Ralf Meyer BSc

Interpolation of potential energy surfaces using force field and semi-empirical quantum chemistry methods

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Technical Physics

submitted to

Graz University of Technology

Supervisor

Ass.Prof. Mag. Dipl.-Ing. DDr. Andreas W. Hauser

Institute of Experimental Physics

Graz, July 2016

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Abstract

Accurate interpolation techniques for potential energy surfaces bear the potential to speed up quantum chemistry calculations by reducing the number of necessary “ab initio” evaluations. The most common interpolation methods, such as high dimensional cubic splines or interpolating moving least squares, are motivated strictly by mathematical arguments and require a large number of ab initio reference points.

In this thesis, alternative interpolation schemes based on physically motivated methods such as force fields or extended Hückel theory, are investigated. Since these methods already predict the functional course of the potential energy surface reasonably well, an interpolation to the same accuracy can be done using fewer reference points than for mathematically motivated interpolation routines.

Kurzfassung

Mit Hilfe präziser Interpolationsmethoden für Potentialenergieflächen können quantenchemische Rechnungen durch Einsparen von “ab initio” Auswertungen beschleunigt werden. Typische Interpolationsmethoden wie mehrdimensionale kubische Splines oder die “Interpolating Moving Least Squares” Methode, sind durch mathematische Argumente motiviert und benötigen große Mengen an Referenzwerten.

In der vorliegenden Arbeit werden alternative Interpolationsroutinen untersucht, die auf physikalisch motivierten Ansätzen wie Kraftfeldern oder der erweiterten Hückel-Theorie basieren. Da diese Methoden den Verlauf der Potentialenergiefläche bereits einigermaßen gut beschreiben, kann eine Interpolation zur selben Genauigkeit mit geringerer Anzahl an Referenzpunkten erreicht werden als für mathematisch motivierte Interpolationsmethoden.

Acknowledgement

First and foremost I would like to thank my supervisor Ass.Prof. Mag. Dipl.-Ing. DDr Andreas W. Hauser, not only for giving me the opportunity to write this thesis but also for his encouragement and support in our, often hour long, discussions.

I want to thank all of my colleagues at the Institute of Experimental Physics as well as my fellow students. Especially Markus Bainschab, Michael Draxler, Paul Maierhofer and Michael Rumetshofer, who wrote their thesis at the same time, helped me through my numerous setbacks and motivation drops.

Last but not least I want to thank my family for their continuing support throughout my university studies.

Contents

1 Introduction

A key objective of quantum chemistry is to solve the Schrödinger equation for the molecular Hamiltonian:

$$H = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{a=1}^{M} \frac{1}{2M_a}\nabla_a^2 - \sum_{i=1}^{N}\sum_{a=1}^{M} \frac{Z_a}{r_{ia}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} + \sum_{a=1}^{M}\sum_{b>a} \frac{Z_a Z_b}{R_{ab}}. \tag{1.1}$$

For the problems discussed in this thesis it will be sufficient to use the time-independent Schrödinger equation:

$$H|\Psi\rangle = E|\Psi\rangle. \tag{1.2}$$

Equation 1.1 is written in atomic units. These simplify the equations by absorbing natural constants that appear in the equation into the definition of new units. An overview of the most important quantities is given in Table 1.1.

Table 1.1: Relation of atomic units to SI units.

physical quantity	conversion factor	value in SI	name of unit
energy	$E_h = \frac{\hbar^2}{m_e a_0^2} = \frac{e^2}{4\pi\epsilon_0 a_0}$	$4.359744650 \cdot 10^{-18}$ J	hartree
length	$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2}$	$0.52917721067 \cdot 10^{-10}$ m	bohr
mass	m_e	$9.10938356 \cdot 10^{-31}$ kg	
charge	e	$1.6021766208 \cdot 10^{19}$ C	

This complicated many-body problem can not be solved without first employing several approximations. Only a selection of typical simplifications is presented in this short introduction. However, note that for the concept of interpolation explored in this thesis the results of a quantum chemical calculation will always be treated as an exact solution, no matter which approximations were employed.

1.0.1 Born-Oppenheimer approximation

A first simplification of the molecular Schrödinger equation is given by the Born-Oppenheimer approximation. It decouples the electronic movement from the motion of the nuclei. A simple argument that justifies such a separation is the ratio of the electron mass and the mass of nucleons $m_p/m_e \approx 1838$. The mathematical derivation of the approximation can be found in chapter 7. The total molecular Hamiltonian from equation 1.1 is simplified by neglecting the second term, the kinetic energy of the nuclei, and by treating the last term, the nuclear repulsion, as a constant value added to the electronic energy. The remaining terms are called the electronic Hamiltonian:

$$H_{elec} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{a=1}^{M} \frac{Z_a}{r_{ia}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \tag{1.3}$$

The solution of 1.2 with 1.3 is the electronic wave function $\Phi(\mathbf{r}, \mathbf{R})$, a function of the $3N$ electronic coordinates $\mathbf{r}$ with the nuclear coordinates $\mathbf{R}$ as parameters. By solving the electronic Schrödinger equation for various $\mathbf{R}$, every nuclear geometry can be assigned a corresponding energy. This leads to the picture of a potential energy surface (PES) on which the nuclear motion takes place.

1.1 Ab initio calculations

The term *"ab intio"* (latin: from the beginning) refers to calculations done without empirical parameters or experimental results. Using Slater determinants as an ansatz to solve the electronic Schrödinger equation leads to the Hartree Fock equations. This set of integrodifferential equations is solved by transforming it into a set of algebraic equations using a basis set expansion for the electronic wave function. In the following section some of the most widely used quantum chemistry methods and their formal scaling with the number of basis functions K are described briefly.

Hartree Fock $\mathcal{O}(\mathbf{K^4})$: The Hartree Fock (HF) method is the foundation of most quantum chemistry calculations. In this method, the electron-electron interaction is simplified, only describing the motion of one electron in the mean field created by the remaining electrons. This requires the HF equations to be solved iteratively, because the positions of the electrons can only be known after solving for the molecular orbitals. The HF method is often referred to as self-consistent field (SCF) method. Despite this mean field approximation, the resulting HF wave function accounts for $\sim 99\%$ of the total energy of the system. Unfortunately, describing chemical phenomena often requires the knowledge of the remaining 1%, typically referred to as electron correlation energy. The other methods presented here, often called post-HF methods, try to improve on the Hartree Fock energy by recovering at least parts of the correlation energy.

Møller-Plesset perturbation theory second order $\mathcal{O}(\mathbf{K^5})$: Møller-Plesset perturbation theory (MPx, where x is the perturbation order) reintroduces the exact electron-electron interaction as a perturbation to the Hartree Fock Hamiltonian. In the context of perturbation theory the HF ansatz corresponds to a first order treatment. Therefore, second order is the first correction to the HF energy. The scaling of the computational effort $\mathcal{O}(K^{3+x})$ depends on the perturbation order x.

Configuration interaction with singles and doubles $\mathcal{O}(\mathbf{K^6})$: In configuration interaction (CI) methods the total wave function is written as a linear combination of slater determinants. This linear combination consists of the HF ground state and a selection of excited state determinants. The formal scaling depends on the number of electrons that are excited from their ground state. The example given here, with single and double excitations (CISD), gives $\mathcal{O}(K^6)$. If triple excitations are included (CISDT), the computational effort scales as $\mathcal{O}(K^8)$.

Coupled cluster with singles and doubles $\mathcal{O}(\mathbf{K^6})$: In perturbation theory, the corrections from all types of excitations (S single, D double, T triple, ...) are calculated up to a specific order. In coupled cluster methods (CC), the contributions of a given number of excitations are included to infinite order. Using singles and doubles (CCSD) the method scales with $\mathcal{O}(K^6)$. If triple excitations are included (CCSDT), the computational effort $\mathcal{O}(K^8)$ is too high for all but the smallest molecules. Therefore, the triples contribution is typically calculated from perturbation theory (MP4) such as in CCSD(T), the current "gold-standard" in quantum chemistry. CCSD(T) scales with $\mathcal{O}(K^7)$.

Density functional theory $\mathcal{O}(\mathbf{K^4})$: Density functional theory (DFT) is based on the Hohenberg-Kohn theorems, which state that the ground state energy can be expressed as a functional of the electron density. This could, in theory, decrease the computational effort significantly, as the density is only a function in the three dimensional space, while the electronic wave function is a function of all $3N$ electronic coordinates. However, to reach comparable accuracy, many DFT methods (hybrids) calculate the kinetic energy as well as parts of the exchange contribution using wave function based methods. These methods, therefore, typically scale $\mathcal{O}(K^4)$ just as HF theory.

1.2 Interpolation techniques

For a given set of reference values f_i at corresponding reference points $\mathbf{x_i}$, a function $f(\mathbf{x})$ is called a interpolation function if it

(a) reproduces the reference values $f(\mathbf{x_i}) = f_i$,

(b) is smooth (continuous and differentiable) on the whole definition area.

The following section presents the mathematical basics for several of the interpolation schemes commonly used for potential energy surfaces.

1.2.1 Shepard interpolation

This interpolation scheme, which is based on weighted averages, was first proposed by Donald Shepard in 1968 [1]. Originally, it was intended for the interpolation of two-dimensional data taken from areas such as meteorology and geography. In its simplest form, often referred to as inverse distance weighting, the interpolation function is defined as follows:

$$f(\mathbf{x}) = \begin{cases} \frac{\sum_i w_i(\mathbf{x}) f_i}{\sum_i w_i(\mathbf{x})}, & \text{if } d(\mathbf{x}, \mathbf{x_i}) \neq 0 \text{ for all i} \\ f_i, & \text{if } d(\mathbf{x}, \mathbf{x_i}) = 0 \text{ for some i} \end{cases} \tag{1.4}$$

with $d(\mathbf{x}, \mathbf{x_i})$ as the distance between two points and a weight function $w_i(\mathbf{x})$ defined as

$$w_i(\mathbf{x}) = \frac{1}{d(\mathbf{x}, \mathbf{x_i})^p}, \tag{1.5}$$

with power $p \geq 1$. Note that for $p < 1$, the resulting interpolation function is not differentiable. Shepard also considered alternative weight functions using a constant number of reference points, cut off distances, and direction dependent weights.

One major problem of the Shepard method is so called “flat-spot” phenomenon. For any exponent $p > 1$, the derivative of the interpolation function is zero at the reference points. Figure 1.1 illustrates this problem for a sine curve.

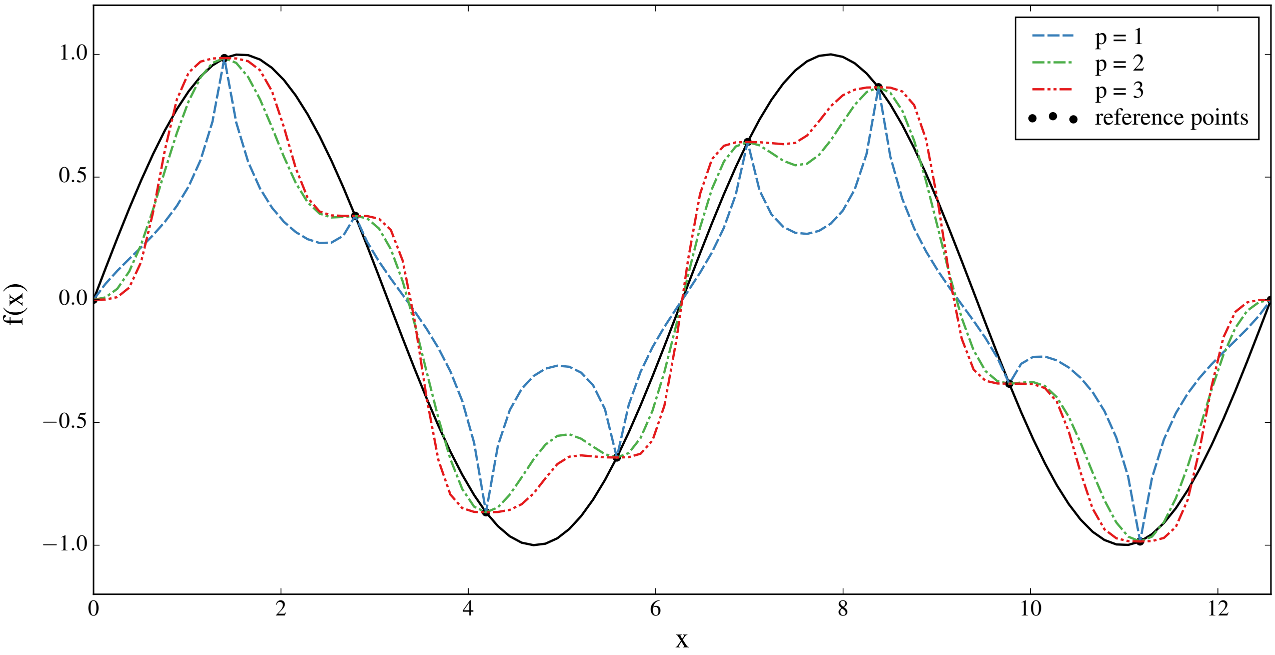


Figure 1.1: Problems of Shepard’s method when interpolating a sine wave.

For $p = 1$ the interpolation function has corners at the reference points. For $p = 2$ the “flat-spot” phenomenon becomes obvious and gets even worse more for $p = 3$. One reason for this phenomenon is the diverging weight function at the reference points. All interpolation functions plotted in figure 1.1 are calculated using normalized weights to eliminate the divergence:

$$v_i(\mathbf{x}) = \frac{w_i(\mathbf{x})}{\sum_j w_j(\mathbf{x})} \tag{1.6}$$

The normalized weights have the following properties:

$$\begin{array}{lll} \text{(a)} & v_i(x_j) = \delta_{ij} & \\ \text{(b)} & 0 \leq v_i(x) \leq 1 & \text{for all x} \\ \text{(c)} & \sum_{i=0}^{N} v_i(x) = 1 & \text{for all x} \\ \text{(d)} & v_i(x) \rightarrow N^{-1} & \text{as } d(x,0) \rightarrow \infty \end{array} \tag{1.7}$$

McLain [2] suggested a shift of the denominator in order to avoid the divergence:

$$w_i(\mathbf{x}) = \frac{1}{d(\mathbf{x}, \mathbf{x_i})^p + \epsilon} \tag{1.8}$$

However, it turns out that the choice of the parameter ϵ has a large impact on the quality of the interpolation. Due to this flaw and the lack of a non-empirical way of determining the parameter the normalized weight function is preferable.

Figure 1.1 shows that the “flat-spot” phenomenon can not be eliminated solely by avoiding the divergence in the weight function. Within the context of the simple Shepard method this problem can only be solved by including first and second derivatives, which are often computationally too expensive for high level ab initio calculations.

1.2.2 Interpolating moving least squares

The interpolating moving least squares (IMLS) method is the generalization of Shepard’s method to higher order polynomials. Shepard’s method can be considered as zeroth order IMLS. This method eliminates the “flat-spot” phenomenon for orders larger than zero, but involves more computational effort, since a system of coupled linear equations has to be solved at every point the interpolation function is evaluated. The derivation presented here, in one dimension for simplicity, follows Lancaster and Salkauskas [3] as well as Maisuradze and Thompson [4].

Using a polynomial of order m as basis for the interpolation function,

$$p(x) = \sum_{n=0}^{m} a_n(x) x^n \tag{1.9}$$

the coefficients $a_n(x)$ are calculated by minimizing the weighted square deviations,

$$\sum_{i=0}^{N} w_i(x) \left[p(x_i) - f_i\right]^2 \overset{!}{=} \min, \tag{1.10}$$

with f_i as the values at the reference points x_i. Taking the derivative of equation 1.10 with respect to the coefficients a_n yields $m+1$ normal equations:

$$\begin{aligned}\left[\sum_i w_i(x)x_i^0\right] a_0 + \cdots + \left[\sum_i w_i(x)x_i^m\right] a_m &= \sum_i w_i(x) f_i \\ \left[\sum_i w_i(x)x_i\right] a_0 + \cdots + \left[\sum_i w_i(x)x_i^{m+1}\right] a_m &= \sum_i w_i(x) x_i f_i \\ &\vdots \\ \left[\sum_i w_i(x)x_i^m\right] a_0 + \cdots + \left[\sum_i w_i(x)x_i^{2m}\right] a_m &= \sum_i w_i(x) x_i^m f_i\end{aligned} \tag{1.11}$$

This set of equations 1.11 can be rewritten in matrix form as:

$$\mathbf{B}^T \cdot \mathbf{W} \cdot \mathbf{B} \cdot \mathbf{a} = \mathbf{B}^T \cdot \mathbf{W} \cdot \mathbf{f}, \tag{1.12}$$

where $\mathbf{B}$ contains the reference points taken to different powers, $\mathbf{W}$ contains the weights, $\mathbf{a}$ is the coefficient vector and $\mathbf{f}$ is a vector of the reference values:

$$\mathbf{B} = \begin{bmatrix} 1 & x_0 & \cdots & x_0^m \\ 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \cdots & x_N^m \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} w_0(x) & 0 & \cdots & 0 \\ 0 & w_1(x) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_N(x) \end{bmatrix} \quad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} \quad \mathbf{f} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_N \end{bmatrix} \tag{1.13}$$

The matrix $\mathbf{B}$, known as Vandermonde matrix, is ill conditioned. Instead of direct inversion of $\mathbf{B}^T \cdot \mathbf{W} \cdot \mathbf{B}$, the equation should be solved by singular value decomposition (SVD). Just as in Shepard's method, the divergence of the weight function has to be avoided. In addition, the weight function has to ensure that at least $m+1$ reference points are included in the calculation at every point. Otherwise, the matrix equation 1.12 is underdetermined, which can lead to problems during interpolation.

Figure 1.2 shows that IMLS can solve the "flat-spot" phenomenon. Both plotted interpolation functions use normalized inverse distance weights with $p = 2$.

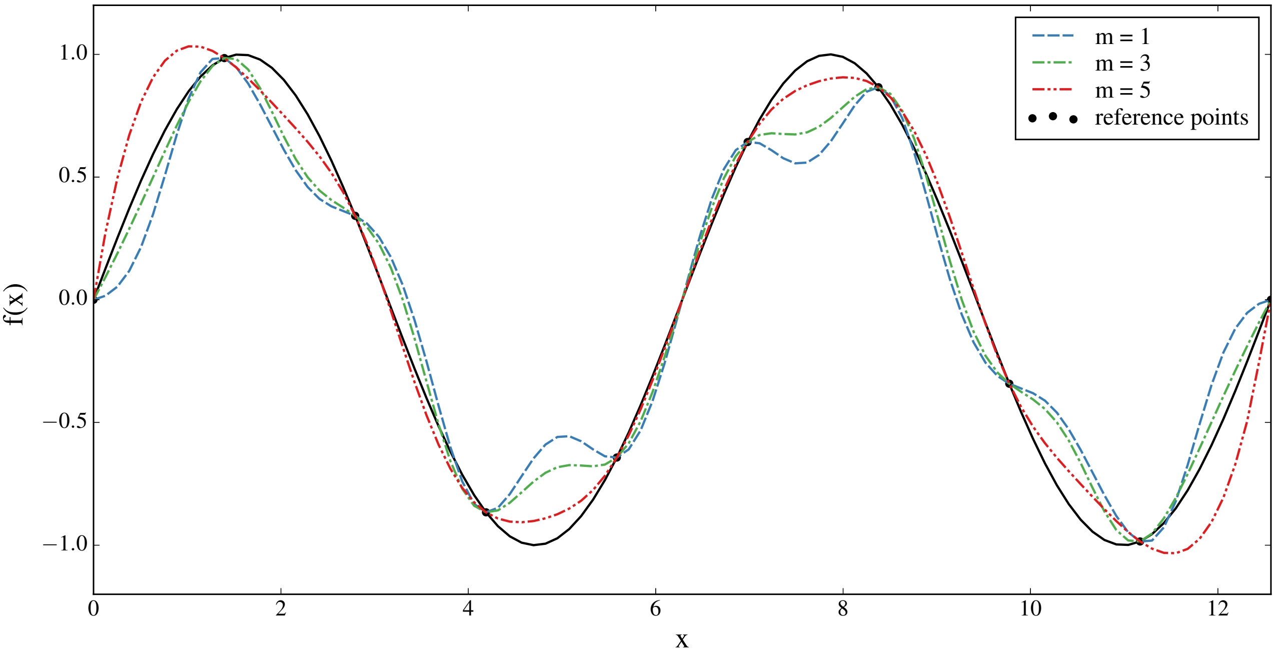


Figure 1.2: Interpolation of a sine wave using first and third order IMLS.

1.2.3 Splines

One of the most widely used interpolation schemes is the spline method. The interpolation function is build piecewise from n-th degree polynomials between the reference functions. The coefficients of the polynomials are determined in a way to ensure (n-1 time) differentiability. The most widespread version is the natural cubic spline interpolation. As the name suggests, it uses cubic polynomials as interpolation functions. Every cubic spline, interpolating $N+1$ reference points $x_0 < x_1 < ... < x_N$, can be written in the form [3]

$$f(x) = \alpha x^3 + \beta x^2 + \gamma x + \delta + \sum_{i=1}^{N-1} a_i |x - x_i|^3. \tag{1.14}$$

This gives a total of $N+3$ free parameters for only $N+1$ reference points. Therefore, two more conditions are needed to calculate a unique interpolation curve. Different solutions have been proposed in the literature, depending on the specific application:

- "natural" cubic splines: the second derivative at the two outermost points is chosen to be zero $f''(x_0) = f''(x_N) = 0$.
- periodic boundary conditions: the first piecewise polynomial f_1 is connected to the last f_N and the usual condition of continuous first and second derivative is enforced $f_1'(x_0) = f_N'(x_N)$, $f_1''(x_0) = f_N''(x_N)$.
- clamped string: The first derivative at the endpoints is set to a fixed value $f_1'(x_0) = f_0'$, $f_N'(x_N) = f_N'$.
- not-a-knot: The third derivative, which is a constant for cubic polynomials, is set to the same value from the second to the penultimate reference point.

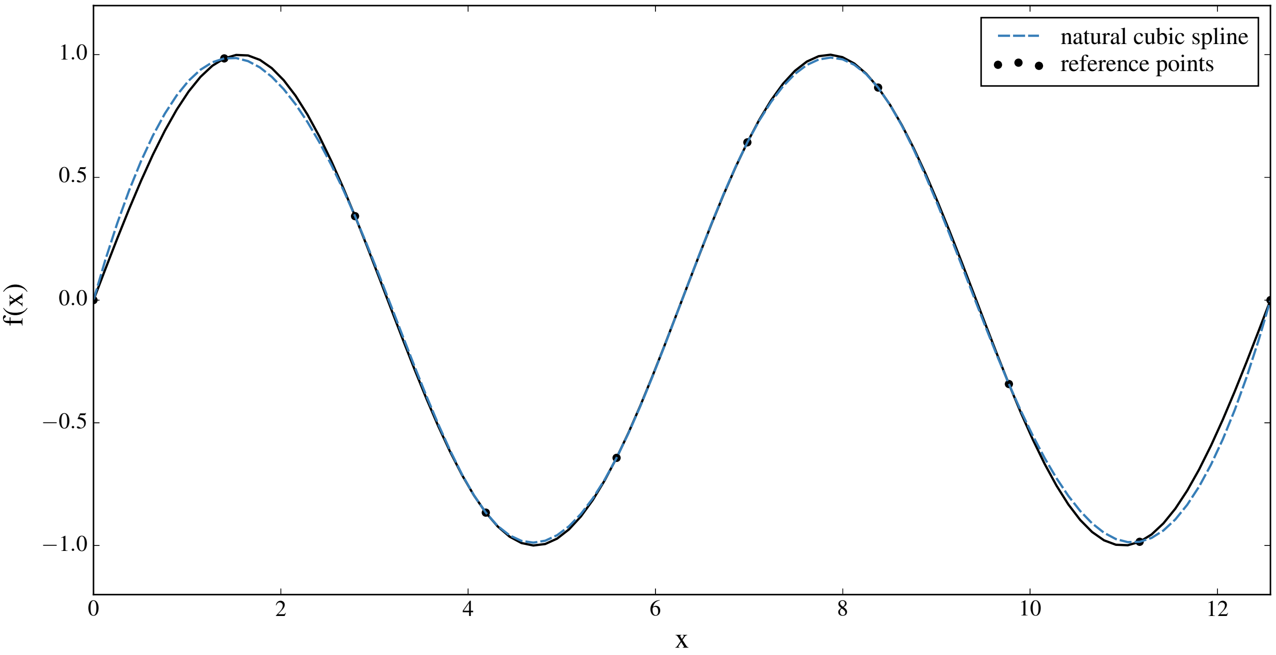


Figure 1.3: Natural cubic spline interpolating a sine wave.

While the natural cubic spline excels at interpolating the simple sine function, it often shows oscillations at points of high curvature in the reference function and has a bad limiting behaviour for low reference point densities. Extrapolation from a given set of reference points can only be done using prior knowledge of the reference function.

1.2.4 Neural networks

It can be shown that using a sigmoid neural network with one hidden layer any function can be fitted to arbitrarily high precision [5]. Figure 1.4 shows a schematic of such a single-hidden-layer network.

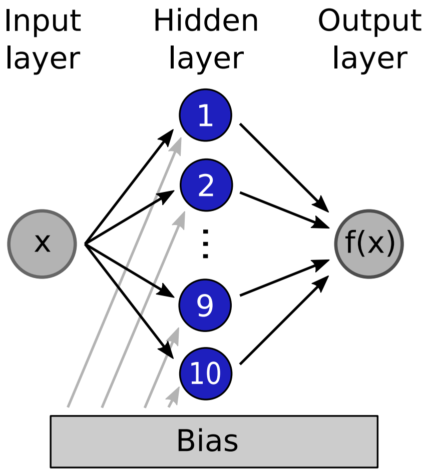


Figure 1.4: Schematic depiction of a single-hidden-layer neural network

Recently, this rapidly growing field of computer science has been applied in various forms to quantum chemistry. One of the possible applications is to train neural networks to completely replace ab initio calculations. In the approach of Behler and Parrinello [6] a whole system of subnets is trained to reproduce the energy contributions of each atom in a local environment. Their goal is to give a generally applicable method without the need for further ab initio evaluations. Even though reference points are needed for the training, this approach goes far beyond what is typically considered as fitting or interpolation. Neural networks can also be used for fitting and interpolating of given potential energy surface data points. Manzhos et al. [7] presented a molecule-independent fitting method based on a single-hidden-layer sigmoid neural network. They do, however, acknowledge that their approach needs large numbers of reference points to determine the many parameters of the network.

Figure 1.5 shows the results of two single-hidden-layer feed forward neural networks with six and ten neurons in the hidden layer respectively trained to interpolate the 10 reference points of the sine curve. Supervised learning by back-propagation with a learning rate $\eta = 0.05$ and a momentum $\alpha = 0.1$ is used. A neural network with six hidden neurons does not reproduce the reference values and therefore only fits the reference curve, but can not be considered as interpolation technique. In order for a neural network to be able to interpolate the N given target values at least N hidden neurons are needed.

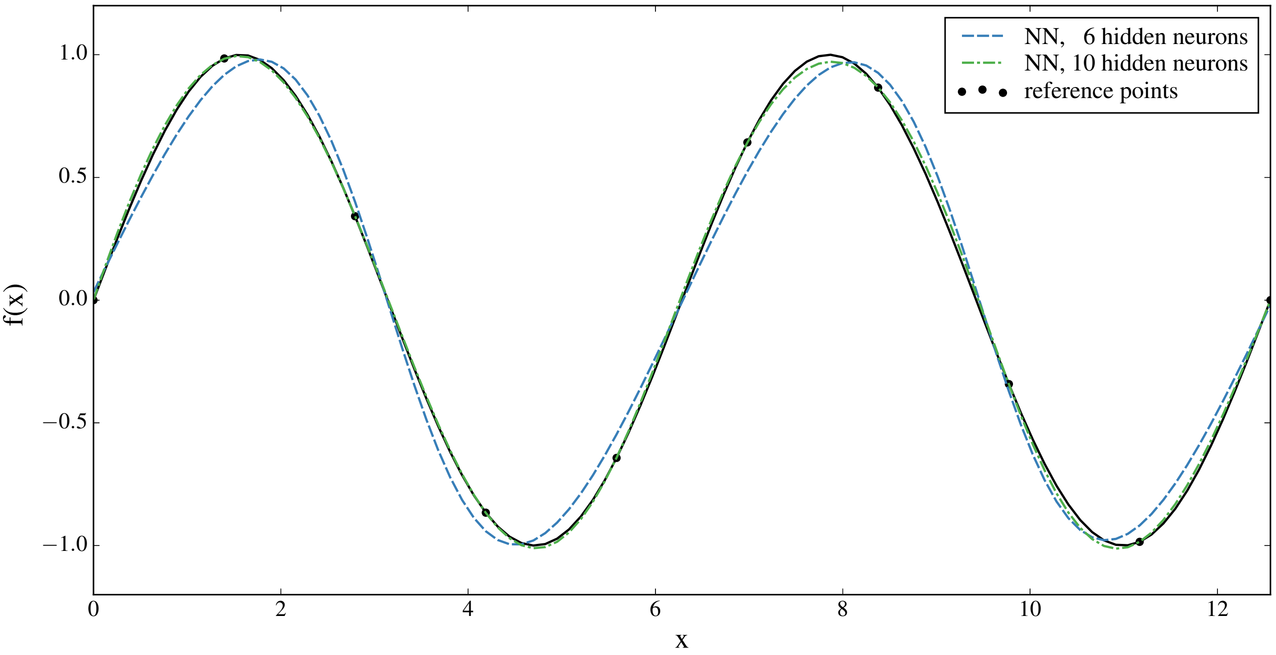


Figure 1.5: Interpolation of the sine curve using single-hidden-layer feed forward neural networks with 6 and 10 hidden neurons.

1.3 Methodology and outline of this thesis

In the course of this thesis, different physically motivated models are used to interpolate the PES. Ideally, the parameters of the model are constant throughout large parts of the PES, because the interpolation function would then only require few reference points. In order to test the parameter dependence of the models, they are first fitted to a reference trajectory (see section 1.3.1). If the model is flexible enough to fit the reference curve, ideally using only few parameters, an IMLS-inspired variation of the linear parameters of the model is investigated. The accuracy of the fit and interpolation functions is quantified by the root mean square deviation (RMSD) from the reference curve:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(f(x_i) - f_i)^2}{N}}. \tag{1.15}$$

The structure of the thesis roughly follows our work chronologically. Starting from the simplest chemically informed models, the force fields, we tried to incorporate increasingly sophisticated molecular orbital theory approaches in order to take advantage of additional information provided during typical "ab initio" evaluations of energies at the reference points. Increasing the level of theory in the models does, however, come with the trade off of larger computational effort.

The goal for the accuracy is set to about 1 kcal/mol, which is typically referred to as "chemical accuracy". Depending on the field of application, different requirements on the accuracy of interpolation functions are set.

1.3.1 Reference curve

Throughout this thesis, the discussed methods are evaluated and compared using a reference curve for the butadiene molecule. This reference curve corresponds to an ab initio molecular dynamics trajectory calculated with the velocity Verlet algorithm. Starting from the equilibrium geometry, the atoms are assigned a velocity based on a Boltzmann distribution for $T = 298.0$ K. A total of 200 time steps of 0.48 femtoseconds each were evaluated based on the ab initio forces from a Hartree-Fock calculation using a cc-pVDZ basis set [8].

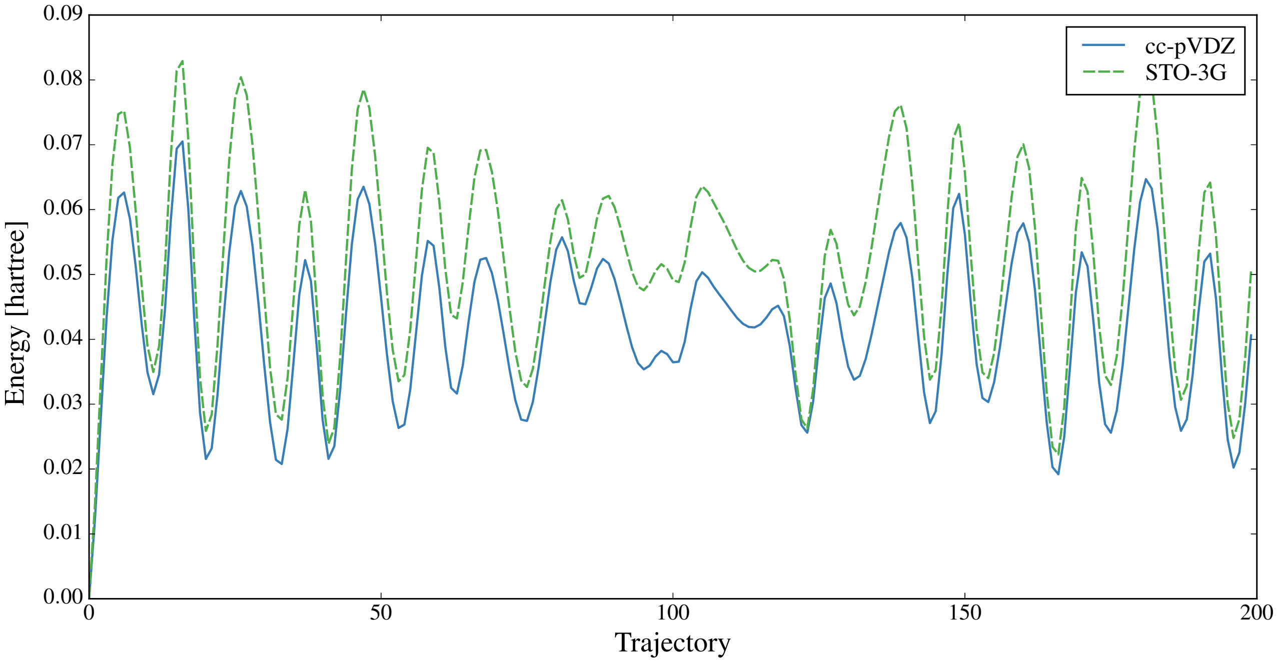


Figure 1.6: Reference trajectory of butadiene evaluated using cc-pVDZ and STO-3G basis sets.

Figure 1.6 shows the reference curve plotted over the trajectory. Note that the x-axis shows the trajectory point index and not the time in femtoseconds as to emphasize the discrete nature of these evaluations. While the potential curve itself is smooth, the reference curve shows jagged peaks due to this discretization.

These calculated reference geometries were also evaluated using a smaller basis (STO-3G [9]). This second reference curve has been used in the extended Hückel interpolation approach (see chapter 3) for the sake of simplified and less costly integral evaluation in the "home-grown" code.

1.3.2 Used programs

The TINKER molecular modeling software [10] and the LAMMPS Molecular Dynamics Simulator [11] are used to verify the correct implementation of the used force fields in the "home-grown" Python based molecular mechanics code. The extended Hückel molecular orbital package YAeHMOP (yet another extended Hückel molecular orbital package) [12] is used as reference for an extended Hückel theory implementation in Python. The ab initio packages Qchem [13] and Molpro [14] are used to generate the Hartree-Fock reference potential energy surfaces. The quantum chemistry programs PyQuante [15] and PySCF [16] (both in the Python language) are used (in modified form) for integral evaluations.

1.4 Potential applications for interpolation

PES interpolation bears the potential to accelerate quantum chemistry calculations by reducing the number of necessary ab initio evaluations. Depending on the particular type of calculation, different interpolation methods prove to be most useful.

1.4.1 Total or partial potential energy surface

Interpolating the total potential energy surface is the ultimate goal of all interpolation techniques. This requires the interpolation function to be flexible enough to represent any possible functional form. Typical applications include the fitting and interpolating of ab initio calculations to experimental values as well as the calculation of spectroscopic data such as vibrational frequencies. Another possible application is the interpolation of the total PES prior to a molecular dynamics simulation. If the PES can be evaluated cheaply, molecular dynamics can be done for large time scales or large systems. However, note that all of these applications require a highly accurate interpolation, i.e. a very large number of reference points.

1.4.2 On-the-fly interpolation

While interpolation is typically done after the calculation of ab initio points, an on-the-fly interpolation may prove useful for certain types of applications. The information gained from the interpolation function could be used to determine the molecular geometry for the next ab initio evaluations. Possible applications include transition state searches, reaction coordinate following and ab initio molecular dynamics, in particular for studying of anharmonic effects such as internal vibrational relaxation via normal mode projection techniques [17]. The number of available reference points in these calculations is significantly smaller (< 50). Also, the requirements on the accuracy are not as demanding. The interpolation function is only supposed to give qualitatively correct estimates of the local shape of the PES.

2 Force fields

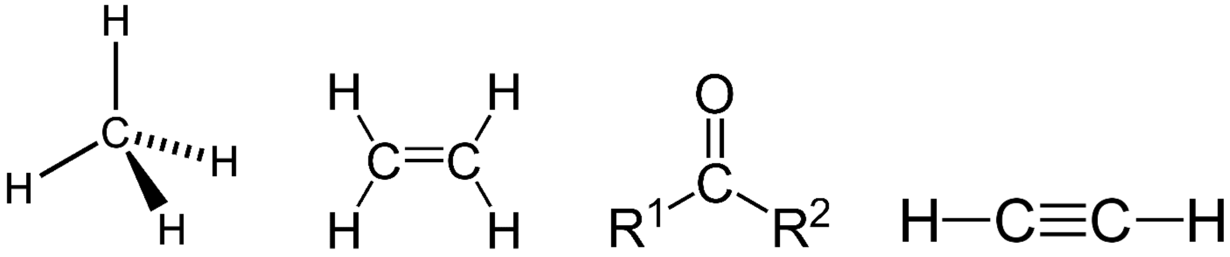
2.1 Introduction

Force field methods describe the energy of a given nuclear configuration as an expansion in the internal coordinates of the molecule. The total energy is constructed by a sum of contributions from distortions of the molecule.

$$E = E_{str} + E_{bend} + E_{tors} + E_{oop} + E_{cross} + E_{elec} + E_{vdw} \dots \tag{2.1}$$

These contributions can be classified as either bonded (directional) or non-bonded terms. The three simplest examples for bonded terms are the bond stretch energy, the angle bend energy and the torsional energy. They are the backbone of any force field and describe the interaction of a limited number of atoms with each other. Examples for non-bonded interactions include the electrostatic energy and the Van der Waals energy. They describe long range intra- as well intermolecular forces.

The individual energy contributions are described by physically motivated expansions in their respective internal coordinates. The expansion coefficients are fitted to high level ab initio data or experimental data and are typically optimized for a whole class of molecules. Depending on the specific force field, atoms are assigned an atom type not only based on their physical nature, but also on their environment and bonding state. While a force field designed for small molecules may, for example, only differentiate between different hybridization states of carbon, a force field developed for the calculation of protein structures might implement different atom types for every possible functional group surrounding a carbon atom.



(a) Type 1: sp^3-carbon

(b) Type 2: sp^2-carbon in alkenes

(c) Type 3: sp^2-carbon in carbonyl group

(d) Type 3: sp-carbon

Figure 2.1: Different carbon types in the MM2 force field

2.1.1 Bond stretch energy

The potential energy curve for bond stretching energy E_{str} is often modeled by a Morse potential [18],

$$E_{morse} = D_e(e^{-2a(l-l_0)} - 2e^{-a(l-l_0)}), \tag{2.2}$$

where l is the bond length and l_0 the equilibrium bond length. This equation not only includes the anharmonicity of the actual potential, but also shows the correct dissociation behaviour. Even though the Morse potential only uses three parameters it outperforms most four-coefficient-models.

Nonetheless, most force fields use power series expansions of the potential curve due to the lower computational effort.

$$E_{str} = \sum_{m=2}^{m_{max}} k_m (l - l_0)^m \tag{2.3}$$

Even a simple harmonic ($m_{max} = 2$) expansion is often sufficient as force fields are typically used only for evaluations near the equilibrium geometry. In order to account for the anharmonicity of the actual potential, cubic and quadratic terms can be included if necessary.

2.1.2 Angle bending energy

The energy contribution E_{bend} due to distortions of the angle between three atoms is typically modeled as a power series expansion around the equilibrium geometry angle:

$$E_{bend} = \sum_{m=2}^{m_{max}} k_m (\theta - \theta_0)^m \tag{2.4}$$

Most force fields truncate the expansion after the second order. The resulting harmonic potential is typically valid for deviations ±30° around θ_0 [19]. In the force fields by Norman Allinger (MM2, MM3) [20] [21], which are frequently discussed throughout this thesis, the angle bending energy is expanded up to sixth order. It is also important to note that the angle θ is measured in plane for planar molecules, for example around a sp^2 hybridized carbon. The energy penalty due to bending out of the plane is accounted for by the out-of-plane energy contribution.

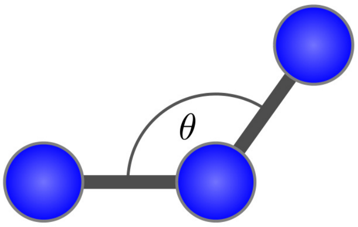


Figure 2.2: Angle between three atoms

2.1.3 Torsional energy

The torsional energy E_{tors} is the contribution due to twisting of the dihedral angle ω between four atoms as shown in Figure 2.3. It was first introduced to explain the different energies of cis and trans (or gauche and anti) conformational isomers of small molecules and is an important contribution in any force field. Due to the periodic nature of this potential it is typically expanded in a Fourier series:

$$E_{tors} = \sum_{m=1} V_m \cos(m\omega). \tag{2.5}$$

Depending on the symmetry of the molecule different terms are used in the expansion. The ethane molecule, for example, has C_3 rotational symmetry with respect to the C-C bond axis. Therefore, only coefficients corresponding to multiples of three ($m = 3, 6, 9$) can enter in the Fourier expansion. For the C_2 symmetric ethene only even numbered V_m are non zero. Since the coefficients V_m depend on all four atom types involved in the dihedral angle, the number of necessary parameters in the force field grows quickly with the number of atom types. To counteract this rapid increase many force field implementations use V_m coefficients that are only dependent on the atom type of the two center atoms.

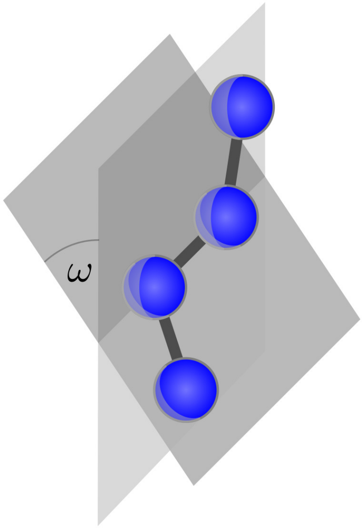


Figure 2.3: Dihedral angle

2.1.4 Out-of-plane energy

The out-of-plane energy E_{oop} gives an energy penalty for the distortion of planar atom arrangements. This is, for example, of importance for sp^2 hybridized carbon atoms.

$$\begin{aligned} E_{oop} &= kd^2 \\ E_{oop} &= k\theta^2 \end{aligned} \tag{2.6}$$

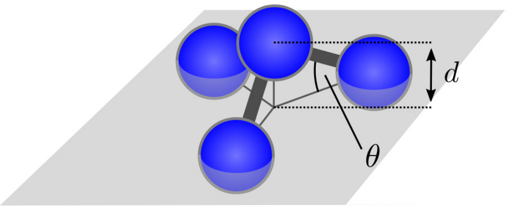


Figure 2.4: Out-of-plane bending

There are two possible ways to quantify out-of-plane bending, either by the distance d, by which one of the atoms protrudes from the plane spanned by the remaining three atoms, or by the angle θ enclosed by the bond and its projection onto the plane. Both methods are depicted in Figure 2.4. Using the Taylor series expansion of the sine function it can be shown that the two descriptions are equal for the harmonic expansion of the out-of-plane energy. Another method to account for the out-of-plane energy is based on improper torsional contributions. However, the definition of this improper torsional angle is not unique and therefore not used in this thesis.

2.1.5 Cross terms

Cross terms are contributions to the force field energy that couple the bonded contributions. One of the simplest examples is the stretch-bend cross term,

$$E_{str-bend} = k(\theta - \theta_0)[(l_{ab} - l_{ab,0}) + (l_{bc} - l_{bc,0})], \tag{2.7}$$

where θ is the angle between two bonds with lengths l_{ab} and l_{bc} and equilibrium bond lengths $l_{ab,0}$ and $l_{bc,0}$. This contribution is motivated by the fact that, at large bond distances, a variation in the angle will not give a large energy contribution. However, for very short bond lengths relatively small angle distortions can lead to a large change in the distance between the outer two atoms.

In theory, any combination of the standard energy contributions is possible for a cross term. Force fields are classified by the number of contributions that are used for cross terms. A class 2 force field, for example, only contains cross terms coupling at most two other contributions. Following Jensen [19], other examples for cross terms include:

$$\begin{aligned} E_{str,str} &= k(l_{ab} - l_{ab,0})(l_{bc} - l_{bc,0}), \\ E_{bend,bend} &= k(\theta_{abc} - \theta_{abc,0})(\theta_{bcd} - \theta_{bcd,0}), \\ E_{str,tors} &= k(l_{ab} - l_{ab,0})\cos(n\omega_{abcd}), \\ E_{bend,tors} &= k(\theta_{abc} - \theta_{abc,0})\cos(n\omega_{abcd}), \\ E_{bend,tors,bend} &= k(\theta_{abc} - \theta_{abc,0})(\theta_{bcd} - \theta_{bcd,0})\cos(n\omega_{abcd}). \end{aligned} \tag{2.8}$$

2.1.6 Electrostatic energy

The electrostatic energy E_{elec} in force fields can be viewed from the point of a multi-pole expansion. The first order expansion, which is used in most force fields, is the interaction of point charges. While the functional form of the coulomb potential of two point charges is simple,

$$E_{elec} = \frac{q_a q_b}{4\pi\epsilon_0\epsilon_r r_{ab}}, \tag{2.9}$$

it remains difficult to assign a partial charge to the atoms involved. The values assigned are typically taken from ab-initio calculations and assumed constant.

Higher order multi-pole expansions would be dipole or quadrupole expansions. In the dipole expansion, the electrostatic energy is a function of three parameters, the distance and the two angles the dipoles enclose with the vector connecting the atoms. Due to this increased computational effort most force fields use the simple point charge expression.

Since the coulomb energy between bonded atoms is already taken into account by bonded energy contributions such as the stretch or the bend energy the electrostatic, the electrostatic energy is scaled depending on the number of bonds between the interacting atoms. This bond distance is denoted as *1,X* where X is the number of bonds in between the atoms plus one. A *1,2* interaction, for example, is the interaction of two atoms bonded to each other. Typically, the calculation of non-bonded interactions is only done for bond distances greater than the *1,4* interaction. The *1,4* interaction itself is scaled down by a factor between 1 and 2 in most force fields. In addition to that, most force field codes implement a cut-off radius for non-bonded interaction to reduce the computational effort.

2.1.7 Van-der-Waals energy

The Van-der-Waals interaction between two atoms can be modeled using different functions. They typically involve an r^{-6} dependence for the attractive part of the potential. This is physically motivated by dipole-dipole interactions which vary as r^{-6} with distance. Additional terms proportional to r^{-8} and r^{-10} occur if dipole-quadrupole and quadrupole-quadrupole interactions are taken into account.

One of the most frequently used functional forms is the Lennard-Jones potential [22]:

$$E_{vdw} = \epsilon \left[\left(\frac{r_0}{r} \right)^{12} - 2 \left(\frac{r_0}{r} \right)^{6} \right] \tag{2.10}$$

The r^{-12} dependence of the repulsive part is not motivated by physical arguments but convenient from an (antiquated) computational point of view. The actual repulsive potential is better modelled using an exponential function. This functional form is often referred to as Buckingham potential [23].

$$E_{vdw} = Ae^{-Br} - \frac{C}{r^6} \tag{2.11}$$

The exponential repulsion describes the Van-der-Waals interaction better than the r^{-12} potential. It does, however, give rise to a new problem: The attractive r^{-6} diverges for short distances. This causes the potential to yield an attractive force for small distances. However, in actual calculations this region of the potential is never reached. This problem can be avoided by using a Morse potential to model the Van-der-Waals interaction. The Morse potential excels at describing the interactions for short distances but does not show the correct r^{-6} asymptotic dependence.

2.1.8 Commonly used force fields

Depending on the application, different force field implementations are used. These vary not only in the energy contributions used but also in the number of atom types that are parameterized.

MM2, MM3: The “Molecular Mechanics” force fields by Norman Allinger are general force fields for small molecules. The relatively simple MM2 force field [20] is used in many of the calculations presented in this thesis. MM3 [21] improves upon MM2 by the introduction of more cross terms and more atom types.

MMFF: The “Merck Molecular Force Field” [24] is a general purpose force field similar to MM3. A notable feature is the untypical 7-14 Van-der-Waals potential used in MMFF.

AMBER: The "Assisted Model Building with Energy Refinement" family of force fields was first introduced by Peter Kollman [25]. There exist many variations for different use cases, most notably the "General AMBER force field" (GAFF).

CHARMM: The "Chemistry at Harvard Macromolecular Mechanics" is a family of force fields for calculations of large systems [26]. Different versions are available for proteins, nucleic acids and lipids.

UFF: Most force fields are parameterized for a certain class of molecules. The "Universal Force Field" [27], on the other hand, is a general force field for all elements on the periodic table. The parameters are derived from atomic properties such as atom radius or electronegativity.

2.2 Comparison of common force fields

Comparing different force field energies is hard to justify, as they often include different contributions and different expansion orders. The test set of molecules the force fields were parametrized with differ widely as well. Figure 2.5 shows such a comparison of MM2, MM3, MMFF and GAFF energies along the reference trajectory as they all were designed for small molecules like butadiene.

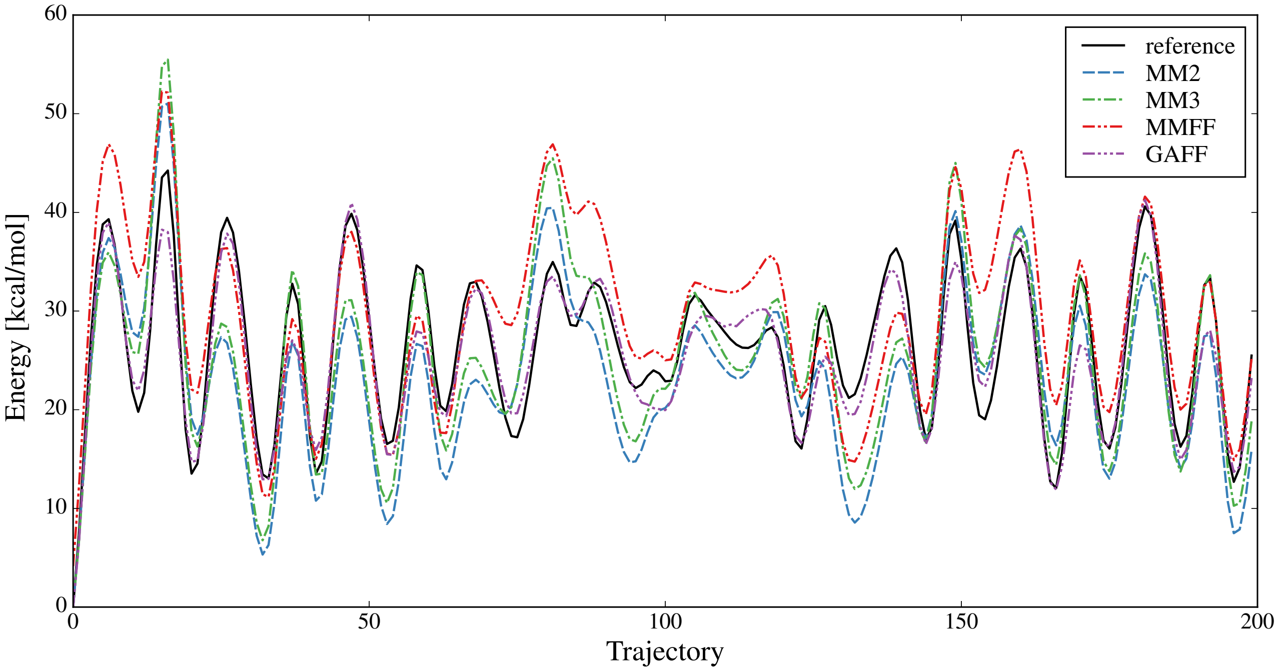


Figure 2.5: Reference energy compared to force field energies from MM2, MM3, MMFF and GAFF.

In order to quantify the accuracy the RMSD from the reference curve is calculated. All energy curves are shifted to the value 0 for the first reference point. MM2 and MMFF give nearly the same accuracy with a RMSD of 6.5 and 6.4 kcal/mol, respectively. MM3, as expected, improves upon the MM2 energy with a RMSD of 5.5 kcal/mol. For this specific reference curve GAFF outperforms the other force fields significantly. The RMSD for GAFF is about 2.4 kcal/mol, even though it has the simplest functional form of all force fields presented here. This indicates that, given an adequate set of parameters, high expansion orders and numerous cross terms are not necessary.

2.3 Contributions to the total energy

In order to investigate the size of the different energy contributions the quadratic average of the energy contributions of the different force fields are calculated for the butadiene reference curve. The results are plotted in a bar chart in Figure 2.6.

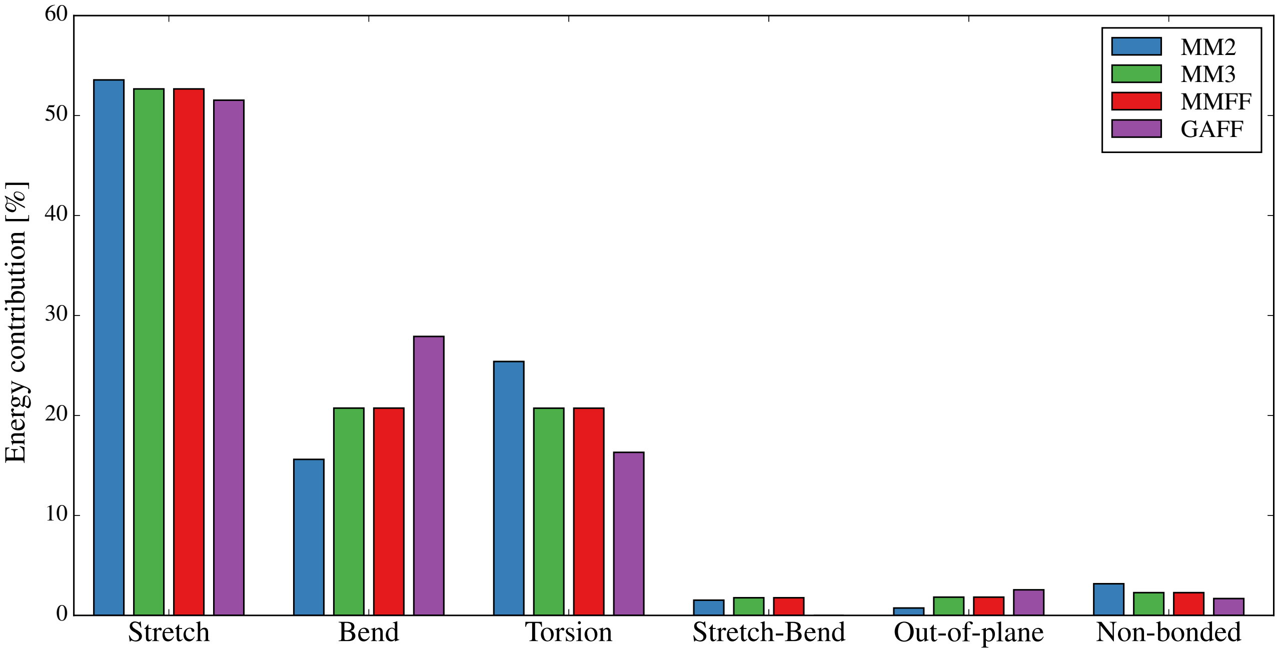


Figure 2.6: Quadratic average of the MM2 force field contributions for the butadiene reference curve

The largest contributions stem from the three most basic force field terms: The bond stretch energy, the angle bend energy, and the torsional energy. The electrostatic and Van-der-Waals energy, here summarized as non-bonded contributions, have a small contribution for small molecules like butadiene, because they are only calculated between pairs of atoms which are separated by at least 3 bonds. In larger molecules or molecule ensembles these contribution play a significantly larger role.

2.4 Improvements by individualization

Standard force fields are parameterized for whole classes of molecules. This in turn means that the description for individual molecules is not as accurate as it could be for a given functional form. The effects of individualization are demonstrated using the MM2 force field due to its relatively simple functional form and yet high expansion orders. This gives the MM2 force field the functional flexibility to be refitted to a specific molecule. For two of the major contributions in the butadiene reference curve, the bond stretch energy and the angle bend coefficients, the expansion coefficients up to sixth order are determined. This refitting was done using additional SCF-calculations, in which the respective internal coordinates are varied near the equilibrium geometry. The remaining internal coordinates are fixed to the optimized values. The contribution of the non-bonded interactions was subtracted from the SCF-energies to avoid double counting of these energy contributions.

Figure 2.7 shows the refitting of the bond stretch parameters. In butadiene there are a total of five different bond types. This distinguishment is an improvement to the standard MM2 force field which only uses two bond types. The comparison with standard MM2 bond energy curves shows that the refitted curves reproduce the SCF-energy better over a great range. Most notable is the energy curve for the CC2-stretch, which is the stretching of the central C-C-bond in butadiene. The standard MM2 parameters underestimate the equilibrium bond length by over 0.1 Å. In addition, the bond potentials for GAFF are plotted. Refitting the simple harmonic expansion to the anharmonic reference potential does not improve the GAFF energy. It does, however, show that the significant advantage of GAFF over the other presented force fields is the better discrimination of different types of bonds.

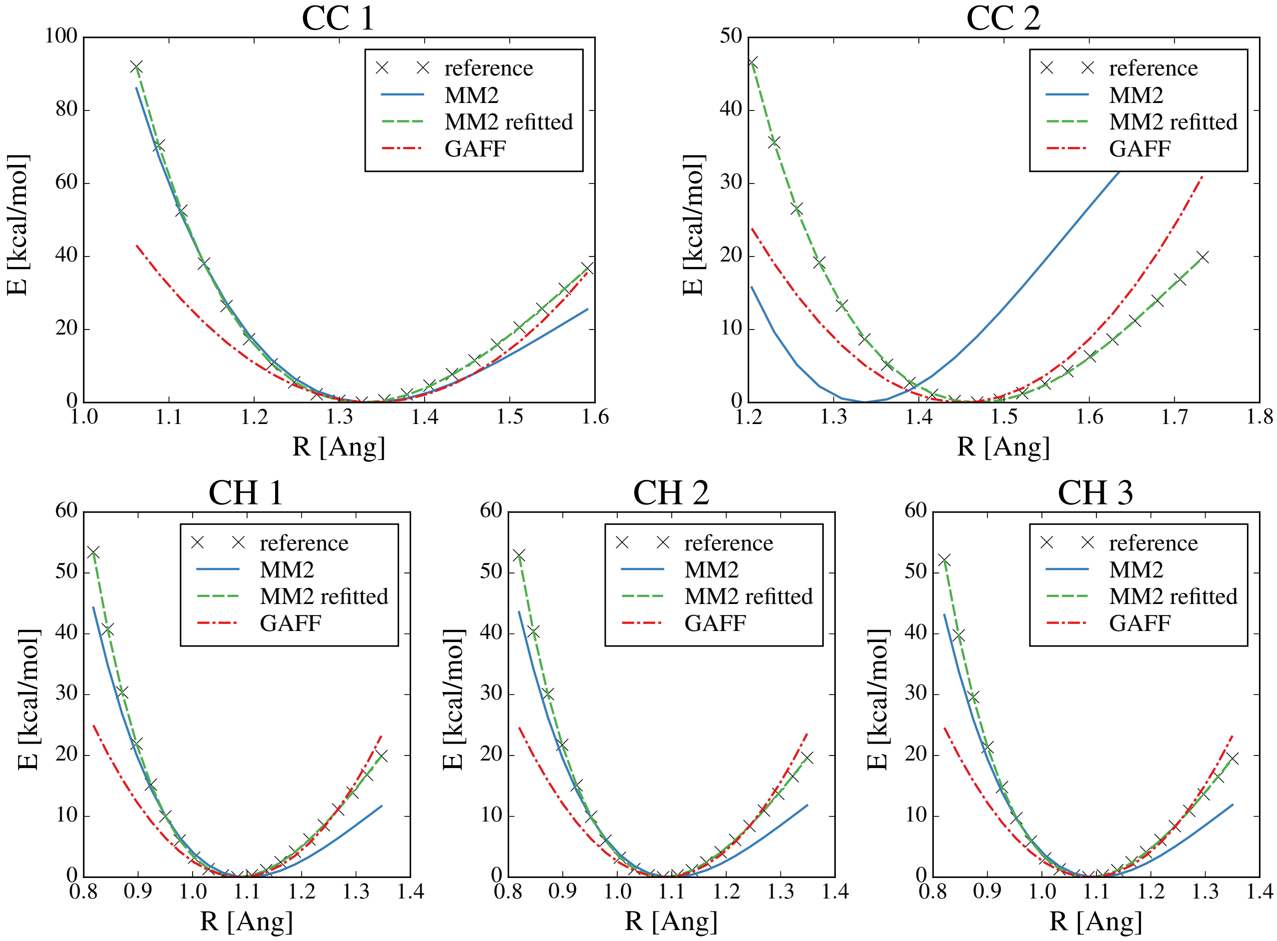


Figure 2.7: Optimization of the bond stretch coefficients by fitting the potential curves.

Refitting the angle bend parameters involves more effort as the angles can not be varied individually. Every carbon atom in butadiene has three bonded partners and three angles enclosed by these bonds. Since the sum of all angles has to be 360°, only two of them have to be varied. Figure 2.8 shows the four angles that are varied in order to extract all angle bend coefficients. In addition, the atom numbering that is used to refer to the individual atoms of the butadiene molecule throughout this thesis, is depicted in Figure 2.8.

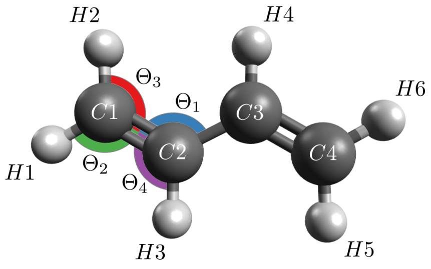


Figure 2.8: Naming of the atoms and the four angles varied for the refitting.

Figure 2.9 shows the fitting of the two-dimensional angle bending potential. The parameters of all six different angle types in butadiene can be determined using these fits.

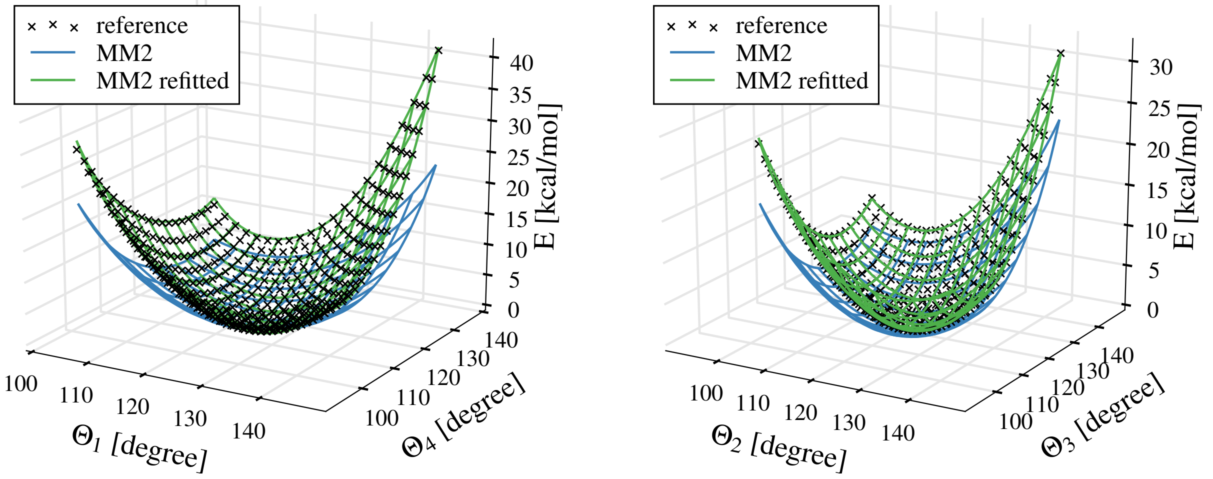


Figure 2.9: Optimization of the angle bend coefficients by fitting a two-dimensional slice of the PES.

Using the coefficients extracted from these fits, the force field energy predictions for the butadiene reference curve can be improved significantly. Figure 2.10 shows a comparison between the standard MM2 force field and its individualized version. The MM2 RMSD of 6.53 kcal/mol is reduced to 2.97 kcal/mol for the refitted force field. This method of individually varying the internal coordinates in order to refit the force field parameters becomes increasingly difficult for the other energy contributions and is therefore not feasible in actual applications. Another approach to extract force field parameters from ab initio calculations would be to use the Hessian matrix that is computed during a frequency calculation. When transformed into internal coordinates, the values of the Hessian matrix provide the force constants in the harmonic expansion.

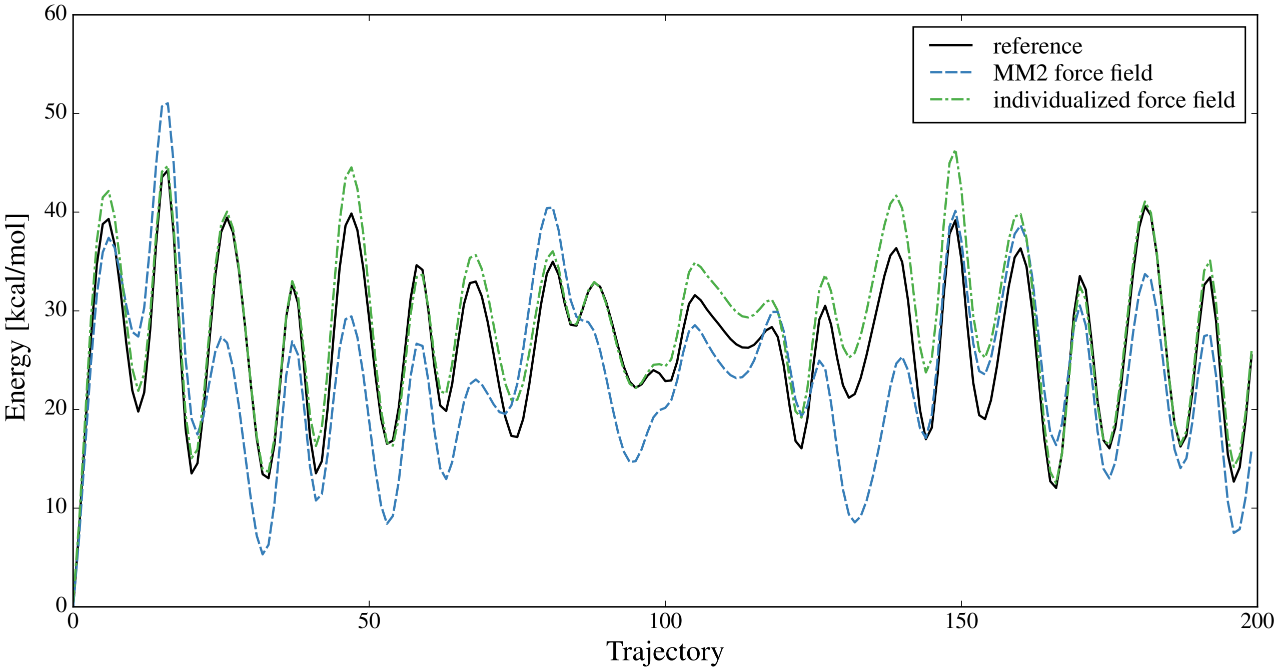


Figure 2.10: Refitted bond stretch and angle bend coefficients in MM2 force field for butadiene in comparison to the reference curve and standard MM2.

The idea of automatically deriving a molecule specific force field from ab-initio calculations as suggested here has recently been the topic of several publications [28] [29] [30] [31] [32]. Depending on the system these algorithms are applied to, different fitting techniques are employed and different force fields are used.

Chapter 5 builds upon the results of this section and investigates the possibility of an IMLS scheme to adjust the MM2 force field along the PES.

2.5 Improvements by additional energy contributions

Ab initio reference points contain information beyond the total energy which can be used to support the parametrization of a force field. Examples for quantities that can be extracted from these ab initio calculations are the electron density and the molecular orbital energies. The difficulty in including them into the force field energy is that large parts of their energy contributions are already taken into account by standard force field contributions, and a clear separation of contributions turns out difficult.

2.5.1 Mulliken charges

In a first attempt of improving the force field energy, we consider the inclusion of Mulliken charges from SCF data. Mulliken charges are partial charges calculated by assigning parts of the electron density to each atom. These charges are used as partial charges in the electrostatic energy contribution of the force field. Note that according to Figure 2.6 non-bonded energy contributions account for less than 5% of the total energy for the butadiene molecule. However, it seems worth to test whether a slight improvement of the energy prediction can be achieved or not. Figure 2.11 shows a comparison of the standard MM2 energy to a MM2 force field calculation augmented by Mulliken charges. The partial charge interaction is only calculated between atoms with at least three bonds between them (in MM2: *1,4* scaling = 1). The resulting correction of the total energy is very small. As shown in Figure 2.6, this is true for all non-bonded energy contributions in a small molecule.

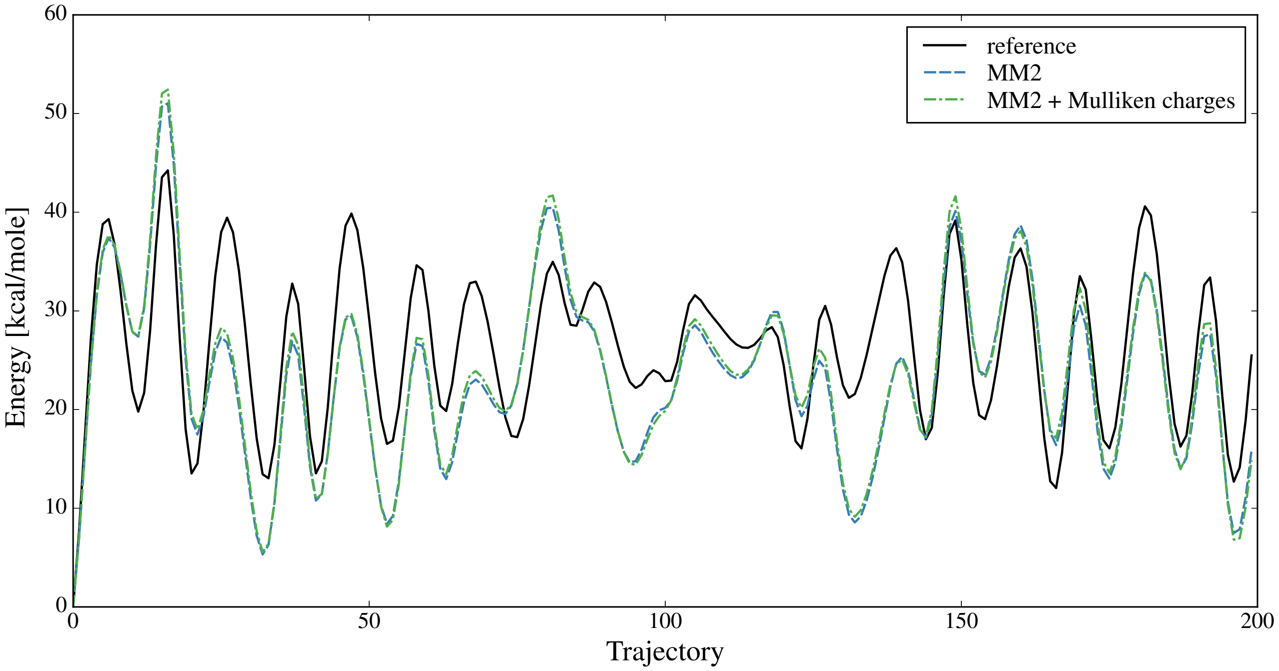


Figure 2.11: Comparison of the standard MM2 force field to an calculation using the ab initio Mulliken charges for the Coulomb interaction

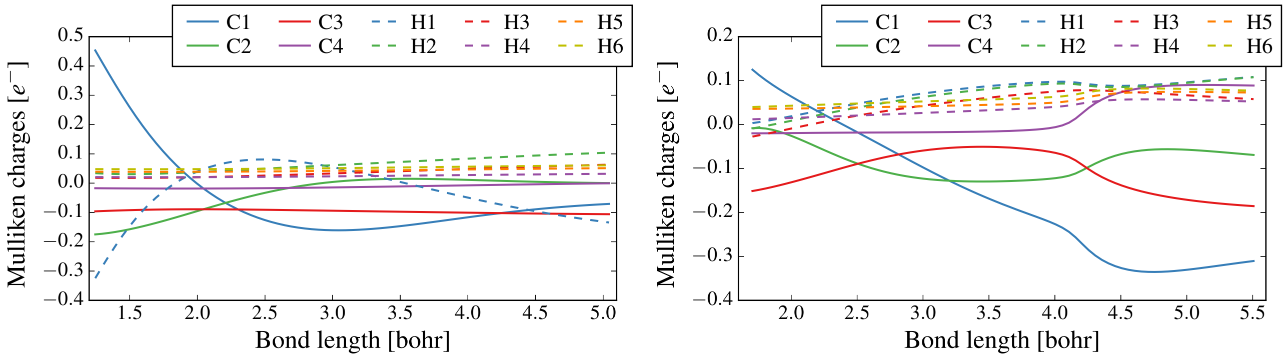
All three curves in Figure 2.11 are shifted to give an energy of 0 kcal/mol at the starting geometry. Indeed, the RMSD for the standard MM2 force field of 6.53 kcal/mol is slightly reduced to a value of

6.39 kcal/mol by inclusion of Mulliken charges, which is an even larger improvement than what was to be expected given the small overall contribution of non-bonded terms.

Incorporating Mulliken charges into the larger force field contributions is not straightforward. One of the investigated schemes, for example, is varying the bond stretch coefficient based on the charge difference of the bonded atoms,

$$E_{str} = (k - c\Delta q)(l - k_0)^2. \tag{2.12}$$

This gives slightly larger improvements even though it is not strictly motivated by physical arguments. The slight improvements, however, do not justify the effort of modeling the change in Mulliken charges at a force field level. For the simple test in Figure 2.11 the Mulliken charges are taken from the ab initio calculation at every geometry. For a possible application, the geometry dependence of the charges would have to be modeled in a force field inspired expansion in the internal coordinates. However, an inconvenient complexity of the geometry dependence of the Mulliken charges appears as is shown in Figures 2.12a and 2.12b.



(a) CH bond stretch

(b) CC bond stretch

Figure 2.12: Mulliken charges of the atoms in butadiene as a function of (a) the CH bond length between atom C1 and atom H1, (b) the CC bond length between atom C1 and atom C2.

Figure 2.12a shows that for small CH bond lengths the electron pair of the covalent bond is assigned in larger part to the hydrogen atom. This results in a negative partial charge. At the equilibrium distance of about 2.1 bohr the carbon has a small negative charge. After a maximal negative charge at a distance of about 3 bohr, the partial charge on the carbon tends towards the equilibrium situation for large distances. In Figure 2.12b, the Mulliken charges of the atoms are plotted as a function of one of the double bonds in butadiene. A sudden change in the behaviour can be seen at about 4.3 bohr. The conjugated p_z orbitals in the molecule break apart at the given distance. This affects the partial charges of all atoms.

From the analytical form of expression 2.12 it follows that any attempt of modeling the changes in the Mulliken charges, which then in turn are used to improve the force field energy, can be absorbed into cross terms in the force field contributions. The gain in information from the ab initio points is annulled by the additional effort that is needed to model these new quantities. This is equally true for the electron density and the molecular orbital energies.

In conclusion, it is difficult to find a physically motivated way to include the partial charge information that can be gained from each reference point, especially because it remains unclear which interactions are already taken into account by the bonded contributions. An additional problem is the modeling of changes in these quantities. Interestingly, it turns out that this most often leads to contributions similar to cross terms in the force field energy.

3 PES derived from Hückel approach

In an attempt to use more information from the SCF data points we leave the force field ansatz for a moment before exploring a possible force field based interpolation scheme in chapter 5.

3.1 Introduction

In this chapter we investigate the question whether ab initio reference points have more to offer than energies and charges. However, this comes at the cost of a deeper entanglement with basic concepts of SCF theory.

The term extended Hückel theory (EHT) was coined by R. Hoffmann [33]. Previously, the method was simply referred to as molecular orbital (MO) method. It generalizes the simple Hückel method, which was intended just for conjugated π-bonds, to include all valence electrons. In the original form, the Wolfsberg-Helmholz [34] formula is used to approximate the off-diagonal elements of the Hamiltonian.

$$H_{ij} = \frac{1}{2} k_{ij} (H_{ii} + H_{jj}) S_{ij} \tag{3.1}$$

The diagonal elements H_{ii} are considered as model parameters. Typically, they are assigned based on the valence state ionization potentials (VSIPs) [35], which can be determined experimentally. For the diagonal elements the proportionality constant k_{ij} is set to 1. A typical choice for the off-diagonal elements is $k_{ij} = 1.75$. The overlap matrix S_{ij} is calculated in an atomic orbital basis using Slater-type orbitals:

$$\phi(\mathbf{r})_{nml} = N r^{n-1} e^{-\zeta r} Y(\mathbf{r})_l^m \tag{3.2}$$

Using this Hamiltonian, the molecular orbitals and their energy are calculated from the Hückel equations, which can be related to the Roothaan-Hall equations [36] of restricted SCF methods:

$$HC = SC\epsilon. \tag{3.3}$$

In the extended Hückel theory the total energy of a system is a simple sum of the molecular orbital energies of the occupied orbitals,

$$E = 2 \sum_i \epsilon_i. \tag{3.4}$$

The approximation of the integrals used in the Wolfsberg-Helmholz formula was first introduced by Mulliken [37]. Blyholder and Coulson conclude from their investigation [38] that the Wolfsberg-Helmholz formula gives a reasonable approximation to the SCF Fock matrix for uniform charge distributions. They do note, however, that the exchange integrals can not be written in the form $c \times S_{ij}$ and that the approximated electron-core-interaction integrals show inaccuracies of about $10 - 20$ %.

There have been numerous attempts to improve the EHT using more elaborate approximations of the Hamiltonian. A few examples are the weighted Wolfsberg-Helmholz formula [39],

$$H_{ij} = \frac{1}{2} \left[k_{ij} + \Delta^2 + \Delta^4 (1 - k_{ij}) \right] (H_{ii} + H_{jj}) S_{ij}, \tag{3.5a}$$

with,

$$\Delta = \frac{H_{ii} - H_{jj}}{H_{ii} + H_{jj}}, \tag{3.5b}$$

the Cusachs formula [40],

$$H_{ij} = \frac{1}{2} S_{ij}(2 - |S_{ij}|)(H_{ii} + H_{jj}), \tag{3.6}$$

the exsin formula [41],

$$\operatorname{sgn}(S_{ij})\frac{1}{2}\{(1 + |S_{ij}|)\left[1 + c\sin(\pi|S_{ij}|)\exp(b|S_{ij}|)\right] - 1\}(H_{ii} + H_{jj}), \tag{3.7a}$$

$$b = -\pi\cot(\pi|S_m|), \tag{3.7b}$$

with parameters c and S_m, and the Calzaferri formula [42],

$$H_{ij} = \frac{1}{2}\left[1 + (\kappa + \Delta^2 - \Delta^4\kappa)\exp(-\delta(R_{ij} - d_0))\right] S_{ij}(H_{ii} + H_{jj}), \tag{3.8}$$

with additional parameters κ and δ. Δ is defined same as in equation 3.7b and d_0 is the sum of the orbital radii $r_n(A) + r_n(B)$ defined by

$$r_n = \frac{1}{\int_0^\infty (1/r)R_{nl}^2(r)r^2 dr}, \tag{3.9}$$

where $R_{nl}(r)$ refers to the radial part of the atomic orbital wave function. Note, however, that none of these parametrizations goes beyond a simple proportionality to the overlap matrix S.

3.1.1 Comparison of different parametrizations

Figure 3.1 shows a comparison of the sum of molecular orbitals energies for the different formulas presented in section 3.1. The Hartree Fock reference is a sum of the molecular orbital energies of the 11 valence orbitals.

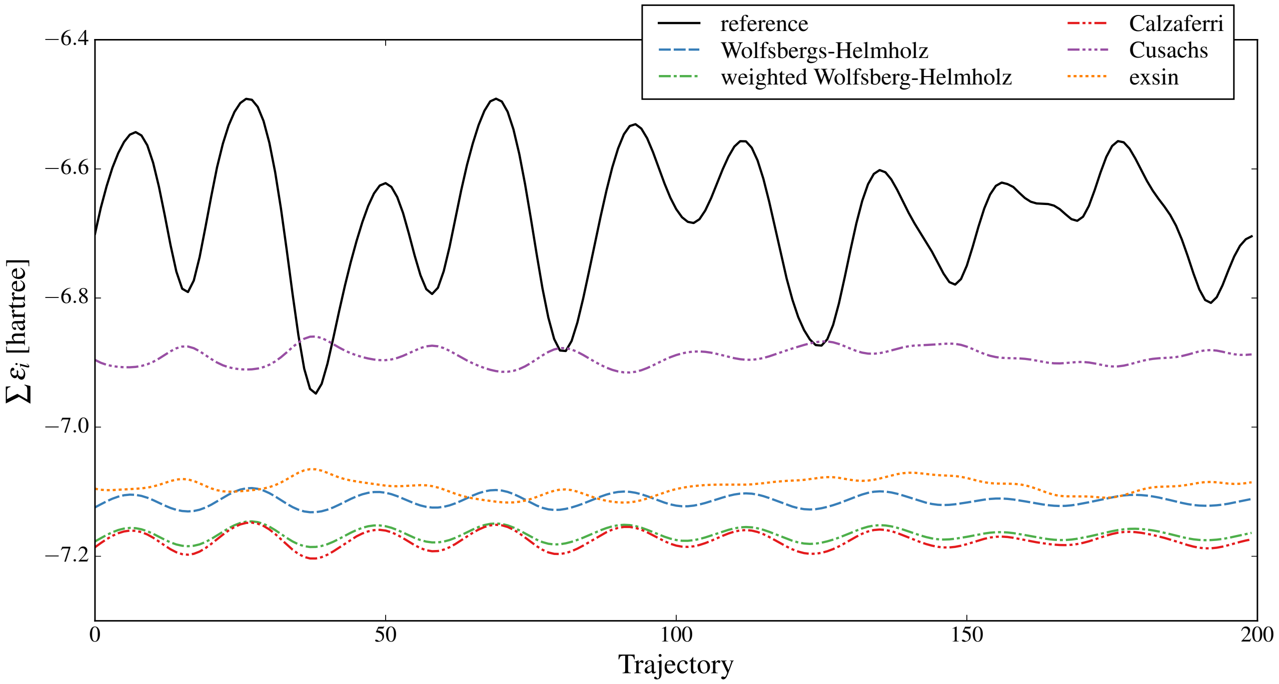


Figure 3.1: Comparison of the sum of molecular orbital energies for the different Hamiltonian parametrizations.

All calculations in Figure 3.1 use 21.4, 11.4, and 13.6 eV as VSIPs for the carbon 2s, the carbon 2p and the hydrogen 1s orbitals respectively [43]. In both the Wolfsberg-Helmholz formula and the weighted Wolfsberg-Helmholz formula a constant k_{ij} of 1.75 is used. For the Calzaferri approach, which is based on the weighted Wolfsberg-Helmholz formula, $\kappa = 0.75$ and $\delta = 0.1$ is used. These three parametrizations are successive improvements and, as a consequence, give similar total energies. The energies calculated from these formulas show the correct functional course but underestimate the relative change in energy. The Cusachs and the exsin formula (which is based on the former) share a similar functional course. For the exsin formula the parameters suggested by Kalman in his original publication of the formula are used [41]. The functional course of these approaches shows opposite behaviour to the reference, exhibiting minima where the reference shows maxima and the other way round.

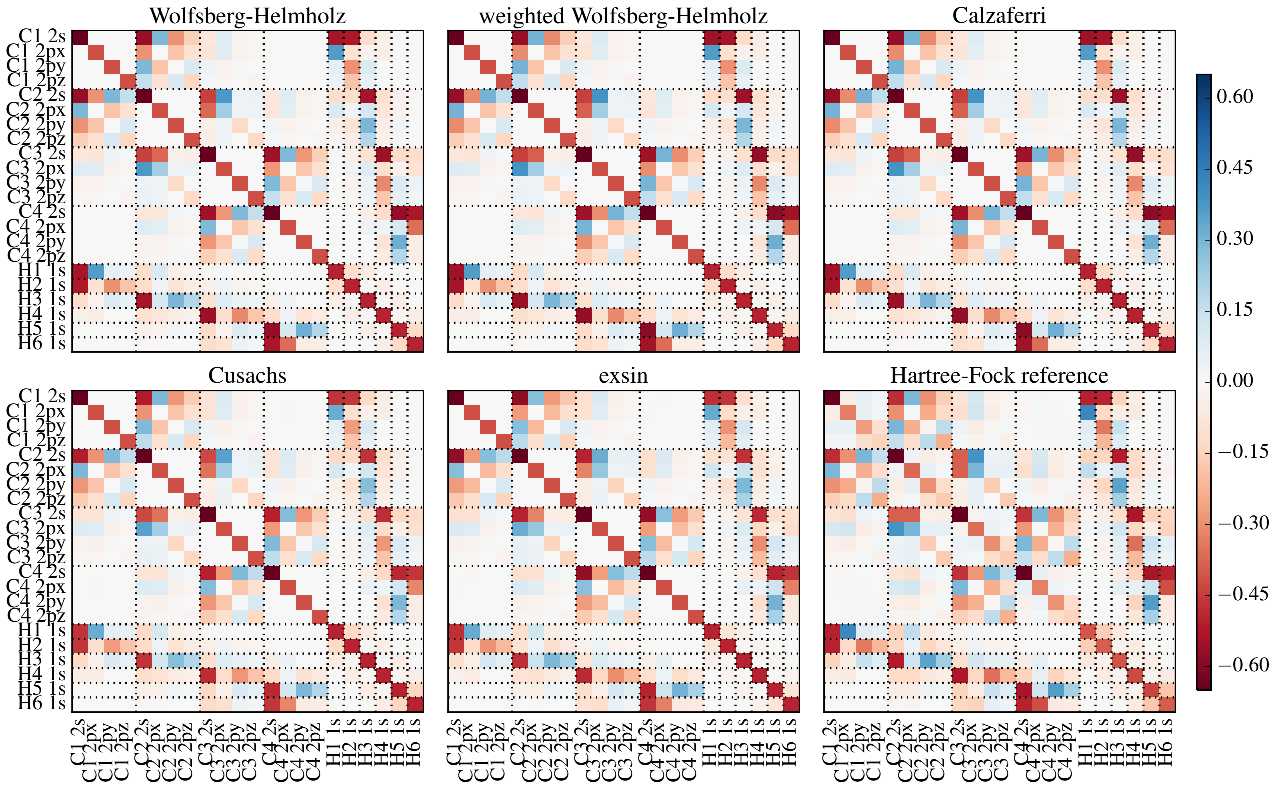


Figure 3.2: Comparison of the Fock matrices for different Hamiltonian parameterizations, calculated for the first geometry of the butadiene trajectory.

Figure 3.2 shows a graphical representation of the Fock matrices for the first geometry of the butadiene trajectory, constructed by the different formulas. The Fock matrix of the HF reference was projected onto the smaller EHT basis. In addition, the contribution of the core electrons on the carbon atoms was removed. A more detailed discussion of this process can be found in section 7.3. The general shape of the Fock matrix is reproduced correctly by all parameterizations. The RMSDs of the matrix elements for all 200 trajectory geometries are 0.0438, 0.0440, 0.0445, 0.0436, and 0.0444 hartree for the Wolfsberg-Helmholz, the weighted Wolfsberg-Helmholz, the Calzaferri, the Cusachs and the exsin method, respectively.

One notable deficiency of all formulas is their simple dependence of the off-diagonal elements on the overlap matrix. Due to the orthogonality of the basis functions on one atom, the Fock matrix elements for those basis functions can not be approximated. Another source of the deviations are the diagonal elements themselves. The VSIPs used in all of the formulas are assumed constant throughout the trajectory and the same for all three p-type orbitals. However, the comparison with the HF reference shows that both these assumptions are wrong for the butadiene molecule. This problem can be addressed by the self-consistent Hückel method (SC-EHT), first introduced by Harris [44]. In its simplest form, the SC-EHT diagonal elements are adjusted based on the calculated charge of the corresponding atom,

$$H_{ii} = H_{ii}^0 - a_i q_A, \tag{3.10}$$

where H_{ii}^0 is the initial value of the diagonal element (the VSIP), a_i is a parameter and q_A is the charge of the atom the basis function i is centered on. Typically, Mulliken charges [45] calculated from the density matrix are used, which makes the method iterative. While the SC-EHT introduces a geometry dependence of the diagonal elements and corrects the assumption that all atoms of a certain atom type, regardless of their environment, share the same ability to attract electrons, it does not differentiate between the different types of p orbitals.

The iterative nature of the SC-EHT makes the formulation of an interpolation scheme based on this method difficult and inefficient. The remaining of this thesis, therefore, focuses on non-iterative variants of the EHT.

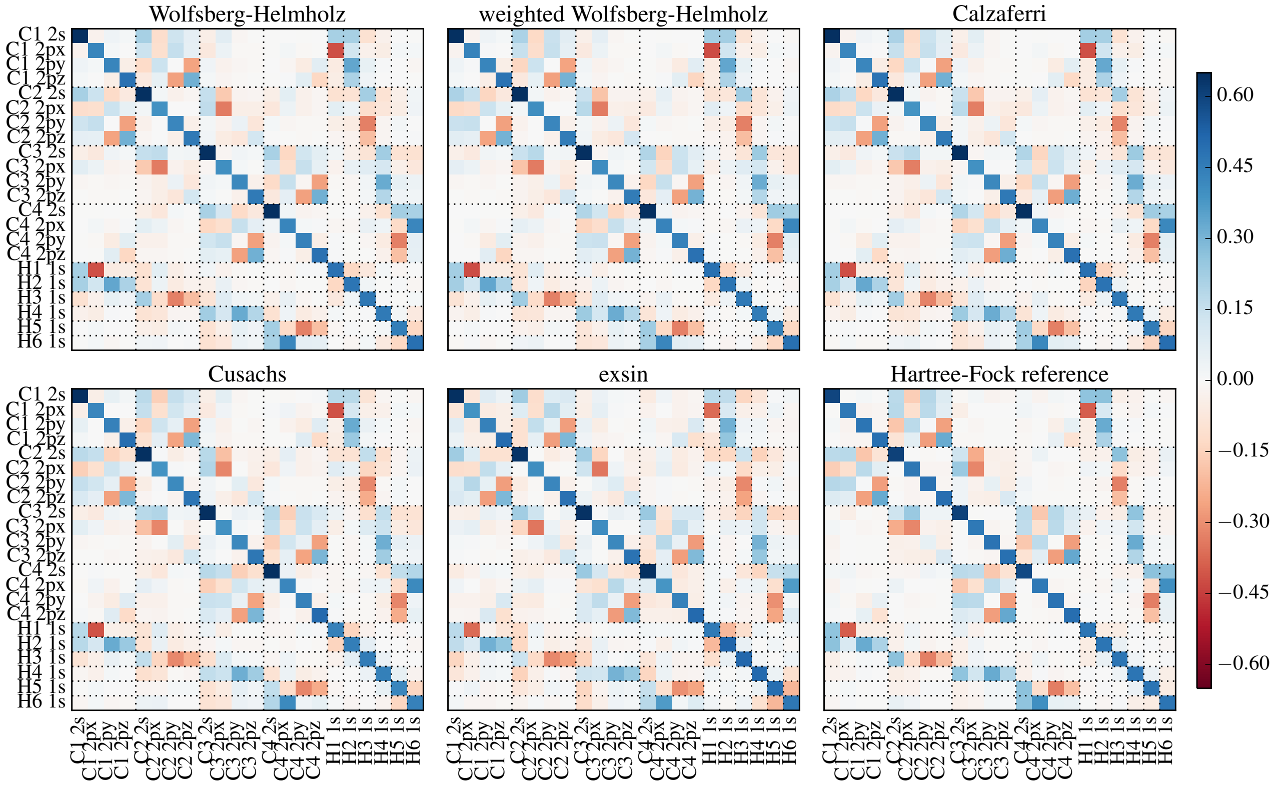


Figure 3.3: Comparison of the density matrices for different Hamiltonian parameterizations, calculated for the first geometry of the butadiene trajectory.

Figure 3.3 shows the density matrices, for the first geometry of the butadiene trajectory, calculated by the different formulas and the HF reference. Again, the qualitative shape of the matrices is reproduced correctly. The RMSDs of the matrix elements for all 200 trajectory geometries are 0.030, 0.027, 0.028, 0.035, and 0.047 for Wolfsberg-Helmholz, weighed Wolfsberg-Helmholz, Calzaferri, Cusachs, and exsin, respectively. Comparing this to the RMSD for the Fock matrix elements shows no correlation. Certain Fock matrix elements seem to have a larger influence on the density than others. This is to be expected, because the density matrix is constructed only from the eigenvectors corresponding to the lowest eigenvalues. Therefore, information that is contained in the eigenvectors of virtual orbitals is not used.

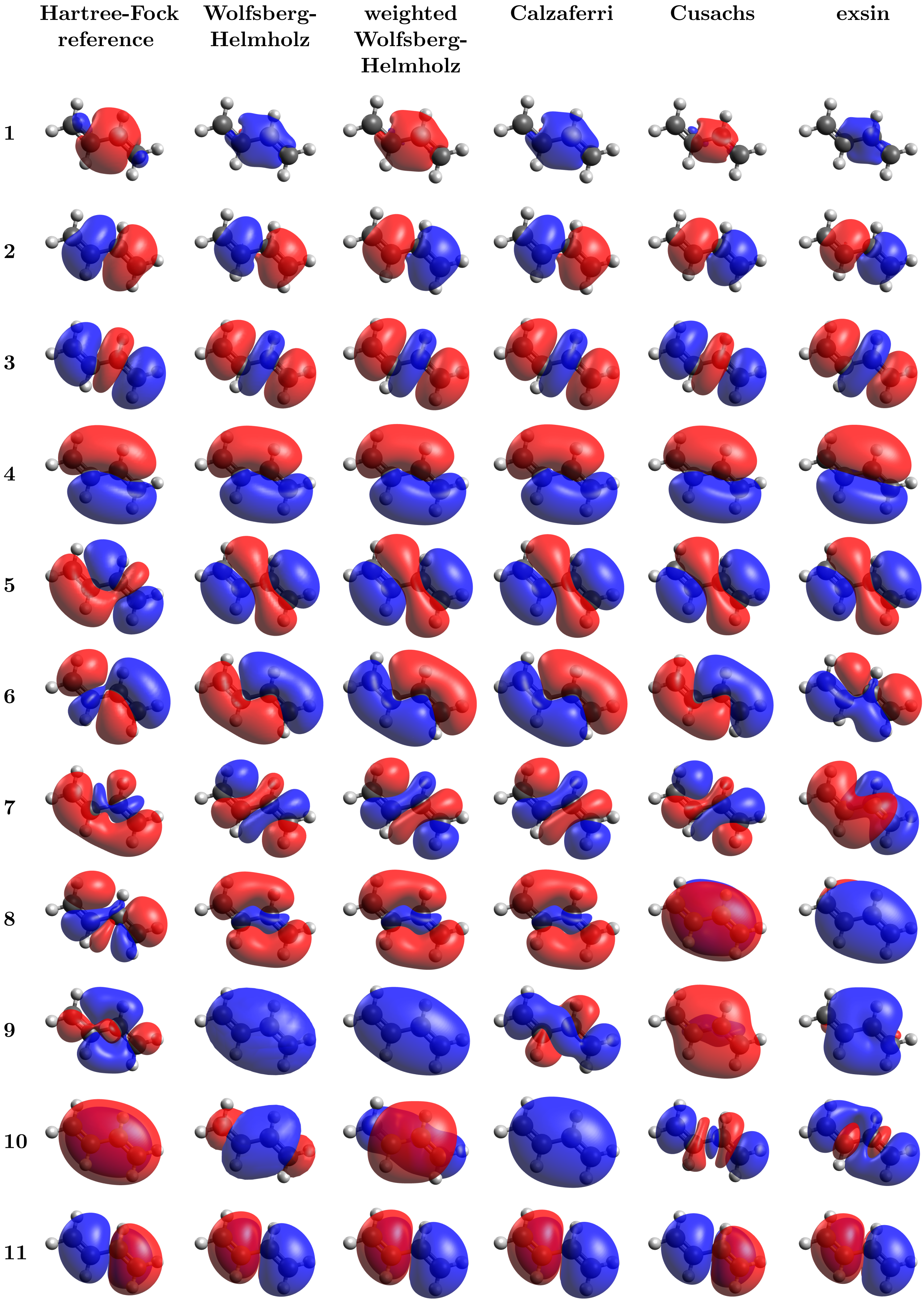


Figure 3.4: Comparison of the 11 valence orbitals of butadiene for the different Hückel parametrization schemes.

Figure 3.4 shows a comparison between the valence molecular orbitals calculated by the different EHT methods and the HF reference. The isosurfaces of the eleven molecular orbitals for the value 0.02 are plotted using the program Avogadro [46]. Energetically lower lying orbitals are described better than the outermost. In the range from molecular orbital number seven to number eleven a clear assignment of nodal structure and energy is not straightforward. However, the total electron density, composed of a sum of all occupied molecular orbitals, is almost indistinguishable from the reference density for all of the parametrizations. A similar trend can be observed for the molecular orbital energies, plotted in Figure 3.5. For the energetically lower lying orbitals the parametrizations give reasonable results, but show increasing deviations for the higher lying valence orbitals.

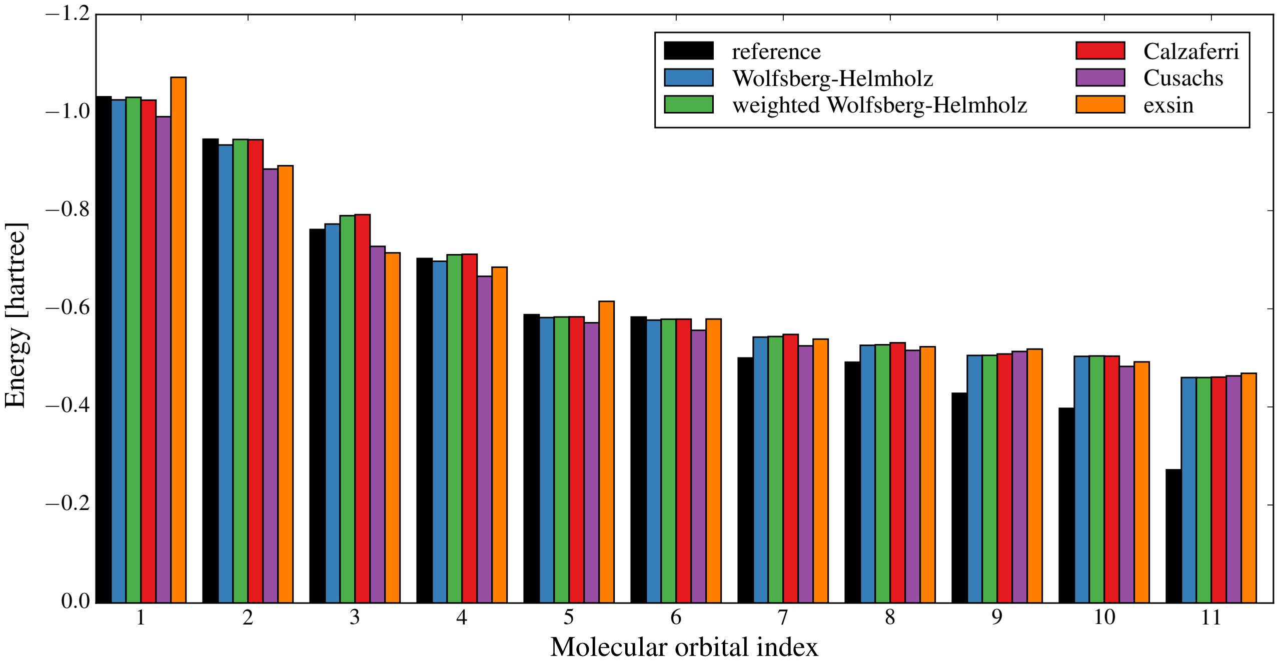


Figure 3.5: Comparison of the molecular orbital energies for different Hamiltonian parameterizations, calculated for the first geometry of the butadiene trajectory.

Concluding this introduction, all of the presented formulas reproduce the valence molecular orbitals of a Hartree-Fock calculation qualitatively, but can not estimate the molecular orbital energies to a sufficient accuracy using the standard parameters. A further complication to be treated later is given by the fact that the total energy of the SCF-calculations can not be replicated at all without additional energy expressions, even if the molecular orbital energies were exact.

3.2 Improvements through parameter refitting

Refitting the parameters of the different formulas to the valence orbital energies of the HF reference gives significant improvements. Table 3.1 shows an overview of the optimized parameters for the various formulas.

Table 3.1: Optimized EHT parameters see equations 3.1 to 3.8.

		default	Wolf.-Helm.	weighted Wolf.-Helm.	Calzaferri	Cusachs	exsin
H_{C2s}	[eV]	21.40	8.34	8.33	5.75	22.12	18.10
H_{C2p}	[eV]	11.40	5.30	5.30	3.50	7.97	3.42
H_{H1s}	[eV]	13.60	6.55	6.53	3.80	17.09	12.76
k	[1]	1.75	5.74	5.73			
κ	[1]	0.75			8.94		
δ	[Å^{-1}]	0.1			0.56		
RMSD	[hartree]		0.07	0.07	0.02	0.13	0.03

The twelve additional parameters of the exsin formula are tabulated separately in Table 3.2 for improved readability.

Table 3.2: Optimized parameters for the exsin formula.

	reference [41]		fit			reference [41]		fit	
Overlap	S_m	c	S_m	c	Overlap	S_m	c	S_m	c
1s1s	0.674	0.030	0.237	1.281	2s2s	0.342	0.480	0.653	0.185
1s2s	0.518	0.161	-0.217	0.797	2s2p	0.255	0.432	0.273	1.674
1s2p	0.334	0.420	0.133	0.524	2p2p	0.500	0.265	0.802	1.132

Figure 3.6 shows the sum of molecular orbital energies for the refitted formulas along the reference trajectory.

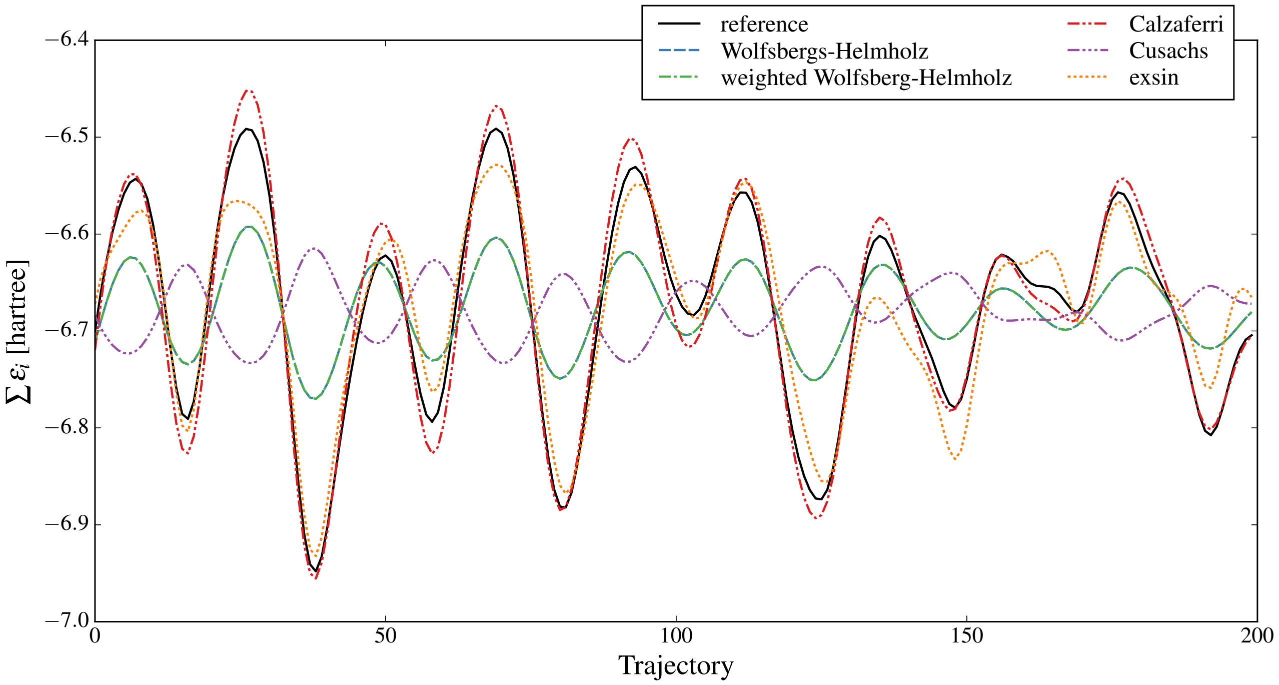


Figure 3.6: Comparison of the sum of molecular orbital energies for the different Hückel-Hamiltonian formulas after refitting of their respective parameters.

The fitted parameters for the different formulas deviate significantly from default values. For strictly mathematical models this would not be a problem. However, in the EHT, the parameters do have a physical meaning and should not deviate too much from the experimentally determined values.

Comparing the formulas is difficult because of the different number of parameters. The functional course is best reproduced by the Calzaferri formula and the exsin formula. The Wolfsberg-Helmholz and the weighted Wolfsberg-Helmholz give similar refitted parameters and, as a consequence, a similar functional course. The Cusachs formula still shows opposite behaviour after refitting.

There are suggestions for most of the formulas to improve the quality of the approximation by introducing additional parameters. The Wolfsberg-Helmholz formula, for example, could be improved by using a different factor k_{ij} for each type of bond [47] [48]. We conclude from our investigation of these suggestions that, while they do improve the accuracy (as can be expected when using more parameters), they can not improve the geometry dependence of the formulas. Therefore, they can not correct the remaining deviations in the functional course.

3.3 Improvements based on ab initio data

In this section we link the extended Hückel approach to our goal of improved fitting of ab initio reference points. The Hückel Hamiltonian can be related to the Fock matrix in the SCF method. In other semi-empirical methods such as MNDO (Modified Neglect of Diatomic Overlap [49]) and AM1 (Austin Model 1 [50]) the parametrization happens at the level of the integral evaluation, while in EHT the Fock matrix itself is parametrized.

3.3.1 Improvements using the Fock matrix diagonal

A simple way to incorporate SCF information into the Hückel formalism is to use the diagonal of the converged Fock matrix to populate the diagonal of the Hückel matrix. Obviously, a full use of the Fock matrix would lead to the exact molecular orbital energies; however, by using only diagonal elements in combination with the parametrization formulas of Section 3.1, larger deviations are to be expected. If one method for the calculation of non-diagonal elements from diagonal elements proves useful, future improvements of the energy predictions as a function of geometry can be achieved by a simple fitting of the main diagonal elements.

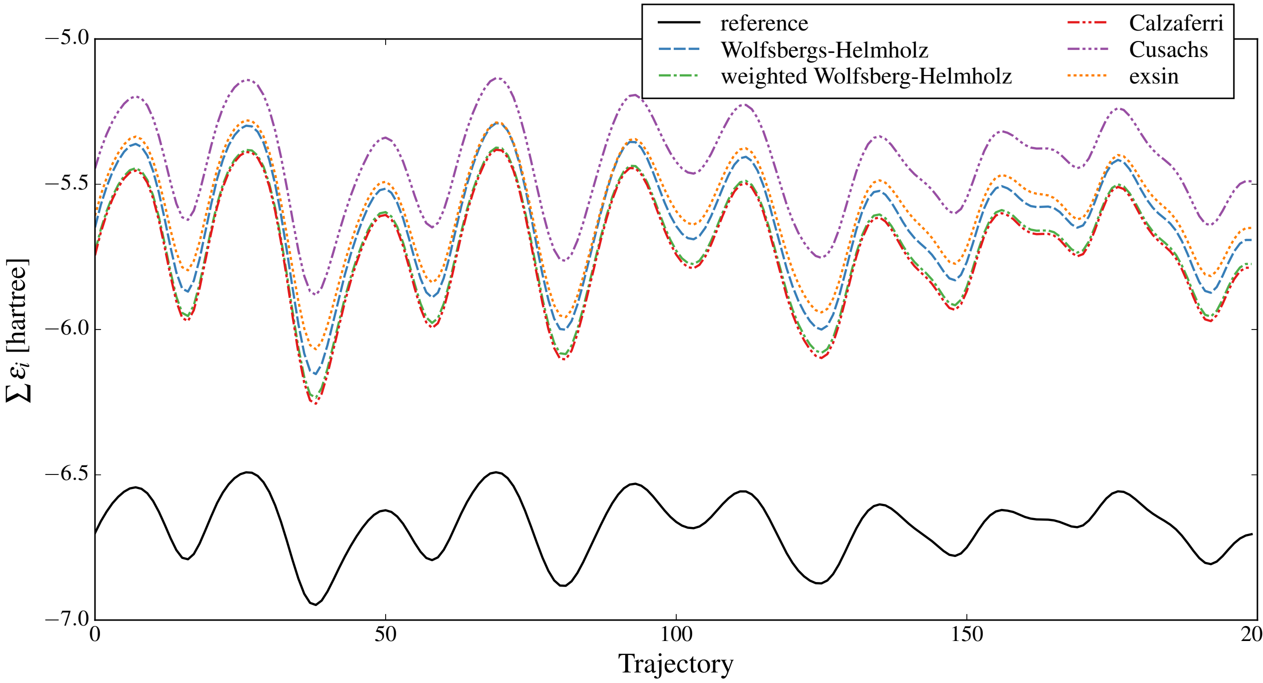


Figure 3.7: Comparison of the sum of molecular orbital energies along the butadiene trajectory for the different Hamiltonian parametrizations using the Fock matrix diagonal instead of VSIPs.

Figure 3.7 shows the sum of molecular orbital energies for the different formulas using the Fock matrix diagonal at every SCF data point. The remaining parameters are set to the same values as in 3.1.1. The new functional course shows that choosing geometry independent Hückel matrix diagonal elements is an incorrect assumption. This is due to the fact that largest contributions to the molecular orbital energies, at least in the atomic orbital basis, are the diagonal elements. The absolute value of the sum of molecular energies is underestimated and the geometry dependent variation overestimated.

In order to correct these deviations, the energies are shifted by their value at the first trajectory point, and the remaining parameters fitted to the reference energy. Figure 3.8 shows the refitted sum of molecular orbital energies.

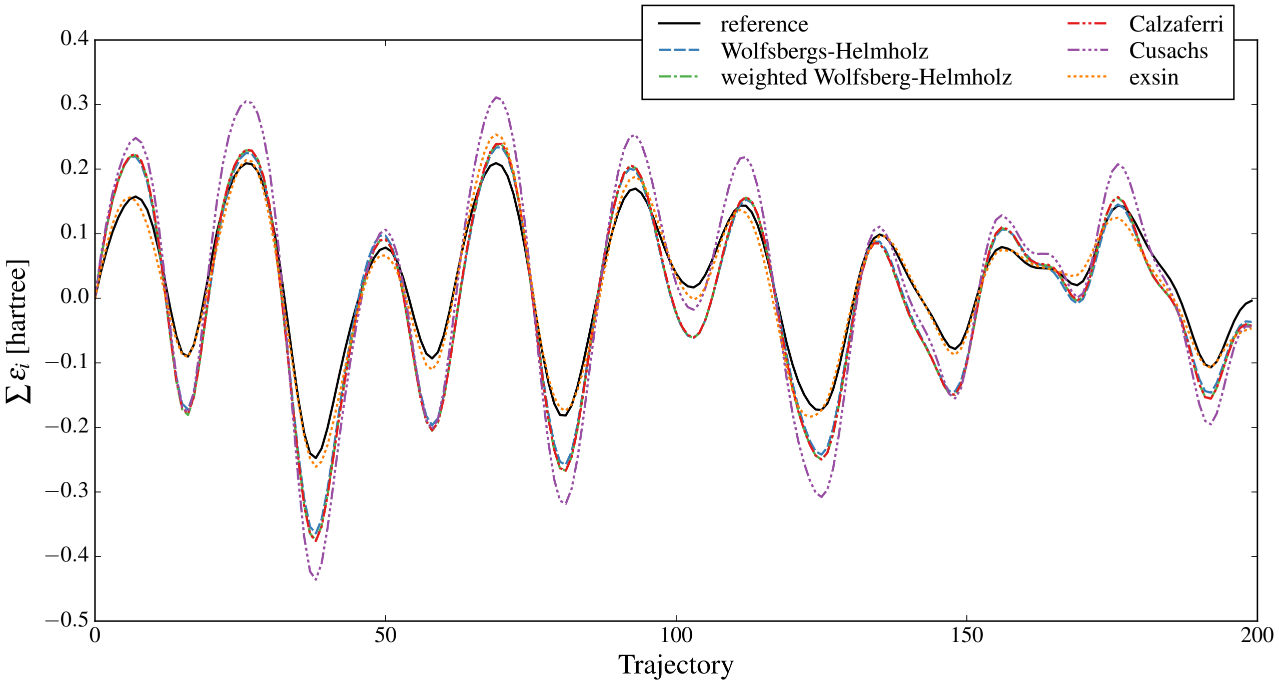


Figure 3.8: Comparison of the sum of molecular orbital energies along the butadiene trajectory. The different Hamiltonian parametrization formulas use the exact Fock matrix diagonal instead of VSIPs along the main diagonal and the parameters for the off-diagonal element approximations are fitted to the reference energy.

The resulting RMSDs are 0.045, 0.048, 0.048, 0.068, and 0.017 hartree for the Wolfsberg-Helmholz, the weighted Wolfsberg-Helmholz, the Calzaferri, the Cusachs, and the exsin formula, respectively. Note that the Cusachs formula does not include any additional parameters into the off-diagonal approximation and therefore refitting this parametrization is not possible. The RMSD is reduced in comparison to the simple fitting in section 3.2 for all approximations with exception of the Calzaferri formula.

3.4 From molecular orbitals to the total energy

Many approaches to solve the problem of linking the molecular orbitals to the total energy in extended Hückel theory have been proposed over the years.

Blyholder and Coulson [38] argue that a total energy expression similar to the one in the SCF-scheme should be used. In order to avoid the calculation of two electron integrals they suggest using

$$E = \sum_i \epsilon_i + \sum_i \vec{c_i}^T h \vec{c_i} + \sum_{a<b} \frac{Q_a Q_b}{R_{ab}}, \tag{3.11}$$

with molecular orbital energies ϵ_i, the molecular orbital coefficients $\vec{c_i}$, the core Hamiltonian h as a $K \times K$ matrix in the atomic basis, and nuclear charges Z_a screened by the number of core electrons n_{ca}: $Q_a = Z_a - n_{ca}$. This method increases the computational effort significantly as the core Hamiltonian has to be evaluated for each geometry. Note that formula 3.11 is a exact expression of the SCF energy if SCF molecular orbitals are used.

Calzaferri, Forss, and Kamber developed a formula that can be interpreted as an approximation to the core Hamiltonian integrals [42].

$$E = \sum_i \epsilon_i + \sum_{a<b} E_{ab}, \tag{3.12}$$

with

$$E_{AB} = \frac{Q_A Q_B}{R_{AB}} - \frac{1}{2}\left[Q_A \int \frac{\rho_B}{|R_{AB} - r|} dr + Q_B \int \frac{\rho_A}{|R_{AB} - r|} dr\right]. \tag{3.13}$$

This two-body interaction term covers the interaction of the electron density with the nuclei, but neglects the kinetic energy part of the core Hamiltonian.

Dixon and Jurs proposed what they call the EHNDO (Extended Hückel Neglect of Differential Overlap) method [51]. Two-body interaction terms between pairs of atoms are added to the total energy expression:

$$E = 2\sum_i \epsilon_i + \sum_{A<B} \left[E_{AB,elec} + E_{AB,nuc}\right]. \tag{3.14}$$

The two-body interaction is split into two terms. The electronic interaction is given by

$$E_{AB,elec} = z_A z_B \frac{e^{-(a_A+a_B)R_{AB}^{b_A+b_B}}}{R_{AB} + c_A + c_B}, \tag{3.15}$$

with a_A, a_B, b_A, b_B, c_A and c_B as atomic parameters of the method. The numbers of valence electrons on the atoms A and B are denoted as z_A and z_B. The nuclear interaction is given by

$$E_{AB,nuc} = z_A z_B \frac{e^{-(\delta_A+\delta_B)R_{AB}^{\epsilon_A+\epsilon_B}}}{R_{AB}}, \tag{3.16}$$

with δ_A, δ_B, ϵ_A and ϵ_B as atomic parameters. Here, z_A and z_B are the net charges not screened by inner-shell electrons. The method was parametrized for H, C, N, O, and F.

Figure 3.9 shows a comparison of the three suggested methods. The HF results projected onto the EHT basis are used as a basis in order to compare solely the influence of the additional energy expressions. Note that using the core Hamiltonian expression of equation 3.11 does not reproduce the exact HF energies due to this projection. The integrals for the Calzaferri formula are evaluated using the formulas given by reference [42]. For the EHNDO method the parameters given in the original publication are used. These parameters were fitted simultaneously with the remaining EHT parameters and probably work best when used together. Refitting the ten parameters reduces the RMSD from 0.045 hartree to a value of 0.013 hartree but worsens the description of the functional course. Including the exact kinetic energy expression in the Calzaferri method does not yield any improvements.

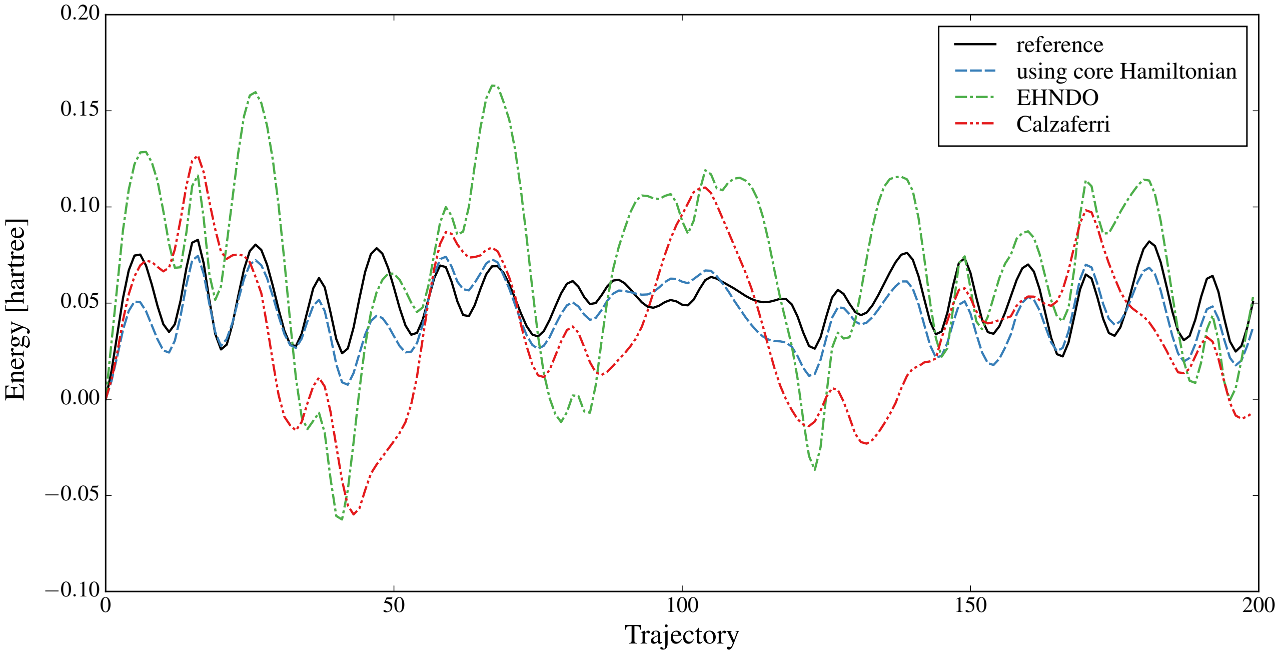


Figure 3.9: Comparison of three approaches for additional EHT energy expressions along the butadiene trajectory.

Despite the fact that both of the approximate formulas give a reasonable functional course, we will be focusing on the approach proposed by Blyholder and Coulson during the rest of this work because of its proximity to the SCF formalism.

3.4.1 Introduction of effective core potentials

The approach of modelling valence electrons only reminds of an other area of quantum chemistry: the treatment of molecules containing heavy atoms via the introduction of effective core potentials (ECPs). In this simplification, the core electrons are replaced by a pseudo-potential and the electronic Hamiltonian takes the form

$$H_{elec} = -\sum_{i=1}^{N_v}\left[\frac{1}{2}\nabla_i^2 + \sum_{a=1}^{M}\left(-\frac{Q_a}{r_{ia}} + V_a^{ECP}(r_{ia})\right)\right] + \sum_{i=1}^{N_v}\sum_{j>i}^{N_v}\frac{1}{r_{ij}} + \sum_{a=1}^{M}\sum_{b>a}\frac{Q_aQ_b}{R_{ab}}. \tag{3.17}$$

The summation runs over the N_v valence electrons, and the nuclear charges are reduced by the number of of core electrons $Q_a = Z_a - N_{c,a}$. The total energy can then be written as

$$E = \sum_i \epsilon_i + \sum_i \vec{c_i}^T(h + V^{ECP})\vec{c_i} + \sum_{a<b}\frac{Q_aQ_b}{R_{ab}} \tag{3.18}$$

A typical one-component effective core potential is a sum of potentials for the different angular momentum numbers:

$$V_a^{ECP}(r_{ia}) = \sum_{l=0}^{L-1} V_{a,l}^{ECP}(r_{ia})\mathcal{P}_{a,l}, \tag{3.19}$$

with the projection operator for the spherical harmonics centered on the nucleus a,

$$\mathcal{P}_{a,l} = \sum_{m_l=-l}^{l} |lm_l\rangle\langle lm_l|. \tag{3.20}$$

For the potentials, an expansion in terms of Gauss-functions is used:

$$V_{a,l}^{ECP}(r_{ia}) = \sum_k a_{lk} r_{ia}^{n_{lk}} e^{-\alpha_{lk} r_{ia}^2}. \tag{3.21}$$

For most ECPs the exponents n_{lk} are set zero. Following the convention of the EHT to set integrals proportional to the overlap matrix S we present two possible approximations of the V^{ECP} matrix:

$$\langle\phi_i|V_a^{ECP}|\phi_j\rangle = a_{ij} \sum_m S_{im} * S_{mj}, \tag{3.22a}$$

$$\langle\phi_i|V_a^{ECP}|\phi_j\rangle = \sum_m a_{im} S_{im} * S_{mj} a_{mj}. \tag{3.22b}$$

In the approach of equation 3.22a, the proportionality factor a_{ij} depends on the type of the valence basis functions i and j. The potentially more complicated approach in equation 3.22b uses a proportionality factor that is also dependent on the type of the core basis function m. For the butadiene molecule there is only one type of core basis function, the 1s basis function on the carbon atoms. This yields three a_{im} parameters: $a_{1s,2s}$, $a_{1s,2p}$, and $a_{1s,1s}$. For the sake of comparability we use three parameters in the approach of equation 3.22a: a_{ss}, a_{sp}, and a_{pp}.

In a simple, preliminary test both approaches are fitted using the butadiene reference curve. The projected HF-calculations are used as a basis, again representing an ideal EHT result. Figure 3.10 shows the results of the fits. The optimized parameters are: $a_{ss} = 12.78$, $a_{sp} = 16.40$, and $a_{pp} = 7.67$ for the first variant and $a_{1s,2s} = 3.53$, $a_{1s,2p} = 3.39$, and $a_{1s,1s} = 3.29$ for the second variant. The RMSD is 0.0029 and 0.0017 hartree, respectively.

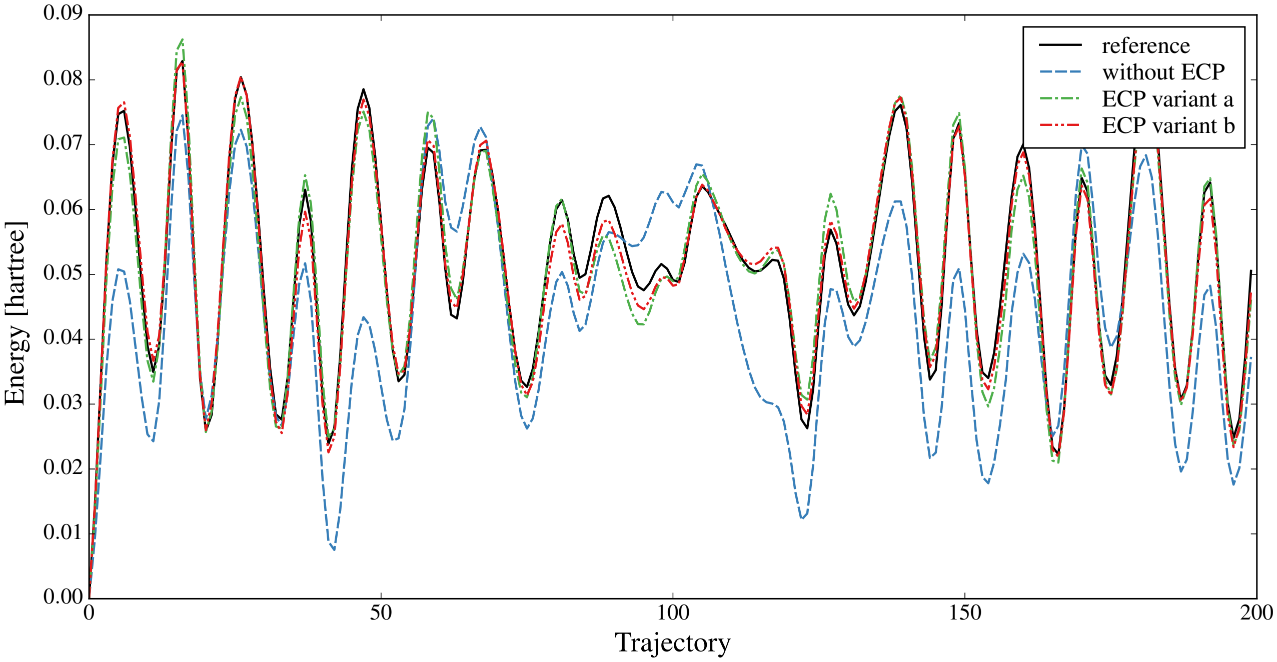


Figure 3.10: Comparison of energies along the trajectory for the two overlap matrix dependent formulations of the ECP.

Note the similarity of both approaches for the simple butadiene molecule: Since there is only one type of core basis function, the a_{im} formally do not depend on the index m. Equation 3.22b can therefore be rewritten as

$$\langle\phi_i|V_a^{ECP}|\phi_j\rangle = a_i a_j \sum_m S_{im} * S_{mj}. \tag{3.23}$$

This expression looks similar to 3.22a, with the additional restriction that the factor a_{ij} is represented by a product $a_i a_j$. The product based approach still outperforms the more general one because of the restriction we made to allow just for three parameters a_{ss}, a_{sp}, and a_{pp} instead of using all six possible combinations of the three basis function types: a_{1s1s}, a_{1s2s}, a_{1s2p}, a_{2s2s}, a_{2s2p}, and a_{2p2p}. Fitting all six parameters gives an RMSD of 10^{-3} hartree. The approach of equation 3.22b might still be a viable alternative using significantly fewer parameters.

3.4.2 Integration of ECPs in an EHT approach

This section focuses on combining the results of the previous investigations to find a EHT model that can be used as the basis for an interpolation scheme. In an effort to keep the number of parameters low, the Calzaferri formula for the Fock matrix is combined with the ECP motivated approach of equation 3.22b. The resulting model is fitted to the total energy for the reference curve. The Calzaferri formula is restricted to two parameters fixing the diagonal elements to the VSIPs of reference [43]. Optimizing the remaining five parameters yields: $a_{1s,2s} = 3.26$, $a_{1s,2p} = -2.11$, $a_{1s,1s} = 3.54$, $\kappa = 2.73$, and $\delta = 0.44$.

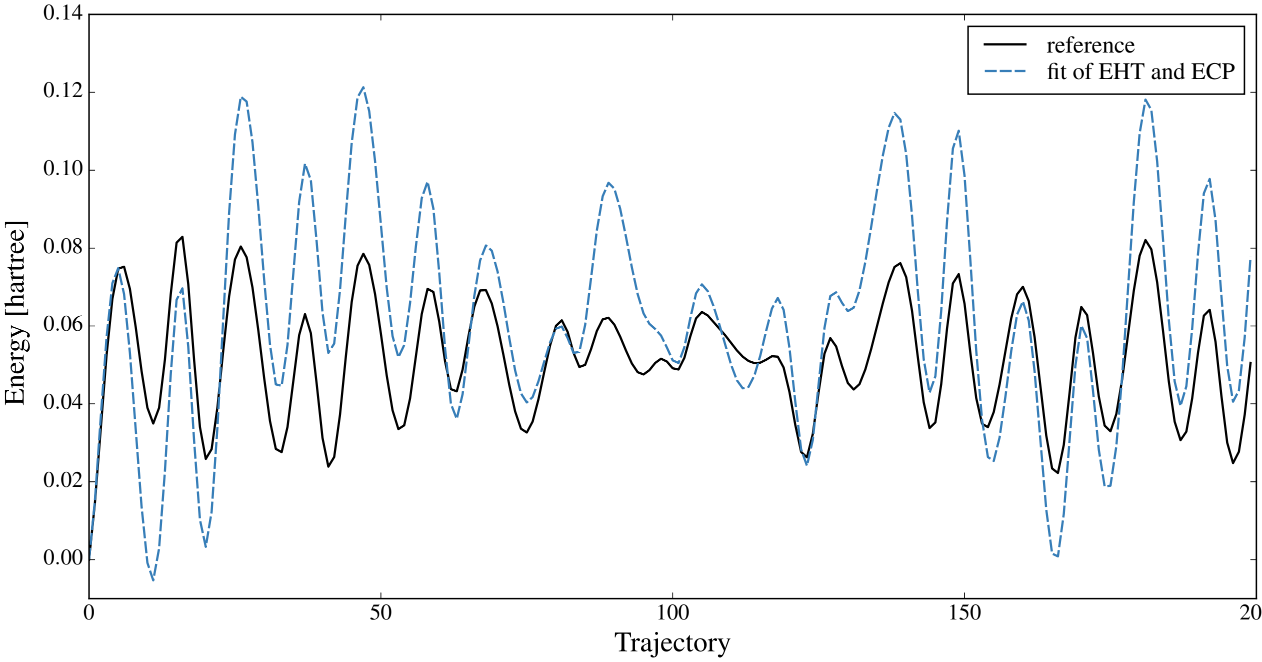


Figure 3.11: Results of the simultaneous fit of the EHT and ECP parameters compared to the reference curve.

Figure 3.11 shows a comparison of the fitted model to the reference curve. The RMSD for the entire trajectory is 0.021 hartree. It is important to note that while the absolute accuracy is not comparable to the force field methods, this model shows the correct functional course using significantly fewer parameters.

4 Direct interpolation of SCF ingredients

In section 3.2 we could demonstrate that the Wolfsberg-Helmholz formula seems promising for evaluation of non-diagonal elements based on the diagonal elements of the Hückel matrix. In section 3.4.1 we showed that a replacement of core electrons by a simplified ECP ansatz (sporting a series expansion in the overlap matrix S) is capable of reproducing the total energies if exact ingredients for the molecular orbital coefficients and energies are provided (taken from the SCF data points). What needs to be shown now in order to obtain a competitive fitting algorithm is that the geometry-dependent variation of the main diagonal of the Hückel matrix can be captured as well. Note the double importance of these matrix elements for a correct evaluation of non-diagonal elements in the EHT approach as well as for the generation of an appropriate eigenset (molecular orbital energies and coefficients) for the evaluation of the ECP-dependent total energy expression.

4.1 A few words on SCF ingredients

The physical information of a molecule is contained in the converged SCF Fock matrix. The total energy of a converged SCF calculation is:

$$E = \sum_i \epsilon_i + \sum_i \vec{c_i}^T h \vec{c_i} + V_{NN}, \tag{4.1}$$

with the molecular orbital energies ϵ_i and the molecular orbital coefficients $\vec{c_i}$. Using the eigenvalue equation and splitting the Fock matrix into the one-electron part h (core Hamiltonian) and a two-electron contribution G ($F = h + G$), the equation can be rewritten as

$$E = \sum_i \vec{c_i}^T \left(F + h\right) \vec{c_i} + V_{NN} = \sum_i \vec{c_i}^T \left(2h + G\right) \vec{c_i} + V_{NN}. \tag{4.2}$$

Introducing the electronic kinetic energy T, the electron-core interaction V, the electron-electron interaction J, and the exchange energy K, we obtain

$$E = \sum_i \vec{c_i}^T \left(2T + 2V + 2J - K\right) \vec{c_i} + V_{NN}. \tag{4.3}$$

The difficulty of interpolating these contributions is illustrated in Figures 4.1 and 4.2, which show evaluations of these quantities along the butadiene reference curve. For simplicity, all calculations presented in this chapter use the STO-3G basis set [9]. Figure 4.1 contains energy contributions that vary on a large scale with the geometry. These large contributions are the three Coulomb interactions: the electron-electron interaction E_J, the electron-core interaction E_V, and the core-core interaction V_{NN}.

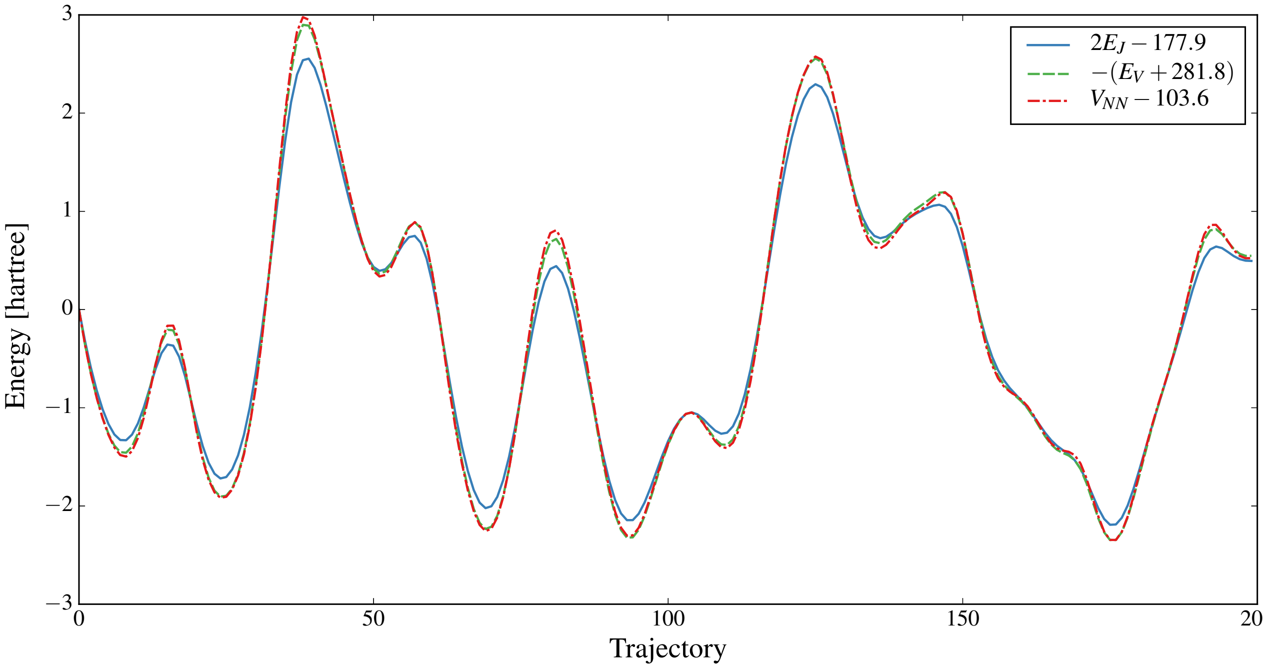


Figure 4.1: Energy comparisons between the three Coulomb contributions along the butadiene reference trajectory.

The three energies are shifted by their respective value at the first trajectory point to make them easier to compare. In addition, the attractive electron-core interaction was flipped to the same sign as the repulsive contributions. The functional course of all three contributions is very similar. Their cancellation at this scale can be argued by fact that the electrons position around the nuclei to screen the nuclear potential. The slight deviations between these energies are nevertheless an important contribution to the course of the total energy. Therefore, the quantities have to be combined ($E_V + 2E_J$ and $E_V + V_{NN}$) in order to compare them to the remaining contributions of the kinetic and the exchange energy.

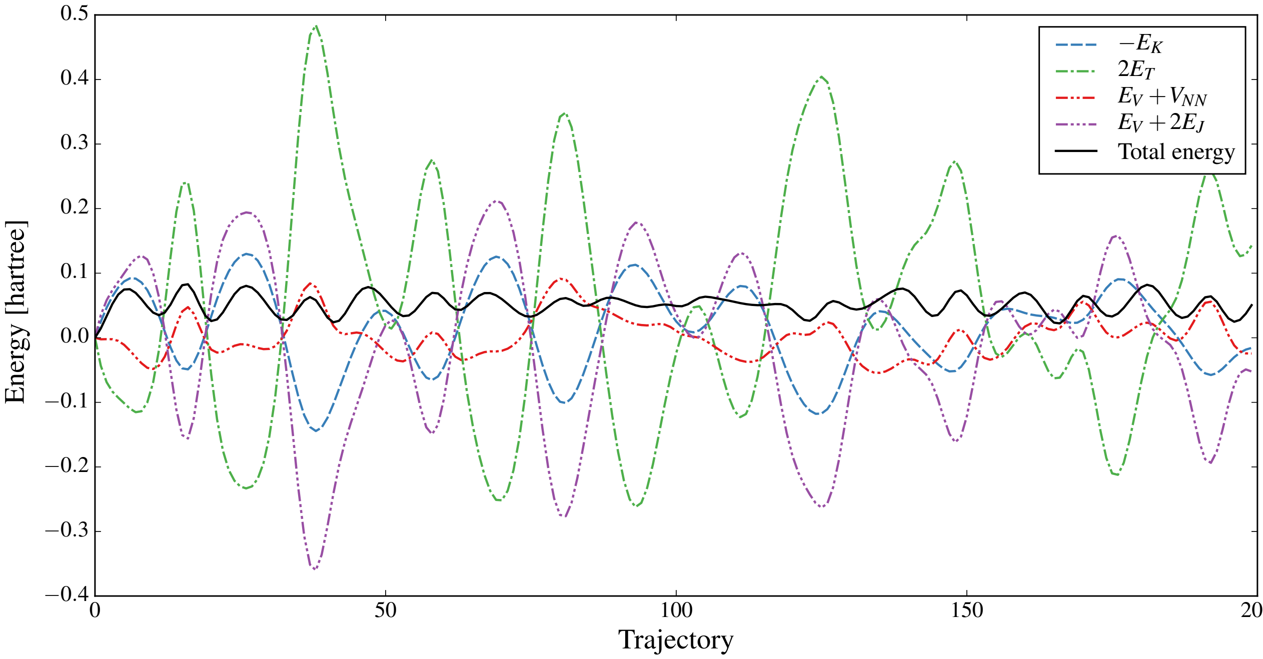


Figure 4.2: Plot of the various contributions to the SCF energy plotted over the butadiene reference curve.

Figure 4.2 shows a comparison between the individual contributions and the total energy along the reference curve. The cancellation of different contributions becomes obvious at this scale as well. The kinetic energy seems to almost cancel the contributions from the exchange energy and the combination of electron-electron and electron-core interaction. The remaining part, the combination of electron-core attraction and core-core repulsion, already shows most of the features of the total energy curve. While it may seem like many of this quantities are not too expensive to compute (with the exception of the core-core interaction), all of them require the knowledge of the converged SCF electron density.

Recently, Martinez et al. [52] presented a scheme (dSCF) to separate the Fock matrix into a part originating from the superposition of atomic densities (SAD) and a deformation due to chemical bonding. They suggest that this method can eliminate the problem of large scale contributions as they are already contained in the SAD contribution. However, this approach requires the calculation of two-electron integrals which is too computationally expensive in the context of PES interpolation.

4.2 Fitting of Fock matrix elements

Section 3.1 showed that the Wolfsberg-Helmholtz formula gives good estimates for the structural form of the Fock matrix. An IMLS formalism for the Fock matrix elements, based on the ideas from extended Hückel theory, will be presented in section 4.3. In this section, different power series expansions are suggested for the diagonal and the off-diagonal elements of the matrix.

4.2.1 Off-diagonal elements

The off-diagonal elements can be expanded in terms of the overlap matrix S.

$$F_{ij} = k_{0,ij} + k_{1,ij}S_{ij} + \sum_l k_{2,ijl}S_{il}S_{lj} + \sum_l \sum_m k_{3,ijlm}S_{il}S_{lm}S_{mj} \tag{4.4}$$

An averaging of diagonal elements similar to the Wolfsberg-Helmholz formula is not used because is restricts the functional freedom and complicates the fitting procedure.

We further simplify equation 4.4 by restricting the coefficients $k_{n,ijl\ldots}$ to be dependent on the expansion order and the position in the matrix given by the indices i, j but not on the indices summed over $(l, m, \ldots)$.

$$k_{2,ijl} = k_{2,ij} \quad k_{3,ijlm} = k_{3,ij} \quad \ldots$$

This is done to reduce the number of parameters for a given order. Without this simplification, the number of parameters would also be dependent on the number of basis functions. Equation 4.4 can be rewritten using the definition of the power of a matrix $A^m = \underbrace{A \cdot \ldots \cdot A}_{m}$. The constant factor is separated out since $A^0 = \mathbf{1}$ is zero for off diagonal elements:

$$F_{ij} = k_{0,ij} + \sum_{m=1} k_{m,ij}(S^m)_{ij}. \tag{4.5}$$

4.2.2 Diagonal elements

The interpolation of the diagonal matrix elements is more difficult due to higher requirements on precision. Two possible expansions are investigated here: The expansion in an inverse power series of internuclear distances and, again, an expansion in terms of the overlap matrix.

As a first approximation only nearest neighbor distances are used for the inverse power series.

$$F_{ii} = \sum_B \sum_{m=0} \frac{c_m}{R_{AB}^m}, \tag{4.6}$$

with A denoting the atom the basis function i is centered on, B as the neighboring atoms, and R_{AB} as the distance between A and B.

The second approach tries to incorporate an angular dependence by using the overlap matrix elements as basis for the expansion,

$$F_{ii} = \sum_{B} \sum_{j \in B} \sum_{m} c_{jm} S_{ij}^{m}. \tag{4.7}$$

In order to compare the functional form of these two approaches the diagonal elements of the butadiene Fock matrix are fitted along the reference trajectory.

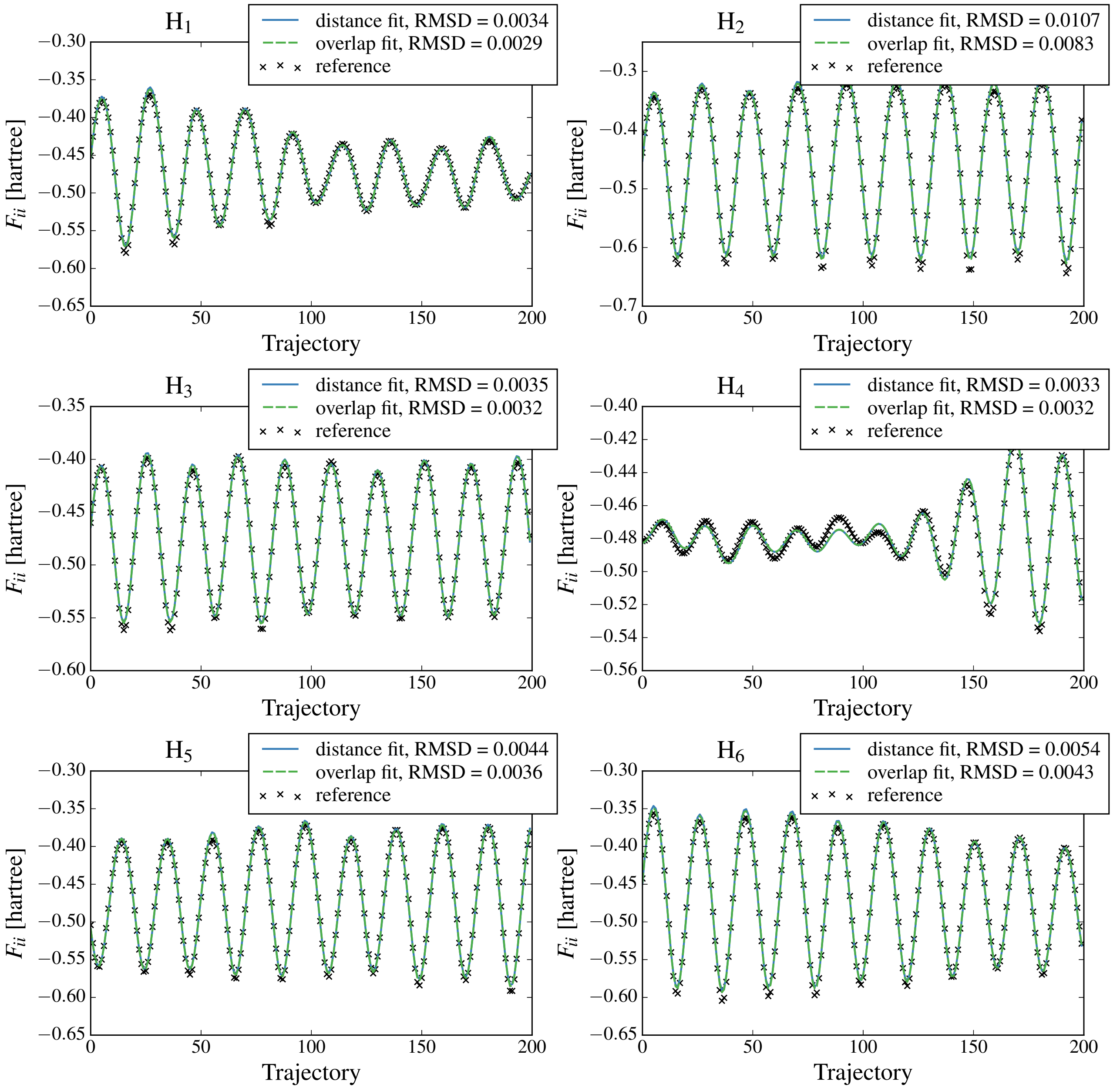


Figure 4.3: Comparison of the two different expansion approaches. Fit of the diagonal elements for the 1s basis functions of the hydrogen atoms in butadiene on the reference curve.

Figure 4.3 shows the fits for the diagonal elements corresponding to the 1s basis functions on the six hydrogen atoms. For these fits only the distance to the neighboring carbon atoms and the overlap between the 1s orbital on the hydrogen and the 2s orbital on the carbon are used. Both fitting functions are of first order and therefore have the same number of degrees of freedom. Judging by the RMSD, the expansion in terms of the orbital overlap gives slightly better results.

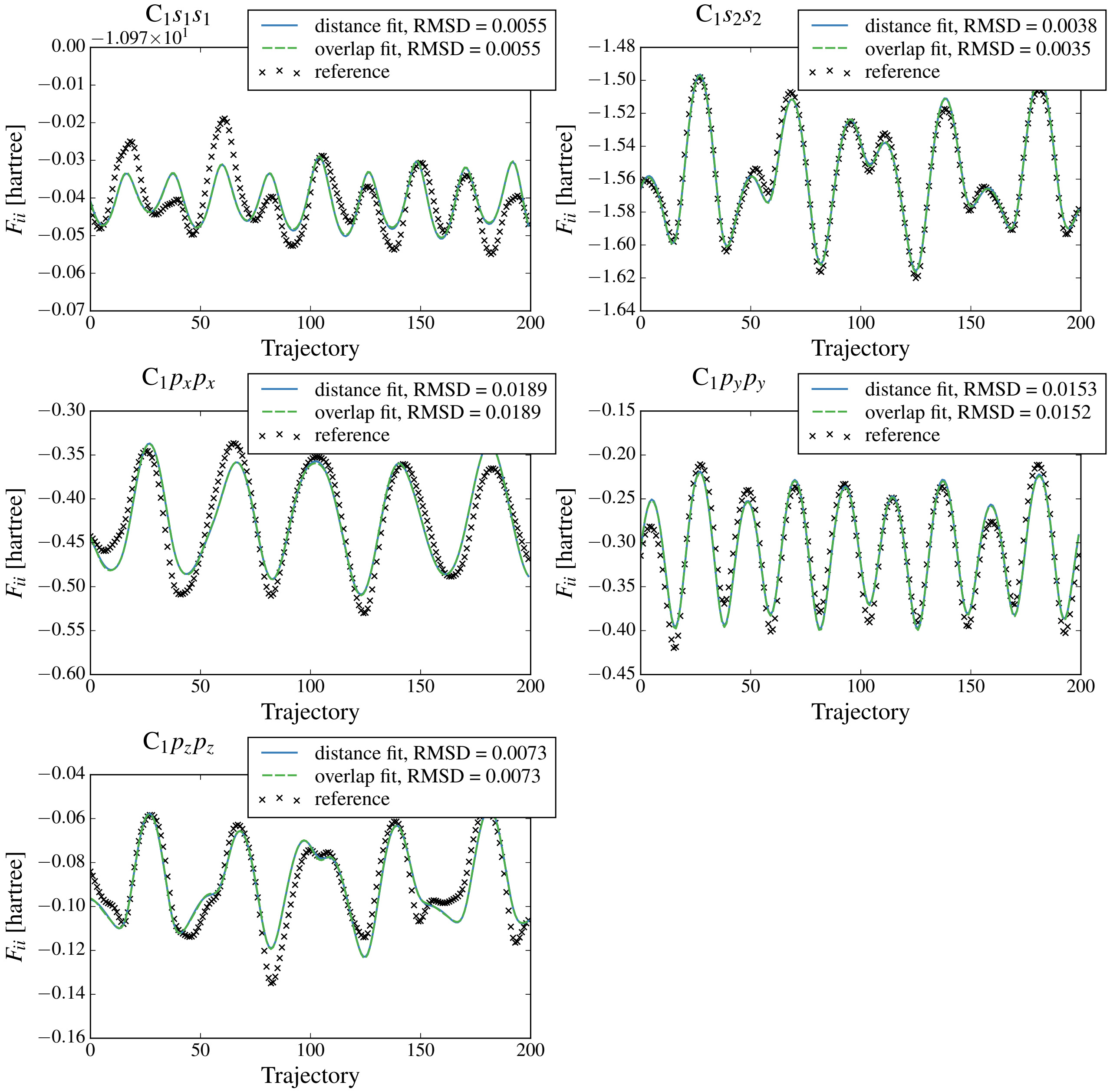


Figure 4.4: Comparison of the two fit approaches for diagonal elements belonging to basis functions on a carbon atom along the butadiene reference curve.

Figure 4.4 shows similar calculations for the diagonal elements corresponding to the basis functions on the first carbon atom. The fits shown are based on first order expansions of the inverse distance to the three neighboring atoms and the overlap of the 2s orbitals with the highest radial symmetric orbital on the neighboring atoms, respectively. Again, the overlap based approach performs slightly better with respect to the RMSD. For the sake of comparability, the possible advantage of the overlap based expansion, which is the addition of an orientational dependence through non-radial symmetric orbitals, was not used. In a tentative implementation the inclusion of these orbitals lead to a unreasonably high number of parameters. Using the p orbitals on the carbon would, for example, increase the number of parameters from 4 to 13 without giving significant improvements in accuracy.

4.3 Fock matrix IMLS

The expansions used in section 4.2 are used in an IMLS scheme for the individual Fock matrix elements. As a first test, the same distance-dependent weight function is used for all matrix elements. In theory, it would be possible to use only local distortions in the definition of the weight function. It is important to note that this method is an interpolation to the Fock matrix. A direct interpolation of the energy would not be possible because of the complicated energy expression occuring in SCF calculations. The computational effort of this method is comparatively high. The overlap matrix and the core Hamiltonian have to be calculated in the full basis (STO-3G) at every geometry.

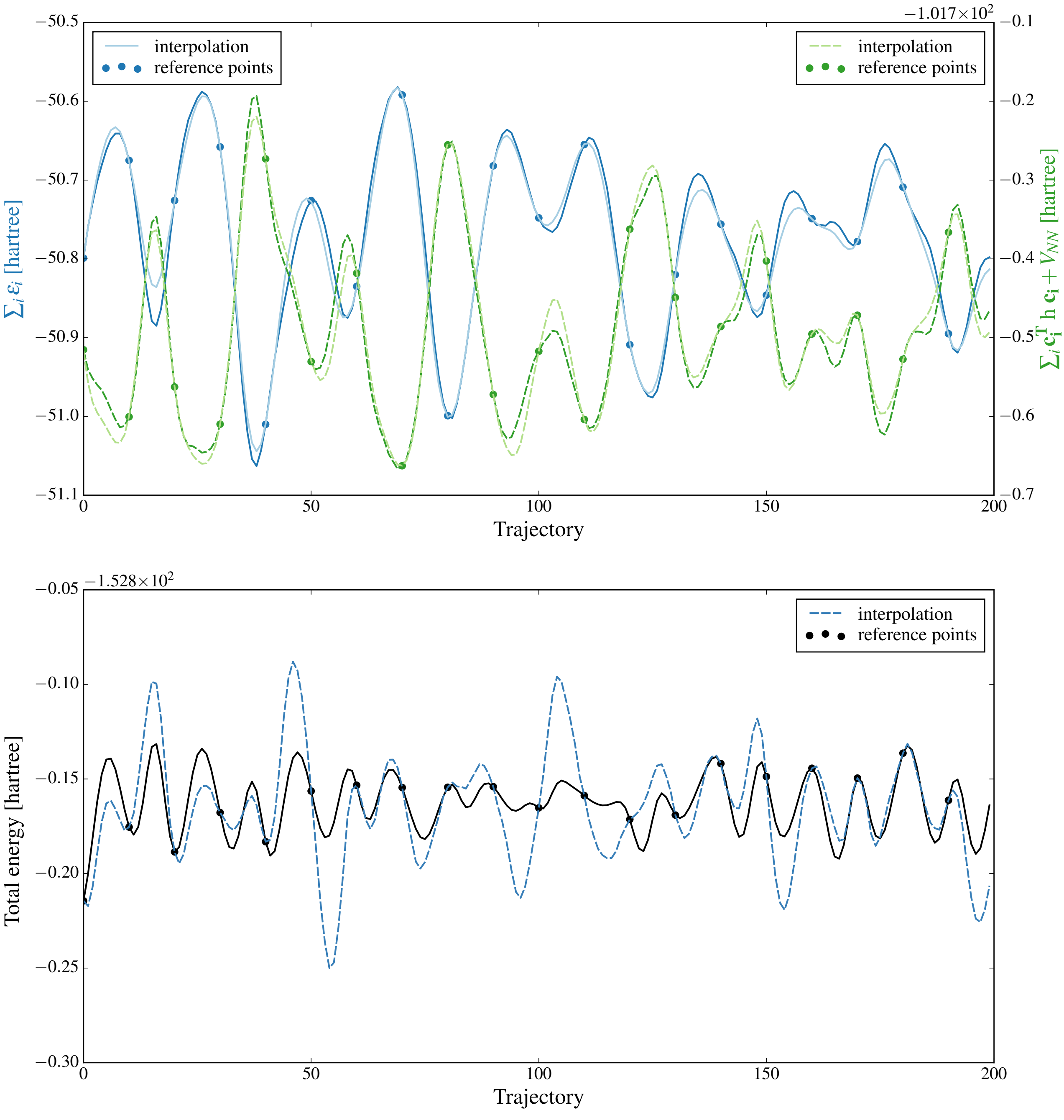


Figure 4.5: Results of the IMLS calculation for the Fock matrix elements. The upper figure shows the sum on molecular orbital energies on the left hand side axis and the remaining energy contributions on the right hand side axis. The lower figure compares the total energies of interpolation and reference. The off-diagonal elements are expanded to sixth order, the diagonal elements to first order.

Figure 4.5 shows the results of an IMLS interpolation of the Fock matrix elements. A total of 20 reference points from the butadiene reference trajectory are used. The off-diagonal elements are expanded to sixth order in terms of the overlap matrix. The diagonal elements are interpolated using a first order inverse nuclear distance expansion. The simple inverse Euclidean distance weight function from equation 1.5 (with $p = 4$) is used. This example again illustrates the problems of fitting SCF-ingredients: The individual contributions to the SCF-energy are large quantities but almost cancel each other. Therefore, the energy changes only by about a thousandth of the absolute energy along the trajectory and very high relative precision is needed for all contributions.

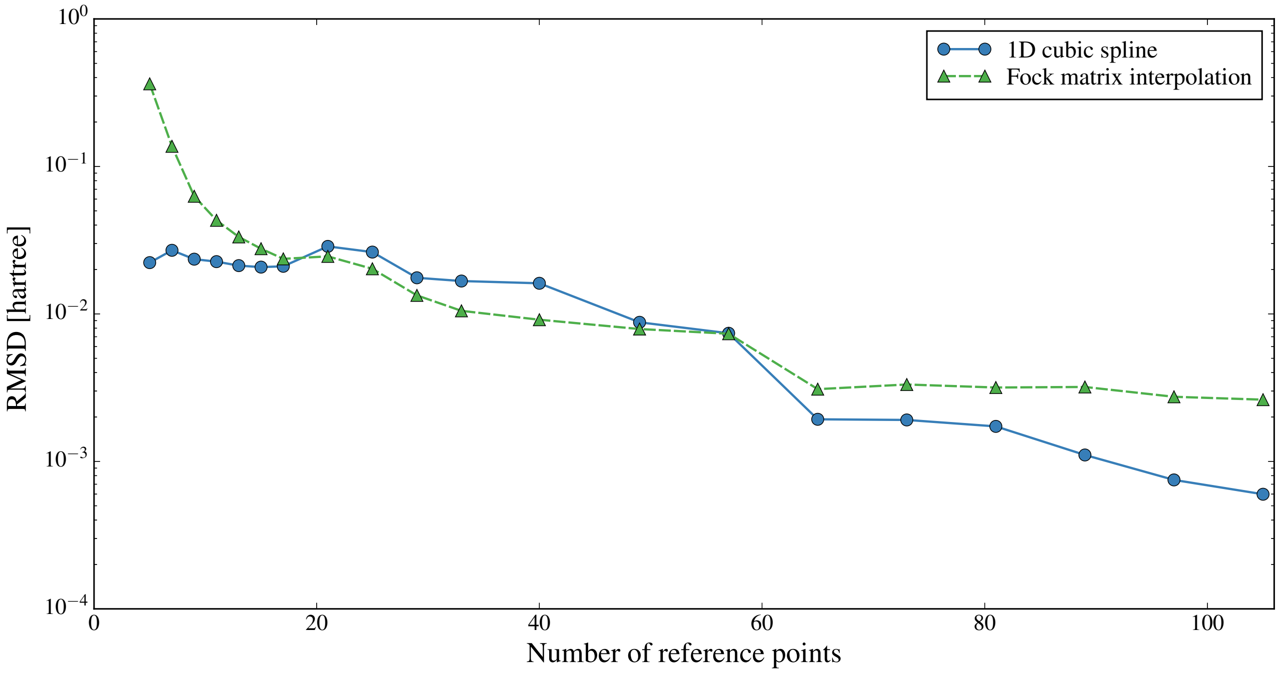


Figure 4.6: Accuracy of the trajectory interpolation using Fock matrix IMLS, plotted as a function of the number of reference points used.

Figure 4.6 shows the accuracy of the interpolation as a function of the number of reference points, measured by the RMSD. Expansion order and weight function are the same as in the prior example. The accuracy is compared to the accuracy of a spline interpolation along the trajectory coordinate. The behaviour of the cubic spline interpolation is discussed in detail in section 5.2.3. The Fock matrix interpolation is slightly more accurate for the region from 20 to 56 reference points. Nevertheless, it can not be considered a viable alternative due to the vastly greater computational effort and the fact that the goal of an accuracy of ≈ 1 kcal/mol seems out of reach.

5 A last promising approach: IMLS + FF

5.1 General concept

In this approach the interpolating moving least squares formalism is combined with a force field expansion of the potential energy surface. Most of the force field contributions are simple power series expansions in the internal coordinates, similar to the polynomial basis used in classic IMLS. The advantage of this approach is that an expansion based on physical arguments can reduce the number of coefficients necessary. For example, the knowledge of a good expansion point for a bond stretching contribution (the equilibrium distance) saves the effort of calculating the linear coefficient in a power series expansion:

$$E = k\ (l - l_0)^2 = kl^2 - 2kl_0 l + l_0^2 \overset{!}{=} al^2 + bl + c. \tag{5.1}$$

Another example is the Fourier series employed in the torsional contribution which includes the periodicity of this internal coordinate by definition:

$$E = \sum_{dihedrals} \sum_{n} V_n\ cos(n\omega) \tag{5.2}$$

Using a sufficiently large number of cross interaction contributions in the force field energy, any arbitrarily shaped PES can be represented and therefore be interpolated. The computational effort (or the necessary number of reference points) can be reduced if further refinement is built on top of already parameterized force fields, giving a good estimate of which cross interactions to include and which contributions to consider most important for interpolation. They also provide starting values for all parameter optimizations. For small numbers of reference points, this approach also offers the possibility to optimize just a selection of the coefficients and use the standard force field parameters for the remaining.

The mathematical derivation for this approach can be found in the appendix 7.4 and does not differ much from the derivation of classic IMLS (section 1.2.2). It uses a force field based trial function

$$E_{trail} = E_0 + \sum_{bonds} k\ (l - l_0)^2 + \sum_{angles} k\ (\theta - \theta_0)^2 + \sum_{dihedrals} \sum_{n} V_n\ cos(n\omega) + E_{non-opt}. \tag{5.3}$$

$E_{non-opt}$ refers to all of the force field contributions that are not optimized during the interpolation. Using the trial function in equation 5.3, the weighted sum of square deviations is minimized as follows:

$$\Delta = \sum_{i=0}^{N} w_i(\mathbf{r})\ [E_{trial}(\mathbf{r_i}) - E_i]^2 \overset{!}{=} \min \tag{5.4}$$

In the presented formulation (matrix equation solved with SVD) only linear coefficients of the force field energy can be adjusted. For this reason, equilibrium lengths of bonds l_0 or parameters of the Van-der-Waals potential can not be optimized this way. They would require a non-linear minimization of the distance function Δ, which would increase the computational effort of the interpolation significantly.

5.2 Tests

As a prove of concept, the force field based IMLS was applied to the butadiene reference curve. For all of these simple tests an inverse distance weight function (with $p = 4$) was used:

$$w_i(\mathbf{r}) = \frac{1}{d(\mathbf{r}, \mathbf{r_i})^4}. \tag{5.5}$$

In classic IMLS the complexity of the interpolation function is given by the degree of the multivariable polynomials used. In the force field based approach one has to differentiate not only by degree of the individual contributions, but also by the type of contributions that are incorporated into the interpolation.

5.2.1 Zeroth order

Analogous to Shepard's method, in zeroth order IMLS-FF only an additional constant is adjusted during the interpolation. By rearranging the optimization function it becomes obvious that this method can be considered as a Shepard's interpolation for the deviations of the force field energy from the reference energy:

$$\begin{gathered} E_{trial} = E_0 + E_{ff} \\ \Delta = \sum_{i=0}^{N} w_i(\mathbf{r}) \left[E_{trial}(\mathbf{r}) - E_i\right]^2 = \sum_{i=0}^{N} w_i(\mathbf{r}) \left[E_0 - (E_i - E_{ff}(\mathbf{r_i}))\right]^2 \end{gathered} \tag{5.6}$$

Zeroth order IMLS-FF therefore suffers from the same problems as the original Shepard's method. The impact of these shortcomings of Shepard's method on the quality of the interpolation are reduced because the deviation of the force field energy from the reference curve varies on a smaller scale than the total molecular energy itself. The presented results are based on the MM2 and GAFF force fields.

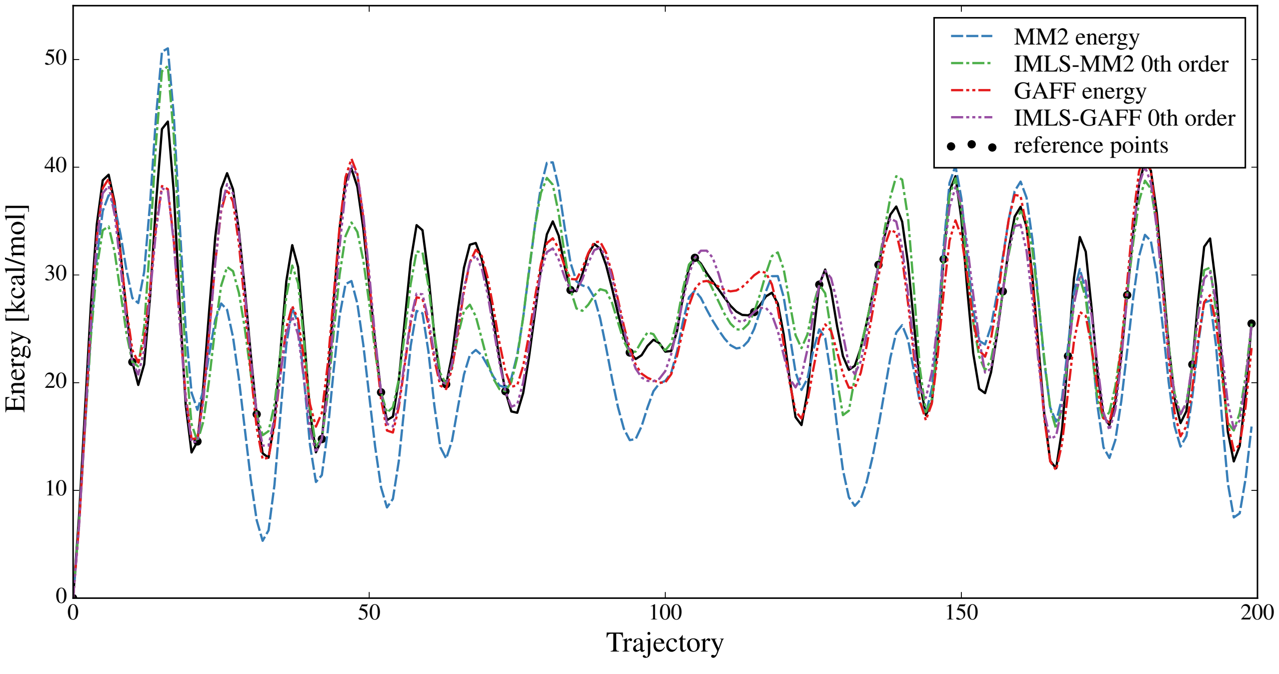


Figure 5.1: Comparison of IMLS-MM2 and IMLS-GAFF zeroth order (using 20 reference points), the MM2 and GAFF force field energy and the butadiene reference curve.

Figure 5.1 shows a comparison of the MM2 and GAFF force field energy and the zeroth order IMLS-MM2 and IMLS-GAFF, both interpolating 20 reference points of the butadiene reference curve. The 'flat-spot' phenomenon of Shepard's method is not as notable in this approach as the interpolation function will have the same derivative as the force field at the reference points. It does, however, still appear in the interpolated deviations of force field energies from the reference values. While there are notable improvements in comparison to the MM2 force field, the interpolation function is restricted by the functional form of the force field. The peak height of features in the potential energy surface can easily be adjusted using this formalism. However, changing the functional course would take a very high number of reference points.

5.2.2 Interpolating the stretching energy

As shown in section 2.3, the largest contribution to the total energy for small molecules typically is the bond stretching energy. Including the bond stretch coefficients into the interpolation scheme is therefore the obvious next step. The trial function is adapted to also incorporate the bond stretching energy as an optimized contribution:

$$E_{trial} = E_0 + \sum_{bonds} \sum_{n=2}^{n_{max}} k_n (l - l_0)^n + E_{non-opt}. \tag{5.7}$$

The $E_{non-opt}$ in equation 5.7 refers to the sum of all force field contributions but the stretch energy. Again, this can be interpreted as fitting the bond energy (and a constant) to the difference between the reference energy and the sum of all the other force field contributions.

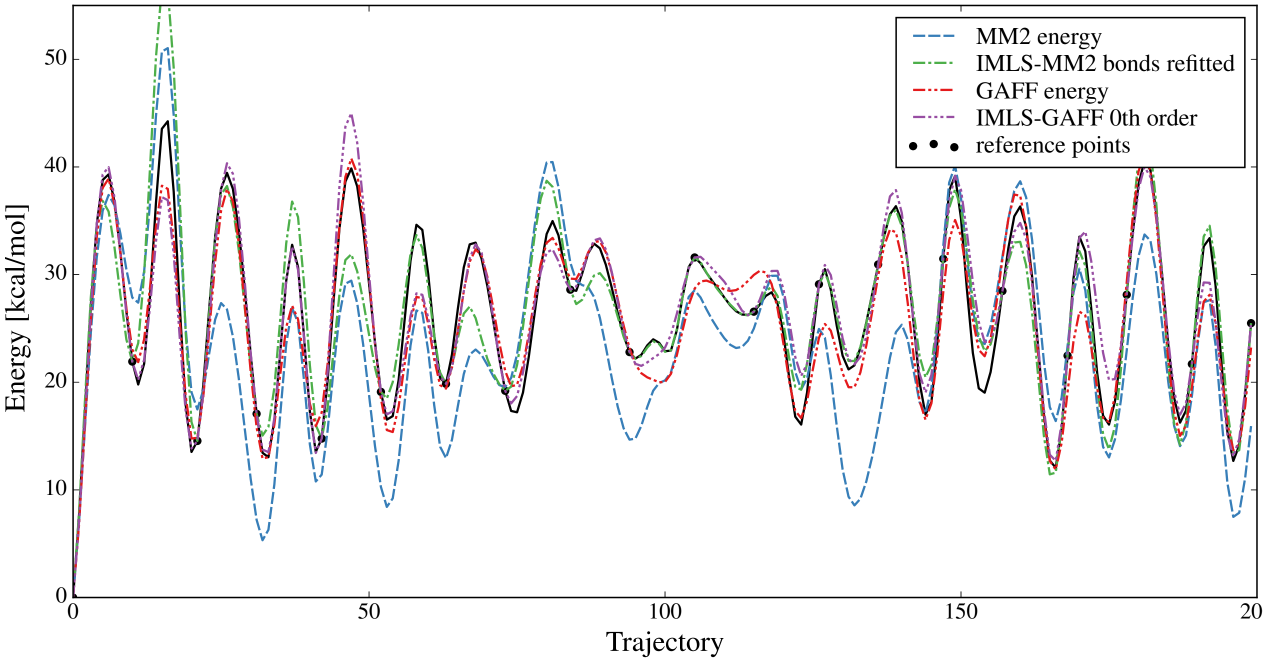


Figure 5.2: Comparison of IMLS-FF with interpolated bond coefficients (using 20 reference points), the MM2 force field energy and the butadiene reference curve.

For the butadiene reference curve using the MM2 force field as a basis this gives a total of four bond parameters to interpolate. The two types of bonds (C-C and C-H) are expanded to third order. GAFF distinguishes four bond types but only expands them to the harmonic order, resulting, again, in 4 parameters. The first order terms for these potentials are not optimized but the equilibrium distances from the respective force field are used. Figure 5.2 shows the result for such an interpolation using 20 equidistant reference points along the trajectory. The comparison with the reference curve and the standard force fields again shows significant improvements. Increasing the expansion order in MM2 does not yield further improvement. This is probably because the bond stretch energy is already sufficiently taken care of, and the remaining error is due to a wrong description of other energy contributions. In order to further improve the bond stretch contribution the bond types would have to be distinguished in more detail.

5.2.3 Limiting behaviour

Using the butadiene reference curve, the accuracy of the two presented approaches was calculated for a given number of reference points. Comparison to any of the classic interpolation techniques is difficult given the low number of reference points. A cubic spline in 24 dimensions would not even be reasonable using all 200 reference points, and first degree IMLS would still involve the optimization of 25 parameters. In order to give a comparison to mathematically motivated interpolation functions we calculate a one-dimensional cubic spline along the trajectory coordinate. The accuracy is measured using the RMSD from the points of the reference curve that were not used as reference points in the interpolation.

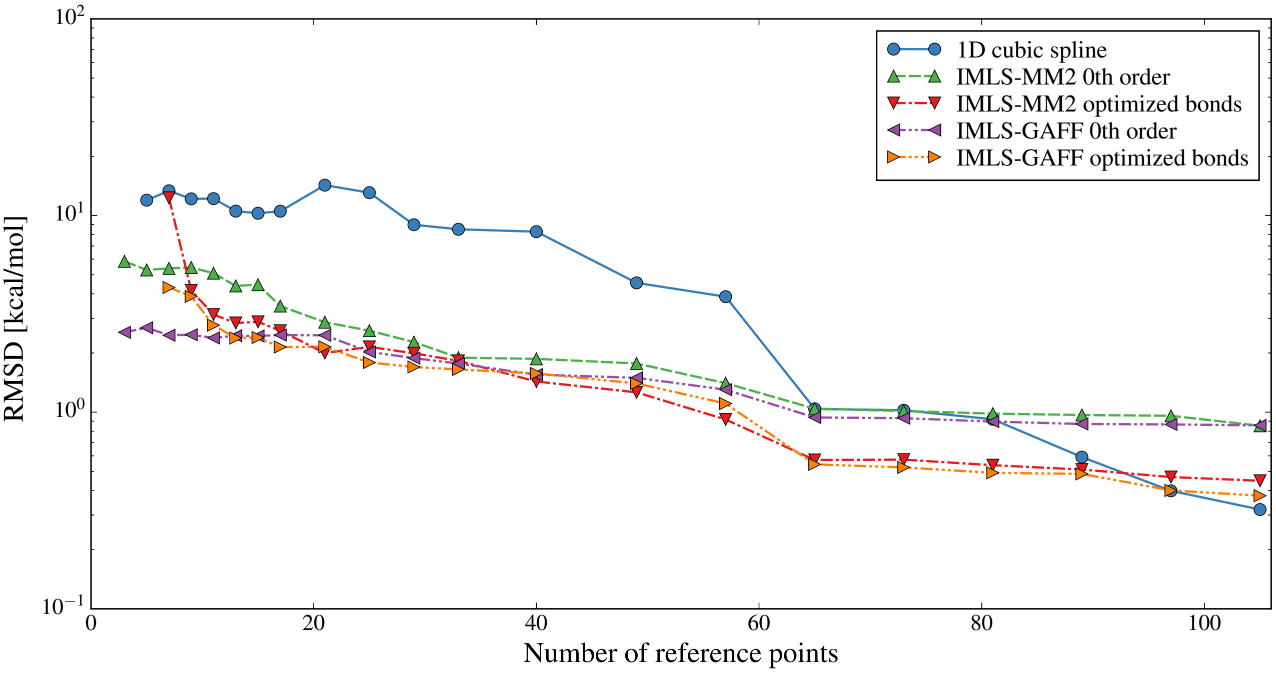


Figure 5.3: The accuracy of different interpolation schemes measured by the RMSD for the butadiene reference curve

In Figure 5.3 the RMSD for different interpolation methods as a function of the number of reference points is compared. For low numbers of reference points the zeroth order IMLS-FF methods show the best accuracy. Both of the other methods, the one-dimensional spline and the bond optimized IMLS-FF, need a certain minimum of reference points in order to work properly. Both bond optimized IMLS-FFs have a total of five parameters and surpass the accuracy of the zeroth order IMLS-FFs at about ten to twenty reference points. The two force field based methods are more accurate than the spline by almost an order of magnitude for up to 60 reference points. The Nyquist–Shannon sampling theorem states that the sampling rate needs to be at least twice the frequency of the fastest vibration in the system in order to capture all information of the signal. This is exactly what can be seen at about 65 reference points, where the accuracy of the spline interpolation drastically increases.

5.3 Outlook

The tests of section 5.2 show that one important addition would be to include additional information in the form of gradients at the reference values. The reformulation of the force field based IMLS approach to incorporate gradients can be done analogous to the modified Shepard's method [53]. Since this would restrict the method to reference calculations where gradient information is easily accessible, the number possible applications would be reduced. In this first investigation no special application was to be preferred.

5.3.1 Reactive force fields

The argument for using force fields as a basis for the interpolation scheme was based on the fact that they can reproduce the PES reasonably well. This in turn restricts the possible applications to regions of the PES where this is known to be true. Classical force fields are an expansion at the equilibrium geometry and can not properly describe bond breaking or large amplitude motions. It would, therefore, be of interest to apply the formalism presented in this work to so called reactive force fields which allow for bond breaking by introducing variable bond order [54].

5.3.2 Adaptive complexity

Figure 5.3 shows that improvements in the accuracy fade out after a certain number of reference points is reached. As discussed, this is caused by the limited flexibility of the interpolation function, which does not allow for a better description. As soon as this density of reference points is reached other contributions of the force field energy would need to be optimized as well. The aim of further investigations will therefore be the implementation of an automatic scheme to adapt the complexity of the interpolation function depending on the density of reference points.

5.3.3 Alternative weight functions

In his original publication Shepard [1] already suggested alternative weight functions. For the IMLS-FF approach new types of weight functions would be of interest. A property of the normalized inverse distance weight function is:

$$v_i(x) \to N^{-1} \qquad \text{as } d(x,0) \to \infty, \tag{5.8}$$

which gives equal weight to all reference points at infinite distance. This corresponds to having no prior knowledge at all. However, in a force field based interpolation, it would be best to reproduce the original coefficient of the optimized contributions at large distances from the reference values.

Another aspect to consider is that the Euclidean distance might not be the best distance function for the weights. Distortions at the one end of the molecule are not of important for bond stretching coefficients at the other end of the molecule. Weights for the interpolation of individual contributions could be based on distortions in the nearby environment.

6 Conclusion

During the course of this thesis several concepts of the wide field of computational quantum chemistry were investigated for their suitability in the context of PES interpolation.

Among them, the simplest ansatz is given by force field methods. The intuitive description of the molecular energy as a sum of contributions proved to be the most viable for interpolation. Its greatest advantage is the simple and straight forward energy expression as a function of the internal coordinates. On the downside, the energy contributions are an abstraction of the physical processes happening at the quantum mechanical level, and therefore incorporating more information of the SCF reference calculations than just the total energy, such as molecular orbital energies or the electron density, proves to be difficult. The lack of a consistent way to incorporate more information from the SCF reference was the motivation to investigate more sophisticated quantum chemistry methods.

It was shown that the energy expression of force fields is flexible enough to reproduce the ab intio PES for geometries near the equilibrium. By refitting the largest energy contributions the deviations of the MM2 force field from the reference values could be reduced significantly.

Building on the investigations of the individualization of force fields for a specific molecule, a force field based IMLS scheme (IMLS-FF) was presented. It allowed to adjust linear parameters of the energy contributions over the course of the PES. First tests showed that the approach excels at low numbers of reference points, where common interpolation techniques could not be employed. However, there are still numerous restrictions that need to be overcome any possible application. In its current implementation the algorithm can, for example, not handle bond breaking and other large amplitude phenomena. A possible solution would be to use reactive force fields. Another challenge is to incorporate adaptive complexity in order get the flexibility needed to interpolate high densities of reference points while maintaining the ability to interpolate low numbers of SCF evaluations. Both problems could possibly be solved by weight functions that are more tailored to the application of IMLS to the force field energy expression.

The extended Hückel theory was chosen from the large variety of semi-empirical methods because of its proximity to the SCF formalism. Our investigations confirmed the predictive power of this theory, which explains its ongoing well-established use for qualitative descriptions of electron densities in modern quantum chemistry. However, the method does not provide a sufficiently accurate enough description of the molecular orbital energies. Refitting the parameters can reduce the deviations significantly, but can not improve the incorrect functional course. Our study revealed that this is largely due to a false assumption of constant Hückel matrix diagonal elements. Unfortunately, a simple fitting the diagonal elements using a series expansion in terms of the overlap matrix does not yield the desired accuracy. In addition, the total energy of an SCF calculation can not be reproduced using the simple energy expression of the Hückel theory.

A few of the numerous improvements to the total energy expression that had been suggested over the years were reviewed in this thesis. However, none of them proved to be competitive to the force field approach for the description of the molecular energy.

Inspired by the concept of effective core potentials a new contribution for the total energy expression was presented. The suggested simplification of ECPs based on the overlap matrix showed promising results if the exact SCF electron densities were used. However, combining this new approach with classic EHT proves unsuccessful as the molecular orbitals are not sufficiently well described by the

Hückel methods.

The investigated approach to interpolate the ingredients of a SCF calculation directly suffers from the problem of different scaling in the energy contributions. Relative accuracies better than 10^{-5} would be necessary in order to achieve the desired interpolation quality. Based on ideas of the EHT, namely the expansion in terms of the overlap matrix, the presented IMLS formalism for the Fock matrix also encounters the limits of the EHT approximations. Relating the integrals which define the Fock matrix to the overlap matrix is only valid for the homogeneous electron densities exhibited by valence electrons. Ultimately, the studied concepts based on the interpolation of SCF ingredients prove to be computationally very demanding but offer only minimal gain in accuracy, which makes them hardly competitive in their current form.

7 Mathematical derivations

7.1 Born-Oppenheimer approximation

The Born-Oppenheimer approximation is a fundamental assumption that is made in most quantum chemistry calculations. It states that the motion of the electron and the nuclei can be treated separately. This derivation follows lecture notes of W. Domcke [55]. The total molecular Hamiltonian in atomic units is:

$$H_{tot} = \underbrace{-\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2}_{T_E} \underbrace{-\sum_{a=1}^{M}\frac{1}{2M_a}\nabla_a^2}_{T_N} \underbrace{-\sum_{i=1}^{N}\sum_{a=1}^{M}\frac{Z_a}{r_{ia}}}_{V_{NE}} \underbrace{+\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}}}_{V_{EE}} \underbrace{+\sum_{a=1}^{M}\sum_{b>a}\frac{Z_aZ_b}{R_{ab}}}_{V_{NN}}. \tag{7.1}$$

The wave functions and the corresponding energies can be calculated from the Schrödinger equation:

$$(H_{tot} - \mathcal{E})\Psi(\mathbf{r},\mathbf{R}) = 0. \tag{7.2}$$

For fixed nuclear positions $\mathbf{R}$ the nuclear kinetic energy T_N vanishes. The nuclear positions $\mathbf{R}$ are parameters in the electronic Hamiltonian,

$$H_{elec} = T_E + V_{NE} + V_{EE} + V_{NN}. \tag{7.3}$$

The solutions of the corresponding electronic Schrödinger equation,

$$(H_{elec} - E_m(\mathbf{R}))\Phi_m(\mathbf{r},\mathbf{R}) = 0, \tag{7.4}$$

are a complete set of functions. The total wave function $|\Psi(\mathbf{r},\mathbf{R})\rangle$ can therefore be expanded in terms of the electronic wave functions:

$$\Psi(\mathbf{r},\mathbf{R}) = \sum_{m=0}^{\infty}\chi_m(\mathbf{R})\Phi_m(\mathbf{r},\mathbf{R}). \tag{7.5}$$

Plugging 7.5 into equation 7.2 yields:

$$(H_{tot} - \mathcal{E})\sum_{m=0}^{\infty}\chi_m(\mathbf{R})\Phi_m(\mathbf{r},\mathbf{R}) = 0. \tag{7.6}$$

Multiplying by Φ_n^* and integrating over the electron coordinates $\mathbf{r}$ gives:

$$\sum_{m=0}^{\infty}\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})\,(T_E + T_N + V_{NE} + V_{EE} + V_{NN} - \mathcal{E})\,\chi_m(\mathbf{R})\Phi_m(\mathbf{r},\mathbf{R}) = 0. \tag{7.7}$$

Using equation 7.4 and the orthogonality of the electronic wave functions one obtains

$$\sum_{m=0}^{\infty}\left[\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})T_N\Phi_m(\mathbf{r},\mathbf{R}) + (E_m(\mathbf{R}) - \mathcal{E})\,\delta_{mn}\right]\chi_m(\mathbf{R}) = 0. \tag{7.8}$$

Applying the differential operator T_k to both $\Phi_m(\mathbf{r},\mathbf{R})$ and $\chi_m(\mathbf{R})$ yields

$$\begin{gathered}\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})T_N\Phi_m(\mathbf{r},\mathbf{R})\chi_m(\mathbf{R}) = -\sum_{a=1}^{M}\frac{1}{2M_a}\sum_{j=1}^{3}\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})\frac{\partial^2\Phi_m(\mathbf{r},\mathbf{R})}{\partial R_{aj}^2}\chi_m(\mathbf{R}) \\ -2\sum_{a=1}^{M}\frac{1}{2M_a}\sum_{j=1}^{3}\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})\frac{\partial\Phi_m(\mathbf{r},\mathbf{R})}{\partial R_{aj}}\frac{\partial\chi_m(\mathbf{R})}{\partial R_{aj}} - \sum_{a=1}^{M}\frac{1}{2M_a}\sum_{j=1}^{3}\int d\mathbf{r}\ \frac{\partial^2\chi_m(\mathbf{R})}{\partial R_{aj}^2}\chi_m(\mathbf{R})\delta_{mn}\end{gathered} \tag{7.9}$$

$$\begin{gathered}\left[-\sum_{a=1}^{M}\frac{1}{2M_a}\sum_{j=1}^{3}\int d\mathbf{r}\ \frac{\partial^2\chi_m(\mathbf{R})}{\partial R_{aj}^2} + E_m(\mathbf{R}) - \mathcal{E}\right]\chi_n(\mathbf{R}) = \\ \sum_{a=1}^{M}\frac{1}{2M_a}\sum_{j=1}^{3}\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})\frac{\partial^2\Phi_m(\mathbf{r},\mathbf{R})}{\partial R_{aj}^2}\chi_m(\mathbf{R}) + \sum_{a=1}^{M}\frac{1}{M_a}\sum_{j=1}^{3}\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})\frac{\partial\Phi_m(\mathbf{r},\mathbf{R})}{\partial R_{aj}}\frac{\partial\chi_m(\mathbf{R})}{\partial R_{aj}}\end{gathered} \tag{7.10}$$

$$\boxed{[T_N + E_n(\mathbf{R}) - \mathcal{E}]\,\chi_n = \sum_m \Lambda_{nm}\chi_m(\mathbf{R})} \tag{7.11}$$

$$\Lambda_{nm} = \sum_{a=1}^{M}\frac{1}{2M_a}\sum_{j=1}^{3}\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})\frac{\partial^2\Phi_m(\mathbf{r},\mathbf{R})}{\partial R_{aj}^2} + \sum_{a=1}^{M}\frac{1}{M_a}\sum_{j=1}^{3}\int d\mathbf{r}\ \Phi_n^*(\mathbf{r},\mathbf{R})\frac{\partial\Phi_m(\mathbf{r},\mathbf{R})}{\partial R_{aj}}\frac{\partial}{\partial R_{aj}} \tag{7.12}$$

This is a system of coupled differential equations. Neglecting the dependency of the Φ_n on the nuclear coordinates $\mathbf{R}$ ($\Lambda_{nm} = 0$) gives the Born-Oppenheimer approximation:

$$\boxed{\begin{aligned}[T_N + E_n(\mathbf{R}) - E_{n\nu}]\,\chi_{n\nu}(\mathbf{R}) = 0 \\ \Psi_{n\nu}(\mathbf{r},\mathbf{R}) = \Phi_n(\mathbf{r},\mathbf{R})\chi_{n\nu}(\mathbf{R})\end{aligned}} \tag{7.13}$$

The total wave function is a product of the nuclear and the electronic wave function. A criterion for the quality of the approximation can be given in the context of perturbation theory:

$$\frac{|\langle\chi_{n\nu}|\Lambda_{nm}|\chi_{m\nu'}\rangle|}{|E_{n\nu} - E_{m\nu'}|} \ll 1 \qquad \text{for every } \nu \neq \nu', n \neq m \tag{7.14}$$

The matrix elements have to be small compared to the energy difference of the states. The approximation will therefore break down if the energy difference of two states is on the order of magnitude of the vibronic frequencies. An exception to this are states for which Λ_{nm} is zero because of symmetry considerations.

7.2 Projection of molecular orbitals

This section explains the method used to project the molecular orbitals obtained from the SCF-calculations onto a smaller basis in order to use them in the context of the EHT. In the larger SCF-basis $\{|\phi_1\rangle, \ldots, |\phi_N\rangle\}$ the molecular orbitals are expanded using the coeffients c_{ai}:

$$|\psi_a\rangle = \sum_{i=1}^{N} c_{ai}|\phi_i\rangle \tag{7.15}$$

In the reduced extended-Hückel-basis $\{|\bar{\phi}_1\rangle, \ldots, |\bar{\phi}_K\rangle\}$ the molecular orbitals are represented using new expansion coefficients d_{ai}:

$$|\bar{\psi}_a\rangle = \sum_{i=1}^{K} d_{ai}|\bar{\phi}_i\rangle \tag{7.16}$$

A suitable projection is found by minimizing the sum of the squared deviations with the constraint of maintaining the orthogonality for the new wavefunctions:

$$\sum_{a}^{M} \big(\langle\psi_a| - \langle\bar{\psi}_a|\big)\big(|\psi_a\rangle - |\bar{\psi}_a\rangle\big) \tag{7.17}$$

$$\langle\bar{\psi}_a|\bar{\psi}_b\rangle - \delta_{ab} = 0 \tag{7.18}$$

The problem is solved using the method of Lagrange multipliers:

$$\begin{aligned}
\mathcal{L} = \sum_{a}^{M} \big(\langle\psi_a| - \langle\bar{\psi}_a|\big)\big(|\psi_a\rangle - |\bar{\psi}_a\rangle\big) - \sum_{a}^{M}\sum_{b}^{M} \lambda_{ab}\big(\langle\bar{\psi}_a|\bar{\psi}_b\rangle - \delta_{ab}\big) &= \\
\sum_{a}^{M} \big(\langle\psi_a|\psi_a\rangle - 2\langle\bar{\psi}_a|\psi_a\rangle + \langle\bar{\psi}_a|\bar{\psi}_a\rangle\big) - \sum_{a}^{M}\sum_{b}^{M} \lambda_{ab}\big(\langle\bar{\psi}_a|\bar{\psi}_b\rangle - \delta_{ab}\big) &= \\
\sum_{a}^{M}\left(\sum_{i}^{N}\sum_{j}^{N} c_{ai}c_{aj}\langle\phi_i|\phi_j\rangle - 2\sum_{i}^{N}\sum_{j}^{K} c_{ai}d_{aj}\langle\phi_i|\bar{\phi}_j\rangle + \sum_{i}^{K}\sum_{j}^{K} d_{ai}d_{aj}\langle\bar{\phi}_i|\bar{\phi}_j\rangle\right)& \\
-\sum_{a}\sum_{b}\lambda_{ab}\left(\sum_{i}^{K}\sum_{j}^{K} d_{ai}d_{bj}\langle\bar{\phi}_i|\bar{\phi}_j\rangle - \delta_{ab}\right)&
\end{aligned} \tag{7.19}$$

the expressions $\langle\phi_i|\phi_j\rangle$, $\langle\phi_i|\bar{\phi}_j\rangle$, and $\langle\bar{\phi}_i|\bar{\phi}_j\rangle$ are replaced by the typical symbol for the overlap matrix S_{ij}, S^*_{ij}, and S'_{ij}, respectively.

$$\begin{aligned}
0 \overset{!}{=} \frac{\partial\mathcal{L}}{\partial d_{ck}} = \sum_{a}^{M}\left[-2\sum_{i}^{N}\sum_{j}^{K} c_{ai}\underbrace{\frac{\partial d_{aj}}{\partial d_{ck}}}_{\delta_{ac}\delta_{jk}} S^*_{ij} + \sum_{i}^{K}\sum_{j}^{K}\left(\underbrace{\frac{\partial d_{ai}}{\partial d_{ck}}}_{\delta_{ac}\delta_{ik}} d_{aj} + d_{ai}\underbrace{\frac{\partial d_{aj}}{\partial d_{ck}}}_{\delta_{ac}\delta_{jk}}\right) S'_{ij}\right]& \\
-\sum_{a}\sum_{b}\lambda_{ab}\sum_{i}^{K}\sum_{j}^{K}\left(\underbrace{\frac{\partial d_{ai}}{\partial d_{ck}}}_{\delta_{ac}\delta_{ik}} d_{bj} + d_{ai}\underbrace{\frac{\partial d_{bj}}{\partial d_{ck}}}_{\delta_{bc}\delta_{jk}}\right) S'_{ij} &= \\
-2\sum_{i}^{N} c_{ci}S^*_{ik} + \sum_{j}^{K} d_{cj}S'_{kj} + \sum_{i}^{K} d_{ci}S'_{ik} - \sum_{b}^{M}\lambda_{cb}\sum_{j}^{K} d_{bj}S'_{kj} - \sum_{a}^{M}\lambda_{ac}\sum_{i}^{K} d_{ai}S'_{ik} &= \\
-2\sum_{i}^{N} c_{ci}S^*_{ik} + 2\sum_{i}^{K} d_{ci}S'_{ik} - 2\sum_{a}^{M}\lambda_{ac}\sum_{i}^{K} d_{ai}S'_{ik} &\overset{!}{=} 0
\end{aligned} \tag{7.20}$$

$$\sum_{i}^{K} d_{ci} S'_{ik} - \sum_{a}^{M} \lambda_{ac} \sum_{i}^{K} d_{ai} S'_{ik} = \sum_{i}^{N} c_{ci} S^{*}_{ik} \tag{7.21}$$

In matrix notation, this can be written as

$$S'd - S'd\Lambda = S^{*}c, \tag{7.22}$$

where the S' is the $K \times K$ overlap matrix in the reduced basis, d is a $K \times M$ matrix consisting of the expansion coefficients in the reduced basis as column vectors, Λ is the $M \times M$ matrix of the Lagrange multipliers, S^{*} is a $K \times N$ overlap matrix and c is a $N \times M$ matrix consisting of the expansion coefficients in the original basis as column vectors.

After rearranging equation 7.22,

$$S'd\,\underbrace{(\mathbf{1} - \Lambda)}_{=:B} = S^{*}c \tag{7.23}$$

$$d = S'^{-1} S^{*} c B^{-1} \tag{7.24}$$

it can be plugged into the orthogonality relation 7.18 to calculate B:

$$d^{T} S' d = \mathbf{1} \tag{7.25}$$

$$\begin{aligned} \left(\underbrace{\left(B^{-1}\right)^{T}}_{=B^{-1}} c^{T} S^{*T} \underbrace{\left(S'^{-1}\right)^{T}}_{=S'^{-1}} \right) S' \left(S'^{-1} S^{*} c B^{-1}\right) &= \mathbf{1} \\ B^{-1} c^{T} S^{*T} S'^{-1} S^{*} c B^{-1} &= \mathbf{1} \\ c^{T} S^{*T} S'^{-1} S^{*} c &= B^{2} \end{aligned} \tag{7.26}$$

$$B = (c^{T} S^{*T} S'^{-1} S^{*} c)^{1/2}. \tag{7.27}$$

Finally, combining equation 7.24 and 7.27 gives

$$\boxed{d = S'^{-1} S^{*} c \left(c^{T} S^{*T} S'^{-1} S^{*} c\right)^{-1/2}.} \tag{7.28}$$

7.3 Projection of the Fock matrix

In order to construct the Fock matrix in a different basis, the spectral representation for the generalized eigenvalue problem,

$$F\vec{d_a} = S'\vec{d_a}\epsilon_a, \tag{7.29}$$

is derived in the following section. The spectral representation in terms of the molecular orbitals is simple because the molecular orbitals are orthogonal:

$$\hat{F} = \sum_{a} \epsilon_a |\psi_a\rangle\langle\psi_a| \tag{7.30}$$

Using the representation of the molecular orbitals in the basis $\{|\bar{\phi_1}\rangle, \ldots, |\bar{\phi_K}\rangle\}$ yields:

$$\begin{aligned} \hat{F} &= \sum_{a} \epsilon_a \sum_{k} d_{ak} |\bar{\phi_k}\rangle \sum_{l} d_{la} \langle\bar{\phi_l}| = \\ &= \sum_{a} \epsilon_a \sum_{k} \sum_{l} d_{ak} d_{la} |\bar{\phi_k}\rangle\langle\bar{\phi_l}| \end{aligned} \tag{7.31}$$

Evaluating the matrix elements in the new basis gives:

$$\begin{aligned} F_{ij} = \langle \bar{\phi_i} | \hat{F} | \bar{\phi_j} \rangle &= \sum_a \epsilon_a \sum_k \sum_l d_{ak} d_{la} \underbrace{\langle \bar{\phi_i} | \bar{\phi_k} \rangle}_{S'_{ik}} \underbrace{\langle \bar{\phi_l} | \bar{\phi_j} \rangle}_{S'_{lj}} = \\ &= \sum_a \epsilon_a \underbrace{\sum_k d_{ak} S'_{ik}}_{=S'^T \vec{d_a}} \underbrace{\sum_l d_{la} S'_{lj}}_{\vec{d}_a^T S'}, \end{aligned} \tag{7.32}$$

which can be rewritten in matrix notation as

$$\boxed{F = \sum_a \epsilon_a S'^T \vec{d_a} \vec{d}_a^T S'.} \tag{7.33}$$

Multiplying (7.33) by $\vec{d_b}$ and using the symmetry of the overlap matrix ($S'^T = S'$) leads to the generalized eigenvalue problem we started with

$$F \vec{d_b} = \sum_a \epsilon_a S'^T \vec{d_a} \underbrace{\vec{d}_a^T S' \vec{d_b}}_{\delta_{ab}} = \epsilon_b S'^T \vec{d_b} = \epsilon_b S' \vec{d_b}. \tag{7.34}$$

Using equation 7.33 the Fock matrix, can be constructed for a subset of the orbitals. In the EHT only valence electrons are considered. To construct a Fock matrix without the core electrons the corresponding eigenvalues can be set to zero. In the case of the four core and eleven valence orbitals of the butadiene molecule this yields

$$F_{red} = \sum_{a=4}^{15} \epsilon_a S'^T \vec{d_a} \vec{d}_a^T S'. \tag{7.35}$$

This Fock matrix has the eleven valence molecular energies as eigenvalues. The remaining eigenvalues are degenerate, having all the value zero.

The information contained in the virtual orbitals is not used throughout this thesis. The corresponding eigenvectors are therefore chosen at random. This is done by creating a random vector and orthogonalizing it with respect to all the other eigenvectors. By introducing the scalar product,

$$\langle \vec{v}, \vec{w} \rangle = \vec{v}^T S \vec{w}, \tag{7.36}$$

the standard Gram-Schmidt algorithm can be used:

$$\begin{aligned} \vec{d'_n} &= \vec{x} - \sum_{i=1}^{n-1} \langle \vec{d_i}, \vec{x} \rangle \vec{d_i}, \\ \vec{d_n} &= \frac{\vec{d'_n}}{\langle \vec{d'_n}, \vec{d'_n} \rangle}, \end{aligned} \tag{7.37}$$

with a random starting vector $\vec{x}$.

7.4 IMLS based on force fields

Analogous to the derivation of IMLS in section 1.2.2 this derivation for IMLS-FF follows Lancaster and Salkauskas [3] as well as Maisuradze and Thompson [4].

The trial function is a sum of force field contributions,

$$E_{trail} = E_0 + \sum_{b \,\in\, bonds} k_b \,(l - l_0)^2 + \sum_{a \,\in\, angles} k_a \,(\theta - \theta_0)^2 + \sum_{d \,\in\, dihedrals} \sum_n V_{d,n} \, cos(n\omega) + E_{non-opt}, \quad (7.38)$$

in which only some of the coefficients (and a constant) are optimized during the moving least square procedure. The other contributions are collected in the $E_{non-opt}$ term. For this trial function the weighted sum of square deviations has to be minimized:

$$\begin{aligned} \Delta &= \sum_{i=0}^{N} w_i(\mathbf{r}) \left[E_{trial}(\mathbf{r_i}) - E_i\right]^2 \\ &= \sum_{i=0}^{N} w_i(\mathbf{r}) \left[E_0 + \sum_{bonds} k_b \;(l_i - l_0)^2 + \sum_{angles} k_a \;(\theta_i - \theta_0)^2 + \sum_{dihedrals} \sum_n V_{d,n} \; cos(n\omega_i) + E_{non-opt} - E_i\right]^2 . \end{aligned} \quad (7.39)$$

E_i are the values at the reference points $\mathbf{r}_i$. In order to simplify the notation all types of parameters are relabeled and numerated continuously:

$$\{E_0, k_{b=1}, k_{b=2}, ..., k_{a=1}, k_{a=2}, ..., V_{d=1,n=1}, V_{d=1,n=2}, ...\} = \{k_0, k_1, ..., k_m\}. \quad (7.40)$$

Deriving equation 7.39 with respect to the coefficients k_i yields $m+1$ normal equations,

$$\begin{aligned} &\sum_i \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_i}} w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_i}} k_0 + \cdots + \sum_i \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_i}} w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_i}} k_m = \sum_i w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_i}} (E_i - E_{non-opt}(\mathbf{r_i})) \\ &\sum_i \left.\frac{\partial E_{trial}}{\partial k_1}\right|_{\mathbf{r_i}} w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_i}} k_0 + \cdots + \sum_i \left.\frac{\partial E_{trial}}{\partial k_1}\right|_{\mathbf{r_i}} w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_i}} k_m = \sum_i w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_1}\right|_{\mathbf{r_i}} (E_i - E_{non-opt}(\mathbf{r_i})) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\vdots \\ &\sum_i \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_i}} w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_i}} k_0 + \cdots + \sum_i \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_i}} w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_i}} k_m = \sum_i w_i(\mathbf{r}) \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_i}} (E_i - E_{non-opt}(\mathbf{r_i})). \end{aligned} \quad (7.41)$$

The equations 7.41 can be rewritten as a matrix equation,

$$\boxed{\mathbf{B}^T \cdot \mathbf{W} \cdot \mathbf{B} \cdot \mathbf{k} = \mathbf{B}^T \cdot \mathbf{W} \cdot \tilde{\mathbf{E}}} \quad (7.42)$$

where $\mathbf{B}$ contains the derivates of the trail function with respect to the coefficients k evaluates at the reference points $\mathbf{r_r}$, $\mathbf{W}$ contains the weights, $\mathbf{k}$ is the coefficient vector and $\tilde{\mathbf{E}}$ is a vector of the energy difference between the non optimized force field contributions and the reference energies:

$$\mathbf{B} = \begin{bmatrix} \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_0}} & \left.\frac{\partial E_{trial}}{\partial k_1}\right|_{\mathbf{r_0}} & \cdots & \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_0}} \\ \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_1}} & \left.\frac{\partial E_{trial}}{\partial k_1}\right|_{\mathbf{r_1}} & \cdots & \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_1}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial E_{trial}}{\partial k_0}\right|_{\mathbf{r_N}} & \left.\frac{\partial E_{trial}}{\partial k_1}\right|_{\mathbf{r_N}} & \cdots & \left.\frac{\partial E_{trial}}{\partial k_m}\right|_{\mathbf{r_N}} \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} w_0(\mathbf{r}) & 0 & \cdots & 0 \\ 0 & w_1(\mathbf{r}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_N(\mathbf{r}) \end{bmatrix}$$

$$\mathbf{k} = \begin{bmatrix} k_0 \\ k_1 \\ \vdots \\ k_m \end{bmatrix} \quad \tilde{\mathbf{E}} = \begin{bmatrix} E_0 - E_{non-opt}(\mathbf{r_0}) \\ E_1 - E_{non-opt}(\mathbf{r_1}) \\ \vdots \\ E_N - E_{non-opt}(\mathbf{r_N}) \end{bmatrix} \quad (7.43)$$

Bibliography

[1] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," *Proceedings of the 1968 23rd ACM national conference*, 1968.

[2] D. H. McLain, "Drawing Contours from Arbitrary Data Points," *The Computer Journal*, vol. 17, p. 318–324, Nov 1974.

[3] P. Lancaster and K. Salkauskas, *Curve and Surface Fitting.* Academic Press, 1986.

[4] G. G. Maisuradze and D. L. Thompson, "Interpolating Moving Least-Squares Methods for Fitting Potential Energy Surfaces: Illustrative Approaches and Applications," *The Journal of Physical Chemistry A*, vol. 107, p. 7118–7124, Sep 2003.

[5] K. Hornik, M. Stinchcombe, and H. White, "Multilayer Feedforward Networks Are Universal Approximators," *Neural Networks*, vol. 2, pp. 359–366, July 1989.

[6] J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," *Physical Review Letters*, vol. 98, Apr 2007.

[7] S. Manzhos, X. Wang, R. Dawes, and T. Carrington, "A Nested Molecule-Independent Neural Network Approach for High-Quality Potential Fits †," *The Journal of Physical Chemistry A*, vol. 110, p. 5295–5304, Apr 2006.

[8] T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen," *Journal of Chemical Physics*, vol. 90, no. 2, p. 1007, 1989.

[9] W. J. Hehre, R. F. Stewart, and J. A. Pople, "Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals," *Journal of Chemical Physics*, vol. 51, no. 6, p. 2657, 1969.

[10] J. W. Ponder and F. M. Richards, "An efficient newton-like method for molecular mechanics energy minimization of large molecules," *Journal of Computational Chemistry*, vol. 8, p. 1016–1024, Oct 1987. http://dasher.wustl.edu/tinker/.

[11] S. Plimpton, "Fast Parallel Algorithms for Short-Range Molecular Dynamics," *Journal of Computational Physics*, vol. 117, p. 1–19, Mar 1995. http://lammps.sandia.gov.

[12] G. A. Landrum and W. V. Glassey. bind (ver 3.0). bind is distributed as part of the YAeHMOP extended Hückel molecular orbital package and is freely available on the WWW at; http://overlap.chem.cornell.edu:8080/yaehmop.html.

[13] Y. Shao, Z. Gan, E. Epifanovsky, A. T. B. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kús, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. DiStasio Jr., H. Dop, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar,

A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. D. Hanson-Heine, P. H. P. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, P. A. Pieniazek, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, N. Sergueev, S. M. Sharada, S. Sharmaa, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. W. Thom, T. Tsuchimochi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, V. Vanovschi, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhou, B. R. Brooks, G. K. L. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xua, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. Van Voorhis, J. M. Herbert, A. I. Krylov, P. M. W. Gill, and M. Head-Gordon, "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package," *Molecular Physics*, vol. 113, pp. 184–215, 2015.

[14] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, W. Györffy, D. Kats, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, Y. Liu, A. W. Lloyd, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, and M. Wang, "MOLPRO, version 2015.1, a package of ab initio programs," 2015. see.

[15] R. P. Muller. PyQuante, Version 1.6.3; from: http://pyquante.sourceforge.net/.

[16] Q. Sun. PySCF, Version 1.1; from: http://sunqm.net/pyscf/.

[17] X. Li, D. T. Moore, and S. S. Iyengar, "Insights from first principles molecular dynamics studies toward infrared multiple-photon and single-photon action spectroscopy: Case study of the proton-bound dimethyl ether dimer," *Journal of Chemical Physics*, vol. 128, no. 18, p. 184308, 2008.

[18] P. M. Morse, "Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels," *Physical Review*, vol. 34, pp. 57–64, jul 1929.

[19] F. Jensen, *Introduction to computational chemistry.* Chichester New York: Wiley, 1999.

[20] N. L. Allinger, "Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms," *Journal of the American Chemical Society*, vol. 99, p. 8127–8134, Dec 1977.

[21] N. L. Allinger, Y. H. Yuh, and J. H. Lii, "Molecular mechanics. The MM3 force field for hydrocarbons. 1," *Journal of the American Chemical Society*, vol. 111, p. 8551–8566, Nov 1989.

[22] J. E. Jones, "On the Determination of Molecular Fields. II. From the Equation of State of a Gas," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 106, p. 463–477, Oct 1924.

[23] R. A. Buckingham, "The Classical Equation of State of Gaseous Helium, Neon and Argon," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 168, p. 264–283, Oct 1938.

[24] T. A. Halgren, “Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94,” *Journal of Computational Chemistry*, vol. 17, p. 490–519, Apr 1996.

[25] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules,” *Journal of the American Chemical Society*, vol. 117, p. 5179–5197, May 1995.

[26] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, “CHARMM: A program for macromolecular energy, minimization, and dynamics calculations,” *Journal of Computational Chemistry*, vol. 4, no. 2, p. 187–217, 1983.

[27] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, “UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations,” *Journal of the American Chemical Society*, vol. 114, p. 10024–10035, Dec 1992.

[28] O. Akin-Ojo, Y. Song, and F. Wang, “Developing ab initio quality force fields from condensed phase quantum-mechanics/molecular-mechanics calculations through the adaptive force matching method,” *Journal of Chemical Physics*, vol. 129, no. 6, p. 064108, 2008.

[29] S. K. Burger, M. Lacasse, T. Verstraelen, J. Drewry, P. Gunning, and P. W. Ayers, “Automated Parametrization of AMBER Force Field Terms from Vibrational Analysis with a Focus on Functionalizing Dinuclear Zinc(II) Scaffolds,” *Journal of Chemical Theory and Computation*, vol. 8, p. 554–562, Feb 2012.

[30] L.-P. Wang, J. Chen, and T. Van Voorhis, “Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data,” *Journal of Chemical Theory and Computation*, vol. 9, p. 452–460, Jan 2013.

[31] S. Grimme, “A General Quantum Mechanically Derived Force Field (QMDFF) for Molecules and Condensed Phase Simulations,” *Journal of Chemical Theory and Computation*, vol. 10, p. 4497–4514, Oct 2014.

[32] L. Vanduyfhuys, S. Vandenbrande, T. Verstraelen, R. Schmid, M. Waroquier, and V. Van Speybroeck, “QuickFF: A program for a quick and easy derivation of force fields for metal-organic frameworks from ab initio input,” *Journal of Computational Chemistry*, vol. 36, p. 1015–1027, Mar 2015.

[33] R. Hoffmann, “An Extended Hückel Theory. I. Hydrocarbons,” *Journal of Chemical Physics*, vol. 39, no. 6, p. 1397, 1963.

[34] M. Wolfsberg and L. Helmholz, “The Spectra and Electronic Structure of the Tetrahedral Ions MnO4-, CrO4–, and ClO4-,” *Journal of Chemical Physics*, vol. 20, no. 5, p. 837, 1952.

[35] R. S. Mulliken, “A New Electroaffinity Scale; Together with Data on Valence States and on Valence Ionization Potentials and Electron Affinities,” *The Journal of Chemical Physics*, vol. 2, no. 11, p. 782, 1934.

[36] C. C. J. Roothaan, “New Developments in Molecular Orbital Theory,” *Reviews of Modern Physics*, vol. 23, p. 69–89, Apr 1951.

[37] R. S. Mulliken, “Quelques Aspects de la Théorie des Orbital Moléculaires,” *Journal de chimie physique*, vol. 46, pp. 497–542, 1949.

[38] G. Blyholder and C. A. Coulson, “Basis of extended Hückel formalism,” *Theoretica chimica acta*, vol. 10, no. 4, pp. 316–324, 1968.

[39] J. H. Ammeter, H. B. Buergi, J. C. Thibeault, and R. Hoffmann, "Counterintuitive orbital mixing in semiempirical and ab initio molecular orbital calculations," *Journal of the American Chemical Society*, vol. 100, p. 3686–3692, Jun 1978.

[40] L. C. Cusachs, "Semiempirical Molecular Orbitals for General Polyatomic Molecules. II. One-Electron Model Prediction of the H-O-H Angle," *The Journal of Chemical Physics*, vol. 43, no. 10, p. S157, 1965.

[41] B. L. Kalman, "Self-consistent extended Hückel theory. I," *The Journal of Chemical Physics*, vol. 59, no. 9, p. 5184, 1973.

[42] G. Calzaferri, L. Forss, and I. Kamber, "Molecular geometries by the Extended Hueckel Molecular Orbital (EHMO) method," *The Journal of Physical Chemistry*, vol. 93, pp. 5366–5371, jul 1989.

[43] A. B. Anderson, "Description of diatomic molecules using one electron configuration energies with two-body interactions," *Journal of Chemical Physics*, vol. 60, no. 11, p. 4271, 1974.

[44] F. E. Harris, "Self-Consistent Methods in Hückel Theory," *Journal of Chemical Physics*, vol. 48, no. 9, p. 4027, 1968.

[45] R. S. Mulliken, "Electronic Population Analysis on LCAO MO Molecular Wave Functions. I," *Journal of Chemical Physics*, vol. 23, no. 10, p. 1833, 1955.

[46] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform," *Journal of Cheminformatics*, vol. 4, no. 1, p. 17, 2012.

[47] F. P. Boer, M. D. Newton, and W. N. Lipscomb, "Extended Hückel Theory and molecular Hartree-Fock SCF Theory," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 52, pp. 890–893, 1964.

[48] M. D. Newton, F. P. Boer, and W. N. Lipscomb, "Molecular Orbital Theory for Large Molecules. Approximation of the SCF LCAO Hamiltonian Matrix 1," *Journal of the American Chemical Society*, vol. 88, p. 2353–2360, Jun 1966.

[49] M. J. S. Dewar and W. Thiel, "Ground states of molecules. 38. The MNDO method. Approximations and parameters," *Journal of the American Chemical Society*, vol. 99, p. 4899–4907, Jun 1977.

[50] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, "Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model," *Journal of the American Chemical Society*, vol. 107, p. 3902–3909, Jun 1985.

[51] S. L. Dixon and P. C. Jurs, "Fast geometry optimization using a modified extended Hückel method: Results for molecules containing H, C, N, O, and F," *Journal of Computational Chemistry*, vol. 15, p. 733–746, Jul 1994.

[52] R. M. Parrish, F. Liu, and T. J. Martínez, "Communication: A difference density picture for the self-consistent field ansatz," *Journal of Chemical Physics*, vol. 144, p. 131101, Apr 2016.

[53] J. Ischtwan and M. A. Collins, "Molecular potential energy surfaces by interpolation," *Journal of Chemical Physics*, vol. 100, no. 11, p. 8080, 1994.

[54] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, "ReaxFF: A Reactive Force Field for Hydrocarbons," *The Journal of Physical Chemistry A*, vol. 105, p. 9396–9409, Oct 2001.

[55] W. Domcke, "Theorie der Molekülschwingungen und der vibronischen Wechselwirkung (1999)." Retrieved from http://theochem.pctc.uni-kiel.de/pdfs/TC/SS99.pdf on 26 July 2016. [lecture notes].