The effect was found to be as small as previously discussed in Section 5.2.3 for the poliglycine condensation reaction.

### 5.3.2 Hydroxylation Reaction

The second example concerns the hydroxylation step in the *p*-Hydroxybenzoate Hydroxylase (PHBH) enzyme catalytic cycle. The cofactor FADOOH hydroxylates the substrate (*para*- hydroxybenzoate) after being oxidized by molecular oxygen. The peroxide moiety is broken and an OH group is moved onto the substract ring. The reaction is represented in Fig. 6.2. QM/MM calculations have been carried out with the cofactor and substrate treated at the QM level, and the remaining enzyme and solvent at the MM level. The MM environment is accounted for by point charges in the QM Hamiltonian. Further details on the reaction and the QM/MM modeling are given in Section 6.2.

The hydroxylation occurs between two aromatic systems, and therefore breaking bonds close to the reaction should lead to large errors. This system is another example where the LMOMO approach allows to divide the system, even with significant electron delocalization. As before, the smallest region included the atoms directly involved in the process (the peroxide moiety and the carbon to which the OH group moves). The first, second, and third neighboring atoms were then added to the list. In the coupled cluster region, the distance criteria for weak pairs was set to 7 Bohr, in order to include triples in the interaction of the substrate and cofactor with the migrating OH group. Further details of these calculations



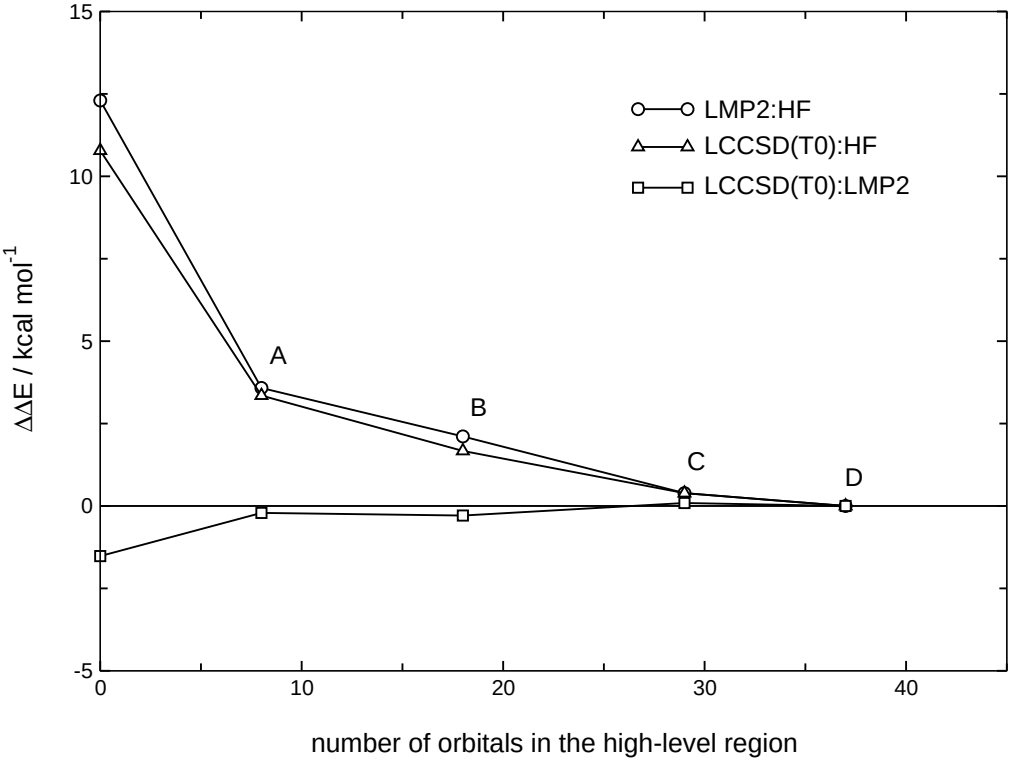




Figure 5.8: Regions calculation error $\Delta\Delta E = \Delta E(\text{high:low}) - \Delta E(\text{high})$ (in kcal mol$^{-1}$) for the proton transfer reaction. The error is given as a function of the orbitals included in the high-level region. The first point is a low-level calculation, the second a regions calculation including the two nitrogens and the transfered proton in the high-level region. The following points are obtained by adding the next neighbouring atoms (see ilustration above).

are given in Section 6.2.

As shown in Fig. 5.9, the region calculation converges to an accuracy of 1 kcal mol$^{-1}$ rather quickly. In contrast to the previous example, quantitative results are already obtained when the first neighbors are included in the high-level region of the LCCSD(T0):HF calculations (region **B**). Again, when the low-level region is treated by LMP2, the convergence with region size is further improved, and as in the previous example it is sufficient to use the smallest region (**A**) for the high-level calculation. Even though our program is not yet fully optimized, this leads to dramatic savings (approximately a factor of 10) as compared to the full LCCSD(T0) calculation. The cost of the correlation calculation is even comparable to the HF calculation. One LCCSD(T0):HF **B** single point calculation takes about 1142 min, compared to 447 min for the HF. The LCCSD(T0):LMP2 **A** takes 690 min, less than double the time. For even larger system sizes the HF calculation should become the bottleneck. It should also be noted that the algorithm is still not optimized, and that further savings should be possible.

## 5.4 Comparison to other partitioning methods

In this Section, a comparison is made between the LMOMO and IMOMO schemes.[71] In IMOMO, a model system is built and the difference between a low and a high-level method is used to extrapolate the total energy, using the low-level estimate for the real system. This allows for the combination of any two (or more) quantum mechanical methods. The local approach can only combine *ab initio* local methods (e.g., DFT is not supported). Also, IMOMO allows for the use of different basis sets. In LMOMO, dual- or multiple basis set approaches could be used. For instance, it may be sufficient to use a smaller basis set for regions of the system that are sufficiently far apart from the correlated region. Even though unphysical polarization artifacts may occur for neighboring atoms with different basis set sizes, it should at least be possible to reduce the number of polarization functions without significantly affecting the accuracy of the method. This will, however, not be discussed in the context of this work.

For the results shown in this Section, a simple `Perl` program was written and interfaced to the `Molpro` program package in order to perform IMOMO calculations. Upon receiving a geometry the program would generate the model system (with parametrized distances for the link atoms, see below) and run the three calculations displayed in Eq. 2.69.

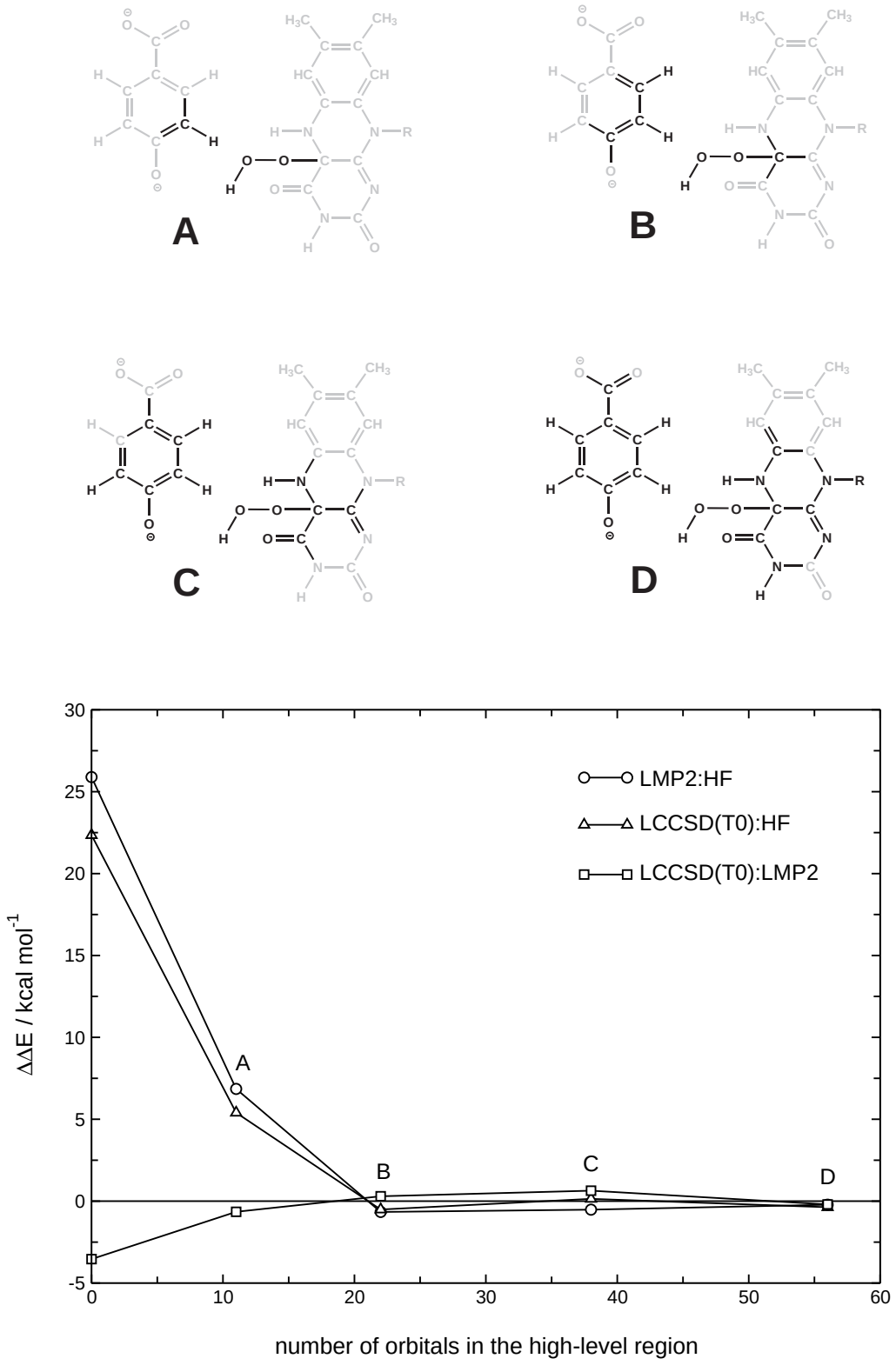



Figure 5.9: Regions calculation error $\Delta\Delta E = \Delta E(\text{high:low}) - \Delta E(\text{high})$ (in kcal/mol) for the hydroxilation reaction. The error is given as a function of the orbitals included in the high-level region. The first point is a low-level calculation, the second a regions calculation including the peroxide moiety and the hydroxilated carbon in the high-level region. The following points are obtained by adding the next neighboring atoms (see ilustration above).

### 5.4.1 Chlorohydrocarbon SN2 reactions

To compare both methods, calculations were performed on a series of chlorohydrocarbons reactions, a system previously studied with IMOMO by Re *et al.*[86] The SN2 reactions with $OH^-$ are described by the following general formula

$$CH_{(4-n)}Cl_n + OH^- \rightarrow CH_{(4-n)}Cl_{(n-1)}OH + Cl^-,$$

with $n = 2, 3$ or 4.

For this set of calculations, LCCSD(T0) and LMP2 were used as high and low-level methods. Hartree-Fock is not discussed since it gives large errors, even when used in combination with the other methods. This had already been mentioned in the above cited work.[86]

The regions used are the same as in the Re and Morokuma work. The carbon, the entering OH group and the leaving Cl are the high-level region, the lower region corresponds to the spectating atoms. The only orbitals to be included at the lower level are therefore the spectating chlorine lone pairs. This is the closest comparison possible, since in any other case the central carbon would be almost excluded from the high-level region (only one C orbital would be included). This choice leads to small differences between full LCCSD(T0) and the regions calculation in the $n$=2 case, but is already a 50/50 partition for the larger system.

The results are shown in Table 5.1. Although for the smaller systems the comparison is quite good (0.2 kcal/mol difference), for the $n$=4 case there is a small increase in the error. This should be connected to the fact that this is the largest approximation made, but it is also interesting to notice that the IMOMO results follow the same trend. The two procedures compare well with each other, and the errors involved are of the same magnitude as expected, from the observations in previous Sections.

### 5.4.2 Aminoacid-water complexes

For a second series of tests, hydrogen bonded systems were considered. A previous IMOMO study of Anderson *et. al.* [87] compared the performance of different high and low-level methods for the prediction of dissociation energies of aminoacids-water complexes. It was found that the combination MP2:HF gave very small errors. Using the same HF/6-31+G(d) optimized structures, single point calculations were carried out with the cc-pVTZ basis set, and a combination of LMP2 and HF (with the IMOMO and local approaches). The aminoacids included in the study were Asparagine (ARG), Glutamine (GLN), Serine (SER) and Threonine (THR) (the water complexes are depicted in Fig.

Table 5.1: SN2 reaction energies (in kcal/mol) for the $CH_{(4-n)}Cl_n$ series, with $n = 2,3,4$. The basis set used was aug-cc-pVTZ throughout.

| | LMP2 | LCCSD(T0):LMP2 | | LCCSD(T0) |
|---|---|---|---|---|
| | | IMOMO | local | |
| $n = 2$ | | | | |
| reactant complex | -22.4 | -22.7 (-0.3) | -23.0 (-0.6) | -22.4 |
| transition state | -12.2 | -14.0 (-0.5) | -13.6 (-0.1) | -13.5 |
| product | -56.5 | -59.1 ( 0.2) | -59.2 ( 0.1) | -59.3 |
| average absolute error | | ( 0.3) | ( 0.3) | |
| $n = 3$ | | | | |
| transition state | -8.5 | -10.7 (-1.3) | -9.9 (-0.5) | -9.4 |
| product | -61.5 | -64.2 ( 0.5) | -64.3 ( 0.4) | -64.7 |
| average absolute error | | ( 0.9) | ( 0.5) | |
| $n = 4$ | | | | |
| reactant complex | -7.8 | -8.1 (-0.5) | -8.0 (-0.4) | -7.6 |
| transition state | 0.9 | -0.7 (-0.9) | -0.3 (-0.5) | 0.2 |
| product | -62.9 | -65.6 ( 0.9) | -65.7 ( 0.8) | -66.5 |
| average absolute error | | ( 0.8) | ( 0.6) | |

5.10). The dissociation energy for the aminoacid + water system was computed, as well as for the tripeptides (built by extending the respective aminoacid with two Glycine (GLY) side chains). The results are shown in Table 5.2.

The difference between the two approaches is minimal. The largest error is seen in the local calculation for the ASN water complex, but even this is below 0.2 kcal $mol^{-1}$. Also, the errors seem to be independent of the system size, with comparable accuracy between the aminoacid and the tripeptides results. The local hybrid scheme could therefore be a useful tool for the study of weak interactions in biological systems. Possible applications are enzyme docking and the study of specific DNA strains interactions.

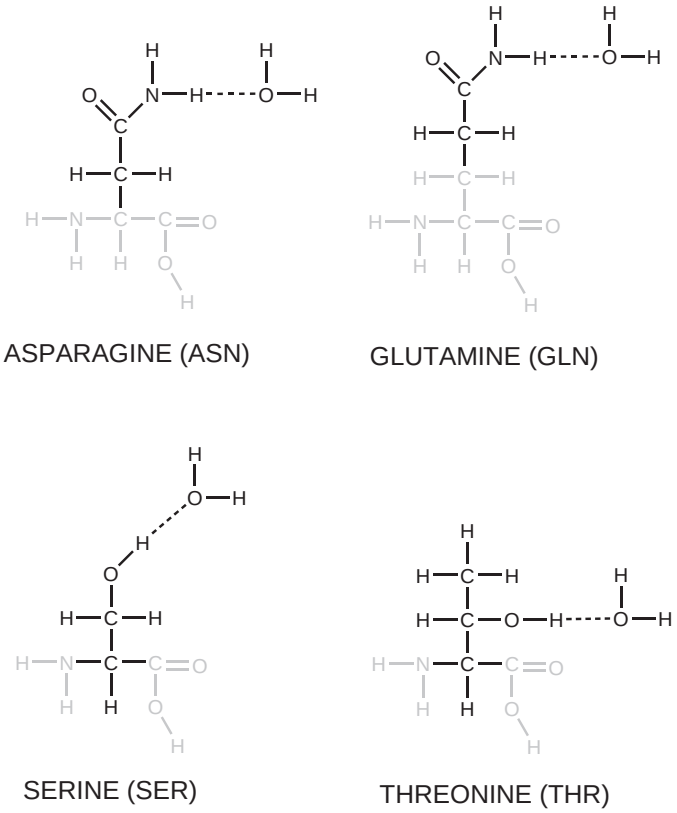



Figure 5.10: Aminoacid-water complexes. The tripeptides are built by expanding the aminoacid from both sides with glycine chains.

Table 5.2: Dissociation energies $\Delta E_d$ (in kcal mol$^{-1}$) computed at the HF/cc-pVTZ and LMP2/cc-pVTZ level, and errors relative to the LMP2 estimate $\Delta\Delta E_d$ (in kcal mol$^{-1}$) for the IMOMO and local hybrid schemes. The same scaling factors as in the Anderson *et. al.* study (0.786011 for N-C bonds and 0.723866 for C-C bonds) were used in the IMOMO calculations.

| | $\Delta E_d$ | | $\Delta\Delta E_d$ (LMP2:HF) | |
|---|---|---|---|---|
| | HF | LMP2 | IMOMO | local |
| ASN | 5.86 | 7.27 | -0.07 | -0.17 |
| GLN | 4.83 | 6.27 | -0.01 | -0.01 |
| SER | 5.00 | 6.15 | 0.05 | 0.00 |
| THR | 3.60 | 6.30 | 0.09 | 0.02 |
| GLY-ASN-GLY | 6.91 | 8.71 | -0.08 | 0.04 |
| GLY-GLN-GLY | 7.52 | 9.44 | -0.08 | -0.03 |
| GLY-SER-GLY | 6.14 | 7.39 | 0.12 | 0.01 |
| GLY-THR-GLY | 5.41 | 6.76 | 0.00 | -0.09 |
| average absolute error | | | 0.06 | 0.05 |

# Chapter 6

# Computation of Activation Barriers in Enzymes

## 6.1 Local Correlation Methods - Tools for Computational Biochemistry

The accurate prediction of enzyme kinetics from first principles is one of the central goals of Computational Biochemistry. Possible applications include the development of mutant analogues of natural ocurring enzymes, with improved reactivity and/or selectivity. Currently there is considerable debate about the applicability of Transition State Theory (TST) to compute rate constants of enzyme-catalyzed reactions. Classical TST is known to fail for some cases, but corrections can be added to include the effects of dynamical recrossing and quantum mechanical tunneling, as discussed in Section 2.4.1. Nevertheless, the framework of TST has been heavily disputed, particularly on the possible role of protein dynamics and conformational effects on the enzyme activity. The following is taken from a Feature Article in the Journal of Physical Chemistry *B* by Prof. Martin Karplus:[88]

> *"Simple behavior [is] defined by two (related) aspects of reactions. The first is that a simple phenomenological rate law with an exponential time dependence for the rate applies and the second is that the temperature dependence of the rate follows the Arrhenius equation. We have seen that although simple behavior is found in some protein reactions, significant deviations from both types of simplicity have been documented and interpreted theoretically."*

Although some emphasis is given to the "complex" cases, the point is made that Arrhenius dependence can be seen even in such elaborate systems as enzymes. In the same article, three requirements are given for this "simple" behavior:

(1) it should be possible to define a reaction coordinate (or other progress variable),

(2) a well-defined barrier with a free energy several times $kT$ separating the reactant and product states should exist along the reaction coordinate,

(3) and the rate of the reaction (as defined by the reaction coordinate) should be slow compared to the elementary collisional events that lead to equilibration of the other degrees of freedom.

All three requirements are interconnected, since (2) and (3) will need the definition of a reaction coordinate, and (3) would be hard to fulfill without a high activation barrier. However, it is still quite hard to make a clear distinction between one and the other case, since all of the points are difficult to prove either theoretically or experimentally, and many arguments arise over the weight of complex behavior on the activated reactions.

The use of computational methods could help to establish whether TST is adequate for the quantitative treatment of enzymatic reactions. Comparison of the experimental Gibbs free energy of activation ($\Delta^{\ddagger}G$), assuming simple behavior, with the computed activation free energy should show the importance of other effects. However, the methods used to date for estimating activation energies are in general unable to give quantitative predictions. Quantum mechanical/molecular mechanical methods are normally used, but the size of the required QM region is often too large for accurate *ab initio* treatment of the active site. These regions include normally from 20 to 100 atoms, depending on the reaction under study. Semiempirical methods, though applicable to large systems, are generally not accurate enough because computed free energies of activation may have an error of ten or more kcal mol$^{-1}$. DFT offers improved accuracy but still lacks key physical interactions (e.g., dispersion). Often, DFT underestimates barrier heights by several kcal mol$^{-1}$, which cannot be systematically improved. Thus, when theoretical barriers do not agree with those from experiment, it is not clear whether the discrepancy arises from deficiencies in the electronic structure theory, in the experimental observations, or in the underlying theoretical framework of QM/MM and TST.

With the development of linear scaling local correlated methods, the amenable system sizes has grown significantly. One can now routinely treat up to 50 atoms with the density fitted LCCSD(T0) algorithm as implemented in `Molpro`. LMP2 calculations have also been reported for system sizes well above 100 atoms. Density-fitting approximations as well as explicit correlation terms can be used together with these methods in order to speed up the calculations with respect to basis set size, or to minimize basis errors in the barrier energies. The use of local methods could improve the quantum mechanical treatment of the active site, approaching a 1-2 kcal mol$^{-1}$ accuracy, as is nowadays routinely possible for small molecular systems. This would allow for a quantitative comparison between theory and experiment, checking the validity of TST.

In this Chapter, applications of local correlation methods for the computation of activation barriers in enzymes are presented. The two systems chosen seem to fulfill the requirements for simple behavior. The reaction coordinate is well defined and decoupled from other movements in the enzyme. The estimated free energy activation barriers are also estimated to be in the range of 13-15 kcal mol$^{-1}$, which fulfills requirement (2).

The same procedure will be followed as in the study of small molecular systems, with the exception of the conformational sampling which is later discussed. In estimating the free energy activation barrier at a given temperature $T$, we consider the different contribu-

tions individually

$$\Delta^{\ddagger}G(T) = \Delta^{\ddagger}E_0 + \Delta^{\ddagger}E_{\mathrm{ZPVE}} + \Delta^{\ddagger}H(0 \rightarrow T) + T\Delta^{\ddagger}S. \tag{6.1}$$

The first term is the electronic activation energy, which is equal to the electronic energy difference between the reactant and the transition state. The $\Delta^{\ddagger}E_{\mathrm{ZPVE}}$ term is the Zero-Point Vibrational Energy (ZPVE) correction, while $\Delta^{\ddagger}H(0 \rightarrow T)$ corrects for temperature effects[1]. The last term is the entropic contribution[2]. For a reaction where a single bond is broken and a new one formed, both $\Delta^{\ddagger}E_{\mathrm{ZPVE}}$ and $\Delta^{\ddagger}H(0 \rightarrow T)$ should be small, and of the order of 1-2 kcal mol$^{-1}$. If the reaction does not involve large changes to the solvation shell and the conformational liberty of the substrate is kept, the entropic correction should also be relatively small. The largest contribution in this case would be given by $\Delta^{\ddagger}E_0$, perhaps even an order of magnitude larger than the other terms. As such, most of the effort should be focused in reducing the error in the electronic energy. However, one should keep in mind that there are several documented cases where the reaction bottleneck is due to a diffusion barrier or an entropic effect. In these cases the discussion above will not apply.

Due to the flexible structure of the enzymatic environment, sampling of the conformational space is needed. This is done by computing a number of reaction trajectories, starting from different initial structures. These structures are obtained from molecular dynamics simulations, and will be referred to as *snapshots*. The activation barriers can then be averaged to obtain the final result. No weighting is used since the conformational space available will be small and this would lead to a strong bias for smaller barriers. Also, the only quantity to be sampled is $\Delta^{\ddagger}E_0$, since the other quantities should not vary much and, as discussed above, they will have smaller contributions[3].

## 6.2 The *p*-Hydroxybenzoate Hydroxylase enzyme

### 6.2.1 Overview

The *p*-Hydroxybenzoate Hydroxylase (PHBH) enzyme is a flavoprotein classified as monooxygenase which catalyzes the transformation of *p*-hydroxybenzoate (pOHB) to 3,4-dihydroxybenzoate (3,4-DOHB). It plays a major role in the oxidative degradation of aromatic compounds, being 3,4-DOHB the substrate for subsequent cathecol ring-cleavage

[1] This is calculated as $\Delta^{\ddagger}H(0 \rightarrow T) = \Delta^{\ddagger}H(T\mathrm{K}) - \Delta^{\ddagger}H(0\mathrm{K})$.

[2] This term is normally computed as the difference between the enthalpy and the Gibbs energy at a given temperature.

[3] The entropic estimate is taken from dynamic runs, so it does include conformational sampling, but not in the same way as $\Delta^{\ddagger}E_0$, since this effect is only included through the $\Delta^{\ddagger}G$ term.

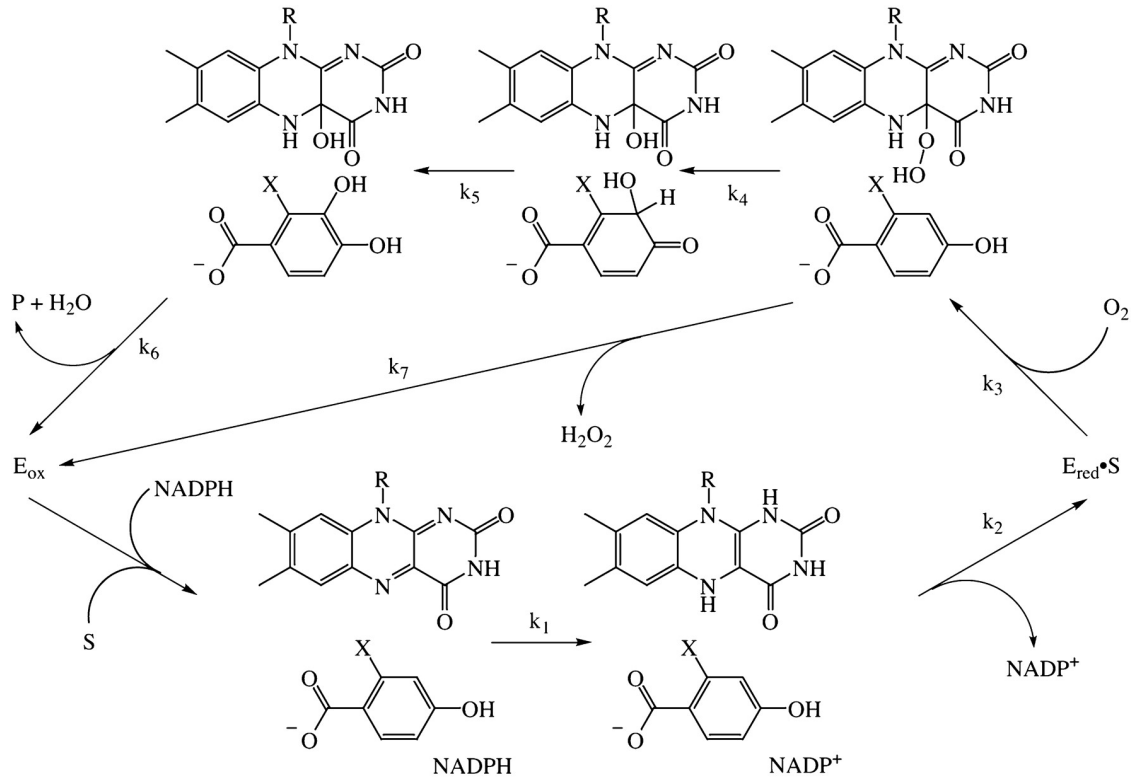


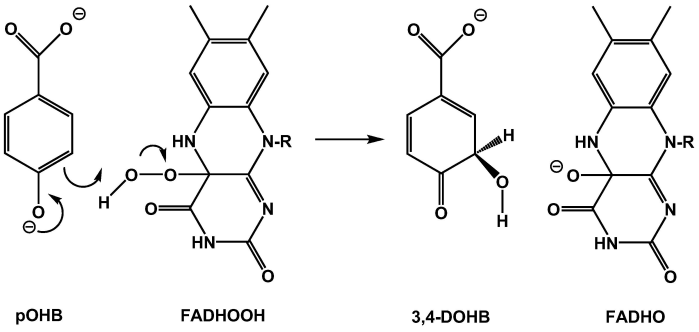

Figure 6.1: Catalytic cycle for the PHBH enzyme.

reactions. The catalytic cycle is shown in Fig. 6.1. It has also been proposed as a biocatalyst for the hydroxylation of fluorinated and chlorinated pOHB derivatives.

The enzyme activity is highest in the pH range 7.5-8.5,[89] where the rate-limiting step is the one depicted in Fig. 6.1 with rate constant $k_4$, and in further detail in Fig. 6.2. The flavin-adenine (FAD) cofactor is at this stage in its hydroperoxide form (FADHOOH), and serves as the active hydroxylation agent. pOHB is hydroxylated at the meta position, resulting in FADHO and a hydroxycyclohexadienone intermediate. It is known that the reaction follows the aromatic electrophilic substitution mechanism, with the FADHOOH cofactor acting as a formal "$OH^+$" donor. The substrate is believed to be in its dianionic (phenolate) form during the process, and both the phenolate as well as the resulting oxidoflavin should be significantly stabilized by a positive electrostatic potential in the enzyme pocket.

The activation enthalpy for this step has been estimated to be around 12 kcal $mol^{-1}$. This value was taken from temperature-dependent measurements of the overall rate between 277 and 298 K, at pH 8.[89] The authors mention that the Arrhenius plot[4] yielded a straight line, but give no further information on the diagram (or error bar). Other mea-

[4]In an Arrhenius diagram the $ln(k)$ is plotted against $1/t$. According to the Arrhenius equation, the slope of the line should be equal to $-\Delta^{\ddagger}H/R$.



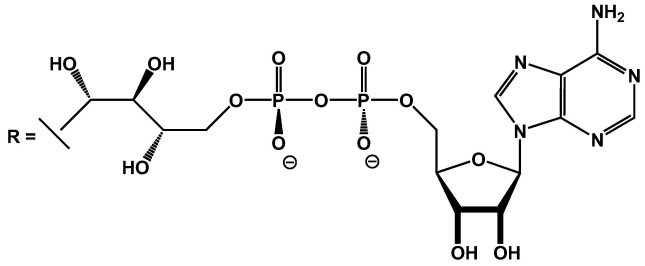




Figure 6.2: Hydroxylation step in the PHBH catalytic cycle, with pOHB as substrate. In the upper picture only the QM region is shown. The "-R" group in the cofactor represents the QM/MM crossing, and is substituted in the QM calculations by a link hydrogen atom. The remaining FADHOOH structure is depicted below.

surements provided rates of hydroxylation as well as turnover rates, which, converted to activation free energies $\Delta^{\ddagger}G$, give a rough estimate of 14-15 kcal mol$^{-1}$ for the hydroxylation step[5]. Both the enthalpy values as well as the Gibbs free energies will be later discussed.

Several theoretical works have also been dedicated to the study of the hydroxylation activation barrier. Already in 1992, a correlation was found between the HOMO energies of fluorosubstituted pOHB derivatives and the experimental turnover rates for the compounds.[90] These calculations were done in the gas phase, and only at the semiempirical AM1 level, but already supported the idea that the hydroxylation step could be rate-determining. The first QM/MM calculation of the system was made by Ridder *et al.*,[91] employing a combination of AM1 for the QM region and the CHARMM force field for the description of the enzymatic environment. A good correspondence was found with the earlier results of Vervoort *et al.*[90] However, the computed values for the activation barrier vary considerably according to the QM level used. The first predictions made at the AM1/CHARMM level gave a value of about 17 kcal mol$^{-1}$ for the enthalpy.[91] The HF results drastically overshooted to 30 kcal mol$^{-1}$, while the higher level corrections at the B3LYP and LMP2 levels gave 11-12 kcal mol$^{-1}$ [92] which compared well to the exper-

[5]According to Eq. (2.71), and setting the transmission coefficient to 1, the Gibbs energy can be estimated from the rate constant $k_4$ as $\Delta^{\ddagger}G = RTln\left(\frac{k_BT}{hk_4}\right)$.

imental estimate. This is however in disagreement with findings in the gas phase, which show that B3LYP and MP2 should underestimate the barrier.[93] In the same work, an AM1 prediction is given with extraordinary agreement to experiment, which is rather puzzling. Only in a recent paper by Senn *et al.* has this value been revised, and the converged AM1 QM/MM estimate was found to lie between 22-26 kcal mol$^{-1}$.[94] These variations are all linked to the size of the QM system treated, how solvation effects are included and the TS modeling approach. As such, there are some doubts about many of the theoretical results previously presented. Our proposal was to combine a well controled approach for the QM/MM treatment of the system with a converged *ab initio* result in order to provide a reliable theoretical estimate on the limit of todays computational tools.

### 6.2.2 Model Setup and Simulation

#### QM/MM Model

The QM region chosen is depicted in Fig. 6.2. It includes the pOHB molecule and the isoalloxazine part of the cofactor up to the first methylene unit of the side chain. The dangling bond was saturated in the QM region by an hydrogen link atom. This gives a total of 49 QM atoms. Previous theoretical works on the same system with larger QM regions showed only small deviations below 1 kcal mol$^{-1}$. The total system size is of about 23 000 atoms. This includes enzyme, substrate and cofactor (FADHOOH) as well as the aqueous solvent layer. Electrostatic embedding was used between the two regions. A charge-shift scheme was applied at the QM/MM boundary, meaning that the atom connecting the two regions will have its charge shifted to the nearest neighbors.[95] Further charges are also added between the atoms in order to correct the dipole moment[6].

#### Reaction Path Modeling

The QM/MM modeling was based on the X-ray structure of a PHBH-substrate complex obtained by Gatti *et. al.* (2.0 Å resolution).[96] Details on system preparation are given in Ref. [93]. The MM forcefield used throughout was GROMOS96,[97] a unified-atom force field.

After system preparation (minimization and MD runs protocols are given in the above cited reference) and equilibration, a MD run of 200 ps was performed in a cubic box of water molecules, from which six snapshots were taken in 40 ps intervals. Another four

[6]Charge neutrality is obtained by shifting the charge, but this leads to an artificial dipole. This can be compensated by distributing the charge in the neighboring region. Higher-order moments are not corrected, since they are considered to have a negligible effect.

snapshots were taken from AM1/MM MD simulations of the snapshots at 120 and 200 ps that were carried out in the course of free energy calculations using thermodynamic integration.[98] The snapshots will be numbered from 1-10 (i.e., 1-6 from MM MD after 0, 40, 80, 120, 160, 200 ps, and 7-8 from QM/MM MD at 120 ps, and 9-10 from QM/MM MD at 200 ps, respectively). The resulting structures were re-optimized at the B3LYP/MM level, including the QM region and all surrounding residues within a distance of 5 Å. The basis set used was TZVP.[99]

In order to ensure a continuous reaction path connecting TS and reactant structures, a reaction coordinate was defined as the difference between the breaking O-O and forming C-O bond lengths. This coordinate was then moved from the TS value up to a reactant minimum, relaxing all other geometry parameters (in the region detailed before) along the path. This part of the project was carried out by the group of Prof. Walter Thiel at the Max-Planck Institute in Mülheim.

### 6.2.3 The Hydroxylation Activation Barrier

The two main sources of error in an *ab initio* estimate of the activation barrier will be both the truncation of the n-particle and of the one-particle expansions. Relativistic (scalar and spin-orbit) effects should be negligible, since all atoms involved are at most second row elements. From small molecule calculations, it is known that the CCSD(T) method can give reaction barrier predictions to within 1 kcal $mol^{-1}$ of the FCI result. With this error estimate in mind, the CCSD(T) method with a complete basis set (CCSD(T)/CBS) will be from now on taken as a reference value. By approaching this result with the use of a local density fitting coupled cluster algorithm, the sources of error to be considered are

(1) the basis set error. The LCCSD(T0) estimate for a certain basis set should be corrected for the finite expansion used. This may involve extrapolation procedures or the use of explicitly correlated methods.

(2) the local approximations. There are several involved, each of these are to be later discussed:

   (a) domain approximations

   (b) pair and triples list approximations

   (c) the non-iterative triples (T0) approximation

(3) the density fitting approximation. This error is estimated to be in the order of 0.1 kcal $mol^{-1}$, and it will therefore be neglected.

Table 6.1: HF and MP2 computed barrier heights (in kcal mol$^{-1}$) for the 0 ps reaction pathway. All reported values include the effect of the MM environment, and were calculated as the energy difference between the B3LYP/def-TZVP pre-optimized reactant and transition state structures.

| | $\Delta^{\ddagger}E_0$ | |
|---|---|---|
| Basis Set | HF | MP2 |
| cc-pVTZ | 35.6 | 12.1 |
| [aug]-cc-pVTZ[a] | 35.8 | 11.8 |
| aug-cc-pVTZ | 36.0 | 11.7 |
| aug-cc-pVQZ | 36.0 | 12.1 |
| DF-MP2-F12[b] | | 12.2 |

a) diffuse functions only on O atoms

b) MP2-F12/**2**$^{*}$A(loc)/aug-cc-pVTZ correction (see Ref. [100])

The following Sections discuss the magnitude of the above mentioned error sources. All of the following calculations were performed with density fitting approximations, and the "DF" prescript will be dropped.

### Basis set Error

To estimate the effect of basis set truncations on the reaction barriers, MP2 calculations were performed for one of the reaction paths. The effect is expected to be similar for all the paths, as the major differences between them lie in the conformational space scanned (to which the basis set should be more or less insensitive). The MP2 values should be enough to give a good estimate of the basis set effect at the coupled cluster level, since CCSD(T) is known to have a similar basis set dependence (or even lower) as Møller-Plesset perturbation theory. The results are shown in Table 6.1.

The convergence of the barrier heights with the basis set size seems to be relatively fast. The cc-pVTZ result is remarkably close to the best estimates (within 0.1 kcal mol$^{-1}$). It seems to be an error compensation effect, due to the use of a small basis and the lack of diffuse functions. The inclusion of diffuse basis functions is, however, recommended due to the anionic nature of the substrate, since they help to describe the diffuse electron cloud of the oxygens. The cc-pVTZ basis with addition of diffuse functions on the oxygens will be denoted as [aug]-cc-pVTZ, and this was the choice for the local coupled cluster calculations. The basis set error is therefore estimated to be around 0.4-0.5 kcal mol$^{-1}$. Significant improvements would only be possible by increasing the basis up to quadruple-zeta quality.

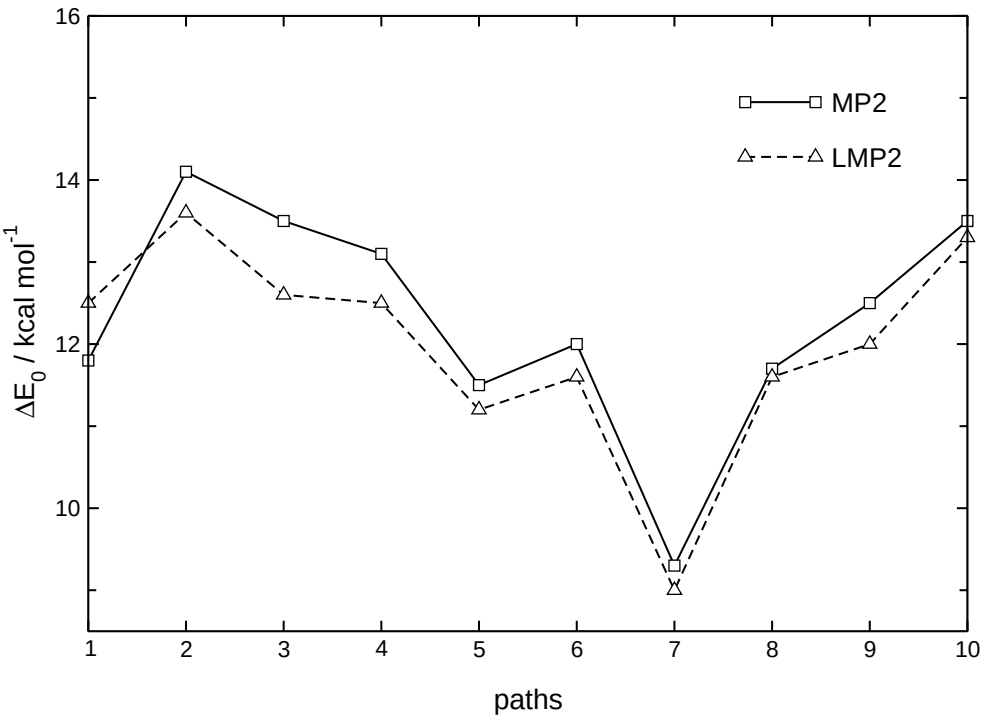



Figure 6.3: LMP2/[aug]-cc-pVTZ and MP2/[aug]-cc-pVTZ computed barrier heights (in kcal $mol^{-1}$). All values are taken from QM/MM single point calculations.

### Local approximations

As already discussed in Section 2.1.4, several approximations are used in the local coupled cluster program. The first source of error to be inspected is the truncation of the virtual space through the use of domains. For this purpose, MP2 and LMP2 calculations were performed for all reaction paths. As already discussed in Chapter 3, the effect should be similar for LMP2 and LCCSD(T0). The results are shown in Fig. 6.3.

The local values underestimate on the average (compared to the canonical counterparts) the reaction barrier by about 0.3 kcal/mol. This is within the expected accuracy for a triple zeta basis set. Test calculations for LMP2 with merged domains (reactant and transition state) give a barrier average of 0.6 kcal $mol^{-1}$ above the canonical estimate. Since there is no domain unbalance in this case, it seems that there is a furtuitous error compensation in the LMP2 result. By merging the domains, the transition state has its energy no longer artificially lowered by its larger domain list. This effect (+0.9 kcal $mol^{-1}$) appears to partly cancel with BSSE. The best local result is the one using the original domains, and this was

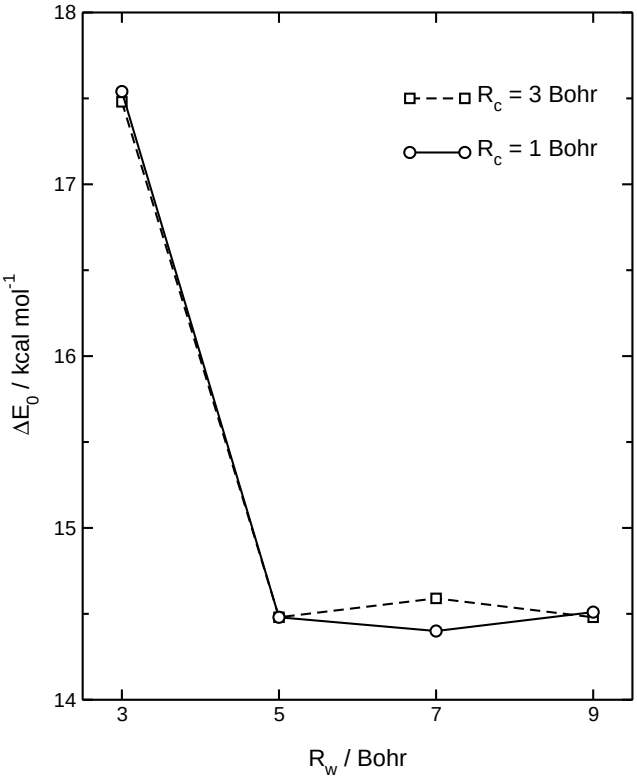



Figure 6.4: Activation barrier energies (in kcal mol$^{-1}$) computed at the LCCSD(T0)/[aug]-cc-pVDZ level as a function of the local distance criteria for weak and close pairs. The domains used were calculated with the [aug]-cc-pVTZ basis set.

chosen also for the LCCSD(T0) calculations.

Regarding the pair approximations in the local coupled cluster program, test calculations were performed using different distance criteria for the pair classification. Two parameters were under study, $R_c$ and $R_w$. As previously discussed, the former defines the maximum distance between orbitals centers for which a pair can still be classified as *strong*. The orbital pairs with distances between the two parameters are classified as *close* and will influence the triples list. Furthermore, close pairs are included in the triples calculation by the use of LMP2 amplitudes. Results are shown in Fig. 6.4.

The first fact one can notice in the graph is that for a given value of $R_w$, the coupled cluster estimate is almost independent of the choice of $R_c$. Increasing the value of $R_c$ from 1 to 5 Bohr (not shown in Fig. 6.4) only changes the activation energy value in about 0.3-0.4 kcal mol$^{-1}$. This indicates that the approximation of computing pair contributions with LMP2, instead of LCCSD, is of little importance. However, the triples pair list seems to be a determining factor in the accuracy. The difference between $R_w = 3$ Bohr and $R_w = 5$

Bohr is quite large. This can be easily explained by considering the reactant and transition state structures. The OH which is transferred in the course of the reaction is found in the reactant complex still bounded to the FADHOOH. In the transition state, it is about 4 Bohr apart from both substrate and cofactor. If the values for $R_w$ are below this distance, the interaction of the OH moiety with the rest of the system will not be treated at the triples level, and the correlation estimate wil be unbalanced. Since the effect can be captured by solely changing $R_w$, this parameter was set to 7 Bohr, while the default value of $R_c = 1$ Bohr was kept [7].

The non-iterative triples (T0) approximation was tested by calculating for one reaction path the converged triples solution. The difference was found to be below 0.1 kcal mol$^{-1}$ and therefore negligible.

In conclusion, the errors associated with the use of the DF-LCCSD(T0)/[aug]-cc-pVTZ level of theory in comparison to CCSD(T)/CBS are individually lower than 0.5 kcal mol$^{-1}$ and should not exceed more than 1.0 kcal mol$^{-1}$ in total. However, one should remember that this estimate is given only for the QM treatment. The MM modeling and QM/MM setup will lead to further errors which have not been accounted for. The reaction path optimization has also been performed at a lower level of theory (B3LYP/TZVP), which should, however, be adequate for the reaction in study.

### Results

The hydroxylation barrier height was computed as the average of ten reaction paths. The $\Delta^{\ddagger}E_0$ values, averages and root mean square deviations are given in Table 6.2. All values shown are QM/MM results (including MM relaxation terms).

[7] At the time, due to convergence problems in the local coupled cluster program, it was not possible to compute all of the values featured in Fig. 6.4. Only values for $R_w = 3,5$ and 7 were available. The series seemed to indicate an almost logarithmic convergence, and the highest value was taken. Now it would seem that a choice of $R_w = 5$ would be more adequate, but both values are equally close to our best estimate.

Table 6.2: Activation barrier energies (in kcal $mol^{-1}$) calculated at different levels of theory and with the [aug]-cc-pVTZ basis set. The results do not include ZPE correction. The LCCSD + (T0)/DZ values correspond to the LCCSD/[aug]-cc-pVTZ results with added triples correction calculated with the smaller [aug]-cc-pVDZ basis (but with the same domains as in the larger basis).

| | HF | B3LYP | MP2 | LMP2 | SCS-LMP2 | LCCSD | LCCSD + (T0)/DZ | LCCSD(T0) |
|---|---|---|---|---|---|---|---|---|
| 1 | 35.8 | 8.6 | 11.8 | 12.5 | 14.8 | 20.4 | 14.1 | 13.8 |
| 2 | 41.1 | 11.6 | 14.1 | 13.6 | 16.4 | 24.0 | 16.6 | 16.1 |
| 3 | 38.6 | 10.1 | 13.5 | 12.6 | 15.2 | 23.2 | 16.0 | 16.2 |
| 4 | 39.2 | 10.3 | 13.1 | 12.5 | 15.2 | 22.7 | 15.4 | 15.9 |
| 5 | 35.8 | 8.4 | 11.5 | 11.2 | 13.7 | 19.9 | 13.4 | 13.4 |
| 6 | 38.7 | 10.1 | 12.0 | 11.6 | 14.4 | 21.5 | 14.5 | 14.4 |
| 7 | 32.1 | 6.9 | 9.3 | 9.0 | 11.4 | 16.8 | 11.1 | 10.9 |
| 8 | 37.7 | 9.1 | 11.7 | 11.6 | 14.3 | 20.8 | 14.0 | 14.0 |
| 9 | 39.8 | 10.7 | 12.5 | 12.0 | 14.8 | 22.3 | 15.1 | 14.5 |
| 10 | 41.2 | 11.2 | 13.5 | 13.3 | 16.1 | 23.6 | 16.3 | 16.4 |
| average | 38.0 | 9.7 | 12.3 | 12.0 | 14.6 | 21.5 | 14.6 | 14.6 |
| RMS | 2.6 | 1.4 | 1.3 | 1.2 | 1.3 | 2.1 | 1.5 | 1.6 |

Taking the LCCSD(T0) values as reference, HF clearly overestimates the activation barriers, as previously discussed. On the other hand, MP2 (local and canonical) results underestimate the barrier, just as B3LYP. These values show in fact a common tendency observed in small molecular systems and confirm the observations made in Ref. [93]. The MP2 method introduces a correlation correction to the HF estimate which due to its perturbative nature overcorrects. The Coupled Cluster methods, LCCSD and LCCSD(T0), show a more univocal type of convergence. The former method still overestimates the barrier height by about 7 kcal $mol^{-1}$. A further comment should be made on the two remaining columns of Table 6.2.

The *spin component scaled* LMP2 method (SCS-LMP2) is an empirically corrected MP2 approach first introduced by Grimme.[16] Since in HF theory the movement of parallel spins is "correlated" due to the Pauli Principle (the two electrons are not allowed to occupy the same orbital), the correlation energy will be different for parallel and antiparallel spins. By introducing two empirical factors which scale both contributions it is possible to obtain more balanced MP2 results. This is reflected in the values in the Table, with the SCS-LMP2 and LCCSD(T0) averages in agreement to within 0.1 kcal $mol^{-1}$. The SCS-LMP2 calculation, however, is performed with only a fraction of the cost of the CC calculation. In this case, the computational time is about an order of magnitude lower.

Another approximation tested in the course of this work was the use of a smaller basis set to obtain the triples correction (T0). These calculations were performed with the cc-pVDZ basis set, with diffuse functions added to the oxygens. It will be refered to as [aug]-cc-pVDZ. In order to minimize the effect of domain approximation errors, the domains calculated with the [aug]-cc-pVTZ basis were used. LCCSD(T0)/[aug]-cc-pVDZ calculations were performed, and the triples correction added to the LCCSD/[aug]-cc-pVTZ values. These results are depicted in Table 6.2 with the denomination LCCSD + (T0)/DZ. The individual values are somewhat different, with a maximum absolute deviation of 0.5 kcal $mol^{-1}$, but agree on the average with the triples result using the larger basis set.

The activation enthalpies and free energies are obtained by adding the missing contributions listed in Eq. (6.1). The simulations carried out in Mülheim provided a zero point energy correction of $-1.1$ kcal $mol^{-1}$ and an enthalpic temperature correction of $-0.2$ kcal $mol^{-1}$. Entropic corrections were only computed using AM1 for a gas phase model and amount to 0.4 kcal $mol^{-1}$. Some selected computed values are shown in Table 6.3, together with the experimental estimates.

The results show how important the treatment of correlation is for the activation barrier value. The only result in agreement with the experimental estimates is the LCCSD(T0)/[aug]-cc-pVTZ, taking into account the error bars of Table 6.2. In fact, the

Table 6.3: Activation barrier enthalpies and free energies computed at 300 K, and collected experimental estimates. The values in parentheses are the root mean square deviations for the 10 computed paths.

| | $\Delta^{\ddagger}H$ | $\Delta^{\ddagger}G$ |
|---|---|---|
| HF | 36.7 (2.6) | |
| B3LYP | 8.4 (1.4) | |
| LMP2 | 10.7 (1.2) | |
| LCCSD(T0) | 13.3 (1.5) | 13.7 (1.5) |
| experiment | 12.0 | 14-15 |

root-mean square deviation of the local coupled cluster results is larger than the error estimates for the QM treatment discussed in the previous subsections. If one would be interested in increasing the precision of the result, a broader sampling should be used. Only then it would be reasonable to improve the QM method.

The remaining thermodynamic correction terms may become a determining factor of the accuracy, if both the sampling and the QM method would be improved. Even if computationally much more demanding than determining the static activation energies, there is still some room for improvement. The enthalpic and ZPVE corrections need two optimized structures (reactant and TS) and the respective Hessians. By todays computational standards it would be impossible to perform these calculations at the coupled cluster level. Since no analytical gradients are available for the local CC algorithm, this would involve $49 * 3 * 2 = 258$ single points to obtain a numerical gradient, needed at each optimization step. A parallel code running on a large cluster could perform this task, but the Hessian calculation would be an impossible challenge. Another possibility would be to use the SCS-LMP2 method, which has readily available analytic gradients. The optimizations could be performed on a single computer, and the Hessian run could be split into parallel calculations. Considering the results in Table. 6.2, this should be a good approximation to the CC result. The entropy correction would have to be tackled in a different way. Since this correction is obtained by computing the difference between the Gibbs energy and the enthalpy, one would need to do an Umbrella Sampling[101] or Thermodynamic Integration[98] with the higher level methods in the QM region. However, this involves thousands of single point calculations, and even using HF it would already be at an enormous cost. But again there are ways by which the cost could be reduced, delivering still something close to the higher level result. One could compute some single points along the reaction path at the higher level and take the difference relative to a lower level (e.g., semiempirical). A contin-

uous function can then be obtained by interpolating these differences, and added to the total energy at each point of the run.[102, 103] This will simulate the higher level run, as long as a sufficient number of points have been used and the low level to high level difference is kept relatively stable.

## 6.3 The Chorismate Mutase enzyme

### 6.3.1 Overview

The Chorismate Mutase (CM) enzyme catalyzes the Claisen rearrangement of chorismate to prephenate, a key step in the shikimic acid pathway that produces aromatic amino acids. It has been the object of extensive experimental and computational research, in part due to its biological significance, but also for being a rare example of an enzymatically catalyzed reaction which keeps the same mechanism in various solvents as well as in the enzyme environment. The chemical step is believed to be largely rate-limiting in *Bacillus subtilis* CM (BsCM), and catalysis proceeds without covalent binding of the substrate to the enzyme. This makes BsCM a particularly convenient target for QM/MM studies, which have focused on aspects such as the structure of the enzyme-substrate complex, reaction pathways, and the role of active-site residues in TS stabilization. However, almost all of the previous reaction modeling has been carried out by using semiempirical or DFT methods, which do not predict barrier heights with chemical accuracy.

This particular CM species has sparked for some years a vivid debate in the community. The $2x10^6$-fold reaction rate enhancement over the uncatalyzed process in water was early on connected to reactant deformation.[104] The enzyme pocket would trap the chorismate in a reactive conformation, in analogy to CMs found in other bacteria and plants. However, Hilvert and coworkers[105] found a strong enthalpic effect of about 8 kcal $mol^{-1}$, and an almost marginal entropic contribution. The Gibbs free energy of activation was estimated to be around 15.4 kcal $mol^{-1}$, and the enthalpy 12.7$\pm$0.4 kcal $mol^{-1}$. These values are to be compared to the ones in solution, 24.5 and 20.7$\pm$0.4 kcal $mol^{-1}$ respectively. Electrostatic stabilization of the TS, together with substrate conformational effects were put forward as an explanation for the strong catalytic effect.

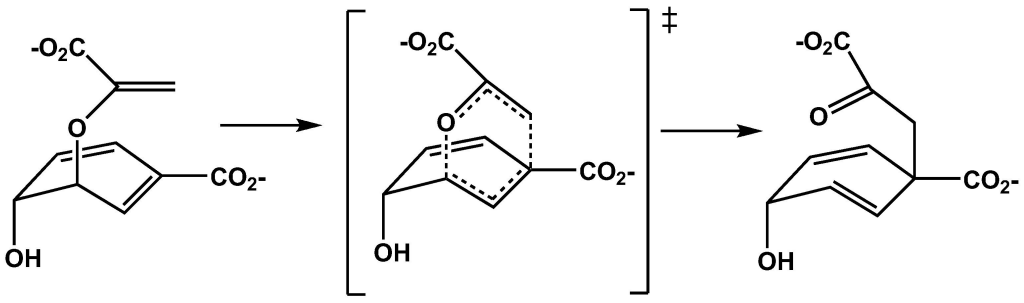



Figure 6.5: The chorismate to prephenate Claisen rearragement catalyzed by BsCM.

Evidence for TS stabilization has been given by several QM/MM studies. Lyne *et. al.*[106] recorded the effect of deleting neighboring amino acid residues on an approximate activation barrier. The level of theory used (AM1) was insufficient to deliver quantitative results on the total effect. However, significant contributions from positively charged residues were observed, in agreement with the TS stabilization proposal. The active site of the wild-type (WT) enzyme is represented in Fig. 6.6. Several positively charged residues are found in the pocket. The residue Arg90 is of special importance. It has close contacts to the breaking C-O bond, and a number of experimental and theoretical works have been dedicated to the analysis of this particular residue in the enzymatic activity. Kienhöffer and coworkers synthesized and studied a BsCM mutant with the important cationic residue Arg90 replaced with the non-coded amino acid citrulline (Cit).[107] The system has potential for a largely focussed alteration of TS effects, as the arginine is mutated to a neutral but isosteric analogue, proposed to form a similar but less stabilizing hydrogen bonding pattern with the substrate. The mutation resulted in a $10^4$-fold reduction in $k_{cat}$, or 5.9 kcal mol$^{-1}$ increase in the overall free-energy barrier. The $K_m$ registered a 2.7-fold increase. This small dissociation constant difference was taken as an evidence of relatively minor complex conformational distortion. The reduced efficiency is therefore interpreted as arising from unfavorable TS stabilization due to loss of the cationic nature of the stabilizing hydrogen bond donor. These results have been supported by theoretical QM/MM investigations of the mutant.[108]

On the other hand, Worthington *et. al.*[109] have argued that cationic stabilization occurs to an equal extent on the transition and reactant states. Another explanation was put forward by Hur *et. al.*[110]. The catalytic effect of BsCM would be due to the enzyme ability to preferentially bind near attack conformations (NACs) of the substrate. This theory is based on the premise that specific conditions are needed for the reaction to take place. The atoms must come together at a given distance and angle, and the enzyme would favor such conformations in comparison to solution. Free energies of NAC formation correlated well with the experimental $\Delta\Delta^{\ddagger}G$, but these results were only based on mole fractions. Mulholland and coworkers[111] later disputed these results, and calculated the NAC contribution with help of a free energy perturbation method. The effect was found to be two times smaller. This result came in support of the TS stabilization theory, since the NACs were insufficient to explain the 9.1 kcal mol$^{-1}$ difference in the Gibbs free energy. Also, the same effects that lead to preferential stabilization of the TS should also lead to a higher NAC population. A related work can be found in the same year, with new estimates for the stabilization effect of residues on the activation barrier.[112] The study of a related CM, as well as of a group of mutants shortly followed by the Bruice group[113], keeping a lively

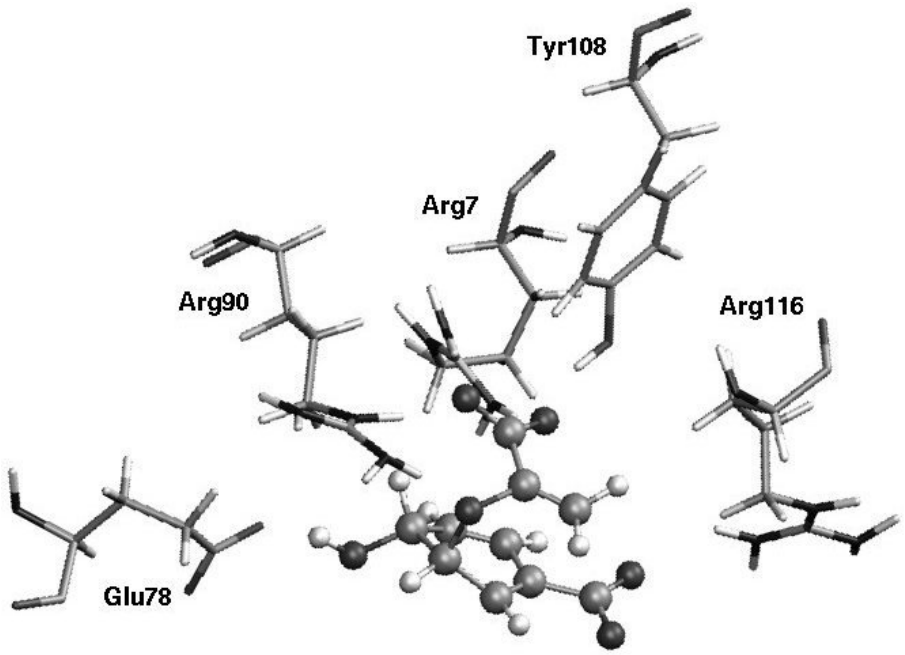



Figure 6.6: Active site of BsCM with substrate, and relevant neighboring residues. The QM region is represented as ball-and-sticks, MM part as sticks, with only the side chains of the residues shown.

discussion for more than 3 years.

In the study of the WT enzyme, my work was based on the modeling studies carried out at the Theoretical Chemistry Group in Bristol. In previous accounts, they had already presented a theoretical barrier height prediction on the basis of QM/MM DFT calculations.[114, 112] A study on the convergence of the QM treatment could solve many of the questions in debate. By calculating an activation barrier at higher levels of theory, the determining factor for catalysis could be identified.

### 6.3.2 Model Setup and Simulation

#### QM/MM Model

The CM enzyme system was taken from previous studies carried out by the Bristol group. The structure is derived from an X-ray structure of an enzyme-transition state analogue complex[8] (PDB code 2CHT).[115]

The QM region consists solely of the chorismate. There is no bonding between the substrate and the environment, and therefore no need to include link atoms. This choice

[8] A transition state analogue is a chemical species with structural properties close to the idealized TS. In this case, an endo-oxabicyclic inhibitor was used.

corresponds to a relatively small region, only 24 QM atoms, but it should be noted that the barrier height has been found to be insensitive to the size of the QM region.[116] The system is made of 7057 atoms in total. It comprises a sphere of radius 25 Å centered on the substrate, containing 4192 protein atoms, 947 water molecules and the substrate. The QM/MM interaction included electrostatic embedding.

For the Arg90Cit mutant, extra parameters were needed for the non-coded aminoacid Cit. Intramolecular interaction and vdW parameters were generated by analogy with those for standard CHARMM residues. Partial atomic charges were taken from the Guimarães *et al.* study,[108] and checked with QM treatment of the amino acid residue. The results compared well, with deviations lying below 1.0 kcal mol$^{-1}$ (which can also be caused by conformational effects).

#### Reaction Path Modeling

Again, a reaction coordinate was chosen as the difference of distances between the breaking and forming bond, evidenced in Fig. 6.5. From previous studies it was known that the TS is found at a a value of $r$= $\delta$(C-C) - $\delta$(C-O) $\approx -0.6$ Å. QM/MM molecular dynamic runs were performed with the BsCM complex, applying a restraint to the coordinate. The AM1 and PM3 semiempirical methods were used for the QM part, the molecular force field used was CHARMM27.[50] Snapshots were taken from 5 to 30 ps for each trajectory, in a total of 16 structures. This modelling work was performed by the Theoretical Chemistry Group in Bristol.[114]

The QM/MM interface was provided by in-house routines from the Bristol group[117] linked to the Jaguar quantum program[118] and the MM program package TINKER.[119] The point charge information was later translated to `Molpro` format. Electrostatic embedding was used throughout.

### 6.3.3 The Claisen Rearrangement Barrier

The same procedure was followed as in the PHBH case. To obtain a converged value for the Gibbs free energy of activation, a high-level result for the $\Delta^{\ddagger}E_0$ value was combined with lower level corrections. The same tests for basis set error and local approximation effects were carried out, and are detailed in the following sections.

#### Basis Set Error

Calculations were carried out at the MP2 level of theory for a series of basis sets. The results are shown in Table 6.4.

Table 6.4: HF and MP2 computed barrier heights (in kcal mol$^{-1}$) for one reaction pathway. All reported values include the effect of the MM environment, and were calculated as the energy difference between the B3LYP/6-31G* pre-optimized reactant and transition state structures.

| | $\Delta^{\ddagger}E_0$ | |
|---|---|---|
| Basis Set | HF | MP2 |
| aug-cc-pVDZ | 29.9 | 10.2 |
| [aug]-cc-pVTZ[a] | 31.1 | 12.3 |
| aug-cc-pVTZ | 30.9 | 11.8 |
| aug-cc-pVQZ | 31.0 | 12.2 |
| aug-cc-pV5Z | 31.0 | 12.3 |
| DF-MP2-F12[b] | | 12.3 |

a) diffuse functions only on O atoms

b) MP2-F12/**2**$^*$A(loc)/aug-cc-pVTZ correction (see Ref. [100])

Except for the aug-cc-pVDZ basis results, all other values are relatively close to our best results (aug-cc-pV5Z or the explicit correlated calculation). At the HF level the near CBS limit is already reached with the triple-zeta basis, the correlation correction is as usual more basis-set dependent. However, the [aug]-cc-pVTZ basis, which was also previously used in the PHBH case actually gives the near CBS result. This is probably an error compensation effect, due to some polarization near the oxygens (due to the diffuse functions added) and the basis set incompleteness. This was also the basis later used in our calculations.

### Local Approximations

The same series of tests were conducted in the CM case as for the PHBH hydroxylation step. The first effect to be under examination was the domain approximation. MP2 and LMP2 calculations with the [aug]-cc-pVTZ basis set were carried out. In the LMP2 runs both regular domains (recomputed at each structure), and merged domains were used (see Chapter 3). The results are plotted in Fig. 6.7 for 10 trajectories[9].

For this reaction the electronic structure changes significantly along the reaction coordinate. The orbitals are more delocalized at the TS and therefore the standard Boughton-Pulay procedure yields larger domains at the TS than for the reactants. This leads to an underestimation of the barrier height if standard domains are used. Fig. 6.7 shows the deviation between canonical and local barrier heights for 10 paths. The merged domains have

[9]The snapshots used for Fig. 6.7 were taken at 10, 12, 16, 18 and 20 ps, and are ordered accordingly. The odd numbers stand for AM1 paths, and the even numbers for the PM3 paths.

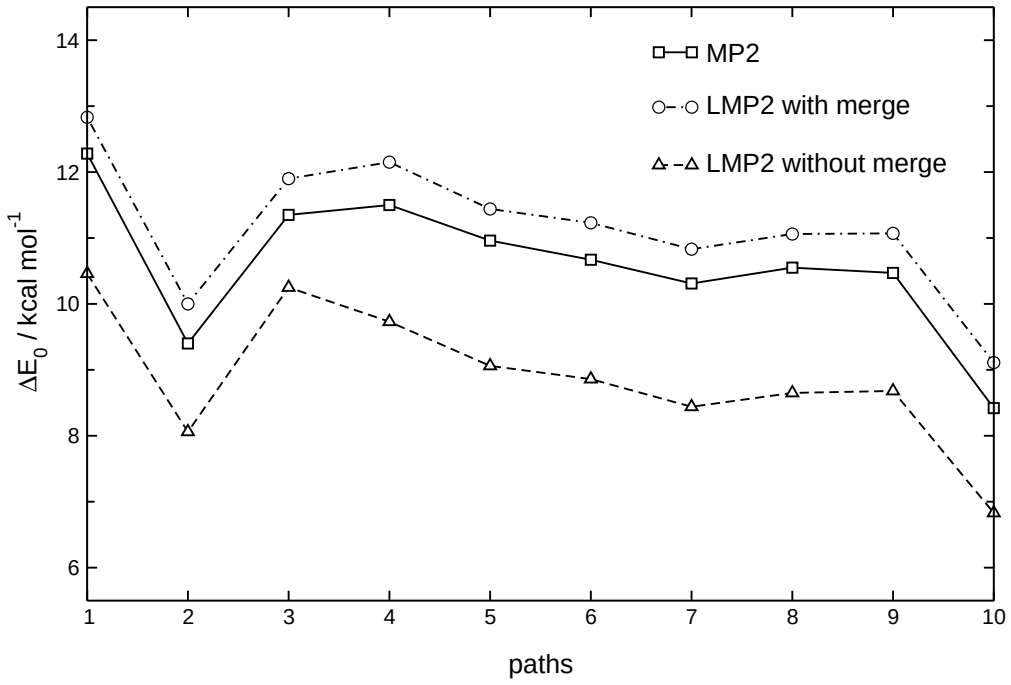



Figure 6.7: LMP2/[aug]-cc-pVTZ and MP2/[aug]-cc-pVTZ computed barrier heights (in kcal $mol^{-1}$). All values are taken from QM/MM single point calculations.

been determined separately for each path, using the reactant and transition state domains as reference. The resulting merged domains are appropriate to describe the whole reaction paths. The effect of the domain merging procedure is found to be very similar for all snapshots. When standard domains are used, the LMP2 barrier heights are about 1.4 kcal/mol lower than the canonical MP2 ones. On the other hand, the LMP2 barrier heights obtained with merged domains are about 0.5 kcal/mol higher than the canonical ones. Probably, this is again at least partly a BSSE effect, which artificially lowers the canonical MP2 barrier heights, as already discussed for the SN2 reactions in Chapter 3.

The pair approximations in the coupled cluster program were again tested. A reaction pathway was chosen (the snapshot at 38 ps, from the AM1/MM dynamics run) and several combinations of parameters were used with LCCSD(T0)/[aug]-cc-pVTZ. However, contrary to the PHBH case, no convergence was observed when increasing the $R_c$ and $R_w$ parameters in the 1-7 Bohr range. In order to effectively choose a combination which would reduce the pair approximation, the calculations were repeated with the [aug]-cc-pVDZ basis. The smaller basis allowed to increase the parameters further and even to completely

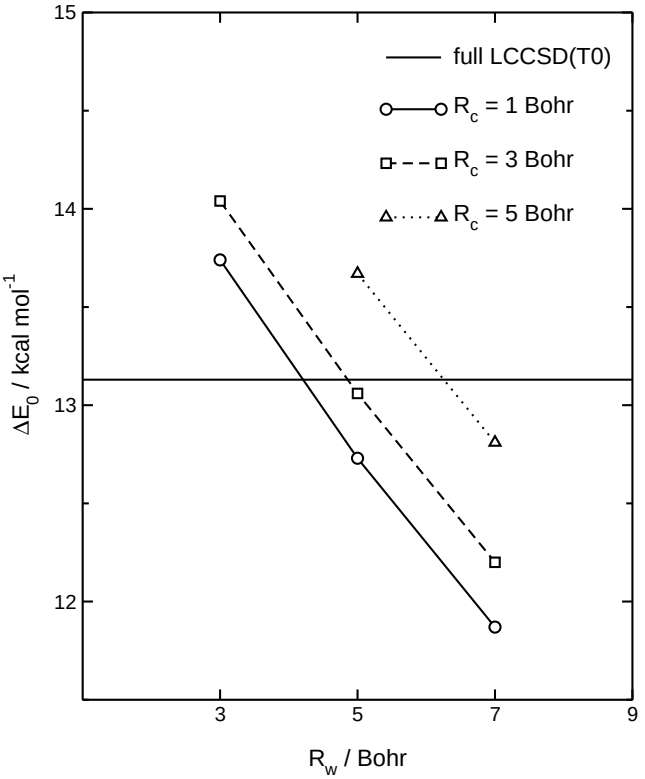



Figure 6.8: Activation barrier energies (in kcal $mol^{-1}$) computed at the LCCSD(T0)/[aug]-cc-pVDZ level as a function of the local distance criteria for weak and close pairs. The domains used were calculated with the [aug]-cc-pVTZ basis set.

turn off the pair approximation. The same procedure as in the PHBH case was applied. The triple-zeta domains were used for the double-zeta local calculations. The results are shown in Fig. 6.8.

The line tagged as full LCCSD(T0) represents the local coupled cluster result without any pair approximations. All pairs are treated at the higher level, and the triples list is full. There are two effects visible in the diagram. For a given $R_c$, increasing the value of $R_w$ leads to a lower value for the activation barrier. Fixing $R_w$ and incrementing $R_c$ leads to the opposite effect. The variations are much smaller than in the PHBH case. However, the fluctuations can be still as large as 1 kcal $mol^{-1}$. In Table 6.5, the LCCSD correlation contributions as well as the triples are shown for several different choices of distance criteria. The reasons behind the strange behavior of Fig. 6.8 become clear. Both values converge rather slowly, but from opposite directions. The LCCSD energy contribution converges from below, since treating pairs at the LMP2 level leads to an underestimation of the barrier. The (T0) contribution converges from above, since the triples will correct the

Table 6.5: LCCSD and LCCSD(T0) correlation contributions (in kcal mol$^{-1}$) to the activation energy ($\Delta E_0$) as a function of the distance criteria $R_c$ and $R_w$ (in Bohr). SD stands for the correlation energy difference between TS and reactant, as computed by LCCSD. (T0) gives the triples effect, and LCCSD(T0) the total correlation contribution. The ($\infty$,$\infty$) result is obtained without any pair approximations (all pairs treated by CCSD and full triples list). At the bottom, the (1,5) values are also shown (the distance criteria used for the final activation energy values). All results were computed with the [aug]-cc-pVDZ basis set, using the domains from [aug]-cc-pVTZ.

| ($R_c$,$R_w$) | SD | (T0) | LCCSD(T0) |
|---|---|---|---|
| (1,3) | -7.74 | -3.46 | -11.20 |
| (3,5) | -7.60 | -4.18 | -11.78 |
| (5,7) | -7.12 | -4.91 | -12.03 |
| ($\infty$,$\infty$) | -6.72 | -4.99 | -11.71 |
| (1,5) | -7.74 | -4.36 | -12.10 |

CC values, lowering the barrier. Both effects tend to cancel out, and it is actually easier to find distance parameters with a good error compensation, than converging both values. The best combination of parameters according to Fig. 6.8 (and Table 6.5) would be $R_c$=3, $R_w$=5 Bohr. However, the combination of $R_c$=1 and $R_w$=5 Bohr also compares quite well. The error is 0.4 kcal mol$^{-1}$ relative to the full coupled cluster result. Since the latter is computationally less demanding, and as the domain approximation error was about the same, but with opposite sign (+0.5 kcal mol$^{-1}$), these were the values chosen. Some error compensation is to be expected.

### Results

Single point calculations were carried out on the reactant and transition state structures for all 16 snapshots. The basis set used was [aug]-cc-pVTZ. The results are shown in Table 6.6 for each pathway individually, together with averages and root-mean square deviations.

Just as in the PHBH case, the LCCSD(T0) values will be taken as reference. The averages vary between 10-30 kcal mol$^{-1}$, again revealing the importance of choosing an adequate QM treatment. Just as before, HF drastically overshoots the barrier, inclusion of correlation lowers the value, but only SCS-LMP2 compares favorably with LCCSD(T0). LCCSD and LMP2 deviate by about 5 kcal mol$^{-1}$. The B3LYP value, which does not include ZPVE or temperature corrections compares well to the enthalpic experimental value of 12.7 kcal mol$^{-1}$. This favorable comparison had already been pointed out by Clayessens *et. al.*[114]. However, the comparison made in the above cited Communication is faulty, as it only takes into account electronic effects (at 0K). The zero point correction to the barrier

has been computed by performing B3LYP frequency calculations on six different optimized reactant and TS structures. The atoms included in the calculation were the substrate, 2 water molecules within 3 Å of chorismate and 4 hydrogen terminated residues close to the substrate (Arg7, Arg63, Arg90 and Glu78). The water molecules and the amino acids were frozen during the optimizations. Rows and columns of the Hessian corresponding to the frozen atoms were deleted prior to diagonalisation. The ZPVE correction to the barrier ammounts to -1.5 kcal $mol^{-1}$ and the average enthalpic temperature correction is -0.1 kcal $mol^{-1}$. Some selected theoretical values for enthalpies and free energies of activation are shown in Table 6.7.

Table 6.6: Activation barrier energies (in kcal mol$^{-1}$) calculated at different levels of theory and with the [aug]-cc-pVTZ basis set. The results do not include ZPE correction.

| | | HF | B3LYP | MP2 | LMP2 | SCS-LMP2 | LCCSD | LCCSD(T0) |
|---|---|---|---|---|---|---|---|---|
| AM1 | 8 ps | 32.5 | 14.2 | 11.8 | 12.4 | 16.9 | 21.4 | 16.2 |
| | 10 ps | 31.2 | 13.4 | 12.4 | 12.9 | 17.1 | 21.5 | 16.5 |
| | 12 ps | 31.3 | 13.4 | 11.5 | 12.0 | 16.3 | 20.7 | 15.7 |
| | 16 ps | 29.9 | 11.7 | 11.1 | 11.5 | 15.7 | 20.2 | 15.0 |
| | 18 ps | 29.3 | 11.3 | 11.0 | 10.9 | 15.1 | 19.6 | 14.4 |
| | 20 ps | 29.9 | 12.0 | 10.6 | 11.2 | 15.5 | 19.8 | 14.8 |
| | 30 ps | 24.7 | 8.1 | 9.5 | 10.0 | 13.7 | 17.6 | 12.9 |
| | 38 ps | 25.7 | 8.6 | 10.1 | 10.5 | 14.3 | 18.4 | 13.5 |
| PM3 | 8 ps | 31.8 | 13.6 | 11.3 | 11.9 | 16.4 | 20.8 | 15.6 |
| | 10 ps | 30.4 | 11.1 | 9.5 | 10.1 | 14.5 | 19.2 | 13.9 |
| | 12 ps | 32.8 | 14.9 | 11.6 | 12.3 | 16.7 | 21.1 | 16.2 |
| | 16 ps | 30.1 | 12.2 | 10.8 | 11.3 | 15.6 | 20.0 | 15.0 |
| | 18 ps | 29.6 | 11.7 | 10.7 | 11.2 | 15.3 | 19.7 | 14.6 |
| | 20 ps | 29.4 | 11.2 | 8.5 | 9.2 | 13.6 | 18.2 | 13.3 |
| | 30 ps | 30.4 | 11.3 | 10.4 | 10.9 | 15.1 | 19.9 | 14.6 |
| | 38 ps | 28.9 | 10.4 | 9.0 | 9.6 | 13.8 | 18.5 | 13.4 |
| average | | 29.9 | 11.8 | 10.6 | 11.1 | 15.4 | 19.8 | 14.7 |
| RMS | | 2.1 | 1.8 | 1.0 | 1.0 | 1.1 | 1.1 | 1.1 |

Table 6.7: Activation barrier enthalpies and free energies computed at 300 K, and collected experimental estimates. The values in parentheses are the root mean square deviations for the 10 computed paths.

| | $\Delta^{\ddagger}H$ | $\Delta^{\ddagger}G$ |
|---|---|---|
| HF | 28.3 (2.1) | |
| B3LYP | 10.2 (1.8) | |
| LMP2 | 9.5 (1.0) | |
| LCCSD(T0) | 13.1 (1.1) | 15.6 (1.1) |
| experiment | 12.7±0.4 | 15.4±0.4 |

The entropic contribution to the activation barrier ($T\Delta^{\ddagger}S$) at 300K has been computed by the Theoretical Chemistry Group in Bristol. Their estimate is 2.5 kcal mol$^{-1}$, to be compared with 2.7 kcal mol$^{-1}$ from experiment. Once more, the only theoretical result within the error bounds of experiment and the pathway sampling is LCCSD(T0). The enthalpy and Gibbs free energy of activation are even within the experimental error. Just as in the PHBH case, these results seem to validate the use of classical TST in such a developed system. Even more surprising is the dependence the values have on the QM treatment, and how this can be converged to reproduce the experimental estimate. This behavior parallels the one found for smaller molecular systems, and the high dimensionality of the problem seems to be well captured by a simple pathway sampling.

Including all corrections to the B3LYP value, it underestimates the reaction barrier by about 2 kcal mol$^{-1}$, a discredit to some of the previous DFT results.[111, 114] The final SCS-LMP2 value is not given, but by comparing the values shown in Table 6.6 it should also reproduce fairly well the LCCSD(T0) enthalpy. This confirms this scaled approach as a reliable improvement to the (L)MP2 method in the computation of activation barriers. Its use in the description of weak interaction forces has recently been discussed.[66]

# Chapter 7

# Summary

The aim of this research was to expand the scope of applications for local correlation methods by means of QM/MM and QM/QM hybrid schemes. The efforts have been fruitful. The QM/MM interface between Chemshell and `Molpro`, which was made available at the beginning of my PhD, has served several applications to date. The interface was coded in collaboration with Stephan Thiel (MPI Mülheim) and was based on an earlier implementation by Paul Sherwood (CLRC Daresbury). Other projects have started since, and the prospects are positive. The QM/MM calculations featured in this thesis have set a new standard for the QM treatment of enzymatic systems, and highlighted some deficiencies in earlier approaches to the description of the active site reactivity. To date, most of the published work in the field still makes use of computed barrier heights at the semiempirical or density-functional levels of theory, although the QM system sizes are similar to those featured in this work (20 to 100 atoms). It has been shown in this thesis that such molecular systems can be treated with high-accuracy through the use of local correlation methods. In particular, that static corrections to the barriers may suffice to significantly improve theoretical values, and with an inexpensive approach. Many of the tests performed to evaluate the errors in the QM treatment were carried out with lower level methods (MP2 for basis set convergence) or smaller basis sets ([aug]-cc-pVDZ for the pair approximations). This approach has the potential to be extended to much larger systems. Most of the tests could be carried for system sizes up to at least 100 atoms. Alternative (and cheaper) approaches have also been tested for obtaining the final barrier values, namely SCS-LMP2 or smaller basis sets for the triples correction. The error in the QM part of the calculation can be reduced to within 1 kcal $\text{mol}^{-1}$. The overall accuracy is thus no longer dependent on the active site treatment, but instead on the modeling and sampling techniques, which are significantly cheaper.

At the heart of this project was also the development of a local QM/QM hybrid scheme. Through the use of local orbital spaces, it has been shown that within a single calculation different levels of theory can be applied to specific parts of the molecule, according to their relevance to the reaction under study. The advantages of the LMOMO method over other proposed hybrid schemes are manifold. Polarization effects are implicitly included, since the HF calculation is performed for the whole system. The approach is flexible enough to allow for extra coupling terms between correlated regions, successfully including environments effects in the high-level region residual equations. However, only for very small-sized high-level regions were these contributions shown to be significant. The approach is also better suited for biomolecular structures, where aromaticity can play a critical role. The regions are defined as groups of orbitals, avoiding the need to cut through bonds and the errors associated with link atoms or frozen localized orbitals.

The method was tested by computing reaction energies for a model system (glycine peptide formation), barrier heights for a proton transfer reaction and an enzymatic reaction, as well by computing SN2 reaction barriers and dissociation energies of some hydrogen bound complexes. In the first three cases it was found that a simple 2-level scheme, like LCCSD(T0):HF or LMP2:HF, requires rather large correlation region sizes (typically 20-30 orbitals) to obtain converged results. However, much smaller high-level regions are sufficient in a 3-level scheme like LCCSD(T0):LMP2:HF; for the tested examples, it was enough to include the atoms directly involved in the reactions into the high-level region. In large systems this leads to a dramatic reduction of the computational effort, and one obtains results of LCCSD(T0) quality with a computational effort that is only slightly larger than that of the initial Hartree-Fock calculation.

The bottleneck in large-scale applications will therefore be the Hartree-Fock step. Even though linear scaling can in principle be achieved, the onset is rather late and the prefactor high. Therefore, for medium size systems, more important than linear scaling is to reduce the prefactor and the total cost of the HF in general. One possibility is to use local density fitting approximations as described in Ref. [84]. Further savings may be possible by using dual- or multiple basis set approaches. For instance, it may be sufficient to use a smaller basis set for the parts of the system that are distant from the correlated region. Even though this is not entirely unproblematic since unphysical polarization artifacts may occur, it should at least be possible to reduce the number of polarization functions. This possibility is still to be explored.

Most of the applications for which the hybrid schemes are best suited revolve around the description of chemical reactivity. The use of local correlation methods, however, has been often criticized for its use of geometry-dependent excitation spaces. The domains defining the virtual space for each occupied orbital may change in the course of bond breaking/formation phenomena, leading to discontinuities in the PES. This is a topic which is still approached in conference talks and several publications. However, in my opinion, some of the criticism is unfounded. In a recent paper, Subotnik *et. al.*[120] state that *"computational chemists cannot always optimize geometries with confidence according to the Pulay-Werner scheme. Geometries have certainly been successfully optimized when the domains are fixed and thus the potential energy surfaces are smooth; but for large and/or subtle changes in geometry, where the best domains are not obvious and should not be held static, geometric optimization is not very practical"*. On the evidence of previous publications and my own experience with local methods, this claim is excessive. The general procedure chosen for optimization with local methods is to use a geometry-dependent domain definition at first, and then to reoptimize the structure with fixed domains. This has

been proven in numerous instances to be a reliable procedure. In fact, it parallels the use of an integral-grid in density functional theory. The grid for integral calculation in DFT is also geometry dependent, and typically it will only be kept fixed for the last steps. These procedures are normally concealed in a black box-like treatment, and the same can be practiced in local methods. Nonetheless, several groups have focused on developing similar local correlation methods, while avoiding the discontinuity problem. These are based on advanced integral prescreening tools and/or different definitions for virtual and occupied spaces. The former methods have been shown to reliably decrease the cost of correlated calculations, but only for MP2 theory, and with small basis sets. The latter approaches also appear problematic with larger basis sets, due to linear-dependencies in the basis, or prescreening deficiencies. The local methods, as proposed by Saebø and Pulay, and further developed by Werner and Schütz, notwithstanding the PES discontinuity problem, are the most effective methods to date.

The question posed was how could one keep the efficiency of the method, while avoiding the problem of a noncontinuous PES. A merging procedure was proposed, which works by comparing domains of different structures, and building a domain definition that encompasses the changes in the sampled space. The method has been applied to several different reaction types, delivering smooth potential surfaces. Moreover, it has been shown that BSSE effects are reduced in local calculations, and this leads to better basis set convergence when computing barrier heights or weak interactions. The effect of the domain approximation was broadly investigated. It has been shown to be similar at the LMP2 and LCCSD(T) levels. It has hence been suggested to test the error of domain approximations by only comparing MP2 and LMP2 results. This has been applied to the study of both QM/MM systems featured in this work.

Since the computational cost of LMP2 and LCCSD(T) calculations increases, respectively, with the third and fourth power of the average domain sizes, the procedure does lead to an increment in the computational cost. However, for all systems under study, the changes were found to be localized at the atoms where the bond breaking/formation is taking place and the linear scaling behavior for larger applications should be maintained. On the other hand, the procedure is limited by the need to define *a priori* the sampling space to be scanned. Notwithstanding, the procedure has been useful for the calculation of reaction path potentials, and as a valid test for the domain approximation error.

The last subject in my work was the use of Natural Localized Orbitals as an occupied space for local correlation methods. While computing the various activation and reaction energies featured in this thesis, a recurrent problem was the basis set dependency of the domain definition. In investigating the effect of different basis sets, it was necessary to adjust

several parameters for the choice of domains, in order to keep the definition constant. This dependency is connected to both the localization procedure (Pipek-Mezey) and the domain definition (Boughton-Pulay). The solution pursued was to use the NLMO orbitals, which are known to be less basis set dependent, and have an associated charge population analysis, more stable than Mulliken or Löwdin. The method is not new (the natural localization algorithm has been proposed for over 20 years) nor is the idea of applying it to local correlation algorithms, Flocke *et. al.*[29] introduced NLCCSD in 2004. My implementation has shown promising results in combination with the Pulay local *ansatz*. By defining a unique parameter for the domain criterion, based on orbital populations, it has been shown that the NLMO orbitals give stable domains with respect to basis set changes. For a test set of 30 molecules not a single domain changed for all 6 basis sets used, even when including diffuse functions. The domains obtained by the new criterion are physically meaningful, with $\pi$-orbitals easily identifiable by their delocalized orbital populations. The degree of delocalization seems to correlate well with chemical intuition. By decreasing the basis set dependency of the domain definition, a further step is taken in approaching the local correlation methods of a model chemistry. Preliminary results reveal that the fraction of correlation energy recovered is not much affected and similar as with the previous methods used. Therefore, it can be expected that previous conclusions regarding the accuracy of local correlation methods will not be much affected. Further systematic studies of reaction energies are in progress, and the initial results support this assertion.

The greatest liability in the use of NLMOs is perhaps in their use for gradient calculations. For the computation of a local correlation method analytic gradient, a minimization criterion is needed. This is available for Pipek-Mezey, but not for NLMOs. Our proposal has been to use the Pipek-Mezey orbitals together with an NPA analysis domain criterion. The greatest fault in the choice of these orbitals is the redundancy problem which occurs when using diffuse or higher-angular momentum basis functions. However, has discussed in Chapter 4, this can be overcome by eliminating some functions from the localization criterion, without great loss to the stability of the method.

Overall, the research presented in these pages has given a strong contribution to a more widespread and reliable use of local correlation methods. By the use of hybrid schemes, the application of these methods has been drastically expanded. During my first PhD year, the possibility (or even the interest) of using Coupled Cluster in the context of enzymatic reactions was disputed in several occasions. Two years later, the first work was published, rewarding the insight and dedication of many involved in the project. I believe that local correlation methods will have a long lasting impact in the field, and am overjoyed to have participated in its first steps.

# Chapter 8

# Zusammenfassung

Die Berechnung von Reaktionsbarrieren molekularer Systeme ist eine der Hauptherausforderungen der Theoretischen Chemie. Reaktionsbarrieren sind von besonderer Relevanz für das Verständnis und die Voraussage von Katalysephänomenen, ebenso wie für die Rationalisierung unserer Kenntnisse der individuellen chemischen Reaktivität im Rahmen der "General Transition State Theory".

Die Hartree-Fock (HF) Näherung ist der Ausgangspunkt für die meisten quantenchemischen Methoden. Sie berücksichtigt die Coulomb Wechselwirkung der Elektronen nur im Mittel, nicht jedoch die Elektronenkorrelation, d.h., die unmittelbare Coulomb Wechselwirkung zwischen den einzelnen Elektronen. Dieser Effekt ist bei der Beschreibung chemischer Reaktivität von großer Bedeutung, da es während eines Bindungsbruches oder einer Bindungsbildung zu starken Änderungen bei der Wechselwirkung der Elektronen kommt. Die HF Methode weist daher normalerweise Fehler im Bereich von 100-500% für Reaktionsbarrieren auf. Die näherungsweise Behandlung der Elektronenkorrelation durch Dichte-Funktional-Theorie (DFT) hingegen stellt eine kosten-günstige Art dar, um diese Effekte zu berücksichtigen. Die Qualität der Ergbnisse hängt jedoch stark davon ab, welche Parametrisierung bei den Funktionalen verwendet wurde. Kein Funktional liefert bisher gleich gute Ergebnisse für alle chemischen Systeme. Konventionelle post-HF *ab initio* Methoden hingegen bieten einen Weg die Ergebnisse systematisch zu verbessern. Der steile Anstieg der Rechenkosten mit wachsender Molekülgröße erlaubt jedoch nur quantitative Berechnungen für kleine Systeme (bis maximal 15 Atome).

Lokale Korrelationsmethoden vermeiden die steile Skalierung konventioneller, kanonischer Methoden durch Verwendung einer lokalen Basis zur Beschreibung des besetzten und des virtuellen Raumes. Die Zahl der Anregungen vom besetzten in den virtuellen Raum kann mittels Entfernungskriterien bezüglich der besetzten Orbitale begrenzt werden.[1] Dies ermöglicht eine hierarchische Behandlung der Korrelation von Elektronen. Dabei werden nahebeieinanderliegende Orbitale mit genaueren Methoden behandelt und entfernte Orbitale vernachlässigt.[2] Lineares Skalierungsverhalten wurde bereits für lokale Møller-Plesset Störungstheorie zweiter Ordnung (LMP2),[3] lokales Coupled-Cluster mit Ein- und Zweifachanregungen (LCCSD) [4] und LCCSD(T0) mit störungstheoretischer Berücksichtigung der Dreifachanregungen[5, 6, 7] nachgewiesen.

In dieser Doktorarbeit sind entscheidende Fortschritte bezüglich der Anwendbarkeit lokaler Korrelationsverfahren bei der akkuraten Berechnung von Reaktionspfaden und -barrieren erzielt worden. In Kapitel 3 und 4 werden die Fortschritte bei der Wahl der Methode für die Domänenberechnung und bei der Lokalisierung der besetzten Orbitale beschrieben. In Kapitel 5 ist eine neue QM/QM-Hybridmethode (Quantum Mechanics/Quantum Mechanics) und ihre Anwendung auf biomolekulare Systeme beschrieben.

In Kapitel 6 werden die Berechnungen der Aktivierungsenergien zweier enzymatischer Reaktionen vorgestellt. In diesen Rechnungen wurden lokale Methoden erstmalig auf QM/MM (Quantum Mechanics/Molecular Mechanics) Systeme angewendet.

**Kapitel 3**

Ein neues Verfahren zur Eliminierung der Geometrieabhängigkeit von Anregungsdomänen ("domains") wurde im `Molpro` Programmpaket [8] implementiert. Diese Abhängigkeit kann zu Unstetigkeiten in der Potentialfläche führen und stellt einen Nachteil lokaler Methoden bei der Berechnung von Reaktionspfaden dar.

Die Orbitaldomänen werden in der Regel durch ein von Boughton und Pulay vorgeschlagene Verfahren bestimmt.[9] Dabei werden zunächst alle Atome gemäß abnehmender Ladung (Mulliken oder Löwdin Gross Population) geordnet. Dann werden Atom für Atom, der Reihe nach, die Basisfunktionen von Atomen der Orbitaldomäne hinzugefügt, bis die Vollständigkeitsbedingung

$$1 - \int |(\phi_i - \hat{\phi}_i)^2| d\tau > T_{\mathrm{BP}}$$

erfüllt ist. Der Parameter $T_{\mathrm{BP}}$ ist ein Vollständigkeitskriterium und nimmt in der Regel Werte um 0.980 an. Die Versuchsfunktion $\hat{\phi}_i$ ist eine Linearkombination von Atomorbitalen (AOs), die sich in der Domäne befinden

$$|\hat{\phi}_i >= \sum_{A \in [i]} \sum_{\mu \in A} |\chi_\mu > \hat{L}_{\mu i}.$$

In den lokalen Korrelationsverfahren bestimmen die Orbital- oder Paardomänen den virtuellen Raum für jedes Orbital und Orbitalpaar. Änderungen in dieser Domänendefinition als Funktion der Geometrie führen zu unstetigen Potentialflächen. Weil das vorgegebene Kriterium geometrieabhängig ist, kann es durchaus passieren, dass Sprünge im Energieprofil vorkommen. Dies ist vor allem bei der Dehnung oder dem Bruch von Molekülbindungen der Fall. Auch die elektronische Struktur des Moleküls wirkt sich auf die Domänen aus. Wenn der Übergangszustand stärker delokalisiert ist als der des Reaktanten, kann dies zur Unterschätzung der berechneten Aktivierungsenergien führen.

Hier soll eine einfache Prozedur vorgeschlagen werden, um diese Probleme zu beheben. Die Domänen zweier Geometrien werden verglichen und anhand dieses Vergleichs neue Domänen definiert. Diese vereinigten Domänen können dann für mehrere Strukturen in einem Pfad verwendet werden. Es wird nachgewiesen, dass die Prozedur stetige Potentialflächen und stabilere Ergebnisse bei der Berechnung von Aktivierungsenergien

delokalisierter Übergangszustände liefert.

**Kapitel 4**

Die Orbitaldomänen in lokalen Korrelationsmethoden werden in der Regel, wie oben gezeigt, mit Hilfe des Boughton-Pulay Verfahrens bestimmt. Diese Methode zeigt jedoch eine starke Abhängigkeit vom verwendeten Basissatz. Die Reihenfolge der Atome wird anhand der Löwdin Ladungen bestimmt. Diese konvergieren mit steigender Basissatzgröße jedoch nicht zu einem Grenzwert. Für kleinere Ladungen erhält man in der Regel keine zuverlässige Vorhersage. Die Vollständigkeitsbedingung selbst ist ebenfalls basissatzabhängig. Für größere Basen ist es einfacher das Kriterium zu erfüllen. Der Parameter $T_{\mathtt{BP}}$ muss folglich der Basis angepasst werden. $T_{\mathtt{BP}} = 0.980$ für eine double-zeta Basis und $T_{\mathtt{BP}} = 0.985$ für triple-zeta sind häufig vorgeschlagene Werte.[10]

Auch das Lokalisierungverfahren ist in der Regel basissatzabhängig. Die Pipek-Mezey Methode[11] hängt von der AO Überlappmatrix ab, die für größere Basen, insbesondere mit diffusen Funktionen, lineare Abhängigkeiten aufweist. Im Benzolmolekül werden die Domänen, die die delokalisierten $\pi$-Orbitalen beschreiben sollen, oft zu groß, und schließen auch die benachbarten Wasserstoffatome mit ein.

Als Alternative zur Pipek-Mezey/Boughton-Pulay (PM/BP) Kombination zur Lokalisierung der besetzten Orbitale, bzw. zur Definition der Domänen, wurden die Natural Localized Molecular Orbitals (NLMO)[12] und die Natural Population Analysis (NPA)[13] vorgeschlagen. Die NPA liefert hierbei eine Ladungsverteilung besetzter Orbitale die als ein Kriterium für die Domänenbildung verwendet wird. Dieses führt zu stabilen Domänen für eine vielzahl verschiedener Bassissätze (von cc-pVTZ bis aug-cc-pVQZ). Hierbei wird kein Überlappkriterium benötigt, was die Basissatzabhängigkeit des Verfahrens deutlich verringert. Die Atome werden in der Domänenliste aufgenommen wenn ihre Ladungen den Parameter $T_{\mathtt{NPA}}$ überschreiten.

Zur Validierung der Methode wurde ein Satz aus 30 Molekülen gewählt. Dieser besteht aus typischen, kleinen, organischen Molekülen, gesättigten und ungesättigten Kohlenwasserstoffen sowie aromatischen Systemen. Es wird gezeigt, dass das neue Kriterium für die Domänenbestimmung sehr stabil in Bezug auf den Basissatz ist. Mit dem in dieser Arbeit empfohlenen Auswahlkriterium $T_{\mathtt{NPA}} = 0.05$ wurde bei allen sechs untersuchten Basissätzen keine einzige Änderung der Domänenstruktur beobachtet. Außerdem sind die Domänen physikalisch deutbar. Die $\pi$-Orbitale sind in der NPA Ladungsverteilung deutlich erkennbar. Der Prozentsatz der kanonischen MP2-Korrelationenergie, der bei LMP2 bisher erhalten wurde, bleibt bei der vorgeschlagenen Methode ähnlich. Das NPA Domänenkriterium kann auch bei PM-Orbitalen verwendet werden, so dass analytische Gradien-

ten weiterhin verfügbar sind.

**Kapitel 5**

Die Anwendung hochgenauer quantenchemischer Methoden ist oft aufgrund der Systemgröße nicht möglich. Die meisten untersuchten Effekte sind jedoch auf einen verhältnismäßig kleinen Teil des Systems beschränkt, nämlich das reaktive Zentrum. Diese kleine Region kann jedoch meist mit genauen lokalen Methoden behandelt werden. Der Effekt der Umgebung sollte aber nicht vernachlässigt werden, denn das reaktive Zentrum kann durch Polarisations- und sterische Effekte beeinflusst werden. Um das reaktive Zentrum und Umgebung beschreiben zu können, wurden in den letzten Jahre mehrere "gekoppelte" (Hybrid) Methoden entwickelt, wobei unterschiedliche Regionen mit verschiedenen Methoden behandelt werden können, je nach Anforderung an die Genauigkeit und je nach Systemgröße.

Im Rahmen dieser Arbeit wird ein neues QM/QM Kopplungsverfahren vorgeschlagen. Dieses stellt eine Erweiterung der lokalen Korrelationsmethoden dar. Da in den lokalen Methoden sowohl der besetzte als auch der virtuelle Raum lokal sind, können die Orbitale und die entsprechenden Domänen so gruppiert werden (Region), dass unterschiedliche Korrelationsmethoden angewendet werden können, ohne dass es des Rückgriffs auf Modellsysteme für Molekülfragmente oder des Bruchs von kovalenter Bindungen bedarf. Die Methode nutzt die Paarnäherung aus, wobei die Orbitalpaare als "strong" (starke), "weak" (schwache) oder "very distant" (sehr entfernte) eingeordnet werden, je nach der Region in der sich die Orbitale befinden. Die Orbitalpaare des reaktiven Zentrums werden als starke Paare klassifiziert und auf möglichst hohem Niveau (z.B. LCCSD(T)) behandelt, während die direkte Umgebung über die schwachen Paare auf MP2 Niveau beschrieben wird.

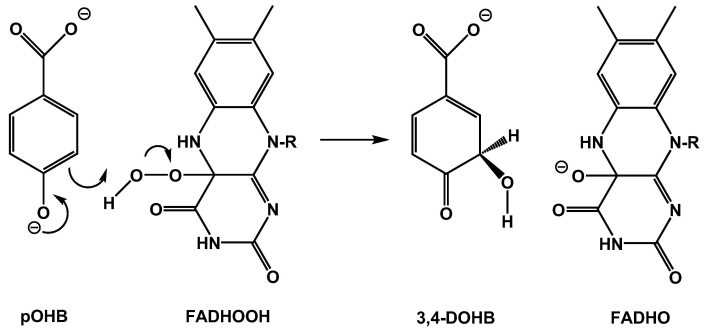
Da es in den lokalen QM Verfahren diese intrinsische Hierarchie gibt, können "gekoppelte" Ergebnisse verschiedener Regionen im Molekül aus einer einziger Rechnung extrahiert werden. Polarisationseffekte sind mit inbegriffen, da das Korrelationsverfahren die HF Orbitale für das gesamte System verwendet. Bei anderen Kopplungsmethoden ist dies nicht möglich, weil die "high-level" Rechnungen an einem kleineren Modellsystem durchgeführt werden müssen und dabei die Korrelationsbeiträge des Fragmentes von der Umgebung unbeeinflusst bleiben.

Rechnungen wurden für mehrere Reaktionen durchgeführt. Mit Hilfe der hier entwickelten LMOMO Methode könnte man in mehreren Fällen die "high-level" Ergebnisse mit reduziertem Rechenaufwand reproduzieren. Die notwendige "high-level"-Regionsgröße, um konvergierte Ergebnisse zu bekommen, ist aber systemabhängig, wenn die direkte Umgebung unkorreliert bleibt. Erkennbare Verbesserungen werden erhalten,

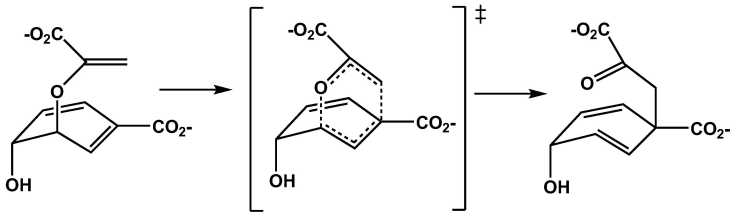
wenn die dem reaktiven Zentrum unmittelbar benachbarte Region auf MP2-Niveau korreliert wird. Das erfordert nur geringem Mehraufwand. Die Skalierung der Methode (für den Korrelationsanteil) kann theoretisch asymptotisch $\mathcal{O}(1)$ erreichen und wird für LMP2, die Lösung der Coupled-Cluster Gleichungen und die "triples" Berechnung bei LCCSD(T0) nachgewiesen.

**Kapitel 6**

Kombinierte QM/MM Methoden gehören heutzutage zum wichtigsten Handwerkszeug der Theoretischen Biochemie. Diese verwenden eine quantenchemische Rechnung zur Beschreibung des reaktiven Zentrums, während die Umgebung mittels einer kostengünstigen Kraftfeldmethode behandelt wird. Die QM-Region ist jedoch meistens zu groß, um eine quantitative QM Methode zu verwenden, und die Diskussion muss oft auf allgemeine Effekte und Tendenzen beschränkt bleiben. Die lokalen Korrelationsmethoden ermöglichen es aufgrund ihres reduzierten Rechenaufwands erstmals quantitative QM/MM Rechnungen durchzuführen. LCCSD(T) Rechnungen können heutzutage für mehr als 50 Atome mit einer triple-zeta oder sogar quadrupel-zeta Basis durchgeführt werden. Die vorliegende Doktorarbeit zeigt die erfolgreiche Anwendbarkeit lokaler Methoden auf Übergangszustände enzymatischer Systeme und beweist gleichzeitig die Gültigkeit der klassischen Übergangszustandstheorie bei den untersuchten enzymatischen Systemen.




Die Hydroxylierungsschritte im Katalysezyklus der *p*-Hydroxybenzoat-Hydroxylase.



Die katalysierte Claisen-Umlagerung von Chorismat zu Prephenat.

Zwei Systeme wurden untersucht. Die *p*-Hydroxybenzoat-Hydroxylase (PHBH) katalysiert die Hydroxylierung des Substrats *p*-Hydroxybenzoat (pOBH). Es spielt eine entscheidende Rolle beim oxidativen Abbau aromatischer Stoffe in Bodenbakterien. Das Chorismat-Mutase (CM) Enzym katalysiert die Claisen-Umlagerung des Chorismats in Prephenat, einem Schritt im Shikimisäureweg für die Produktion aromatischer Aminosäuren. Für beide Systeme liegen experimentelle Werte für die Aktivierungsenthalpie und Gibbs-Energie vor. Durch QM/MM Modellierung der Systeme wurden mehrere Reaktionspfade auf DFT-Niveau optimiert und die Vorhersage der Reaktionsbarriere mit Hilfe lokaler Korrelationsverfahren verbessert. Die Konvergenz mit steigender Basissatzgröße und die lokalen Näherungen wurden überprüft. Die berechneten Barrieren auf das DF-LCCSD(T0)/[aug]-cc-pVTZ-Niveau befinden sich nach Mittelwertbildung innerhalb der Fehlergrenze des Experimentes und der Rechnung (etwa 1,5 kcal $mol^{-1}$ Genauigkeit). Die aus der quantenmechanischen Beschreibung herrührende Abweichung vom CCSD(T)/CBS Limit wurde auf $\leq$ 1 kcal $mol^{-1}$ abgeschätzt. Die mittlere quadratische Abweichung der Aktivierungsenergien der Reaktionspfade liegt oberhalb dieser Abschätzung, und damit ist die QM Methode nicht der entscheidende Faktor für die Genauigkeit.

# Appendix A

# Natural Localized Molecular Orbitals

## A.1 Notation

The following notation will be used throughout

| | |
|---|---|
| $\chi_{\mu,\nu,\ldots}$ | AO |
| $\tilde{\chi}_{r,s,t,\ldots}$ | NAO |
| $\varphi_{r,s,t,\ldots}$ | NHO |
| $\phi_{r,s,t,\ldots}$ | NBO |
| $\psi_{i,j,k,\ldots}$ | occupied NLMO |

## A.2 General Structure

This Appendix complements the description made in Chapter 4, on the construction of Natural Localized Molecular Orbitals. The NBO method performs the analysis of a many-electron molecular wave function in terms of localized electron-pair "bonding" units. This involves the construction of Natural Atomic Orbitals (NAOs), Natural Bond Orbitals (NBOs) and Natural Localized Molecular Orbitals (NLMOs). These may be used to perform Natural Population Analysis (NPA) and other tasks pertaining to the locality of wave function properties. Each natural localized set forms a complete orthonormal set of one-electron functions for expanding the delocalized canonical orbitals.

To obtain the final NLMO set from the AOs, a series of stepwise transformations are required:

$$\text{AO} \xrightarrow{\text{T}^{\text{NAO}}} \text{NAO} \xrightarrow{\text{T}^{\text{NBO}}} \text{NBO} \xrightarrow{\text{T}^{\text{NLMO}}} \text{NLMO}.$$

The next sections detail how each transformation is obtained, starting from the SCF molecular orbitals coefficients and the nonorthogonal AO basis.

## A.3 NAO Transformation

source files : `nbo.f`
`nao.f`

The aim of this procedure is to find the transformation $\mathbf{T}^{\mathrm{NAO}}$ from the nonorthogonal AO basis set $\{\chi_\mu\}$ to the orthonormal NAO basis set $\{\tilde{\chi}_r\}$

$$|\tilde{\chi}_r >= \sum_\mu T_{r\mu}^{\mathrm{NAO}} |\chi_\mu > . \tag{A.1}$$

One starts by computing the first-order reduced density matrix $\mathbf{D}$, obtained from the SCF coefficients as

$$D_{\mu\nu} = 2\sum_i^{\mathrm{occ}} C_{\mu i}^* C_{\nu i} \tag{A.2}$$

and builds the matrix representation for the density

$$\mathbf{\Gamma} = \mathbf{SDS}, \tag{A.3}$$

where $\mathbf{S}$ is the AO overlap matrix. Considering the matrix in block form (each block pertaining to an atom)

$$\begin{pmatrix} \mathbf{\Gamma}^{(AA)} & \mathbf{\Gamma}^{(AB)} & \mathbf{\Gamma}^{(AC)} & \dots \\ \mathbf{\Gamma}^{(BA)} & \mathbf{\Gamma}^{(BB)} & \mathbf{\Gamma}^{(BC)} & \dots \\ \mathbf{\Gamma}^{(CA)} & \mathbf{\Gamma}^{(CB)} & \mathbf{\Gamma}^{(CC)} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

the NAO orbitals will be the eigenvectors of the diagonal density blocks

$$\Gamma^{(AA)} |\tilde{\chi}_r >= \gamma_r |\tilde{\chi}_r >, \tag{A.4}$$

where the index $A$ stands for an atom label, $r$ the NAO label, and $\gamma_r$ is the occupation of the $r$th NAO belonging to center $A$.

The steps to obtain the transformation matrix $\mathbf{T}^{\mathrm{NAO}}$ are as follows:

(1) *Partitioning and symmetry averaging of* $\mathbf{\Gamma}$ *and* $\mathbf{S}$.

Each of these matrices is partitioned into $(Al)$ blocks, $l$ being the angular momentum of the basis function $(s,p,d,...)$. Each of the blocks are then averaged over $m$ (the magnetic quantum number)

$$\Gamma_{\mu\nu}^{(Al)} = \frac{1}{2l+1}\sum_{m=1}^{2l+1} P_{\mu\nu}^{(Alm)}, \tag{A.5}$$

$$S_{\mu\nu}^{(Al)} = \frac{1}{2l+1}\sum_{m=1}^{2l+1} S_{\mu\nu}^{(Alm)}. \tag{A.6}$$

(2) *Formation of pre-NAOs*

For each $(Al)$ block, one solves the generalized eigenvalue problem

$$\mathbf{\Gamma}^{(\mathbf{Al})}\tilde{\mathbf{T}}^{(\mathbf{Al})} = \mathbf{S}^{(\mathbf{Al})}\tilde{\mathbf{T}}^{(\mathbf{Al})}\tilde{\mathbf{W}}^{(\mathbf{Al})}, \tag{A.7}$$

where $\tilde{\mathbf{T}}^{(\mathbf{Al})}$ is the pre-NAO transformation matrix for the $(Al)$ block, and $\tilde{\mathbf{W}}^{(\mathbf{Al})}$ a diagonal matrix with the symmetry-averaged pre-NAOs occupancies.

(3) *Orthogonalization of the high- and low-occupancy NAO spaces*

(a) *Selection of NMB orbitals*

The orthogonalization of the pre-NAOs is done taking into account the occupancy of each orbital. Orbitals with higher occupation should be less distorted in the process, in order to preserve maximum locality of the electron density. With this in mind, a group of orbitals is taken as the Natural Minimal Basis (NMB) set. This selection is made according to the ground state configuration of each atom. For each hydrogen one $s$-type pre-NAO should be taken, for carbon two $s$-type functions and three $p$-type, and so on. The remaining orbitals are tagged to belong to a Natural Rydberg Basis (NRB) set. These are the pre-NAOs of lower occupancy.

(b) *Weighted interatomic orthogonalization of the NMB space.*

The NMB orbitals are orthogonalized among themselves. The first step is to find the largest occupation number in the set, and then to divide all values by this reference to obtain the weighting numbers $\gamma(i)$. The transformation vector for each NMB orbital is weighted

$$\hat{T}_{\mu r}^{\mathrm{NMB}} = \gamma(r).\tilde{T}_{\mu r}^{\mathrm{NMB}}. \tag{A.8}$$

The weighted NMB overlap matrix is built

$$\hat{\mathbf{S}} = (\hat{\mathbf{T}}^{\mathrm{NMB}})^{\dagger}\mathbf{S}\hat{\mathbf{T}}^{\mathrm{NMB}} \tag{A.9}$$

and the new NMB vectors are obtained as

$$\mathbf{T}^{\mathrm{NMB}} = \hat{\mathbf{T}}^{\mathrm{NMB}}\hat{\mathbf{S}}^{-\frac{1}{2}}. \tag{A.10}$$

For orbitals with the same occupancy this procedure would reduce to a simple Löwdin orthogonalization.

(c) *Schmidt interatomic orthogonalization of NRB to NMB orbitals.*

Each NRB orbital is Schmidt orthogonalized to each NMB set orbital:

$$T^{\mathrm{NRB}}_{\mu r} = \hat{T}^{\mathrm{NRB}}_{\mu r} - \sum_{s} T^{\mathrm{NMB}}_{\mu s}\left[(T^{\mathrm{NMB}})^{\dagger} S \hat{T}^{\mathrm{NRB}}\right]_{sr} \tag{A.11}$$

(d) *Restoring the natural character of the NRB space.*

Due to the Schmidt orthogonalization, one needs to repeat steps (1) and (2), but only for the NRB space. This is done by transforming both density and overlap matrices to the NRB basis and rediagonalizing. This is again done in a symmetry averaged way.

(e) *Weighted interatomic orthogonalization of the NRB space.*

The NRB space is divided into two sets. One of low occupation (with weights below $10^{-4}$) and a high occupation set. The latter set is orthogonalized with occupancy weighting. Then, the two sets are Schmidt orthogonalized relative to each other and, at last, the low occupancy orbitals are orthogonalized without occupancy weighting (in order to avoid numerical instability).

(f) *Final diagonalization*

The density and overlap matrices are transformed into the NAO basis, the density blocks diagonalized (as in steps (1) and (2)), giving the final NMB and NRB orbitals and occupancies.

## A.4 NBO Transformation

source files : `nbo.f`
`nho.f`

The final NAO density matrix, in the case of a bonded species, will have significantly large off-diagonal elements. One can therefore expand the search to 2-atom blocks to include a bond description into the set. The highest occupation orbitals are referred to as Lewis-type NBOs. These will be the orbitals with occupation exceeding the value of a parameter `thrnbo`. This parameter is initialized at 1.90. If the number of orbitals found is below half the number of electrons in the system, the search is repeated with `thrnbo` decremented by 0.10. This procedure is repeated until enough NBOs are found or the parameter is below 1.50. In the latter case, the search could be expanded to 3-center bonds, but this option has not yet been implemented.

The orbitals are divided into the following types: core, valence lone pairs and two-centers bonds (all three classes are Lewis-type), Rydberg and two-center antibonding. For searching two-center NBOs, it is useful to build a list of probable pairs. The Wiberg Bond Indices are used. These indices are calculated by summing the square of the off-diagonal elements of the NAO density matrix

$$W_{AB} = \sum_{r \in A, s \in B} \tilde{D}_{rs}^2 \tag{A.12}$$

where $A$ and $B$ are atom labels, $r$ and $s$ NAO labels, belonging to the respective centers. The centers with the largest index will be the first to be considered in the 2-center NBO bond search. If the first cycle fails to find an appropriate Lewis structure, the ordering is then defined according to the z-matrix row numbers.

### A.4.1 Core and Valence lone pair NBOs

The first NBOs in the search are the core and lone pair orbitals. They are both built in the same manner, as eigenvectors of the one-center blocks of the NAO density matrix. The core orbitals are only a convention. The first few eigenvectors for each atom will be taken to be *core*, the number depending on the atom involved. The number of core orbitals will be 1 for the second row elements, 5 for the third row and so on. Both sets of orbitals are obtained by checking the occupancy of the NAOs. If they are above `thrnbo`, new NHO and NBO orbitals are built out of the respective NAO.

After each core/lone pair orbital $\phi_r$ is found, the density matrix block is depleted from their contribution:

$$\Gamma^{(A)} = \Gamma^{(A)} - \gamma_{Ar}|\phi_r >< \phi_r| \tag{A.13}$$

where $\gamma_{Ar}$ is the occupation of the $r$th NBO at center A.

### A.4.2 Two-center Bond NBOs

If the number of core and lone pair NBOs found is below the number of electron pairs, a search is started for two-center bonds. One takes the depleted density matrix, and forms subblocks $\Gamma^{(AB)}$ of the centers with largest Wiberg Bond indices. The block is then diagonalized

$$\mathbf{U}^{\dagger}\mathbf{\Gamma}^{(\mathbf{AB})}\mathbf{U} = \mathbf{W}^{(\mathbf{AB})} \tag{A.14}$$

where $\mathbf{W}^{(\mathbf{AB})}$ is a diagonal matrix with the occupations for each eigenvector. For each eigenvalue above `thrnbo` a new NBO of the form

$$\phi_r = \alpha_{Ar}|\varphi_{Ar} > +\alpha_{Br}|\varphi_{Br} > \tag{A.15}$$

is built, where $\alpha$ is a polarization coefficient and the $\{\varphi_{Ar}\}$ functions directed NHOs. The NHOs are also kept in a NHO transformation matrix, after normalization.

### A.4.3 Rydberg NBOs

A projection matrix is built for each NHO

$$\mathbf{R_{Ar}} = \mathbf{1} - |\varphi_{Ar} >< \varphi_{Ar}|, \tag{A.16}$$

and a full projection operator for atom $A$ is assembled by multiplying together the projectors

$$\mathbf{R_A} = \prod_r \mathbf{R_{Ar}}. \tag{A.17}$$

The significant elements of $\mathbf{R_A}$ are taken and normalized. The resulting matrix $\bar{\mathbf{R}}_\mathbf{A}$ is then used to transform the density subblock for center $A$

$$\bar{\mathbf{\Gamma}}_\mathbf{A} = \bar{\mathbf{R}}_\mathbf{A}^{\dagger}\mathbf{\Gamma}^{(\mathbf{A})}\bar{\mathbf{R}}_\mathbf{A}. \tag{A.18}$$

The density subblock is then diagonalized, its eigenvectors are the Rydberg NHOs/NBOs.

### A.4.4 Orthogonalization of the NHOs

At this point, the number of NHOs should be equal to the total number of basis functions. Although normalized, they are not orthogonal, and therefore they are at this stage symmetrically orthogonalized.

### A.4.5 Antibonding NBOs

The remaining low-occupation NBOs are antibonding orbitals, which may be found by calculating the density in the basis of the constituent NHOs of each 2-center NBO and diagonalizing. In this way one obtains the occupations as the eigenvalues and the polarization coefficients as the eigenvectors for both bonding and anti-bonding NBOs.

## A.5 NLMO Transformation

source files : `nbo.f`
`nlmo.f`

This procedure consists in transforming the Lewis-type orbitals so that they span the occupied space. Two sets of orbitals are considered

- high-occupancy orbitals (NBO set $A$) - which consist of core, lone pair and bond orbitals,
- low-occupancy orbitals (NBO set $B$) - the remaining Rydberg and anti-bonding orbitals.

In order to separate occupied from virtual space, the density matrix should be diagonal in the $\mathbf{D}^{(\mathbf{AA})}$ block (with $D_{ii} = 2$) and 0 elsewhere. There are various ways to diagonalize the NBO density matrix, but to maintain the symmetry properties, the procedure suggested by Reed *et. al.*[12] is followed:

(1) Find the element $D_{ij}$ in $\mathbf{D}^{(\mathbf{AB})}$ of largest magnitude. If $|D_{ij}| < e_1$, the threshold for zeroing the elements of $\mathbf{D}^{(\mathbf{AB})}$, go to step (6) ($e_1 = 5.10^{-9}$)

(2) Find all elements $D_{kl}$ in $\mathbf{D}^{(\mathbf{AB})}$ which are $(1-e_2)|D_{ij}|$ or greater in magnitude, and for which the conditions $(1-e_3)D_{ii} < D_{kk} < (1+e_3)D_{ii}$ and $(1-e_3)D_{jj} < D_{ll} < (1+e_3)D_{jj}$ are true for $k$ and $l$. Here, $e_2$ is the criterion for degeneracy or near

degeneracy, and we have set it to $1.10^{-3}$. The additional criterion $e_3$ should be set to a value safely larger than $e_2$ (to ensure that elements of $D(AB)$ that are symmetry equivalent by the $e_2$ criterion are not rejected), and we have set it to $5.10^{-3}$. The number of near-degenerate off-diagonal elements so found is denoted as $n_{\text{off}}$.

(3) Diagonalize the $n_{\text{off}}$ matrices (2x2 Jacobi rotation)

$$\begin{pmatrix} D_{ii} & |D_{ij}| \\ |D_{ij}| & D_{jj} \end{pmatrix}$$

(4) Find the symmetrized transformation $\mathbf{T}^{\text{sym}}$ that reduces the magnitudes of the $n_{\text{off}}$ off-diagonal elements of $\mathbf{D}$ in an optimal, yet symmetric manner

   (a) Multiply together the $n_{\text{off}}$ rotations to yield $\mathbf{T}^{\text{jac}}$

   (b) Average the elements in $\mathbf{T}^{\text{jac}}$ which are equal in magnitude in $\mathbf{D}$ to within a multiplicative factor of $(1 \pm e_3)$, to give the initial $\mathbf{T}^{\text{sym}}$. (All elements in the *AB* block of $\mathbf{T}^{\text{sym}}$ are made negative, and all elements in the *BA* block are made positive.)

   (c) Multiply the elements in the *AB* and *BA* blocks of $\mathbf{T}^{\text{sym}}$ by the sign of the corresponding element in $\mathbf{D}$ so that all rotation directions are correct.

   (d) Normalize the columns of $\mathbf{T}^{\text{sym}}$

   (e) Perform a Löwdin symmetric orthogonalization of the column vectors of $\mathbf{T}^{\text{sym}}$ to ensure that $\mathbf{T}^{\text{sym}}$ is unitary

(5) Transform $\mathbf{D}$ by $\mathbf{T}^{\text{sym}}$ and return to step (1) for the next rotation.

(6) The NLMO procedure is finished. $\mathbf{T}^{\text{NLMO}}$ is just the product of all the $\mathbf{T}^{\text{sym}}$ transformations and thus will retain the symmetry present in $\mathbf{D}$ in the NBO basis by virtue of step (4b).

### A.5.1 Exclusion of core orbitals

To exclude the core orbitals from the localization, the same procedure is followed, but with the valence density as input. This means that the running index in Eq. (A.2) only goes through the occupied valence orbitals. Also

- the number of NMB orbitals per atom should be accordingly reduced.

- the search for core NBOs is turned off.
- the NLMO procedure is done as detailed before. To build the full orbital set (including the canonical core orbitals)
  - the $n_{\text{virt}}$ virtual NLMO orbitals are Schmidt orthogonalized relative to the $n_{\text{core}}$ core orbitals
  - the new virtual set should be linearly dependent, with $n_{\text{core}}$ redundant functions. These are removed by building the overlap matrix and diagonalizing. The eigenfunctions corresponding to the lower $n_{\text{core}}$ eigenvalues are removed.
  - the core orbital vectors are then added to the coefficient matrix.

# Appendix B

# Domain Merging - Quick Guide

## B.1 General Procedure

In order to create a merged domain, as a junction of domain lists from two (or more) different geometries, the following steps should be taken:

(1) Perform an Hartree-Fock energy run for the first structure, followed by orbital localization and domain list build. The domain list should be saved. It is, however, not necessary to run the full local calculation, only the domains are needed.

**Example**:
```
hf
{lmp2,domonly=1,save=5400.2}
```

(2) Perform the same calculation on the second structure, but this time the previously saved domain list is read, and the merging procedure is applied. The merged domains can then be saved.

**Example**:
```
hf
{lmp2,domonly=1,save=5500.2
mergedom,start=5400.2}
```

(3) Any number of local calculations can be run by reading the merged domain list

**Example**:
```
hf
{lmp2,start=5500.2}
```

or step (2) can be repeated in order to include more points.

Some examples on the use of this procedure can be found in the `Molpro2006.1` testjobs: `hf_loc_merg.test` and `loc_eom.test`.

## B.2 A step-by-step example: ketene

In this Section, a worked example for domain merging is discussed - the ketene dissociation from Chapter 3. Geometries were optimized for fixed C-C distances, ranging from 1.2 Å to 2.5 Å (as depicted in Fig. B.1). The objective of this procedure is to read the domain list at r(C-C)=1.2 Å (the first point in the path) and merge it with the domain list at r(C-C)=2.5 Å. The resulting domains can then be used for the other structures along the path.

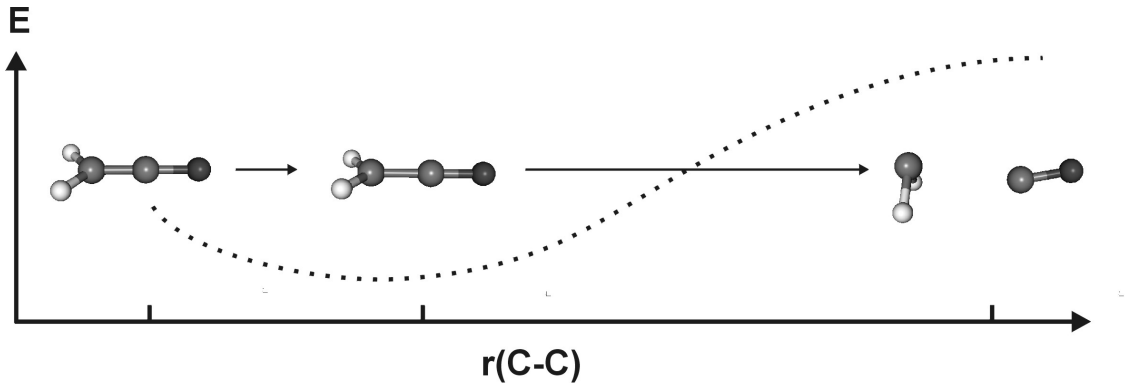



Figure B.1: Diagrammatic representation of the ketene dissociation path.

(1) The following input will perform an SCF calculation for the structure with r(C-C)=1.2 Å, build the LMOs and PAOs, and save this information on record 5400.2:

```
geometry
nosym
5

C1   -1.22038  0.00000  0.00000
C2   -0.02038  0.00000  0.00000
O3    1.15610  0.00000  0.00000
H4   -1.78291  0.93638  0.00000
H5   -1.78291 -0.93638  0.00000
endg

hf
{lmp2,domonly=1,save=5400.2}
```

The domain information as well as a message confirming the save can be found in the output:

```
(...)

 Orbital domains

   Orb.    Atom     Charge      Crit.
   4.1     1 C1      1.11       0.000
           5 H5      0.75       0.992
   5.1     1 C1      1.11       0.000
           4 H4      0.75       0.992
   6.1     1 C1      1.17       0.000
           2 C2      0.78       0.992
   7.1     1 C1      0.98       0.000
           2 C2      0.98       0.997
   8.1     3 O3      1.28       0.000
           2 C2      0.69       0.999
   9.1     3 O3      1.39       0.000
           2 C2      0.57       0.991
  10.1     3 O3      1.68       0.952
           2 C2      0.28       0.982
  11.1     3 O3      1.82       0.997

(...)

 Domain information saved on record     5400.2
```

(2) A calculation for the second structure (at a 2.5 Å distance) is performed, restoring the saved domains and merging both sets

```
geometry
nosym
5

C1   -2.07692  0.00000  0.14319
C2    0.40614  0.00000 -0.14734
O3    1.51994  0.00000  0.07573
H4   -2.10841  0.86712 -0.57628
H5   -2.10841 -0.86712 -0.57628
endg

hf
{lmp2,domonly=1,save=5500.2;mergedom,start=5400.2}
```

The output should give the following information:

```
(...)

 Orbital domains

   Orb.    Atom     Charge      Crit.
   4.1     1 C1      1.17       0.000
           5 H5      0.81       0.998
   5.1     1 C1      1.17       0.000
           4 H4      0.81       0.998
   6.1     1 C1      1.89       0.995
   7.1     2 C2      1.84       0.988
   8.1     3 O3      1.29       0.000
           2 C2      0.70       0.999
   9.1     3 O3      1.44       0.000
           2 C2      0.56       1.000
  10.1     3 O3      1.43       0.000
           2 C2      0.57       1.000
  11.1     3 O3      1.84       0.998

 Domain list read from record     5400.2

 Augmented orbital domains

   Orb.    Atoms
   6.1     1 C1    2 C2
   7.1     1 C1    2 C2

(...)

 Domain information saved on record     5500.2
```

In this case, only orbitals 6.1 and 7.1 are changed. The merged domains are coincident with the orbital domains for the structure r(C-C)=1.2 Å.

(3) If the same procedure would be repeated, this time starting with the structure at r(C-C)=1.33 Å, which bears the domain list:

```
Orbital domains

  Orb.    Atom     Charge      Crit.
  4.1     1 C1      1.13       0.000
          5 H5      0.77       0.994
  5.1     1 C1      1.13       0.000
          4 H4      0.77       0.994
  6.1     1 C1      1.22       0.000
          2 C2      0.72       0.990
  7.1     2 C2      1.01       0.000
          1 C1      0.96       0.997
  8.1     3 O3      1.29       0.000
          2 C2      0.70       0.998
  9.1     3 O3      1.39       0.000
          2 C2      0.59       0.994
 10.1     3 O3      1.65       0.945
          2 C2      0.31       0.980
          1 C1      0.04       1.000
 11.1     3 O3      1.82       0.997
```

the domain merge output (by 2.5 Å) would be as following:

```
Augmented orbital domains

  Orb.    Atoms
  6.1     1 C1    2 C2
  7.1     1 C1    2 C2
  8.1     1 C1    2 C2    3 O3
  9.1     1 C1    2 C2    3 O3
 10.1     1 C1    2 C2    3 O3
```

As before, orbitals 6.1 and 7.1 are slightly augmented, becoming bond orbitals. Orbital 10.1, however, includes C1 in its domain. Comparing both domain sets, the program recognizes that an orbital with domain {O3, C2} should be augmented, but does not distiguish between orbitals 8.1, 9.1 and 10.1, merging the whole set. This procedure might seem somewhat wasteful in this case, but it protects the algorithm from orbital transformations. This has been discussed in further detail in Chapter 3.

# Appendix C

# LMOMO - Quick Guide

## C.1 General Procedure

The use of LMOMO is controlled by the `REGION` directive

`REGION,METHOD=`*method*`,[DEFAULT=`*default_method*`],`
`[TYPE=INCLUSIVE|EXCLUSIVE],` *atom1, atom2 ...*

The list of atoms defines the orbitals which will be treated at the level defined by *method*. If `TYPE=INCLUSIVE`, any orbital containing one of the atoms in its domain centre list will be included. This is the default and has been used throughout this work. For `TYPE=EXCLUSIVE`, only orbitals whose domains are exclusively covered by the given atoms will be added. This is in general **not** advised. With the use of this option the most delocalized orbitals (usually the $\pi$-orbitals) will be in general excluded, even when the major charge centers are inside the region. The potential energy surface between two atoms is also affected in an unpredictable way. The use of the `EXCLUSIVE` option should be restricted to cases where specific orbitals need to be added, and the `INCLUSIVE` option fails to give the desired selection. The remaining atoms, if no further region is assigned, will be treated at the level given by *default_method*.

Any local correlation treatment can be given as method, with the restriction that only MP2 and HF can be used as *default_method*. Up to two `REGION` directives may be included in a single calculation, ordered according to the correlation level (*method*) specified for the region. The highest level region should be given last.

A simple LMP2:HF LMOMO calculation can be invoked with the use of

```
lmp2
region,method=mp2,default=hf,type=inclusive,...
```

A three region LCCSD(T0):LMP2:HF calculation can be called as

```
lccsd(t)
region,method=mp2,default=hf,type=inclusive,...
region,method=ccsd(t),default=hf,type=inclusive,...
```

Some examples on the use of this procedure can be found in the `Molpro2006.3` [8] testjobs: `lmp2_regions.test`, `h2odim_regions.test` and `form_atoml.test`.

## C.2 A step-by-step example: SN2 reaction

As an example for the use of the LMOMO approach (and for clarity) a small molecular system was chosen, the SN2 reaction of ethylchloride with $Cl^-$, a case already discussed in Chapter 3 in the context of domain merging. Two structures will be used in this Section, the van der Waals complex and the transition state (both depicted in Fig. C.1).

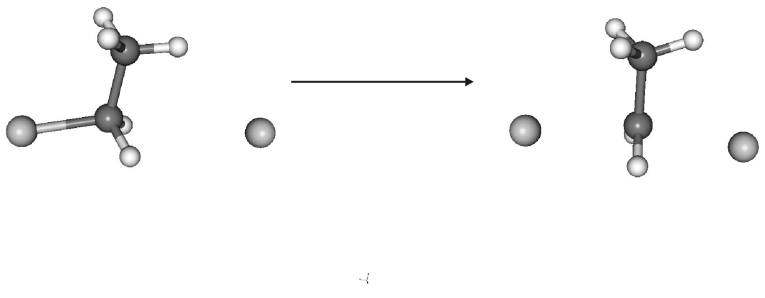


Figure C.1: Depiction of the van der Waals complex (left side) and transition state (right side) structures for the SN2 reaction of ethylchloride with $Cl^-$.

Below is the input for an LMP2:HF calculation on the complex, including only the chlorines in the LMP2 region

```
set,charge=-1

geomtyp=xyz
geometry
   9
C1    -0.017510  0.060386  0.000000
C2     1.489731 -0.001094  0.000000
CL1   -0.596632  1.800801  0.000000
CL2    0.180908 -3.259591  0.000000
H1    -0.445001 -0.403897 -0.876785
H2    -0.445001 -0.403897  0.876785
H3     1.756600 -1.056243  0.000000
H4     1.898238  0.484172 -0.883681
H5     1.898238  0.484172  0.883681
endg

hf
lmp2
region,mp2,type=inclusive,Cl1,Cl2
```

Notice that the atoms **must be numbered**, unless there is only one atom of the given type. The input, as can be seen in this example, can be quite compact. There is no need to use the `METHOD` keyword as long as the order is respected. Also, if the low level region is HF, the `DEFAULT` keyword may be skipped.

In the output, after the orbital domains information, the LMOMO data is displayed:

```
 ========================
 Local Regions
 ========================

 Region= 1  Method=MP2     Type=1  Class= 2  Atoms:CL1   CL2
 Region= 2  Method=HF      Type=1  Class= 1  All remaining atoms

 Ordering localized MOs according to center regions

 Orbital domains and regions

   Orb.    Atom     Region
  13.1     4 CL2    MP2
  14.1     4 CL2    MP2
  15.1     4 CL2    MP2
  16.1     4 CL2    MP2
  17.1     1 C1     MP2
           3 CL1
  18.1     3 CL1    MP2
  19.1     3 CL1    MP2
  20.1     3 CL1    MP2
  21.1     2 C2     HF
           7 H3
  22.1     1 C1     HF
           5 H1
  23.1     1 C1     HF
           6 H2
  24.1     1 C1     HF
           2 C2
  25.1     2 C2     HF
           8 H4
  26.1     2 C2     HF
           9 H5

 Region= 1  Method=MP2     Type=1  Class= 2  Orbitals 13.1  14.1  15.1  16.1  17.1
                                                      18.1  19.1  20.1
 Region= 2  Method=HF      Type=1  Class= 1  All remaining orbitals
```

The orbitals might have to be reordered, so that the higher level ones are numbered first. This has to do with the local correlation program internal structure. The orbitals to be treated at the MP2 level are as expected the chlorine lone pairs and any C-Cl bonds which may be present (since this is the reactant complex, there is only one bond, connecting the carbon to the leaving chlorine).

Using the same input, but this time for the transition state geometry, the following output should be obtained

```
=========================
Local Regions
=========================

Region= 1  Method=MP2     Type=1  Class= 2  Atoms:CL1   CL2
Region= 2  Method=HF      Type=1  Class= 1  All remaining atoms

Ordering localized MOs according to center regions

Orbital domains and regions

  Orb.    Atom     Region
 13.1     4 CL2    MP2
 14.1     4 CL2    MP2
 15.1     4 CL2    MP2
 16.1     1 C1     MP2
          4 CL2
 17.1     1 C1     MP2
          3 CL1
 18.1     3 CL1    MP2
 19.1     3 CL1    MP2
 20.1     3 CL1    MP2
 21.1     2 C2     HF
          7 H3
 22.1     1 C1     HF
          5 H1
 23.1     1 C1     HF
          6 H2
 24.1     1 C1     HF
          2 C2
 25.1     2 C2     HF
          9 H5
 26.1     2 C2     HF
          8 H4
```

```
 Region= 1  Method=MP2     Type=1  Class= 2  Orbitals 13.1  14.1  15.1  16.1  17.1
                                                 18.1  19.1  20.1
 Region= 2  Method=HF      Type=1  Class= 1  All remaining orbitals
```

In calculating relative energies, the number of orbitals in each region should be consistent. One of the previous lone pairs is now a C-Cl bond, which is also included in the LMP2 region.

In a three region LCCSD:LMP2:HF calculation, with both chlorines in the LCCSD region and the carbon in the LMP2 (leaving only the three methyl C-H bonds uncorrelated), the input should be given as:

```
(...)

hf
lccsd
region,mp2,type=inclusive,C1,Cl1,Cl2
region,ccsd,type=inclusive,Cl1,Cl2
```

Notice that the first region (LMP2) contains the second. This is not necessary, but advisable, in order to avoid that orbitals connecting both sets of atoms are left out. The local correlation program to be invoked was also changed. It should always be the highest correlation level.

The output will show the three regions, with the orbitals reordered in a consistent way relative to the levels of theory. The transition state output would be similar, and will be skipped. The complex structure output should be as follows:

```
=========================
Local Regions
=========================

Region= 1  Method=MP2     Type=1  Class= 2  Atoms:C1
Region= 2  Method=CCSD    Type=1  Class= 4  Atoms:CL1   CL2
Region= 3  Method=HF      Type=1  Class= 1  All remaining atoms

Ordering localized MOs according to center regions

Orbital domains and regions

  Orb.    Atom     Region
 13.1     4 CL2    CCSD
 14.1     4 CL2    CCSD
 15.1     4 CL2    CCSD
 16.1     4 CL2    CCSD
 17.1     1 C1     CCSD
          3 CL1
 18.1     3 CL1    CCSD
 19.1     3 CL1    CCSD
 20.1     3 CL1    CCSD
 21.1     1 C1     MP2
          5 H1
 22.1     1 C1     MP2
          6 H2
 23.1     1 C1     MP2
          2 C2
 24.1     2 C2     HF
          7 H3
 25.1     2 C2     HF
          8 H4
 26.1     2 C2     HF
          9 H5

Region= 1  Method=MP2     Type=1  Class= 2  Orbitals 21.1  22.1  23.1
Region= 2  Method=CCSD    Type=1  Class= 4  Orbitals 13.1  14.1  15.1  16.1  17.1
                                                    18.1  19.1  20.1
Region= 3  Method=HF      Type=1  Class= 1  All remaining orbitals
```

# Appendix D

# Electrostatic embedding - the polarized QM Hamiltonian

The use of electrostatic embedding in a QM/MM calculation involves minor changes to the QM program. The total Hamiltonian has already been presented in Section 2.3. The QM Hamiltonian will include the MM atoms in the form of point charges, so that the extra terms to be computed will only involve one-electron integrals. In the HF case, a polarization term is included in the Fock operator

$$< \mu | h^{\text{pol}} | \nu >= -\sum_x q_x \int \frac{\chi^*_\mu(\mathbf{r}_i)\chi_\nu(\mathbf{r}_i)}{\mathbf{r}_{ix}} d\mathbf{r}_i, \tag{D.1}$$

where $q_x$ is the point charge, and $r_{ix}$ the distance between electron $i$ and the point charge $x$.

In the DFT case, the operator is added to the potential $v(\mathbf{r})$ in Eq. (2.56). In post-HF calculations, no further changes are needed since the point-charge effect is induced through a one-electron operator[1]. The nuclei interactions are trivial to include, a classical Coulomb energy term between pairs of point charges is needed.

The extra terms to be added to the gradient are also relatively straightforward, and are given below. The first set of equations give the terms necessary for the gradient relative to the movement of a point charge $y$ in the direction $\lambda^y$. The second set refers to the movement of a nucleus $m$ in the $\lambda^m$ direction.

[1] The polarization will however have an effect on the correlation energy estimate since the reference function is changed.

**MM movement**

- **point charge-nuclei interaction**

$$\frac{\partial}{\partial \lambda^y}\left(\sum_m \sum_x \frac{Z_m q_x}{|\mathbf{r}_{mx}|}\right) = \sum_m -\frac{Z_m q_y}{r_{my}^3}(\lambda^y - \lambda^m) \tag{D.2}$$

- **point charge-electron interaction**

$$\frac{\partial}{\partial \lambda^y}\left(\sum_i \sum_x -2 < i|\frac{q_x}{|\mathbf{r}_{1x}|}|i>\right) = \sum_{\mu\nu} -q_y D_{\mu\nu} < \mu|\frac{\partial}{\partial \lambda^y}\frac{1}{|\mathbf{r}_{1y}|}|\nu> \tag{D.3}$$

**QM movement**

- **point charge-nuclei interaction**

$$\frac{\partial}{\partial \lambda^m}\left(\sum_n \sum_x \frac{Z_n q_x}{|\mathbf{r}_{mx}|}\right) = \sum_x -\frac{Z_m q_x}{r_{mx}^3}(\lambda^x - \lambda^m) \tag{D.4}$$

- **point charge-electron interaction**

$$\frac{\partial}{\partial \lambda^m}\left(\sum_i \sum_x -2 < i|\frac{q_x}{|\mathbf{r}_{1x}|}|i>\right) = \sum_{\mu\nu}\sum_x -q_x D_{\mu\nu}\,( < \frac{\partial}{\partial \lambda^m}\mu|\frac{1}{|\mathbf{r}_{1x}|}|\nu> + < \mu|\frac{1}{|\mathbf{r}_{1x}|}|\frac{\partial}{\partial \lambda^m}\nu>\Big) \tag{D.5}$$

# Appendix E

# Optimized stationary points structures

In this Appendix, the structures for the reactions discussed in Chapter 3 are provided. All stationary points were optimized at the MP2 level of theory, with the cc-pVTZ basis set for C and H atoms, and aug-cc-pVTZ for the halogens (F and Cl). The basis was truncated, removing *d* functions for H and *f* functions for all other atoms, and will be refered to as [aug]-cc-pVTZ(d/p).

## E.1 SN2 Reactions

The following structures have been used in Section 3.3.2. The starting structures for the ethylchloride and propylchloride reactions were based on the geometries given in Ref. [70].

**ethylchloride**

```
   8
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -538.684753678278
C          0.0000491284       -0.0281725670        0.0000000000
C          1.5133326805        0.0056340682        0.0000000000
CL        -0.6820565678        1.6397425625        0.0000000000
H         -0.3919835253       -0.5193607804       -0.8823819824
H         -0.3919835253       -0.5193607804        0.8823819824
H          1.9005397634       -1.0117021252        0.0000000000
H          1.8858362939        0.5190142077       -0.8811946822
H          1.8858362939        0.5190142077        0.8811946822
```

### ethylchloride + $Cl^-$ vdWC

```
   9
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -998.454049715442
C         -0.0175103714         0.0603860676         0.0000000000
C          1.4897305750        -0.0010944576         0.0000000000
CL        -0.5966320198         1.8008005483         0.0000000000
CL         0.1809079349        -3.2595906730         0.0000000000
H         -0.4450013087        -0.4038966857        -0.8767845791
H         -0.4450013087        -0.4038966857         0.8767845791
H          1.7566002000        -1.0562428347         0.0000000000
H          1.8982384203         0.4841717569        -0.8836807219
H          1.8982384203         0.4841717569         0.8836807219
```

### ethylchloride + $Cl^-$ TS

```
   9
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -998.423735700402
C          0.0893485352        -0.1932118058         0.0057894157
C          1.5814783688         0.0070626598        -0.0156402249
CL        -0.4216609371         2.1052282761        -0.0474182077
CL        -0.0042796436        -2.5201671935         0.0790955645
H         -0.4662212878        -0.2650957137        -0.9035760411
H         -0.4501853038        -0.2130344848         0.9269505376
H          2.0856719055        -0.9528148166        -0.0324667740
H          1.8604883980         0.5816952596        -0.8928959171
H          1.8895841831         0.5645728239         0.8631194862
```

### propylchloride

```
  12
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -577.89491330
C             0.0000000000         0.0000000000         0.0000000000
C             0.0000000624        -0.0000001170         1.5162567214
C             1.4008302125         0.0000001225         2.1195011244
CL            0.7081221925        -1.5208117776        -0.6669099573
H            -1.0043124474         0.0638302770        -0.4018910141
H             0.5985937500         0.8135720688        -0.3968188980
H            -0.5390379766         0.8944022377         1.8331771643
H            -0.5668032213        -0.8587898726         1.8715903014
H             1.3542487590         0.0422231295         3.2044425429
H             1.9436674789        -0.8974559929         1.8370533557
H             1.9682241024         0.8620285201         1.7727769374
```

### propylchloride + $Cl^-$ vdWC

```
  12
MP2/[aug]-cc-pVTZ(d/p) ENERGY=  -1037.66265603
C              0.0000000000        0.0000000000        0.0000000000
C              0.0000000000        0.0000000000        1.5114596785
C              1.4084641222       -0.0000001231        2.0950450613
CL             0.6097966673       -1.5929584568       -0.6710061629
CL             0.2490790090        3.3144472800        0.5492214989
H             -0.9962715091        0.1164498208       -0.4079608971
H              0.6398742353        0.7820538958       -0.3888053057
H             -0.4957622007        0.9273630017        1.7956943671
H             -0.5804929134       -0.8473090282        1.8800303572
H              1.3759068871        0.0121340360        3.1831720329
H              1.9678927783       -0.8797163406        1.7804518974
H              1.9268347699        0.8964758697        1.7611463879
```

### propylchloride + $Cl^-$ TS

```
  12
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -1037.63374396
C1             0.0000000000        0.0000000000        0.0000000000
C2             0.0000000000        0.0000000000        1.5064064098
C3             1.4192325487       -0.0000001241        2.0608496716
CL1            0.0549055380       -2.3222706517       -0.2941727203
CL2            0.0879100877        2.3221864185       -0.2902258106
H1            -0.9207877163        0.0070733658       -0.5451807825
H2             0.9138502473       -0.0060923696       -0.5538840186
H3            -0.5372256635        0.8813345344        1.8451491264
H4            -0.5373399246       -0.8811820150        1.8451907852
H5             1.4170810028       -0.0014231805        3.1500005476
H6             1.9488514668       -0.8859701922        1.7165555767
H7             1.9477997256        0.8872098880        1.7184256460
```

### butylchloride

```
  14
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -617.10200758
C              0.0000000000        0.0000000000        0.0000000000
C              0.0000000559       -0.0000001202        1.5158116543
C              1.3940291449        0.0000001219        2.1343585968
C              1.3392543805        0.0436024024        3.6594206895
CL             0.6899485897       -1.5294819789       -0.6666230012
H             -1.0035365571        0.0748891822       -0.4021697762
H              0.6073576460        0.8065702406       -0.3977134665
H             -0.5369486485        0.8959010217        1.8371999470
H             -0.5670927045       -0.8590815060        1.8745173703
H              1.9315911091       -0.8884746593        1.8086078717
H              1.9493207110        0.8612927435        1.7595217749
H              2.3367364716        0.0441231661        4.0916867761
H              0.8229643015        0.9375045390        4.0052341616
H              0.8058426583       -0.8212480411        4.0496180485
```

### butylchloride + $Cl^-$ vdWC

```
  15
MP2/[aug]-cc-pVTZ(d/p) ENERGY= -1076.87090560
C              0.0000000000        0.0000000000        0.0000000000
C              0.0000000000        0.0000000000        1.5102996325
C              1.3993774885        0.0000000000        2.1126834621
C              1.3520694936        0.1690244225        3.6292142616
CL             0.6259591969       -1.5827916892       -0.6746671929
CL             0.5168054290        3.2941666423        0.7798249609
H             -0.9972164751        0.1061576928       -0.4096078020
H              0.6356838704        0.7890547645       -0.3825330823
H             -0.4887433721        0.9326254731        1.7946175203
H             -0.5838312039       -0.8435322269        1.8866176692
H              1.9182017316       -0.9234098419        1.8522705114
H              1.9420707639        0.8367533400        1.6749623563
H              2.3502516815        0.1583711492        4.0647332951
H              0.8851403606        1.1197022957        3.8789304041
H              0.7739878753       -0.6300857020        4.0955058584
```

**butylchloride + $Cl^-$ TS**

```
  15
MP2/[aug]-cc-pVTZ(d/p) ENERGY= -1076.84161729
C            0.0000000000        0.0000000000        0.0000000000
C            0.0000000000        0.0000000000        1.5055518053
C            1.4137589632        0.0000000000        2.0707031481
C            1.4181593540        0.0132275139        3.5973812005
CL           0.0612762866       -2.3219666382       -0.2876101794
CL           0.0833007687        2.3214847949       -0.2840617995
H           -0.9205252782        0.0045991210       -0.5458359741
H            0.9142263414       -0.0041700513       -0.5534474357
H           -0.5349924361        0.8825268537        1.8493052029
H           -0.5354847452       -0.8820325399        1.8497749264
H            1.9306747206       -0.8862718504        1.7031168063
H            1.9359012460        0.8769870653        1.6886484156
H            2.4312406659        0.0111462233        3.9972333714
H            0.9104910629        0.9001815777        3.9752340513
H            0.9004927407       -0.8608602334        3.9913745284
```

## E.2 Hydrogen fluoride addition to double bonds

The following structures have been used in Section 3.3.3. The starting structures for the ethene reaction were based on information available in Ref. [79]. The other geometries were obtained by replacing terminal hydrogen atoms by methyl groups, followed by reoptimization.

**hydrogen fluoride**

```
   2
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -100.313071238678
F          0.0000000000        0.0000000000       -0.0942777040
H          0.0000000000        0.0000000000        0.8270777040
```

### ethene

```
   6
MP2/cc-pVTZ(d/p) ENERGY=    -78.369790996442
C          0.0000000000        0.0000000000       -0.6666936911
C          0.0000000000        0.0000000000        0.6666936911
H          0.9222825072        0.0000000000       -1.2271017049
H         -0.9222825072        0.0000000000       -1.2271017049
H         -0.9222825072        0.0000000000        1.2271017049
H          0.9222825072        0.0000000000        1.2271017049
```

### ethene + FH RC

```
8
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -178.691016864483
C          0.0000000000        1.1405443988       -0.6692956743
C          0.0000000000        1.1411092880        0.6681435357
H          0.9229383890        1.1449049029       -1.2292129521
H         -0.9229383890        1.1449049029       -1.2292129521
H         -0.9229383890        1.1459543476        1.2280571943
H          0.9229383890        1.1459543476        1.2280571943
F          0.0000000000       -1.9162868319        0.0008815253
H          0.0000000000       -0.9847713560        0.0004381288
```

### ethene + FH TS1

```
8
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -178.608060439451
C         -0.0084871764        0.6715812017       -0.3943440342
C         -0.0098235336        0.5764014180        1.0010527421
H          0.9082502150        0.8137640312       -0.9413457645
H         -0.9246095913        0.8098870104       -0.9433636556
H         -0.9345397342        0.7356728856        1.5307205818
H          0.9130462100        0.7395881124        1.5327594007
H         -0.0066977492       -0.6548863139        0.5337084285
F         -0.0042205246       -1.2154386090       -0.6491483445
```

### ethene + FH TS2

```
8
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -178.702797959265
C           0.1970312769       -0.3539038725       -0.4071188308
C           0.2201399492       -0.3789610152        1.1181260054
H           1.1947926557       -0.3158045932       -0.8325369728
H          -0.2909227813        0.5049570907        1.4875600513
H          -0.2892459650       -1.2519156096        1.5134097233
H           1.2340844985       -0.3675880330        1.5048574520
H          -0.3346534891       -1.2033511313       -0.8239320528
F          -0.4722736819        0.7950856841       -0.8580567482
```

### ethene + FH P

```
8
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -178.708176825049
C           0.2001382437       -0.3525527560       -0.4004451323
C           0.2108755926       -0.3630964289        1.1086621882
H           1.2054197262       -0.3234125438       -0.8117892700
H          -0.8038122366       -0.3705003532        1.4952015949
H           0.7267293939        0.5127828871        1.4904852409
H           0.7255539226       -1.2530244047        1.4645909926
H          -0.3299286931       -1.2094898421       -0.8070550089
F          -0.4635939493        0.7950874416       -0.8571466054
```

### propene

```
8
MP2/cc-pVTZ(d/p) ENERGY=   -117.577927665734
C           0.4747049479        0.0000000000       -0.1252372511
C           0.4980393138        0.0000000000        1.2099256872
H           1.4179165064        0.0000000000       -0.6564432295
C          -0.7917587362        0.0000000000       -0.9477485273
H          -0.4165819966        0.0000000000        1.7864720803
H           1.4301228919        0.0000000000        1.7529492961
H          -0.5723099707        0.0000000000       -2.0109550090
H          -1.3965729782       -0.8783059603       -0.7332590234
H          -1.3965729782        0.8783059603       -0.7332590234
```

### propene + FH RC

```
  11
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -217.903717406146
C          -0.2949464958       -0.7589248509       -1.1081130073
C          -0.2635734587       -0.9397181034        0.2191262125
C           0.9819761214       -0.8877121470        1.0521140392
F          -0.5663562167        2.1178713673       -0.0298859660
H          -1.1944155203       -1.1399213606        0.7369851660
H          -1.2201884136       -0.8188655108       -1.6606605649
H           0.6121512452       -0.5671181860       -1.6642307205
H          -0.5174308060        1.2076764453       -0.2320019781
H           0.9024154192       -0.1193064773        1.8186892781
H           1.1380683803       -1.8373237473        1.5605454831
H           1.8528567451       -0.6756304293        0.4375340580
```

### propene + FH TS1

```
  11
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -217.825579045524
C          -0.2520535390       -0.3785729341       -1.3034813408
C          -0.2214244016       -0.5815695909        0.0847879002
C           1.0415252692       -0.5567653211        0.8669271772
F          -0.8226019684        1.2507959262        0.4262276543
H          -1.1014975198       -0.9709355986        0.5736881160
H          -1.0949189532       -0.7505047492       -1.8634137020
H           0.6888968550       -0.2672857925       -1.8201607574
H          -0.5975120282        0.7535134663       -0.7860344691
H           0.8731873015       -0.2346912486        1.8868217680
H           1.4357862877       -1.5749913382        0.8757715055
H           1.7711817743        0.0902096717        0.3906489038
```

**propene + FH TS2**

```
  11
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -217.916808102417
C          -0.2241902530       -0.2764932915       -1.3955578473
C          -0.2045409651       -0.2936131155        0.1325450575
C           1.1882295601       -0.2807821668        0.7196778762
F          -0.8787850632        0.8547575488        0.6046268602
H          -0.7717793383       -1.1359468955        0.5210222778
H           0.7825935587       -0.2996466233       -1.8037003764
H          -0.7778500136       -1.1177759860       -1.8012112166
H          -0.7050142577        0.6375539137       -1.7305519203
H           1.1481783900       -0.2457911070        1.8043458472
H           1.7220651182       -1.1787765824        0.4157372256
H           1.7335177096        0.5869206760        0.3557511258
```

**propene + FH P**

```
  11
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -217.922109481487
C          -0.1696038150        0.3275921246       -0.1071077965
F          -0.1509333087        0.3721646301        1.3030434555
C           1.2623276789        0.3325607362       -0.5923944839
H          -0.6851625589        1.2321667456       -0.4254022341
C          -0.9450396962       -0.9015173180       -0.5241650741
H           1.7865328801        1.2084487119       -0.2217333192
H           1.2861003530        0.3456294229       -1.6798326474
H           1.7755215727       -0.5596677732       -0.2420503504
H          -1.9471817401       -0.8789590111       -0.1063553546
H          -1.0197403579       -0.9435331697       -1.6085615897
H          -0.4372540079       -1.7967630993       -0.1736336056
```

### butene

```
  12
MP2/cc-pVTZ(d/p) ENERGY=   -156.791951913676
C          -0.5510587595       -0.0001209307       -1.8789764162
C          -0.5510458547       -0.0001192664       -0.3783321672
C           0.5510458007       -0.0001191812        0.3783321662
C           0.5510587509       -0.0001210601        1.8789763937
H          -1.5165651447       -0.0001191996        0.1169172580
H           1.5165651188       -0.0001188211       -0.1169172037
H          -1.0656507082        0.8759157516       -2.2711936625
H           0.4654374272       -0.0001053773       -2.2652667474
H          -1.0656221770       -0.8761758660       -2.2711911162
H           1.0656505574        0.8759156607        2.2711937377
H          -0.4654373995       -0.0001057734        2.2652668224
H           1.0656223886       -0.8761759367        2.2711909354
```

### butene + FH RC

```
  14
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -257.115559652543
C          -0.5466759865        0.6792339680       -1.8871234670
C          -0.5486678474        0.6780042622       -0.3864015332
C           0.5555376227        0.6802587566        0.3764479020
C           0.5536107437        0.7017739885        1.8770048881
F          -0.0026055517       -2.2937916224        0.0178696977
H          -1.5144247833        0.6896518164        0.1088748132
H           1.5213190374        0.6792701136       -0.1188655978
H          -0.0008669764       -1.3582134212        0.0104802135
H          -1.0477127078        1.5671786978       -2.2686725491
H           0.4687125148        0.6631776348       -2.2747762999
H          -1.0795347601       -0.1852285375       -2.2795759357
H           1.0840693276       -0.1586308269        2.2810689810
H           1.0560066876        1.5938788482        2.2469456243
H          -0.4620253205        0.6931123219        2.2644622629
```

### butene + FH TS1

```
  14
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -257.032805714759
C           0.0773171907       -0.4340726131       -0.8542952823
C           0.0524864538       -0.2504657896        0.5371407928
C           1.2826425288       -0.2723483405        1.3703233945
F          -0.1474525052        1.7035606259        0.3704032663
H          -0.8985178132       -0.2947979617        1.0487770724
C          -1.1571152549       -0.8742472565       -1.6121890503
H           1.0462925297       -0.6612567966       -1.2810279888
H           0.0306118065        0.8262969425       -0.6570131286
H           1.1791234270        0.3524634437        2.2484172053
H           1.4445385698       -1.3058208512        1.6840184037
H           2.1417399774        0.0487366203        0.7898199000
H          -1.1444473421       -0.4951361724       -2.6300248254
H          -1.2269401153       -1.9590752362       -1.6610349935
H          -2.0543288407       -0.4955729848       -1.1282671178
```

### butene + FH TS2

```
  14
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -257.123434043630
C           0.0205319390       -0.0090883416       -1.0040613205
C           0.0263806831       -0.0218617481        0.5267186038
F           1.3641941862       -0.0235428597        0.9808127934
C          -0.6675814339       -1.2209181222        1.1305745389
H          -0.4046510268        0.9010879085        0.9128627468
H          -0.5016500446       -0.8901351757       -1.3740724739
C          -0.6153912129        1.2533204915       -1.5830250484
H           1.0542654744       -0.0918337729       -1.3319321074
H          -0.6074158466       -1.1960779705        2.2147476800
H          -1.7154163547       -1.2225112006        0.8364262873
H          -0.2046115734       -2.1363397265        0.7695218162
H          -0.6070177767        1.2364410237       -2.6698049293
H          -1.6501738300        1.3505028779       -1.2591857766
H          -0.0769838260        2.1395834925       -1.2541716464
```

**butene + FH P**

```
  14
MP2/[aug]-cc-pVTZ(d/p) ENERGY=   -257.128435872515
C          -0.2738290972        0.0343659040       -1.0005303569
C          -0.2917278645        0.0396897469        0.5150389864
C           1.0768993369        0.0671411131        1.1571113462
F          -0.9850971752        1.1953665918        0.9334373942
H          -0.8707730775       -0.8073238386        0.8841789265
H          -1.2996311468        0.1490498233       -1.3461190644
C           0.3309226137       -1.2460136671       -1.5752191071
H           0.2833531426        0.9067752546       -1.3408738285
H           0.9811164648        0.1667731485        2.2343514038
H           1.6163249151       -0.8520106690        0.9432459299
H           1.6501412670        0.9090995505        0.7754443890
H           0.2426092643       -1.2584132076       -2.6580031277
H           1.3854828144       -1.3363913315       -1.3283466821
H          -0.1823514576       -2.1256644189       -1.1900502092
```

# Bibliography


[1] P. Pulay, *Chem. Phys. Lett.* **100**, 151 (1983).

[2] S. Saebø, P. Pulay, *J. Chem. Phys.* **86**, 914 (1987).

[3] M. Schütz, G. Hetzer, H.-J. Werner, *J. Chem. Phys.* **111**, 5691 (1999).

[4] M. Schütz, H.-J. Werner, *J. Chem. Phys.* **114**, 661 (2001).

[5] M. Schütz, H.-J. Werner, *Chem. Phys. Lett.* **318**, 370 (2000).

[6] M. Schütz, *J. Chem. Phys.* **113**, 9986 (2000).

[7] M. Schütz, *J. Chem. Phys.* **116**, 8772 (2002).

[8] H.-J. Werner, P. J. Knowles, R. Lindh, F. R. Manby, M. Schütz, P. Celani, T. Korona, G. Rauhut, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, C. Hampel, G. Hetzer, A. W. Lloyd, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, P. Palmieri, R. Pitzer, U. Schumann, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, Molpro, version 2006.3, a package of ab initio programs (2007). See http://www.molpro.net.

[9] J. W. Boughton, P. Pulay, *J. Comput. Chem.* **14**, 736 (1993).

[10] H.-J. Werner, K. Pflüger, *Ann. Reports in Comput. Chem.* **2**, 53 (2006).

[11] J. Pipek, P. G. Mezey, *J. Chem. Phys.* **90**, 4916 (1989).

[12] A. E. Reed, F. Weinhold, *J. Chem. Phys.* **83**, 1736 (1985).

[13] A. E. Reed, R. B. Weinstock, F. Weinhold, *J. Chem. Phys.* **83**, 735 (1985).

[14] J. C. Slater, *Phys. Rev.* **34**, 1293 (1929).

[15] A. Szabo, N. S. Ostlund, *Modern Quantum Chemistry* (Dover Pulications, New York, 1996).

[16] S. Grimme, *J. Chem. Phys.* **118**, 9095 (2003).

[17] P. Pulay, S. Saebø, W. Meyer, *J. Chem. Phys.* **81**, 1901 (1984).

[18] P. J. Knowles, H.-J. Werner, M. Schütz, *Modern Methods and Algorithms of Quantum Chemistry*, J. Grotendorst, ed. (NIC-Directors, Jülich, 2000), pp. 97–161.

[19] C. Hampel, K. A. Peterson, H.-J. Werner, *Chem. Phys. Lett.* **190**, 1 (1992).

[20] M. Urban, J. Noga, S. J. Cole, R. J. Bartlett, *J. Chem. Phys.* **83**, 4041 (1985).

[21] J. A. Pople, M. Head-Gordon, K. Raghavachari, *J. Chem. Phys.* **87**, 5968 (1987).

[22] K. Raghavachari, G. W. Trucks, J. A. Pople, M. Head-Gordon, *Chem. Phys. Lett.* **157**, 479 (1989).

[23] W. Kutzelnigg, *Localization and Delocalization in Quantum Chemistry*, O. Chalvet et al., ed. (D. Deidel, Dordrecht, 1975), p. 143.

[24] P. Y. Ayala, G. E. Scuseria, *J. Chem. Phys.* **110**, 3660 (1999).

[25] G. E. Scuseria, P. Y. Ayala, *J. Chem. Phys.* **111**, 8330 (1999).

[26] P. E. Maslen, M. Head-Gordon, *Chem. Phys. Lett.* **283**, 102 (1998).

[27] P. E. Maslen, M. Head-Gordon, *J. Chem. Phys.* **109**, 7093 (1998).

[28] S. Saebø, P. Pulay, *Chem. Phys. Lett.* **113**, 13 (1985).

[29] N. Flocke, R. J. Bartlett, *J. Chem. Phys.* **121**, 10935 (2004).

[30] S. F. Boys, *Quantum Theory of Atoms, Molecules, and the Solid State*, P. O. Löwdin, ed. (Academic Press, New York, 1966), pp. 253–262.

[31] C. Edmiston, K. Ruedenberg, *Rev. Mod. Phys.* **34**, 457 (1963).

[32] S. Saebø, W. Tong, P. Pulay, *J. Chem. Phys.* **98**, 2170 (1993).

[33] C. Hampel, H.-J. Werner, *J. Chem. Phys.* **104**, 6286 (1996).

[34] M. Schütz, G. Rauhut, H.-J. Werner, *J. Phys. Chem. A* **102**, 5997 (1998).

[35] B. Hartke, M. Schütz, H.-J. Werner, *J. Chem. Phys.* **239**, 561 (1998).

[36] N. Runeberg, M. Schütz, H.-J. Werner, *J. Chem. Phys.* **110**, 7210 (1999).

[37] L. Magnko, M. Schweizer, G. Rauhut, M. Schütz, H. Stoll, H.-J. Werner, *Phys. Chem. Chem. Phys.* **4**, 1006 (2002).

[38] T. Korona, K. Pflüger, H.-J. Werner, *Phys. Chem. Chem. Phys.* **6**, 2059 (2004).

[39] T. Hrenar, G. Rauhut, H.-J. Werner, *J. Phys. Chem. A* **110**, 2060 (2006).

[40] P. Hohenberg, W. Kohn, *Phys. Rev.* **136**, B864 (1964).

[41] A. D. Becke, *J. Chem. Phys.* **98**, 5648 (1993).

[42] A. D. Becke, *Phys. Rev. A* **38**, 3098 (1988).

[43] J. C. Slater, *Phys. Rev.* **81**, 385 (1951).

[44] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **37**, 785 (1988).

[45] S. H. Vosko, L. Wilk, M. Nusair, *Can. J. Phys.* **58**, 1200 (1980).

[46] M. J. S. Dewar, W. Thiel, *J. Am. Chem. Soc.* **99**, 4899 (1977).

[47] W. Thiel, A. Voityuk, *Theor. Chim. Acta* **81**, 391 (1992).

[48] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **107**, 3902 (1985).

[49] J. J. P. Stewart, *J. Comput. Chem.* **10**, 221 (1989).

[50] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R.Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, *J. Chem. Phys. B* **102**, 3586 (1998).

[51] P. A. Bash, M. J. Field, M. Karplus, *J. Am. Chem. Soc.* **109**, 8092 (1987).

[52] M. Sierka, J. Sauer, *Farad. Discuss. Chem. Soc.* **106**, 41 (1997).

[53] S. Dapprich, I. Komáromi, K. S. Byun, K. Morokuma, M. J. Frisch, *J. Mol. Struct. THEOCHEM* **461**, 1 (1999).

[54] T. Vreven, K. S. Byun, I. Komaromi, S. Dapprich, J. A. Montgomery, K. Morokuma, M. J. Frisch, *J. Chem. Theory Comput.* **2**, 815 (2006).

[55] G. E. Briggs, J. B. S. Haldane, *Biochem. J.* **19**, 339 (1925).

[56] A. ElAzhary, G. Rauhut, P. Pulay, H.-J. Werner, *J. Chem. Phys.* **108**, 5185 (1998).

[57] G. Rauhut, H.-J. Werner, *Phys. Chem. Chem. Phys.* **3**, 4853 (2001).

[58] M. Schütz, H.-J. Werner, R. Lindh, F. R. Manby, *J. Chem. Phys.* **121**, 737 (2004).

[59] G. Rauhut, A. E. Azhary, F. Eckert, U. Schumann, H.-J. Werner, *Spectrochim. Acta* **A 55**, 647 (1999).

[60] G. Rauhut, H.-J. Werner, *Phys. Chem. Chem. Phys.* **5**, 2001 (2003).

[61] N. J. Russ, T. D. Crawford, *Chem. Phys. Lett.* **400**, 104 (2004).

[62] J. Gauss, H.-J. Werner, *Phys. Chem. Chem. Phys.* **2**, 2083 (2000).

[63] N. J. Russ, T. D. Crawford, *J. Chem. Phys.* **121**, 691 (2004).

[64] W. D. Allen, H. F. S. III, *J. Chem. Phys.* **89**, 329 (1988).

[65] Q. Cui, K. Morokuma, *J. Chem. Phys.* **107**, 4951 (1997).

[66] J. G. Hill, J. A. Platts, H.-J. Werner, *Phys. Chem. Chem. Phys.* **8**, 4072 (2006).

[67] P. Valtazanos, K. Ruedenberg, *Theor. Chim. Acta* **69**, 281 (1986).

[68] W. Quapp, M. Hirsch, D. Heidrich, *Theor. Chim. Acta* **100**, 285 (1993).

[69] A. L. L. East, *J. Chem. Phys.* **108**, 3574 (1998).

[70] F. Jensen, *Chem. Phys. Lett.* **196**, 368 (1992).

[71] S. Humbel, S. Sieber, K. Morokuma, *J. Chem. Phys.* **105**, 1959 (1996).

[72] I. Lee, C. K. Kim, D. S. Chung, B.-S. Lee, *J. Org. Chem.* **59**, 4490 (1994).

[73] J. J. Blavins, D. L. Cooper, P. B. Karadakov, *J. Phys. Chem. A* **108**, 914 (2004).

[74] A. Streitwieser, G. S.-C. Choy, F. Abu-Hasanayan, *J. Am. Chem. Soc.* **119**, 5013 (1997).

[75] T. H. Dunning, Jr., *J. Chem. Phys.* **90**, 1007 (1989).

[76] R. A. Kendall, T. H. Dunning, R. J. Harrison, *J. Chem. Phys.* **96**, 6796 (1992).

[77] H.-J. Werner, F. R. Manby, P. Knowles, *J. Chem. Phys.* **118**, 8149 (2003).

[78] F. Bernardi, S. F. Boys, *Mol. Phys.* **19**, 553 (1970).

[79] D. Cremer, A. Wu, E. Kraka, *Phys. Chem. Chem. Phys.* p. 674 (2001).

[80] F. Jensen, *Introduction to Computational Biochemistry* (John Wiley & Sons, England, 1999).

[81] J. P. Foster, F. Weinhold, *J. Am. Chem. Soc.* **102**, 7211 (1980).

[82] A. E. Reed, L. A. Curtiss, F. Weinhold, *Chem. Rev.* **88**, 899 (1988).

[83] E. D. Glendening, J. K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, F. Weinhold, Nbo 5.0 (2001). Theoretical Chemistry Institute, University of Winsconsin, Madison.

[84] R. Polly, H.-J. Werner, F. R. Manby, P. J. Knowles, *Mol. Phys.* **102**, 2311 (2004).

[85] K. Wermann, M. Walther, W. Gunther, H. Gorls, E. Anders, *J. Org. Chem.* **66**, 720 (2001).

[86] S. Y. Re, K. Morokuma, *Theor. Chim. Acta* **112**, 59 (2004).

[87] J. A. Anderson, B. W. Hopkins, J. L. Chapman, G. S. Tschumper, *J. Mol. Struct. THEOCHEM* **771**, 65 (2006).

[88] M. Karplus, *J. Phys. Chem. B* **104**, 11 (2000).

[89] W. J. H. van Berkel, F. Müller, *J. Chem. Phys.* **179**, 307 (1989).

[90] J. Vervoort, I. M. C. M. Rietjens, W. J. H. van Berkel, *Eur. J. Biochem.* **206**, 479 (1992).

[91] L. Ridder, A. J. Mulholland, J. Vervoort, I. M. C. M. Rietjens, *J. Am. Chem. Soc.* **120**, 7641 (1998).

[92] L. Ridder, J. N. Harvey, I. M. C. M. Rietjens, J. Vervoort, A. J. Mulholland, *J. Phys. Chem. B* **107**, 2118 (2003).

[93] S. R. Billeter, C. F. W. Hanser, T. Z. Mordasini, M. Scholten, W. Thiel, W. F. van Gunsteren, *Phys. Chem. Chem. Phys.* **3**, 688 (2001).

[94] H. M. Senn, S. Thiel, W. Thiel, *J. Chem. Theory Comput.* **1**, 494 (2005).

[95] P. Sherwood, A. H. de Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. S. voll, A. Fahmi, A. Schäfer, C. Lennartz, *J. Mol. Struc.* **632**, 1 (2003).

[96] D. L. Gatti, B. Entsch, D. P. Ballou, M. L. Ludwig, *Biochemistry* **35**, 567 (1996).

[97] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, I. G. Tironi, *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Biomos b.v., Zürich and Groningen, VdF Hochschulverlag, ETH Zürich, Zürich, 1996).

[98] J. G. Kirkwood, *J. Chem. Phys.* **3**, 300 (1935).

[99] A. Schäfer, C. Huber, R. Ahlrichs, *J. Chem. Phys.* **100**, 5829 (1994).

[100] F. R. Manby, H.-J. Werner, T. B. Adler, A. J. May, *J. Chem. Phys.* **124**, 094103 (2006).

[101] G. M. Torrie, J. P. Valleau, *Chem. Phys. Lett.* **28**, 578 (1974).

[102] F. P.-D. Martin, R. Dumas, M. J. Field, *J. Am. Chem. Soc.* **122**, 788 (2000).

[103] J. J. Ruiz-Pernía, E. Silla, I. Tuñón, S. Martí, *J. Phys. Chem. B* **110**, 17663 (2006).

[104] A. Y. Lee, J. D. Stewart, J. Clardy, B. Ganem, *Chem. Biol.* **2**, 195 (1995).

[105] P. Kast, M. Asif-Ullah, D. Hilvert, *Tetrahedron Lett.* **37**, 2691 (1996).

[106] P. D. Lyne, A. J. Mulholland, W. G. Richards, *J. Am. Chem. Soc.* **117**, 11345 (1995).

[107] A. Kienhöfer, P. Kast, D. Hilvert, *J. Am. Chem. Soc.* **125**, 3206 (2003).

[108] C. R. W. Guimarães, M. Udier-Blagović, I. Tubert-Brohman, W. L. Jorgensen, *J. Chem. Theory Comput.* **1**, 617 (2005).

[109] S. E. Worthington, A. E. Roitberg, M. J. Krauss, *J. Phys. Chem. B* **105**, 7087 (2001).

[110] S. Hur, T. C. Bruice, *Proc. Natl. Acad. Sci.* **100**, 12015 (2003).

[111] K. E. Ranaghan, A. J. Mulholland, *Chem. Comm.* pp. 1238–1239 (2004).

[112] K. E. Ranaghan, L. Ridder, B. Szefczyk, W. A. Sokalski, J. C. Hermann, A. J. Mulholland, *Org. Biomol. Chem.* **2**, 968 (2004).

[113] X. Zhang, T. C. Bruice, *Proc. Natl. Acad. Sci.* **102**, 18356 (2005).

[114] F. Clayessens, K. E. Ranaghan, F. R. Manby, J. N. Harvey, A. J. Mulholland, *Chem. Comm.* pp. 5068–5070 (2005).

[115] Y. Chook, H. Ke, W. Lipscomb, *Proc. Natl. Acad. Sci.* **90**, 8600 (1993).

[116] A. Crespo, D. A. Scherlis, M. A. Martí, P. Ordejon, A. E. Roitberg, D. A. Estrin, *J. Phys. Chem. B* **107**, 13728 (2003).

[117] J. N. Harvey, *Faraday Discuss.* **127**, 165 (2004).

[118] I. Schödinger, Jaguar, 4.0 (1996-2001).

[119] J. W. Ponder, Tinker: Software tools for molecular design, v4.0 (2003).

[120] J. E. Subotnik, A. Sodt, M. H. Gordon, *J. Chem. Phys.* **125**, 074116 (2006).