Extension and Parametrization of an Approximate Density Functional Method for Organic and Biomolecules

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

(Dr. rer. nat.)

Von der

Fakultät für Chemie und Biowissenschaften

Karlsruher Institut für Technologie (KIT) – Universitätsbereich

genehmigte

DISSERTATION

von

Michael Gaus

aus

Worms

Dekan: Prof. Dr. Stefan Bräse
Referent: Prof. Dr. Marcus Elstner
Korreferent: Prof. Dr. Willem Klopper
Tag der mündlichen Prüfung: 15.07.2011

Druckjahr: 2011

MICHAEL GAUS

Extension and Parametrization of an Approximate Density Functional Method for Organic and Biomolecules

Zusammenfassung

Diese Arbeit befasst sich mit der Weiterentwicklung eines genäherten quantenchemischen Rechenverfahrens namens „density functional tight binding“(DFTB). Der Fokus liegt auf einer Erweiterung dieser Methode, die eine genauere Beschreibung großer molekularer Systeme für ein breites Spektrum chemischer Umgebungen ermöglichen soll, ohne dabei den Rechenaufwand bedeutend zu erhöhen.

Zunächst werden die Dichtefunktionaltheorie, ihre heutigen Implementierungen und deren Stärken und Schwächen in Kürze dargestellt. Es folgt eine Beschreibung der Näherungen, die zum DFTB-Formalismus führen. Daran anknüpfend wird eine Erweiterung vorgeschlagen, die konsistent mit dieser Herleitung ist. Damit wird die Änderung der Elektron-Elektron-Wechselwirkung berücksichtigt, die durch die Umverteilung der Elektronendichte bei der Bildung eines Moleküls verursacht wird. Die Einbeziehung dieses vorher noch nicht berücksichtigten Effektes und einiger weiterer Verbesserungen führt zu einer verbesserten Beschreibung von Protonenaffinitäten, Bindungsenergien von Wasserstoffbrücken und Barrieren von Protonentransferreaktionen. Außerdem wird auch die Phosphatchemie mit dieser Methode zugänglich, wie an einigen Beispielen gezeigt wird.

In vielen Anwendungen ist eine quantitative Genauigkeit von äußerster Bedeutung. Diese kann nur durch eine geeignete Parametrisierung von DFTB erreicht werden. Bislang konnte eine solche nur durch aufwendige manuelle Arbeit erzeugt werden. In der vorliegenden Arbeit wird ein Schema vorgestellt, das diesen Prozess zum größten Teil automatisiert. Verschiedene Bedingungen werden in einem linearen Gleichungssystem zusammengefasst, das dann im Sinne der Methode der kleinsten Fehlerquadrate gelöst werden kann. So ist es möglich, optimale Parameter zu finden und zugleich Zielkonflikte zu entdecken, die in den Unzulänglichkeiten der DFTB-Methode begründet sind. Mit diesem Schema wird ein Parametersatz für Kohlenwasserstoffe erzeugt, der die Genauigkeit für Bildungsenthalpien und Schwingungsfrequenzen gegenüber dem alten Parametersatz erhöht. Für spezifische Anwendungen wird ein alternativer Parametersatz vorgeschlagen. Dieser produziert etwas größere Fehler für Bildungsenthalpien, erreicht aber Genauigkeiten für Schwingungsfrequenzen, die sonst nur mit wesentlich rechenzeitaufwendigeren *ab-initio*-Methoden möglich sind.

In einem folgenden Kapitel wird eine allgemeine Parametrisierung für die Elemente Kohlenstoff, Wasserstoff, Stickstoff und Sauerstoff vorgestellt, die auf die neuen methodischen Erweiterungen abgestimmt ist. Eine nachfolgende Evaluierung an Atomisierungs-, Isomerisierungs- und Reaktionsenergien, Gleichgewichtsgeometrien und Schwingungsfrequenzen für einen großen Satz organischer Moleküle zeigt eine Verbesserung gegenüber den bereits vorhandenen Parametern. Besonders von Bedeutung bei dieser Parametrisierung ist die einfache Einbeziehbarkeit weiterer Elemente.

MICHAEL GAUS

Extension and Parametrization of an Approximate Density Functional Method for Organic and Biomolecules

Abstract

This work deals with further developments of an approximate quantum chemical method called "density functional tight binding" (DFTB). It focusses on extensions that allow a more accurate description of large molecular systems for a broader range of chemical environments while maintaining the computational speed of that method. After an outline of density functional theory including its today's implementations, their strengths, and limitations, the appoximations made to derive the DFTB formalism are reviewed.

An extension is suggested that is formally consistent within this derivation. It considers the change of electron-electron interaction due to the reorganization of the electron density when forming a molecule, an effect that had not been included before. Together with further methodological refinements this leads to an improved description of hydrogen binding energies, proton affinities, and proton transfer barriers. Moreover, phosphate chemistry becomes feasible as demonstrated on a few examples.

Quantitative accuracy is very important for many applications, but can only be achieved by proper parametrization of such an approximative method. A scheme is presented that automatizes large parts of this formerly manual process. Several constraints can be arranged in a system of linear equations which can then be solved in a least squares sense. In this manner, parameters for optimal performance are deduced, and also conflicts of objectives are detected which illustrate the limitations of the underlying formalism. Along these lines a parameter set for hydrocarbons is produced that improves the accuracy for heats of formation and vibrational frequencies. For specific applications, an alternative parameter set is presented that increases the error for heats of formation on the one hand, but yields vibrational frequencies with an accuracy of high-level calculations on the other.

A more general parametrization specifically designed for the new methodological extensions including the elements carbon, hydrogen, nitrogen and oxygen is suggested in a following chapter. An evaluation on energies of atomization, reactions, and isomerizations, equilibrium geometries, and vibrational frequencies for a large set of organic molecules shows an improvement over yet existent parameters. Most important is the simplicity of the formalism to extend the parameter set to further elements.

Contents

Chapter 1

Introduction

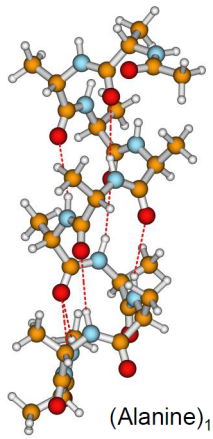
Quantum mechanics is a physical theory describing occurrences on the atomic and subatomic level. One mathematical formulation of quantum mechanics is given by the Schrödinger equation which turns out to be soluable only for very simple problems. Approximations are necessary in order to determine the chemical behavior of molecules. Two main quantum chemical approaches are established, wave function based *ab initio* methods and density functional theory (DFT).

The Hartree-Fock (HF) method is a simple wave function based approach. It considers the Pauli principle in so called exchange integrals, i.e. two electrons of same spin symmetry cannot occupy the same quantum state (orbital) simultaneously. However, HF principally allows two electrons of opposite spin to "move" independently from each other. To prevent this unphysical behavior one has to include electron correlation. Therefore, post-HF methods such as configuration interaction and perturbation theoretical methodologies are necessary (for review see e.g. ref [1]). Unfortunately, these methods require tremendously increased computer resources. While HF is often used for determining structural and vibrational properties, post-HF methods are necessary for reliably describing chemical reactions, energetics and structural properties of hydrogen bonding, and van der Waals complexes.

For these reasons, DFT has become very popular [2]. Even though an exact functional that correctly describes all properties is not known, already simple approximations contain the physics of electron exchange and correlation. This leads to results similar in quality of those from post-HF methods for many molecular properties, but in remarkably reduced computational time. Despite the success, DFT implementations reveal several problems, e.g. in describing van der Waals complexes, transition states, and charge transfer excitations. Therefore, further development of density functionals remains an important issue of current investigation.

For many applications a further reduction of computational costs is required for two reasons: First, to be able to handle large molecular systems, and second to investigate dynamic properties such as conformational changes in biological systems. Even with rapidly increasing computational speed it is not foreseeable that HF or DFT based methods will meet the demand of the scientists for "sufficiently" large systems and "sufficiently" long simulation times.

For accessing experiments that lie beyond these limitations of HF and DFT, semiempirical methods are well established. These methods treat electronic systems in a quantum mechanical sense, but explicit calculation of integrals is avoided by either neglect or parametrization. That way, computational time is reduced by a factor of about 1000 as illustrated for a few examples in Figure 1.1. For a molecular dynamics simulation, this means that a 1000 times



Timings – Single Point Energy Calculation

System	# atoms	B3LYP	SCC-DFTB
$(Alanine)_{10}$	112	110 min	4 s
$(H_2O)_{48}$	144	58 min	3 s
$(H_2O)_{123}$	369	8:30 h	15 s

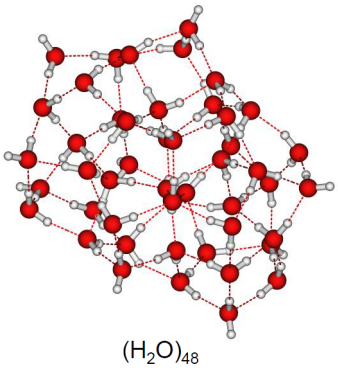


Figure 1.1: A popular DFT implementation B3LYP in comparison to the semiempirical SCC-DFTB method with respect to computational time for a single point energy calculation. The calculations were carried out on a standard desktop PC, B3LYP with the 6-31G(d) basis set utilizing the program package TURBOMOLE [3], SCC-DFTB using DFTB+ [4].

longer simulation time is possible in comparison to HF and DFT based methods. The simplifications in these schemes usually lead to a reduction of reliability and accuracy. Thus, the essence of the development of semiempirical methods is to provide schemes that maintain the accuracy of HF and DFT based methods for as many properties as possible.

Again, two main developmental directions can be distinguished. The first class consists of HF based semiempirical methods which originate from work of the Nobel laureate John Pople back in the 1950s and 60s. As in HF theory, electron correlation is not explicitly considered in the formalism. However, it is captured to some degree implicitly via parameters that are fitted to experimental data. The strengths and weaknesses of such methods are well documented (see e.g. ref [5]). Modern and popular representatives are PDDG/PM3 [6], PM6 [7], and the OMx methods [8, 9] to name only a few.

The second class of semiempirical methods is comprised of tight binding approaches which can be understood as approximate density functional theory [10]. Particularly the density functional tight binding (DFTB) method has become very successful in recent years. Almost all parameters are calculated from DFT and only few empirical ones remain. Like DFT, the conceptually very simple DFTB method contains electron correlation explicitly.

The difference between the first and second class of semiempirical methods is its origin. However, computational costs and applications are rather similar. While methods like PDDG/PM3 and PM6 have proven to give excellent results for heats of formation of small molecules [7, 11], the strengths of DFTB are applications to large biological systems, e.g. structure and relative energies of peptides are described very well [12, 13, 14]; furthermore, DFTB turns out to be very reliable and robust in molecular dynamics simulations.

The aim of this thesis is a further development of DFTB. This method can be derived by a Taylor series expansion of the DFT total energy around a given reference density. Originally, all terms higher than first order were neglected [15]. Therefore, reasonable results were only found for systems with small charge transfer between the atoms.[1] Typical applications concerned the chemistry of hydrocarbons. For describing systems with significant charge transfer the method was extended by including also the second order terms of the Taylor

[1]Later it was found that also systems with very large intramolecular charge transfer, as e.g. in KCl and NaCl molten salts, can be well described by this version of DFTB [16].

series leading to the self-consistent charge density functional tight binding method SCC-DFTB [17, 18]. Together with the so called MIO parametrization, SCC-DFTB became very popular because intramolecular charge transfer as occuring for many organic molecules could be successfully described. A special issue from the Journal of Physical Chemistry A in 2007 underlines the importance of DFTB for the scientific community [19].

To include also terms up to third order of the Taylor series, is the first major topic of the present work. This also allows to accurately treat systems with highly localized atomic net charges as is relevant for properties in biochemistry such as proton affinities and proton transfer barriers. Moreover, highly negatively charged species as often appear in phosphate chemistry become treatable. Further extensions allow to improve hydrogen binding energies, another property of ultimate importance in biology. Due to the perturbational characteristics within the Taylor series expansion, the DFTB variants were (re-)named to DFTB, DFTB2 (formerly SCC-DFTB), and DFTB3 [20].

As already mentioned, the success of DFTB2 is linked to the MIO parametrization for the elements C, H, N, and O [17, 18]. Parameters for various further elements exist, including P [20], S [21], Mg [22], and Zn [23, 24]. Nevertheless, the parametrization procedure is very cumbersome due to the lack of a clearly defined protocol and requires tedious manual work. Therefore, the second task of this thesis is to develop a new and well defined parametrization procedure including a software that automatizes technical issues as far as possible.

A third aim of present work is the derivation of new parameters for DFTB3. This lifts DFTB3 to a self-contained new generation of DFTB exhibiting characteristics similar to the step from DFTB to DFTB2 [18]:

- The new scheme is consistently and transparently derived from density functional theory along the ideas of DFTB and is therefore a general extension of DFTB.
- A new, clearly defined parametrization procedure allows an easy derivation of parameters for further elements.
- The new scheme is superior to the old scheme with respect to accuracy of the results and transferability between various chemical situations for all systems tested so far.

This thesis is organized as follows. In chapter 2, density functional theory is summarized including limitations and shortcomings of current density functionals. Chapter 3 introduces the formalism of DFTB and DFTB2. Furthermore, the MIO parameters are discussed including details which have not yet been documented. The new formalism of DFTB3 is presented in Chapter 4 including extensive benchmark tests. A detailed description of a new parametrization procedure follows in chapter 5 and is applied to the elements carbon and hydrogen at the DFTB2 level. A comprehensive evaluation of their performance is summarized. In chapter 6, parameters for carbon, hydrogen, nitrogen, and oxygen at the DFTB3 level are presented including a more general procedure for deriving new parameters. In the final part, general conclusions are drawn and an outlook of possible further developments of DFTB3 is given.

Chapter 2

Density Functional Theory

Density functional theory (DFT) is a quantum mechanical method for describing the electronic structure of a many-body system. DFT is formally exact, however, no exact functional is known and approximations are necessary. DFT implementations are widely used in chemistry and physics mainly due to its beneficial accuracy-to-computational-time ratio. The overwhelming progress in accuracy due to gradient corrected and hybrid functionals such as PBE and B3LYP curing deficiencies of older models contributed fundamentally to its popularity.

For understanding the underlying topic of this work, the DFTB method, basic knowledge of DFT is necessary and is reviewed in the first section of this chapter. An introduction to several functionals in a second section helps to classify the benchmark and reference calculations carried out in the following chapters. Because DFTB is derived from DFT it profits from the benefits but also incures the limitations of the approximate density functionals. In the last section of this chapter limitations of today's functionals and also recent extensions are discussed, latter being possible starting points for improvements of the DFTB methodology.

2.1 Basics of Density Functional Theory

Density functional theory is based on two theorems by Hohenberg and Kohn [25]. The first theorem states that the ground state properties of a many-electron system are uniquely determined by an electron density $\rho(\boldsymbol{r})$ ($\boldsymbol{r}$ being a vector specifying the spatial coordinates). The second theorem proves that the energy $E[\rho(\boldsymbol{r})]$ of this system takes its minimum value when the density is the ground-state density $\rho^{\mathrm{GS}}(\boldsymbol{r})$.

Usually, the energy of a molecular system is calculated within the Born-Oppenheimer approximation [26], i.e. due to the great difference in mass the motion of the nuclei and the motion of the electrons are examined separatly. It is common to describe the nuclei as classical particles and the electrons moving within their Coulomb potential.

The DFT total energy can be written as a functional of the electron density,

$$E^{\mathrm{DFT}}[\rho(\boldsymbol{r})] = T[\rho(\boldsymbol{r})] + E^{\mathrm{ne}}[\rho(\boldsymbol{r})] + E^{\mathrm{J}}[\rho(\boldsymbol{r})] + \tilde{E}^{\mathrm{xc}}[\rho(\boldsymbol{r})] + E^{\mathrm{nn}}, \tag{2.1}$$

where T is the kinetic energy of the electrons, E^{ne} the energy of the nucleus-electron interaction, E^{J} is the classical portion of the electron-electron energy (as defined in the following), $\tilde{E}^{\mathrm{xc}}$ the exchange-correlation energy (the non-classical portion of the electron-electron in-

teraction as will be described below), and E^{nn} the nucleus-nucleus interaction. The energy contributions are defined as

$$\begin{aligned} E^{\mathrm{ne}}[\rho(\boldsymbol{r})] &= \int V^{\mathrm{ne}}(\boldsymbol{r})\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} &&= -\sum_a \int \frac{Z_a\rho(\boldsymbol{r})}{|\boldsymbol{R}_a-\boldsymbol{r}|}\mathrm{d}\boldsymbol{r} \\ E^{\mathrm{J}}[\rho(r)] &= \frac{1}{2}\iint V^{\mathrm{J}}[\rho(\boldsymbol{r}')]\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r}'\mathrm{d}\boldsymbol{r} &&= \frac{1}{2}\iint \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}'\mathrm{d}\boldsymbol{r} \\ E^{\mathrm{nn}} &= \frac{1}{2}\sum_{a,b\neq a}\frac{Z_aZ_b}{|\boldsymbol{R}_a-\boldsymbol{R}_b|} \end{aligned} \tag{2.2}$$

where a and b are indices for atoms, Z is the nuclear charge, $\boldsymbol{R}$ the coordinates of the nucleus, and V the potential of the respective energy. The exact functionals $T[\rho(\boldsymbol{r})]$ and $\tilde{E}^{\mathrm{xc}}[\rho(\boldsymbol{r})]$ are not known such that several approximation were developed. A direct approach of Thomas and Fermi [27, 28], i.e. developing an approximate expression for these functionals, and also several modifications could not reach reasonable quantitative predictions in atomic or molecular physics [29]. However, Kohn and Sham introduced the nobel-prize winning idea of approximating the kinetic energy by introducing so called Kohn-Sham orbitals $\psi_i(\boldsymbol{r})$ (which are similarly as in HF theory expanded in a basis, see e.g. [2, 30]). The Schrödinger equation is separated for an imaginary non-interacting system of one-electron equations of the form

$$\left[-\frac{1}{2}\nabla^2 + V(\boldsymbol{r})\right]\psi_i(\boldsymbol{r}) = \epsilon_i\psi_i(\boldsymbol{r}) \qquad \forall i \tag{2.3}$$

with the requirement that

$$\rho(\boldsymbol{r}) = \sum_i n_i\psi_i^*(\boldsymbol{r})\psi_i(\boldsymbol{r}) \quad \text{and} \quad \int \rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = \sum_i n_i = N \tag{2.4}$$

where $-1/2\nabla^2$ is the kinetic energy operator of the Schrödinger equation, $V(\boldsymbol{r})$ is the one-electron potential and n_i is the electron occupation number of an orbital and N the total number of electrons in the system [31]. Eq 2.3 can be interpreted as one electron moving within an effective potential $V(\boldsymbol{r})$. That electron, however, is not interacting directly with other electrons. Therefore, the kinetic energy of such a non-interacting system $T^{\mathrm{s}}[\rho(\boldsymbol{r})]$ can be written as

$$T^{\mathrm{s}}[\rho(\boldsymbol{r})] = \sum_i n_i \int \psi_i^*(\boldsymbol{r})(-\frac{1}{2}\nabla^2)\psi_i(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = \sum_i n_i\epsilon_i - \int V(\boldsymbol{r})\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r}. \tag{2.5}$$

In order to have a formally exact theory the difference between T^{s} and the kinetic energy of a system of interacting electrons T is added to the exchange-correlation energy

$$E^{\mathrm{xc}} = \tilde{E}^{\mathrm{xc}} + (T - T^{\mathrm{s}})\,. \tag{2.6}$$

The total energy is then given by

$$E^{\mathrm{DFT}}[\rho(\boldsymbol{r})] = T^{\mathrm{s}}[\rho(\boldsymbol{r})] + E^{\mathrm{ne}}[\rho(\boldsymbol{r})] + E^{\mathrm{J}}[\rho(\boldsymbol{r})] + E^{\mathrm{xc}}[\rho(\boldsymbol{r})] + E^{\mathrm{nn}}, \tag{2.7}$$

In order to yield the ground state energy, the energy $E^{\mathrm{DFT}}[\rho]$ is minimized in a variational search in the space of ρ with the constraint that[1]

$$\int \delta\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = 0 \tag{2.8}$$

[1]The variational search for the minimum of $E^{\mathrm{DFT}}[\rho]$ can be equivalently performed in the space of orbitals $\{\psi_i\}$, now applying the constraint that $\int \psi_i^*(\boldsymbol{r})\psi_j(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = \delta_{ij}$ [29].

to find

$$\left.\frac{\delta E^{\mathrm{DFT}}[\rho]}{\delta\rho}\right|_{\rho=\rho(\boldsymbol{r})} = -V(\boldsymbol{r}) + \left[V^{\mathrm{ne}}(\boldsymbol{r}) + V^{\mathrm{J}}[\rho] + V^{\mathrm{xc}}[\rho]\right]_{\rho=\rho(\boldsymbol{r})} + \mathrm{const} = 0, \tag{2.9}$$

where

$$V^{\mathrm{xc}}[\rho(\boldsymbol{r})] = \frac{\delta E^{\mathrm{xc}}[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})}. \tag{2.10}$$

Solving for $V(\boldsymbol{r})$ yields

$$V(\boldsymbol{r}) = \left[V^{\mathrm{ne}}(\boldsymbol{r}) + V^{\mathrm{J}}[\rho] + V^{\mathrm{xc}}[\rho]\right]_{\rho=\rho(\boldsymbol{r})} + \mathrm{const}. \tag{2.11}$$

This forms a self-consistent condition which can be solved iteratively using a proper starting condition: The potential $V(\boldsymbol{r})$ can be evaluated by eq 2.11 for a given $\rho(\boldsymbol{r})$ which in turn is generated by solving eqs 2.3 and 2.4 where $V(\boldsymbol{r})$ enters.

With the idea of Kohn and Sham, the dominant part of the kinetic energy, i.e. the kinetic energy of a non-interacting system T^{s} can be calculated indirectly but exactly. Inserting eqs 2.2 and 2.5 into eq 2.7 the DFT total energy can be written as

$$\begin{aligned}
E^{\mathrm{DFT}}[\rho(\boldsymbol{r})] &= \sum_i n_i \int \psi_i^*(\boldsymbol{r}) \left(-\frac{1}{2}\nabla^2\right) \psi_i(\boldsymbol{r})\mathrm{d}\boldsymbol{r} + \int V^{\mathrm{ne}}(\boldsymbol{r})\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} \\
&\quad + \frac{1}{2}\int V^{\mathrm{J}}[\rho(\boldsymbol{r})]\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} + E^{\mathrm{xc}}[\rho(\boldsymbol{r})] + E^{\mathrm{nn}} \\
&= \sum_i n_i \int \psi_i^*(\boldsymbol{r}) \left(-\frac{1}{2}\nabla^2 + V^{\mathrm{ne}}(\boldsymbol{r}) + V^{\mathrm{J}}[\rho(\boldsymbol{r})] + V^{\mathrm{xc}}[\rho(\boldsymbol{r})]\right) \psi_i(\boldsymbol{r})\mathrm{d}\boldsymbol{r} \\
&\quad - \frac{1}{2}\int V^{\mathrm{J}}[\rho(\boldsymbol{r})]\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} - \int V^{\mathrm{xc}}[\rho(\boldsymbol{r})]\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} + E^{\mathrm{xc}}[\rho(\boldsymbol{r})] + E^{\mathrm{nn}} \\
&= \sum_i n_i\epsilon_i - \frac{1}{2}\int V^{\mathrm{J}}[\rho(\boldsymbol{r})]\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} - \int V^{\mathrm{xc}}[\rho(\boldsymbol{r})]\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} + E^{\mathrm{xc}}[\rho(\boldsymbol{r})] + E^{\mathrm{nn}}.
\end{aligned} \tag{2.12}$$

For the rearrangements eqs 2.3, 2.11 (setting the constant equal to zero), and 2.4 are used.

2.2 Exchange Correlation Functionals and Their Performance

Kohn-Sham DFT is formally exact. However, an exact expression for the exchange-correlation energy is not known. In the following some widely used approximations for the exchange-correlation energy E^{xc} are briefly summarized. Within the local density approximation (LDA) E^{xc} is calculated as

$$E^{\mathrm{xc}}[\rho(\boldsymbol{r})] = \int \rho(\boldsymbol{r})\epsilon^{\mathrm{xc}}[\rho(\boldsymbol{r})]\mathrm{d}\boldsymbol{r} \tag{2.13}$$

where ϵ^{xc} indicates the exchange-correlation energy per particle of a homogeneous electron gas of density ρ. ϵ^{xc} is usually decomposed into an exchange part ϵ^{x} and a correlation part ϵ^{c}.

The former can be derived for a homogeneous electron gas and takes a simple functional form [32] of

$$\epsilon^{\mathrm{x}}[\rho] = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\rho(\boldsymbol{r})^{1/3}. \tag{2.14}$$

For ϵ^{c} also accurate values are available. In 1980 Ceperley and Alder carried out numerical quantum Monte Carlo calculations [33] which allowed Vosko, Wilk and Nusair to interpolate between those values and provide an analytical form of ϵ^{c} building the popular VWN functional [34]. There are also more recent expressions available, e.g. [35, 36]).

LDA is only adequate for systems of slowly varying electron density and it is not realistic for e.g. molecules. Nevertheless, many molecular properties such as geometries and vibrational frequencies can be determined with great success. Drawbacks are found for the desciption of hydrogen bonds, binding energies are usually overestimated [2].

To include also variations of the density within an electronic system the generalized gradient approximation (GGA) was developed. The exchange-correlation energy density now is a functional of the density and the gradient of the density:

$$E^{\mathrm{xc}}[\rho(\boldsymbol{r})] = \int \rho(\boldsymbol{r})\epsilon^{\mathrm{xc}}[\rho(\boldsymbol{r}), \nabla\rho(\boldsymbol{r})]\mathrm{d}\boldsymbol{r} \tag{2.15}$$

The GGA improves the overall performance in comparison to LDA. A quite popular GGA functional is PBE from Perdew, Burke and Ernzerhof [37].

A further step along this approximation is meta-GGA, where the functional also includes the second derivative of the electron density and/or a dependence on the Kohn-Sham kinetic energy density (for an overview of a development along this line see e.g. ref [38]). Meta-GGA functionals usually give only a slight improvement over GGA.

A different approximation are the so called hybrid functionals. These include a portion of Hartree-Fock (exact) exchange using a few empirical parameters. The popular B3LYP functional for example includes Hartree-Fock exchange $E^{\mathrm{x,HF}}$, LDA exchange and correlation, the gradient correction (to LDA) $\Delta E^{\mathrm{x,B88}}$ from Becke [39] for exchange and the correlation functional $E^{\mathrm{c,LYP}}$ from Lee, Yang, and Parr [40] as

$$E^{\mathrm{xc,B3LYP}} = E^{\mathrm{xc,LDA}} + a(E^{\mathrm{x,HF}} - E^{\mathrm{x,LDA}}) + b\Delta E^{\mathrm{x,B88}} + c(E^{\mathrm{c,LYP}} - E^{\mathrm{c,LDA}}) \tag{2.16}$$

using the three parameters of $a = 0.2$, $b = 0.72$, and $c = 0.81$.

A very helpful classification of these functionals was given by Perdew and named "Jacob's ladder" which can be reviewed in ref [38]. Starting from the "Hartree world" there are five rungs, LDA, GGA, meta-GGA, hyper-GGA or hybrid functionals, and so called random phase approximation like functionals heading towards the heaven of chemical accuracy [38, 41]. GGA and meta-GGA are called semilocal functionals. They include energetical contributions of the electron density at point $\boldsymbol{r}$ and also mimic its surroundings via the gradients of the density. Non-local contributions, i.e. the exchange-correlation energy at point $\boldsymbol{r}$ being dependent on the density at any point $\boldsymbol{r}'$, are then given by the exact exchange as (partially) included in hybrid functionals and higher rungs of Jacob's ladder. While computational effort increases only by a small amount from LDA to GGA and meta-GGA, a considerable increase is neccessary for higher rungs. There have only been developed a few "fifth rung" functionals in the last years but these functionals are not yet widely used nor have been extensively benchmarked.

In the last years a vast amount of density functionals have appeared ranging from purely physically motivated ones to very empirical ones containing a large number of parameters

fitted to experimental or post-HF calculated data. Choosing the most adequate functional for a certain problem is not an easy task. Many benchmark tests have been carried out in literature for selected molecule sets; most popular are the G2/97 [42], G3/99 [43], and G3/05 [44] molecule sets (for a broad review see e.g. refs [2, 45], for a more recent one ref [46]). Another benchmark for several functionals, basis sets and molecular properties on large molecular test sets has recently been carried out [47]. Generally, for a medium sized basis set such as 6-31G(d), hybrid and hybrid-meta-GGA functionals yield best results for bond lengths and angles with deviations of less than 0.01 Å and 1.5° being slightly superior to pure GGA functionals (0.015 Å, 1.5°). For vibrational frequencies pure GGA functionals perform better than the ones containing exact exchange with average errors of about 50 and 80 cm^{-1}, respectively. The mean absolute deviation of ionization potentials and electron affinities are usually within 4–5 kcal/mol independent on the level of functional. For heats of formation a large variation on the performance is found depending on the choice of functional and basis set but not depending on the class of functional (hybrid, GGA, meta-GGA, hybrid-meta-GGA). For reaction barrier heights hybrid functionals show a substantial improvement over GGA and meta-GGA. In the following chapters more benchmarks will be presented.

2.3 Current Limitations

Besides the appreciably well performance of current implementations of DFT there are some general systematic limitations known [2, 29, 38, 48]. Such limiations on Kohn-Sham DFT always depend on the approximations made for the E^{xc} functional. In this section the main focus lies on limitations of LDA, GGA, and hybrid functionals for ground state properties. For functionals on a higher rung in sense of Jacobs ladder comments will be provided.

Three roots of errors are presented. First, the spurious fractional charge behavior is discussed being responsible for the underestimation of barriers of chemical reactions, the band gaps of solids, the energies of dissociating molecular ions, the charge transfer excitation energies, and the overestimation of charge transfer complexes and polarizabilities. Second, the failure to describe static correlation is pointed out leading to errors for degenerate or nearly degenerate states and the breaking of chemical bonds. Finally, the erroneous description of van der Waals interaction is briefly summarized.

Fractional Charge Behavior

Already in 1982 Perdew, Parr, Levy, and Balduz [49] found an energy dependence of fractional charge to reveal the following characteristics (as schematically shown in Figure 2.1; for details see refs [50, 51]):

- The energy difference between integer numbers of electrons (N being the number of electrons for the neutral system)

$$E_{N+1} - E_N = -A \quad \text{and} \quad E_N - E_{N-1} = -I \tag{2.17}$$

  are the electron affinity A and the ionization potential I.

- In the range between the integer numbers, the exact energy is a function of a *straight line*. It can be shown that the slope is given by the change of the energy with respect to the occupation number of the HOMO (between $N-1$ and N) and LUMO (between N

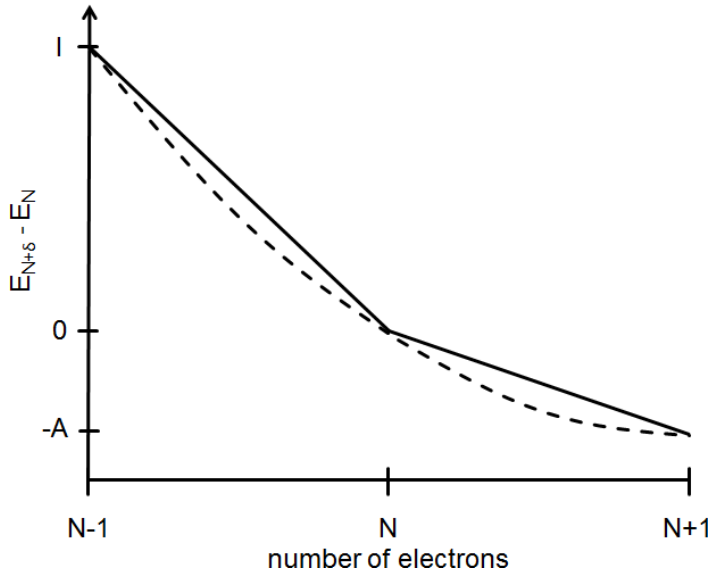


Figure 2.1: Energy in dependence of the number of electrons N. The solid line presents the exact straight line behavior, the dashed line a convex behavior as is typical for LDA and GGA functionals.

and $N+1$). Using Janaks theorem [52] that gives the constraint for an exact functional as

$$\epsilon^{\mathrm{HOMO}} = -I \quad \text{and} \quad \epsilon^{\mathrm{LUMO}} = -A. \tag{2.18}$$

- At an integer number two straight lines with different slopes ($-I$ and $-A$) meet. Thus, directly at the electron number N there is a *derivative discontinuity*.

The current GGA functionals do not show these characteristics [48, 49]. Namely, instead of the *straight line* behavior mostly a convex behavior is found (see Figure 2.1). As consequence reaction barriers are underestimated if fractional occupation numbers appear at the transition states. To illustrate this Yang et al. [48] considered the dissociation of the H_2^+ molecule. Due to the convex behavior partial charges on one H atom are energetically lowered more than for a straight line behavior. Thus, both hydrogen atoms with half an electron are more stable than a single localized electron even in the case of dissociation. The charge is delocalized and within the DFT-GGA description H_2^+ would not dissociate to H and H^+. Yang and coworkers called this finding *delocalization error* and also showed that it is connected to the so called *self-interaction error* [50].

In a system with only one electron there is no electron-electron interaction. Thus, in equation 2.7 the classical electron-electron interaction

$$E^{\mathrm{J}} = \frac{1}{2} \iint \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} \mathrm{d}\boldsymbol{r}' \mathrm{d}\boldsymbol{r} \tag{2.19}$$

must be compensated by the exchange-correlation energy

$$E^{\mathrm{J}} + E^{\mathrm{xc}} = 0. \tag{2.20}$$

However, the current exchange-correlation functionals do no fulfill this constraint. Literally speaking, spuriously one electron is interacting with itself due to the *self-interaction error*. This error of course not only occurs for one electron systems but for any system. Facing this error Perdew and Zunger have suggested to substract these self-interaction orbital by orbital

[53]. The Kohn-Sham equations become orbital dependent which complicates to find their solution. While this scheme amends the description of charge transfer and transition states of chemical reactions a more sophisticated scheme is necessary in order not to deteriorate other properties such as thermochemistry and equilibrium bond lengths [54, 55, 56]. Also other implementations of self-interaction correction have successfully shown improved orbital description and charge transfer (e.g. ref [57]). Alternatively, one can construct an exchange-correlation functional that itself is self-interaction free, as e.g. suggested in ref [58]. It is worth mentioning at that point that the LYP correlation functional [40] is self-interaction free.

Another issue of eq 2.20 is that E^{xc} should compensate E^{J} for all electron-electron distances $|\boldsymbol{r}-\boldsymbol{r}'|$. Inspection of E^{J} (eq 2.2) shows an $1/|\boldsymbol{r}-\boldsymbol{r}'|$ dependence at large interelectronic distances. Thus, E^{xc} should show a $-1/|\boldsymbol{r}-\boldsymbol{r}'|$ *asymptotic behavior*. This however, cannot be found when considering the LDA functional eq 2.13 (or GGA functionals). Since for two fragments being far apart the electron density of each fragment decays exponentially, also E^{xc} decays exponentially and likewise V^{xc} (eq 2.10). This severely affects the virtual Kohn-Sham orbitals leading to errors in electron affinities, response to electromagnetic fields, i.e. polarizabilities, and excitation energies, in particular Rydberg states [2]. Hybrid functionals amend these errors to some degree due to the inclusion of a certain amount of exact exchange which shows the correct asymptotic behavior. Lately several functionals have been developed meeting the correct asymptotic behavior (for a review see e.g. [2]). Another concept is to introduce a range separation of E^{xc}, for small interelectronic distances using GGA and for large $|\boldsymbol{r}-\boldsymbol{r}'|$ applying Hartree-Fock exchange. For the tests done so far these corrections seem to ameliorate at least some of the problems (see e.g. [2] and references therein).

Another problem that can be explained by the wrong behavior of the DFT functionals in Figure 2.1 is the wrong description of the *band gap* in solids (number of electrons $\rightarrow \infty$) which is given by

$$E^{\mathrm{gap}} = I - A. \tag{2.21}$$

While for small molecules I and A can well be calculated using equation 2.17, large errors occur when computing I and A from equation 2.18 due to the fact that the LDA and GGA functionals do not show the straight line behavior. Because a convex behavior is found, band gaps in solids are dramatically underestimated [51].

Static Correlation

The bond length and energy of the H_2 molecule at equilibrium can be well described by GGA functionals. However, when stretching the bond towards dissociation the energy is severly overestimated. This well known problem is attributed to static correlation. In analogy to the same problem in Hartree-Fock theory the question arises whether one Slater determinant as is used in standard DFT implementations is enough for describing degenerate or nearly degenerate states [59]. For H_2 it has been shown that indeed a potential can be constructed using one Slater determinant that describes the equilibrium structure as well as the dissociation behavior. For other examples, however, a linear combination of a few determinants is necessary [60].

In the dissociation limit of H_2 each H atom forms an exotic system of half a spin-up and half a spin-down electron. Yang and coworkers have analyzed that problem and formulated the *constancy condition*, saying that the exact functional requires systems with fractional spin to have an energy equal to that of normal spin states [48]. In addition, they found that LDA and GGA functionals do not fulfill this condition and overestimate the energy

for the correlation of fractional spin in comparison to integer spin states. Moreover, this energy exactly matches the error of the dissociated H_2 molecules. Therefore, correlation functionals fulfilling the constancy condition are matter of current research. Along this line Yang and coworkers also found that this condition can be connected to the above mentioned requirement for a linear fractional charge behavior, the exact energy functional turning out to be a plane [61].

In current practice, a correct energetic behavior is enforced within LDA and GGA by introducing spin densities,

$$\rho = \rho^{\uparrow} + \rho^{\downarrow}. \tag{2.22}$$

At equilibrium of the H_2 example spin-unpolarized densities are produced, however, for a certain bond length symmetry breaks by gradually localizing the spin-up electron density $\rho^{\uparrow}$ at one nucleus and the spin-down electron density $\rho^{\downarrow}$ on the other. While the dissociation curve is then correctly described the spin-density does not yield the correct spatial $D_{\infty h}$ symmetry.

Van der Waals Interaction

Long range interactions necessarily fail using GGA functionals due to its semilocal nature. This failure occurs for van der Waals forces of intermolecular systems, e.g. π-stacking of benzene or DNA base pairs. It can be shown with perturbation theory that there should be (in lowest order) an $-1/r^6$ attraction within the electron correlation energy contribution, where r is the interatomic distance. This cannot be reproduced within the exchange-correlation energy functionals which decay exponentially as discussed above. A correction is only given within the fifth rung functionals of Jacobs ladder [62, 63]. A rather practical approach was carried out in the last years, where an empirical $-1/r^6$ term was added to the total energy specifically for the long range van der Waals interactions [64, 65].

Chapter 3

Density Functional Tight Binding

In this chapter the density functional tight binding method (DFTB) is derived from the DFT total energy as introduced in the last chapter. The starting point is the use of a reference density ρ^0, which is calculated from a superposition of precalculated neutral atomic densities and deviates from the DFT total energy as introduced by $\Delta\rho$. After expanding the exchange-correlation energy in a Taylor series and some rearrangements the DFT total energy can be rewritten in a different form. In that form one can approximate the DFT total energy just by the assumption of a reference density (DFTB) and also by an approximation of the density fluctuation (DFTB2).

DFTB describes only valence electrons in minimal atomic basis sets explicitly; chemical cores are treated in an effective manner via additive two-center potentials. To further reduce computational effort crystal field and three-center integrals are neglected. The remaining two-center Hamilton and overlap matrix elements are precalculated for a dense mesh of interatomic distances in an atomic orbital (AO) basis and oriented during a calculation using the Slater-Koster sin/cos-rules [66], i.e. no integral evaluation is necessary during runtime of the calculation. The remaining contributions to the total energy are then approximated and thus, no further computational cost arises beyond the dominant step, which is the diagonalization of the Hamilton matrix. This and the use of the minimal valence basis set leads to huge computational savings (2–3 orders of magnitude) compared to full DFT.

Chemists require an accuracy within a few kcal/mol. Therefore, the parametrization of DFTB is of paramount interest leading from a qualitative to a quantitative methodology. The successful MIO parametrization is described in a third section followed by a brief overview of the performance and limitations of the resulting DFTB2 method.

3.1 DFTB

The derivation of DFTB starts by using the following definition of the electron density:

$$\rho(\boldsymbol{r}) = \sum_i n_i \psi_i^*(\boldsymbol{r})\psi_i(\boldsymbol{r}) = \rho^0(\boldsymbol{r}) + \Delta\rho(\boldsymbol{r}). \tag{3.1}$$

The DFT total energy (eq 2.12) can then be written as[1]

$$\begin{aligned} E[\rho^0+\Delta\rho] \;=\; & \sum_i n_i \int \psi_i^* \left(-\frac{\nabla^2}{2} + V^{\mathrm{ne}} + \int' \frac{\rho^{0\prime}+\Delta\rho'}{|\boldsymbol{r}-\boldsymbol{r}'|} + V^{\mathrm{xc}}[\rho^0+\Delta\rho] \right) \psi_i \\ & - \frac{1}{2} \iint' \frac{(\rho^{0\prime}+\Delta\rho)(\rho^0+\Delta\rho)}{|\boldsymbol{r}-\boldsymbol{r}'|} - \int V^{\mathrm{xc}}[\rho^0+\Delta\rho](\rho^0+\Delta\rho) \\ & + E^{\mathrm{xc}}[\rho^0+\Delta\rho] + E^{\mathrm{nn}}. \end{aligned} \tag{3.2}$$

In the next step the first term of eq 3.2 is modified such that it is only dependent on the reference density, i.e. the $\Delta\rho$ term of the classical electron-electron interaction (V^{J}) is extracted from this integral and written in a separate term using eq 3.1. Further the exchange-correlation term is removed from the first term and inserted separately; $V^{\mathrm{xc}}[\rho^0]$ (an exchange-correlation contribution dependent only on the reference density) is added into the first term and again subtracted in an additional term, respectively.

$$\begin{aligned} E[\rho^0+\Delta\rho] \;=\; & \sum_i n_i \int \psi_i^* \left(-\frac{\nabla^2}{2} + V^{\mathrm{ne}} + \int' \frac{\rho^{0\prime}}{|\boldsymbol{r}-\boldsymbol{r}'|} + V^{\mathrm{xc}}[\rho^0] \right) \psi_i \\ & + \iint' \frac{\Delta\rho'(\rho^0+\Delta\rho)}{|\boldsymbol{r}-\boldsymbol{r}'|} - \int V^{\mathrm{xc}}[\rho^0](\rho^0+\Delta\rho) + \int V^{\mathrm{xc}}[\rho^0+\Delta\rho](\rho^0+\Delta\rho) \\ & - \frac{1}{2} \iint' \frac{\rho^{0\prime}\rho^0}{|\boldsymbol{r}-\boldsymbol{r}'|} - \frac{1}{2} \iint' \frac{\rho^{0\prime}\Delta\rho}{|\boldsymbol{r}-\boldsymbol{r}'|} - \frac{1}{2} \iint' \frac{\Delta\rho'\rho^0}{|\boldsymbol{r}-\boldsymbol{r}'|} - \frac{1}{2} \iint' \frac{\Delta\rho^{0\prime}\Delta\rho^0}{|\boldsymbol{r}-\boldsymbol{r}'|} \\ & - \int V^{\mathrm{xc}}[\rho^0+\Delta\rho](\rho^0+\Delta\rho) + E^{\mathrm{xc}}[\rho^0+\Delta\rho] + E^{\mathrm{nn}} \\ = \; & \sum_i n_i \int \psi_i^* \left(-\frac{\nabla^2}{2} + V^{\mathrm{ne}} + \int' \frac{\rho^{0\prime}}{|\boldsymbol{r}-\boldsymbol{r}'|} + V^{\mathrm{xc}}[\rho^0] \right) \psi_i \\ & - \frac{1}{2} \iint' \frac{\rho^{0\prime}\rho^0}{|\boldsymbol{r}-\boldsymbol{r}'|} + \frac{1}{2} \iint' \frac{\Delta\rho'\Delta\rho}{|\boldsymbol{r}-\boldsymbol{r}'|} \\ & - \int V^{\mathrm{xc}}[\rho^0]\rho^0 - \int V^{\mathrm{xc}}[\rho^0]\Delta\rho + E^{\mathrm{xc}}[\rho^0+\Delta\rho] + E^{\mathrm{nn}} \end{aligned} \tag{3.3}$$

With this rearrangement it is not clear anymore how to obtain the exact Kohn-Sham wave functions ψ_i. However, within the DFTB formalism one can obtain approximate ψ_i as will be shown in the following subsections. In a last step the exchange-correlation energy is expanded in a Taylor series expansion as

$$\begin{aligned} E^{\mathrm{xc}}[\rho^0+\Delta\rho] \;=\; & E^{\mathrm{xc}}[\rho^0] + \int \left[\frac{\delta E^{\mathrm{xc}}[\rho]}{\delta\rho} \right]_{\rho^0} \Delta\rho + \frac{1}{2} \int' \int \left[\frac{\delta^2 E[\rho]}{\delta\rho\delta\rho'} \right]_{\rho^0,\rho^{0\prime}} \Delta\rho\Delta\rho' \\ & + \frac{1}{6} \int'' \int' \int \left[\frac{\delta^3 E^{\mathrm{xc}}[\rho]}{\delta\rho\delta\rho'\delta\rho''} \right]_{\rho^0,\rho^{0\prime},\rho^{0\prime\prime}} \Delta\rho\Delta\rho'\Delta\rho'' + \ldots \end{aligned} \tag{3.4}$$

[1] For an easier readability in some cases the abbreviations $\int = \int \mathrm{d}\boldsymbol{r}$, $\int' = \int \mathrm{d}\boldsymbol{r}'$, $\int'' = \int \mathrm{d}\boldsymbol{r}''$, $\rho = \rho(\boldsymbol{r})$, $\rho' = \rho(\boldsymbol{r}')$, $\rho'' = \rho(\boldsymbol{r}'')$, and $\psi_i = \psi_i(\boldsymbol{r})$ are used.

Applying eq 2.10 the DFT total energy (eq 2.12) can be written as

$$\begin{aligned} E[\rho^0+\Delta\rho] \;=\; & \sum_i n_i \int \psi_i^* \left(-\frac{\nabla^2}{2} + V^{\mathrm{ne}} + \int' \frac{\rho^{0\prime}}{|\boldsymbol{r}-\boldsymbol{r}'|} + V^{\mathrm{xc}}[\rho^0] \right) \psi_i \\ & - \frac{1}{2} \int' \int \frac{\rho^0 \rho^{0\prime}}{|\boldsymbol{r}-\boldsymbol{r}'|} - \int V^{\mathrm{xc}}[\rho^0]\rho^0 + E^{\mathrm{xc}}[\rho^0] + E^{\mathrm{nn}} \\ & + \frac{1}{2} \int' \int \left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} + \left. \frac{\delta^2 E^{\mathrm{xc}}[\rho]}{\delta\rho\delta\rho'} \right|_{\rho^0,\rho^{0\prime}} \right) \Delta\rho\Delta\rho' \\ & + \frac{1}{6} \int'' \int' \int \left. \frac{\delta^3 E^{\mathrm{xc}}[\rho]}{\delta\rho\delta\rho'\delta\rho''} \right|_{\rho^0,\rho^{0\prime},\rho^{0\prime\prime}} \Delta\rho\Delta\rho'\Delta\rho'' + \ldots \end{aligned} \tag{3.5}$$

Approximations of different levels of sophistication can be introduced by truncation of the Talyor series [67].[2] Standard (nonself-consistent) DFTB [15, 68] neglects second and higher order terms. This leads to a non self-consistent scheme, i.e., the generalized eigenvalue problem has to be diagonalized only once. The DFTB2 method approximates the second order terms in the density fluctuations [17]. Both of these levels are summarized in the following subsections. One objective of the present work is to include also third order terms. This topic is treated in chapter 4.

Total Energy

As discussed by Foulkes and Haydock [10], standard tight-binding (TB) methods like non self-consistent DFTB assume the second and higher order terms in the density fluctuations $\Delta\rho$ to be negligible, i.e., only the first two lines of eq 3.5 contribute to the total energy.

After introducing a minimal basis set of atomic valence orbitals ϕ_μ within the linear combination of atomic orbitals (LCAO) ansatz

$$\psi_i(\boldsymbol{r}) = \sum_a \sum_{\mu\in a} c_{\mu i}\phi_\mu(\boldsymbol{r}-\boldsymbol{R}_a), \tag{3.6}$$

where ϕ_μ is a basis function of orbital μ centered at atom a, $\boldsymbol{R}_a$ is the coordinate of the nucleus of atom a, and the $c_{\mu i}$ are the basis set coefficients[3], the first line of eq 3.5 can be written as

$$E^{\mathrm{H0}} = \sum_i n_i \sum_{ab} \sum_{\mu\in a} \sum_{\nu\in b} c_{\mu i} c_{\nu i} H^0_{\mu\nu}, \tag{3.7}$$

with Hamilton matrix elements $H^0_{\mu\nu} = \int \phi^*_\mu \hat{H}^0 \phi_\nu \mathrm{d}\boldsymbol{r}$ being only dependent on the reference density. a and b are indices for atoms, μ and ν for atomic orbitals. The determination of the coefficients $c_{\mu i}$, the functional form of the basis functions ϕ_μ and the reference density are described in detail below.

The terms in the second line of eq 3.5 consist of the DFT double-counting contributions, the nucleus-nucleus repulsion and exchange-correlation contributions. In tight-binding the-

[2] Very often it is stated that the DFTB total energy is based on a Taylor series expansion of the DFT total energy. This is not a very precise statement, however, one can derive eq 3.5 by expressing the kinetic energy as eq 2.5 and expanding the remaining part of the energy in a Taylor series. This again, gives no solution of how to obtain the exact Kohn-Sham wave functions ψ_i, however they can be approximated within the DFTB formalism.

[3] Real numbers are used for the basis set coefficients $c_{\mu i}$.

ory, they are usually approximated as a sum of one-center terms and short ranged two-center potentials V_{ab}^{rep}[10]

$$-\frac{1}{2}\int'\int\frac{\rho^0\rho'^0}{|\boldsymbol{r}-\boldsymbol{r}'|}+E^{\mathrm{xc}}[\rho^0]-\int V^{\mathrm{xc}}[\rho^0]\rho^0+E^{\mathrm{nn}}\;\approx\;\sum_a V_a^{\mathrm{rep}}[\rho_a^0]+\frac{1}{2}\sum_{ab}V_{ab}^{\mathrm{rep}}[\rho_a^0,\rho_b^0,r_{ab}]. \quad (3.8)$$

The indices a and b denote the atoms, ρ_a^0 and ρ_b^0 atomic reference densities, and $r_{ab} = |\boldsymbol{R}_b - \boldsymbol{R}_a|$ is the interatomic distance. The atomic contributions sum up to a constant energy shift which cancel out when considering energy differences. For DFTB the atomic contributions are neglected and a repulsive energy E^{rep} is defined as

$$E^{\mathrm{rep}} \;=\; \frac{1}{2}\sum_{ab}V_{ab}^{\mathrm{rep}}[\rho_a^0,\rho_b^0,r_{ab}]. \quad (3.9)$$

The pair-potentials V_{ab}^{rep} are fitted atom type and distance dependent to several reference systems. Details will be described in chapter 5.

Finally, the total DFTB energy can be written as

$$E^{\mathrm{DFTB}} \;=\; \sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b}n_i c_{\mu i}c_{\nu i}H^0_{\mu\nu}+E^{\mathrm{rep}}. \quad (3.10)$$

Determination of the Coefficients $c_{\mu i}$

To determine the molecular orbital coefficients $c_{\mu i}$ (eq 3.6) Kohn-Sham equations are derived by taking the derivative of the total energy with respect to the normalization constraints

$$\int\psi_i^*(\boldsymbol{r})\psi_i(\boldsymbol{r})\mathrm{d}\boldsymbol{r}=1 \qquad \forall i \quad (3.11)$$

to get (δ is an index for an orbital, d for an atom):

$$\frac{\partial}{\partial c_{\delta i}}\left[E^{\mathrm{DFTB}}-\sum_j n_j\epsilon_j\left(\sum_{ab}\sum_{\mu\in a}\sum_{\nu\in b}c_{\mu j}c_{\nu j}S_{\mu\nu}-1\right)\right] \;=\; 0, \qquad \forall d,\delta\in d. \quad (3.12)$$

The Kohn-Sham equations are approximate in the sense that an approximated energy E^{DFTB} being dependent on a reference density is considered. Inserting eq 3.10 yields a set of algebraic equations

$$\sum_b\sum_{\nu\in b}c_{\nu i}\left(H^0_{\mu\nu}-\epsilon_i S_{\mu\nu}\right) \;=\; 0, \qquad \forall a,\mu\in a,i, \quad (3.13)$$

where $S_{\mu\nu}=\int\phi_\mu^*(\boldsymbol{r})\phi_\nu(\boldsymbol{r})\mathrm{d}\boldsymbol{r}$ is the overlap matrix.

The total energy in eq 3.10 can also be expressed in terms of the ϵ_i as

$$E^{\mathrm{DFTB}} \;=\; \sum_i n_i\epsilon_i+E^{\mathrm{rep}}. \quad (3.14)$$

Atomic Orbitals ϕ_μ

The atom-centered pseudoatomic wave functions are written in terms of Slater type orbitals and spherical harmonics as

$$\phi_\mu(\boldsymbol{r})=\sum_{n,\alpha,l_\mu,m_\mu}a_{n\alpha}r^{l_\mu+n}\exp(-\alpha r)Y_{l_\mu m_\mu}\left(\frac{\boldsymbol{r}}{r}\right), \quad (3.15)$$

where $r = |\boldsymbol{r}|$, l and m are the azimuthal and magnetic quantum number; five different values for α and $n = 0, 1, 2, 3$ form a converged basis set [15] for elements up to third period of the system of elements. The atomic coefficients $a_{n\alpha}$ are obtained by solving modified self-consistent atomic Kohn-Sham equations:

$$\left[-\frac{1}{2}\nabla^2 + V^{\text{psat}}(\boldsymbol{r})\right]\phi_\mu(\boldsymbol{r}) = \epsilon_\mu^{\text{psat}}\phi_\mu(\boldsymbol{r}) \qquad \forall\mu, \tag{3.16}$$

where the pseudoatomic potential $V^{\text{psat}}(\boldsymbol{r})$ is given by

$$V^{\text{psat}}(\boldsymbol{r}) = V^{\text{ne}}(\boldsymbol{r}) + V^{\text{J}}[\rho(\boldsymbol{r})] + V^{\text{xc}}[\rho(\boldsymbol{r})] + \left(\frac{r}{r^{\text{c}}}\right)^p \tag{3.17}$$

As exchange-correlation potential the local density approximation as parametrized by Perdew and Zunger [53] or the PBE functional [37] is often used. The additional term $(r/r^{\text{c}})^p$ causes a compression of the atomic potential improving the description of interatomic interaction, i.e. the uncompressed atomic potentials would be too diffuse [69, 70]. The so called compression radius r^{c} may be optimized to yield best result. A good starting point for optimization was found to be twice the covalent radius of that element [15]. The parameter p was found to have a rather small influence on the result and is usually chosen as $p = 2$.

Calculation of Matrix Elements $H^0_{\mu\nu}$

During the development of DFTB two ways of calculating the Hamilton matrix elements have been established. The older way is the determination using an overlap of potentials in the form (in Dirac notation)

$$H^0_{\mu\nu} = \begin{cases} \epsilon^{\text{free atom}} & \text{if } \mu = \nu \\ \left\langle \phi_\mu \middle| \hat{T} + V[\rho^0_a] + V[\rho^0_b] \middle| \phi_\nu \right\rangle & \text{if } a \neq b \\ 0 & \text{if } a = b, \mu \neq \nu \end{cases} \tag{3.18}$$

where $\hat{T} = -\nabla^2/2$ is the kinetic energy operator and $\epsilon^{\text{free atom}}$ is the eigenvalue of eq 3.16 when using the potential $V(r)$ of eq 3.17 without the additional term $(r/r^{\text{c}})^p$ [15]. For the off-diagonal terms a two-center approximation is applied, crystal field and three-center terms are neglected. Details for the approximations involved can be found in [71]. $V[\rho^0_a]$ is an atomic Kohn-Sham potential as given in eq 3.17 including the compressed density

$$\rho^0_a = \sum_{\mu\in a} \phi_\mu(\boldsymbol{r})\phi_\mu(\boldsymbol{r}). \tag{3.19}$$

but without the additional term $(r/r^{\text{c}})^p$.

The second way to approximate the Hamilton matrix elements is similar to the first one but using density overlap of the form

$$H^0_{\mu\nu} = \begin{cases} \epsilon^{\text{free atom}} & \text{if } \mu = \nu \\ \left\langle \phi_\mu \middle| \hat{T} + V[\rho^0_a + \rho^0_b] \middle| \phi_\nu \right\rangle & \text{if } a \neq b \\ 0 & \text{if } a = b, \mu \neq \nu. \end{cases} \tag{3.20}$$

The potential $V[\rho_a^0+\rho_b^0]$ is identical to eq 3.17 but again without the additional term $(r/r^{\mathrm{c}})^p$. However, it has been shown for DFTB2 (which will be introduced in the following section) on small molecules, that for accurate and transferable two-center integrals ($a \neq b$) one parameter for the compression of ϕ_μ and ρ_a^0 is not sufficient [18]. Therefore, the ϕ_μ are obtained by eq 3.16 using a wave function compression $r^{\mathrm{c}} = r^{\mathrm{wf}}$ and for determining the atomic density ρ_a^0 (eq 3.19) eq 3.16 is solved again for a density compression $r^{\mathrm{c}} = r^{\mathrm{dens}}$.

The parameter r^{wf} can be considered as the choice of the basis set and a correlation to the Pauli-repulsion seems obvious. For strongly compressed wave functions the Pauli-repulsion sets in for smaller distances in comparison to weakly compressed wave functions. As stated before a good choice of r^{wf} seems to be twice the covalent radius of that element. The parameter r^{dens} on the other hand is rather empirical. Usually, for reasonable results it is found that r^{dens} needs to be larger than r^{wf}. Both parameters are optimized to yield best results as described in detail in sections 3.3.1, 5.2.1, and 6.1.

The overlap matrix elements $S_{\mu\nu}$ mentioned earlier and the Hamilton matrix elements $H^0_{\mu\nu}$ can be precomputed and tabulated (the two-center approximation simplifies this tabulation!) for every atom type pair and a dense mesh of interatomic distances. Thus, it is not necessary to compute the matrix elements during the runtime of the calculation, but it is essential to account for the orientation of the orbitals within a molecule which is done using the Slater-Koster sin/cos-rules [66].

While for applications on physics of solids (e.g. band structure calculations) the DFTB Hamiltonian is often used calculated by potential overlap, for applications on molecules (and biological systems) the density overlap is the standard (usually also in connection with the self-consistent charge extension DFTB2). A single parametrization yielding good results for both fields seems to be not possible at the present status of DFTB. In present work the aim is to extend and reparametrize DFTB for applications on molecular and biological systems. Thus, in the following chapters the Hamilton matrix elements are assumed to be calculated using the density overlap. As long as not stated differently it is always referred to the MIO parametrization [18] as is described in detail in section 3.3.

3.2 DFTB2

In this section the DFTB2 formalism is described, which was formerly called SCC-DFTB. For systems with sizable charge reorganization, the second order terms in the density fluctuation (eq 3.5) become important [17, 67, 72]

$$E^{\mathrm{2nd}} = \frac{1}{2}\int'\int\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} + \left.\frac{\delta^2 E^{\mathrm{xc}}}{\delta\rho\delta\rho'}\right|_{\rho^0,\rho^{0\prime}}\right)\Delta\rho\Delta\rho'. \tag{3.21}$$

To preserve computational efficiency, this term is approximated in a way that avoids explicit integration during the calculation. First, the density fluctuation is written as a superposition of atomic contributions

$$\Delta\rho = \sum_a \Delta\rho_a. \tag{3.22}$$

The main approximation for this term consists of neglecting higher order multipole interactions, i.e. approximating the atomic density fluctuations $\Delta\rho_a$ by charge monopoles, truncating the expansion in spherical harmonics after the monopole term[4]

[4] A detailed discussion about these approximations can be found in ref [18] p. 33.

$$\Delta\rho_a \approx \Delta q_a F_a^{00} Y^{00}. \tag{3.23}$$

In analogy to $\Delta\rho_a = \rho_a - \rho_a^0$, the charge fluctuation is the Mulliken charge $\Delta q_a = q_a - q_a^0$, where q_a^0 is the charge of the valence electrons of a neutral atom a, and q_a is the charge of that atom within the molecule which is calculated as

$$q_a = \sum_i n_i \sum_{\mu\in a} \sum_b \sum_{\nu\in b} c_{\mu i} c_{\nu i} S_{\mu\nu}. \tag{3.24}$$

Therefore, the change of density in eq 3.23 is restricted to charge transfer between the atoms as estimated by the net charge Δq_a on atom a, and the deformation of the charge density is neglected in this approximation ($Y^{00} = \frac{1}{2}\sqrt{\frac{1}{\pi}}$). Inserting eq 3.23 and eq 3.22 into eq 3.21 and assuming exponentially decaying spherical charge densities with coefficients τ_a,

$$\frac{F_a^{00}}{2\sqrt{\pi}} = \frac{\tau_a^3}{8\pi}\exp(-\tau_a|\boldsymbol{r} - \boldsymbol{R}_a|) \tag{3.25}$$

an analytic function γ can be derived [17, 73], which represents the interaction between the charge density fluctuations and further approximates the second order terms (eq 3.21). The functional form of γ is given in appendix B. The second order term becomes

$$E^{\text{2nd}} \approx E^{\gamma} = \frac{1}{2}\sum_{ab} \Delta q_a \Delta q_b \gamma_{ab}. \tag{3.26}$$

Two main properties of γ_{ab} are highlighted which are described in detail in ref [17]. For large interatomic distances r_{ab}, γ_{ab} basically reduces to $1/r_{ab}$; i.e., it describes a pure Coulomb interaction of the partial charges Δq_a and Δq_b. For $a = b$, γ_{ab} describes the on-site self-repulsion,

$$\gamma_{aa} = U_a \tag{3.27}$$

introducing the Hubbard parameter U_a which is the second derivative of the total energy of a neutral atom with respect to the occupation number of the highest occupied atomic orbital. In DFTB it is estimated using Janak's theorem [52] by numerically calculating the first derivative of the energy of the highest occupied atomic orbital with respect to its occupation number for a neutral atom. The Hubbard parameter can also be related to the chemical hardness which is half of the difference of ionization potential and electron affinity (ref [18] p. 39ff).

The exponential coefficient τ_a is determined by [73]

$$\tau_a = \frac{16}{5}U_a, \tag{3.28}$$

which imposes an inverse relationship between the Hubbard parameter and the covalent radius [17][5]

$$U_a \propto \frac{1}{R_a^{\text{cov}}}. \tag{3.29}$$

Therefore, the Hubbard parameter affects two physical properties, the electron-electron interaction within one atom, i.e., the diagonal elements γ_{aa}, and the size of the atoms for estimating the two-center terms γ_{ab}. This estimated atomic size determines the deviation of γ_{ab} from $1/r_{ab}$, as shown in eq Figure 3.1.

[5] One can easily recognize this inverse relation by determining the average distance of electron to nucleus, i.e. integrating $R_a^{\text{avg}} = \int F_a^{00} Y^{00} r d\boldsymbol{r} = 3/\tau_a = 15/(16U_a)$ over all space.

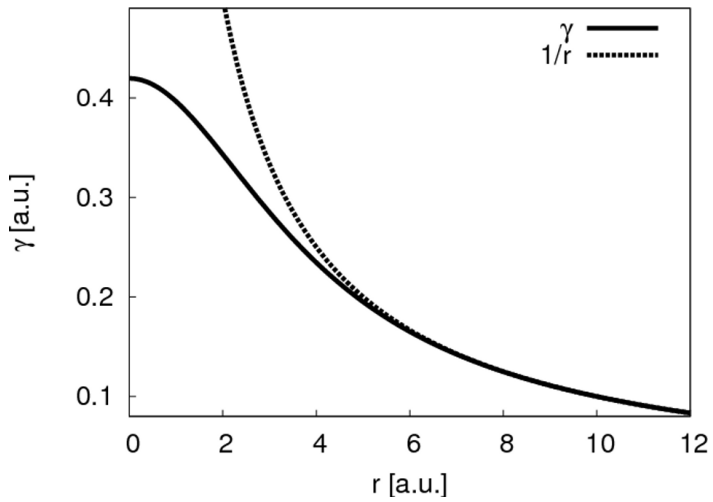


Figure 3.1: The γ-function (solid line) plotted for the hydrogen-hydrogen interaction deviates from $1/r$ (dashed line) at short distances and yields the value of the Hubbard parameter $U_{\mathrm{H}} = 0.4195$ a.u. at $r = 0$ a.u..

The coefficients $c_{\mu i}$ can be obtained by solving the Kohn-Sham equations in a similar way as in eqs 3.12 and 3.13,

$$\sum_b \sum_{\nu \in b} c_{\nu i} \left(H_{\mu\nu} - \epsilon_i S_{\mu\nu} \right) \; = \; 0, \qquad \forall a, \mu \in a, i, \tag{3.30}$$

where the Hamilton matrix elements are

$$H_{\mu\nu} = H^0_{\mu\nu} + \frac{1}{2} S_{\mu\nu} \sum_c \Delta q_c \left(\gamma_{ac} + \gamma_{bc} \right). \tag{3.31}$$

The DFTB2 total energy finally reads

$$\begin{aligned} E^{\mathrm{DFTB2}} &= \sum_{iab} \sum_{\mu \in a} \sum_{\nu \in b} n_i c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{ab} \Delta q_a \Delta q_b \gamma_{ab} + \frac{1}{2} \sum_{ab} V^{\mathrm{rep}}_{ab} \\ &= \sum_i n_i \epsilon_i - \frac{1}{2} \sum_{ab} \Delta q_b (q_a \gamma_{ba} + q^0_a \gamma_{ab}) + E^{\mathrm{rep}}, \end{aligned} \tag{3.32}$$

the recasting in the second line is derived in detail in appendix B.

3.3 The MIO Parametrization

In the last section the DFTB2 formalism was explained and also several parameters were introduced. The idea of DFTB2 lies in a robust scheme being approximate to DFT that is transferable to a wide range of chemical environments. All parameters in DFTB2 are well-defined and an estimate of these parameters is given by physical arguments. In that way DFTB2 is helpful for qualitative answers. However, for many applications the so called "chemical accuracy" is desired, meaning e.g. energetics deviating from experiment only by a few kcal/mol. For this accuracy it is necessary to tune the parameters and fit them to thoroughly selected reference systems. One should be aware of the fact that fitting of parameters covers errors that are intrinsically preset by the formalism. The strategy is to

Table 3.1: Overview of the MIO Parameter Set (in Atomic Units if not Unitless)

Parameter	H	C	N	O	P	S
$l_{\max}$	0	1	1	1	2	2
$n_{\max}$	2	2	2	2	2	2
α_0	0.50	0.5	0.50	0.50	0.50	0.50
α_1	1.00	1.14	1.20	1.26	1.19	1.19
α_2	2.00	2.62	2.90	3.17	2.83	2.83
α_3	—	6.00	7.00	8.00	6.73	6.73
α_4	—	—	—	—	15.00	16.00
p	2	2	2	2	2	2
r^{wf}(s,p)	3.0	2.7	2.2	2.3	3.8	3.8
r^{wf}(d)	—	—	—	—	4.4	4.4
r^{dens}	2.5	7.0	11.0	9.0	9.0	9.0
ϵ_{s}	–0.2386004	–0.5048917	–0.64	–0.8788325	–0.5106391	–0.6300087
ϵ_{p}	—	–0.1943551	–0.2607280	–0.3321317	–0.2027650	–0.2580265
ϵ_{d}	—	—	—	—	0.5204370	0.0214077
E^{spin}	–0.0330000	–0.0439000	–0.1112000	–0.0541400	–0.0588513	—
U	0.4195	0.3647	0.4309	0.4954	0.2894	0.3288

choose the parameters within a physical reasonable range. Parameters that pass out of such a boundary during an optimization indicate the failure of the formalism (at least for a certain property under investigation) and underline the importance of extending the methodology by so far missing physical concepts.

The MIO parameter set was developed for C, H, N, and O [17, 18] and are the basis of the success of DFTB2 for accurately describing organic compounds in chemical and biological applications. Sulfur [21] and phosphorus [20] were parametrized later, extending the MIO set.

The MIO parametrization turned out to be very successful making DFTB2 applicable to a wide range of chemical/biological applications. The parameters that are described in the following are published on the website www.dftb.org under the name “mio-0-1”. In this section the choice of the atomic electronic parameters is dicussed. Further the fitting procedure and its most important parameters of the repulsive diatomic parameters are summarized. In a final subsection a brief overview of the performance and deficiencies is discussed also showing special parametrization for selected cases.

3.3.1 Electronic Parameters

Table 3.1 gives an overview of all atomic electronic parameters. In the following each parameter is commented.

The parameters $l = 0, 1, 2, ..., l_{\max}$, $n = 0, 1, 2, ..., n_{\max}$, and α are used to define the atomic basis set as given by eq 3.15. The use of a minimal basis for DFTB2 means, that for hydrogen only an s-orbital like function ($l_{\max} = 0$) is considered, for carbon, nitrogen, and oxygen an s-orbital and a p-orbital like function ($l_{\max} = 1$). For phosphorus and sulfur it turned out to be essential for the performance to also include d-orbitals ($l_{\max} = 2$). The parameters α are necessary to form a converged basis set [70], it has been found that a good

choice of values is the following: choose α_0 as a number smaller than 1.0, set the largest value $\alpha_{\max}$ equal to the atomic number and determine all other α_i in a geometrical progression as

$$\alpha_i = \alpha_0 \cdot \left(\frac{\alpha_{\max}}{\alpha_0} \right)^{\frac{i}{\max}} . \tag{3.33}$$

There is an exception within the MIO parametrization, for phosphorus the same α values have been used as for sulfur, only the highest coefficient was changed to the atomic number of phosphorus.

Parameter p is the exponent for the confining potential of eq 3.17 and it was found to have only very small influence on the result. r^{wf}(s,p) is the compression (or confining) radius for s- and p-orbital like functions, and r^{wf}(d) is the one for d-orbital like functions. For determining the atomic reference density a different compression radius r^{dens} is introduced (see comments around eqs 3.17 and 3.20). Within the MIO set both parameters r^{wf} and r^{dens} per element were fitted in a "brute force" manner in order to achieve good performance on bond angles, vibrational frequencies, hydrogen bond lengths and hydrogen binding energies as well as reaction energies.

The eigenvalues ϵ_{s}, ϵ_{p}, and ϵ_{d} for the s-, p-, and d-orbitals enter as diagonal elements into the Hamilton matrix (eq 3.20). They are calculated from eq 3.16 without the additional confining potential using the PBE exchange-correlation functional in order to yield the correct dissociation limit. Two exceptions are made, first for nitrogen and second for phosphorus.

For nitrogen the ϵ_{s} is calculated to be –0.6819693 Hartree. While properties such as geometries, vibrational frequencies (bending modes), reaction energies (of the CHNO set) and inversion barrier of ammonia can be fairly well described, it fails for the dihedral angle of peptide bonds. This failure could not be eliminated by adjusting the confining radii without loosing the good performance for other properties of the CHNO-parameter set. Supposing the hybridization of nitrogen is not described correctly, ϵ_{s} was set to –0.64 Hartree within the MIO parameter set.

For phoshorous the ϵ_{d} is calculated to be 0.02043783 Hartree. Because for that value d-orbitals are excessively involved in chemical binding situations the value was rised to 0.5204370 Hartree redeeming this effect.

Within DFTB2 the electron spin is usually not considered. This gives good results for closed-shell systems.[6] However, when calculating for example atomization energies of a molecule as

$$E^{\mathrm{at}} = E^{\mathrm{DFTB2}} - \sum_a E_a^{\mathrm{el}}, \tag{3.34}$$

where the atomic electronic energies are usually computed as

$$E_a^{\mathrm{el}} = \sum_i n_i \epsilon_i^a + E_a^{\mathrm{spin}}, \tag{3.35}$$

and E_a^{spin} is the spin polarization energy, i.e. the energy difference of a neutral atom between having zero total electron spin and an electron spin according to the lowest possible atomic energy. For the MIO parameter set these atom type dependent values were calculated within the local density approximation. As will be shown in detail in section 5.2, PBE values give improved results, which are –0.04106143, –0.04547908, –0.11476560, –0.05577610, –0.06868858, and –0.03121074 Hartree for H, C, N, O, P, and S, respectively.

[6] For radicals spin symmetry is of importance. Therefore, a DFTB method has been developed in ref [74] extending the formalism to spin densities.

The Hubbard parameter U describes the electron-electron repulsion within a neutral atom according to eq 3.27 and is also used for the γ-function to determine two-center electron-electron interactions (eq 3.26). It is the only parameter being introduced for the self-consistent charge extension. Within MIO, U is determined for neutral atoms by numerically taking the first derivative of the energy of the highest occupied atomic orbital with respect to its occupation number using the PBE functional.

To sum up, $l_{\max}, n_{\max}$, α_i, and p are technical parameters (choice of basis set; a different choice does not have much impact on the results). The only atomic parameters that are optimized to yield best performance are the compression radii r^{wf}(s,p), r^{wf}(d), and r^{dens}. The eigenvalues ϵ_{s}, ϵ_{p}, and ϵ_{d} as well as E^{spin} and U are derived from DFT and computed using these technical parameters (exceptions are ϵ_{s} for nitrogen and ϵ_{d} for phosphorus in the MIO parameter set).

The two-center integrals of the Hamilton matrix elements $H^0_{\mu\nu}$ as defined in eq 3.20 as well as the overlap matrix $S_{\mu\nu}$ could be in principle calculated during runtime. However, for saving computational resources they are precomputed and tabulated for a dense mesh of distances (one table for each atom type pair).[7]

For calculating ϵ_{s}, ϵ_{p}, ϵ_{d}, E^{spin} and U as well as the Hamilton and overlap matrix elements non-published in-house programs TWOCENT and SCFATOM are used. All these values are stored in parameter files that are necessary for running DFTB, for the MIO parameter set they are available at www.dftb.org.

3.3.2 Repulsive Parameters

The repulsive energy is given by eq 3.9. Each repulsive potential V^{rep}_{ab} is only dependent on the atomic reference densities of atoms a and b, thus, a parametrization is needed for each pair of atom types (e.g. H–C, C–C, and so on). Within the MIO parametrization the potentials are determined by calculating the total energy of selected reference systems using a reference method (B3LYP [39, 40, 75] with basis set 6-31G(d) [76]) and subtracting the DFTB electronic energy. For example, the H–C potential was determined as

$$V^{\mathrm{rep}}_{\mathrm{HC}}(r_{\mathrm{HC}}) = \{E^{\mathrm{ref}}(r_{\mathrm{HC}}) - E^{\mathrm{el}}(r_{\mathrm{HC}})\}_{\text{reference system}} + E^{\mathrm{shift}}_{\mathrm{HC}} \tag{3.36}$$

for different interatomic distances r_{HC}. E^{el} stands for the DFTB2 total energy without the contribution of the repulsive energy. Methane—as a most basic chemical relevant example containing an H–C bond—served as reference system. Practically, one H–C bond of methane was stretched and compressed. For a sufficient number of geometries the reference energies E^{ref} as well as the electronic energies E^{el} were recorded. $E^{\mathrm{shift}}_{\mathrm{HC}}$ accounts for constant differences of the total energy between the reference method and DFTB (e.g. due to the consideration of only valence electrons within DFTB and the neglect of atomic contributions in eqs 3.8 and 3.9). The choice of $E^{\mathrm{shift}}_{\mathrm{HC}}$ will be explained below.

In the following a few comments are provided for the determination of the repulsive potentials within the MIO parametrization:

[7] The maxtrix elements are computed for the MIO parameter set at 500 equidistant points starting from 0.02 bohr with an increment of 0.02 bohr. Exceptions are the pairs, PP, HP, CP, NP, and OP which are calculated at 619 points (same starting point and increment). For generating a matrix element for a distance in between two precomputed points during runtime a second order polynomial interpolation is used. Beyond the second last point a spline extrapolation is used, the value at these distances being so small that it is negligible for practical purposes.

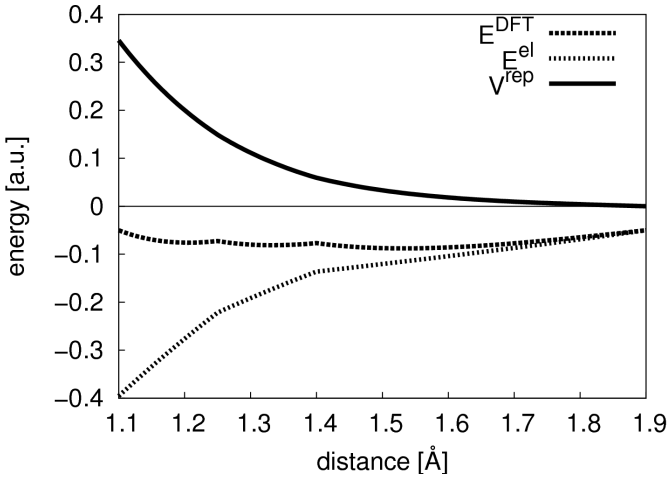


Figure 3.2: V_{CC}^{rep} as difference of the reference total energy E^{DFT} and the DFTB2 electronic energy E^{el} along the C–C distance. Both, E^{DFT} and E^{el} were shifted differently for the different reference systems HC≡CH, $H_2C{=}CH_2$, and $H_3C{-}CH_3$ in order to yield continuous energy curves.

- The C–C potential should be chosen to properly describe different bond types such as single, double, and triple bond. Therefore, ethane, ethene, and ethyne were included in the fit. For each molecule the C–C bond was stretched and compressed. For a continuous repulsive potential the resulting curves had to be shifted relative to each other and were connected as schematically shown in Figure 3.2. Also gradient and curvature were interpolated at these points in order to have a transferable potential that can reproduce geometries and vibrational frequencies. Similar complications arose for other potentials as well, e.g. C–N and C–O.

- A cut-off radius for each repulsive potential was introduced. It was chosen to be shorter than typical second neighbor distances (e.g. in the water molecule one hydrogen is a second neighbor to the other hydrogen). This makes the repulsive potential directly influencing the properties of a covalent bond but not interactions between non-bonded atoms, i.e. bond lengths are directly influenced by the repulsive potential, bond angles and dihedrals angles are not. This was found to be extremely helpful in order to reduce the parameter space of the repulsive potentials.

- E^{shift} of a potential can in principle be chosen to match the reference atomization energy for a molecule (compare eq 3.36 for the example of methane). However, the potential would need to be very repulsive such that an artificial large gradient would be necessary to reach zero energy contribution at the cut-off distance. Therefore, rather than aiming for correct atomization energies the repulsive potentials were energetically shifted to yield consistent reaction energies (benchmarks are shown in section 6.2.3). This caused an overbinding for all bonds.

  Overbinding denotes the difference between the reference atomization energy (for MIO it was calculated from B3LYP/6-31G(d)) and the one calculated using DFTB2 (after geometry optimization with DFTB2). It can be estimated per bond, e.g. the H–C overbinding is calculated from methane as the difference of atomization energies divided by four. For the C–C overbinding in ethane, six times the H–C overbinding (as calculated from methane) is subtracted from the atomization energy difference.

- Besides the repulsive potentials, also the atomic electronic parameter r^{wf} and r^{dens} were optimized. The electronic parameters were chosen in order to reproduce properties that are only marginally influenced by the repulsive potentials due to its cut-off, e.g. bond angles, dihedral angles, hydrogen bond lengths, hydrogen binding energetics. On the other hand the electronic parameters influence the repulsive parameters. For large r^{wf} and small r^{dens} the electronic energy rises, thus, the repulsive potentials need to decrease to lower values in the binding region, for small r^{wf} and large r^{dens} the repulsive potentials should be more repulsive. The choice of electronic parameters helped to properly adjust the overbindings.

In that way every potential was hand-constructed.[8] Moreover, specific refinements for each potential had been carried out. The detailed fitting procedure and all parameters for reconstructing these potentials have never been documented, however, a summary of the molecules that were fitted to, the cut-off radii, and the amount of overbinding for each potential is given in Table 3.2.

In the following a few more details about some specific potentials are described, graphs of all MIO potentials and their first and second derivatives are shown in appendix A:

- H–H-mod: In the process of constucting the H–H potential it has been found that it becomes negative in the binding region. Because an attractive potential might lead to unphysical H–H bonds in a molecular dynamics simulation the potential has been shifted to positive values. A strong underbinding is the result as can be seen in Table 3.2. In contrast, the H–H-mod potential is the unshifted potential leading to a small overbinding.

- H–N-mod: Nitrogen hybridization seems to pose a problem for minimal basis set methods like DFTB as well as for NDDO type semiempirical methods [78]. This problem, which may be related to the neglect of d-orbitals in the basis set, leads to dramatic errors when computing deprotonation energies. In previous studies [72, 79], consistent errors of about 10 kcal/mol were found specifically for proton affinities of sp^3-hybridized nitrogen atoms. Therefore, a modified parameter set "H–N-mod" was introduced in which the H–N repulsive potential was shifted to correct for these errors. However, since sp^2-hybridized nitrogen atoms are described correctly, this correction is needed only for sp^3-hybridized configurations of N. Therefore, similar to the situation in force fields, different "atom types" for N have to be introduced at the moment, which clearly limits DFTB's applicability since these atom types are not allowed to change during a reaction. Thus, H–N-mod usually only comes into play for specific applications, e.g. when calculating proton affinities.

- The shifting E^{shift} of a single potential is done for assuring correct energetics. On the other hand this potential needs to vanish at the cut-off radius. In some cases, both criteria could only be met by introducing a "hump". That means for short distances the curve is at a high energy even though the gradient is quite small (see also the illustrations in appendix A). In order to yield zero at the cut-off an artificial large gradient is introduced. This "hump" is most prominent for the following potentials:

  - C–C: One of the longest C–C single bond is found for the (biologically important) acetate anion with a length of about 1.57 Å using B3LYP/cc-pVTZ. For not letting

[8] Details on the representation of the MIO repulsive potential are documented in ref [77] and on the website www.dftb.org.

Table 3.2: Overview of Parameters Characterizing the MIO Repulsive Potentials

Potential[a]	Cut-off [a.u.]	Reference system	Overbinding/bond [b] [kcal/mol]	
			$E^{\mathrm{spin,LDA}}$	$E^{\mathrm{spin,PBE}}$
H–H	2.08	H_2	–27.1	–37.2
H–H-mod	2.64	H_2	3.4	–6.7
H–C	3.50	CH_4	4.1	–1.3
H–N	3.96	NH_3	9.4	3.6
H–N-mod	3.40	NH_3	21.2	15.4
H–O	3.47	H_2O	14.7	9.1
H–P	3.20	PH_3	–6.3	–13.5
H–S	3.61	H_2S	2.8	–3.5
C–C	4.30	HC≡CH	20.1	18.6
		$H_2C{=}CH_2$	8.4	7.6
		$H_3C{-}CH_3$	7.6	7.1
C–N	4.20	HC≡N	35.4	32.5
		$H_2C{=}NH$	16.8	14.8
		$H_3C{-}NH_2$	8.5	7.4
C–O	4.20	$H_2C{=}O$	20.8	19.3
		$H_3C{-}OH$	9.3	8.7
C–P	5.20	HC≡P	9.4	2.6
		$H_2C{=}PH$	4.9	0.4
		$H_3C{-}PH_2$	5.4	3.3
C–S	4.67	CS_2	18.4	15.7
		$S(CH_3)_2$	11.9	10.5
N–N	4.20	HN=NH	16.9	13.9
		$H_2N{-}NH_2$	11.7	10.2
N–O	3.88	HN=O	41.8	39.3
		$H_2N{-}OH$	21.7	20.5
N–P	5.30	P≡N	17.6	9.1
		HP=NH	5.9	0.4
		$H_2P{-}NH_2$	–2.2	2.2
N–S	3.67	SN_2	40.1	36.2
O–O	4.20	O_2 [c]	37.6	35.5
		HO–OH	8.3	7.3
O–P	5.00	HP=O	53.3	48.3
		H_3PO_4	35.8	33.6 [d]
		$H_4PO_5^-$	25.6	23.8 [d]
O–S	5.23	SO_2	49.5	47.3
P–P	5.70	P≡P	3.0	–9.4
		HP=PH	4.1	–4.0
		$H_2P{-}PH_2$	6.8	3.0
P–S	6.20	HP=S	3.7	–2.6
		$H_2P{-}SH$	2.5	–0.4
S–S	5.44	S_2O	31.2	27.9
		HS–SH	21.6	19.6

[a] Repulsive potentials are symmetric with respect to their order, e.g. the C–O potential is the same as O–C. [b] With respect to B3LYP/6-31G(d) atomization energies. Numbers for $E^{\mathrm{spin,LDA}}$ refer to the original MIO set ("mio-0-1" on www.dftb.org), overbindings using PBE spinpolarization values are illustrated here only for completeness as they will be used in the discussion of the following chapters. [c] Singlet dioxygen. [d] Average overbinding/bond, differences of single and double bond are not considered.

acetate dissociate when using DFTB2 the gradient of the C–C potential should be very small in that region. Additionally, the energy contribution should still be repulsive at that bond length. On the other hand a cut-off at 4.2 a.u. (2.2 Å) is necessary because otherwise the good description of molecules with short C–C second neighbor distances such as furan (C–C second neighbor distance is at 2.2 Å) would be destroyed.

– H–O and H–N: In these cases the "hump" is necessary for the proton barriers $[H_2O–H–H_2O]^+$ and $[NH_3–H–NH_3]^+$ which require the potentials to have a small gradient at around 1.3 Å and a larger gradient at 1.5 Å.

- Due to technical details in the fitting procedure numerical artifacts are present in almost all potentials which cause some oscillations in the second derivative. In practice, no problems have been reported because these oscillations do not lie in relevant binding regions, however, this finding is kept in mind for developing a new parametrization procedure (in the next chapter) to prevent these artifacts.

3.4 Performance and Deficiencies of DFTB2

The success of DFTB2 for biochemical applications is linked to the MIO parametrization. The electronic parameters are calculated using the PBE functional [37]. This means, however, that the well known DFT-GGA deficiencies are inherited by DFTB2. Of particular relevance is the DFT-GGA tendency to overpolarize extended π-conjugate systems [80], the problems of ionic and charge transfer excited states [81], and the missing dispersion interactions, which have been included by augmenting DFTB2 using an empirical extension [64]. This empirical correction has shown to be crucial for the description of nucleic acid base stacking interactions [64] and relative stability of α and 3_{10} helices in proteins [82].

DFTB2 has been tested for various properties of small organic molecules like heats of formations, geometries, vibrational frequencies, dipole moments, etc. in several recent publications. In general, DFTB2 is excellent in reproducing geometries. Also reaction energies are reproduced reasonably well on average [17, 21, 83], while heats of formation are overestimated owing to the overbinding tendency of DFTB2. Recently, the DFTB2 heats of formation have been systematically tested. It turned out, that reparametrization of atomic contributions can improve the performance for heats of formation significantly, however, the refined NDDO methods like OM2 [9] or PDDG-PM3 [6] are still superior to DFTB2 in this respect [14, 11]. The performance of DFTB2 for vibrational frequencies [14, 21, 83, 84, 85, 86, 87], although reasonable on average, is less satisfactory than for geometries. However, also vibrational frequencies could be improved significantly after re-parametrization of the repulsive parameters [88]. It should be noted that all these test sets contain a large number of molecules representative of many chemical bonding situations. However, good performance for small molecules does not guarantee a good description of larger molecules. Good examples are the structures and relative energies of peptides, which pose significant problems for semiempirical models like AM1 [89] and PM3 [90], but are well described at the DFTB2 level [12, 13, 91] or more elaborate NDDO methods like OM1 [8], OM2 [9, 92]. Similarly, relative energies of peptide conformers [12, 13, 91] are nicely reproduced by DFTB2 in comparison to higher level methods as well as hydrogen binding energies [14].

There are deficiencies for some particular cases that are of utmost interest for biological applications:

- Proton affinities are not described correctly. Errors ranging from –10 to +15 kcal/mol and more in comparison to high-level calulations are found.
- Hydrogen binding energies are reasonably described, typically underestimated by 1–2 kcal/mol. Hydrogen bonding is of ultimate importance for biological applications and therefore even this rather small but systematic error severely encounters the applicability of DFTB2.
- Proton transfer barriers of negative charged systems, e.g. $[HO–H–OH]^-$, are drastically underestimated.
- Phosphate chemistry: Reaction energies for phosphorus containing molecules have to be evaluated carefully before applying to a certain type of reactions. While reasonable results are found for proton transfer reactions, large deviations are found for hydrogenation reactions [20]. Proton affinities are again not predictable using DFTB2.

These deficiencies are faced by further methodological development as will be derived in the next chapter.

Chapter 4

DFTB3

It has been shown that extensions of DFTB2 can improve the performance of DFTB2 for hydrogen bonded complexes and molecules with localized charges significantly, thereby improving the transferability of DFTB [67, 72, 93, 94]. These activities concern basically two recent developments, an improvement of the effective electron repulsion term in the DFTB2 formalism, the γ-function, and the extension to include third order terms.

The γ-function describes the Coulomb interaction between atomic partial charges Δq_a. The functional form chosen for this interaction presupposes an inverse relation of atomic size and chemical hardness [17], which is true for elements within one row of the periodic table [67, 72] but not for elements of different periods. A particularly large deviation occurs for the hydrogen atom. A newly introduced γ^{h}-function corrects this incorrect assumption by an extra term including one additional parameter and as a result, systematically improves hydrogen bonding interactions [93].

The inclusion of approximate third order terms leads to a new degree of self-consistency [67, 72]. In DFTB2, the Coulomb repulsion resulting from the charge density fluctuations as described by the second order terms is computed in a monopole approximation utilizing a newly introduced parameter, the Hubbard parameter (chemical hardness). This parameter is computed from DFT for neutral atoms and is a constant for all charge states of the atom. While this approximation seems to be unproblematic for many covalently bound systems, it is insufficient for molecules that contain large localized net charges. As has been shown, these systems require additional flexibility in the model; i.e., the Hubbard parameters have to become charge dependent, which is achieved by including the approximated third order terms [67, 72].

The third order terms can be split up into two parts, a diagonal and a off-diagonal one. The diagonal terms lead to a charge dependent on-site self-interaction, the off-diagonal terms modify the second order Coulomb repulsion between sites. The diagonal contributions significantly improve the proton affinities of CHNO containing molecules, since in these calculations strongly localized net charges occur [93]. They also improve the proton affinities of phosphorus containing molecules [94]. However, a reasonable accuracy was only achieved by adding an empirical energy contribution in a rather *ad hoc* fashion, which still did not lead to an acceptable transferability; i.e., different parameter sets had to be developed for different properties. Although these extensions have been shown to be important for describing proton affinities and hydrogen binding energies in various applications [95, 96, 97, 98, 99, 100, 101, 102], further improvement is required to obtain a more transferrable method for general applications.

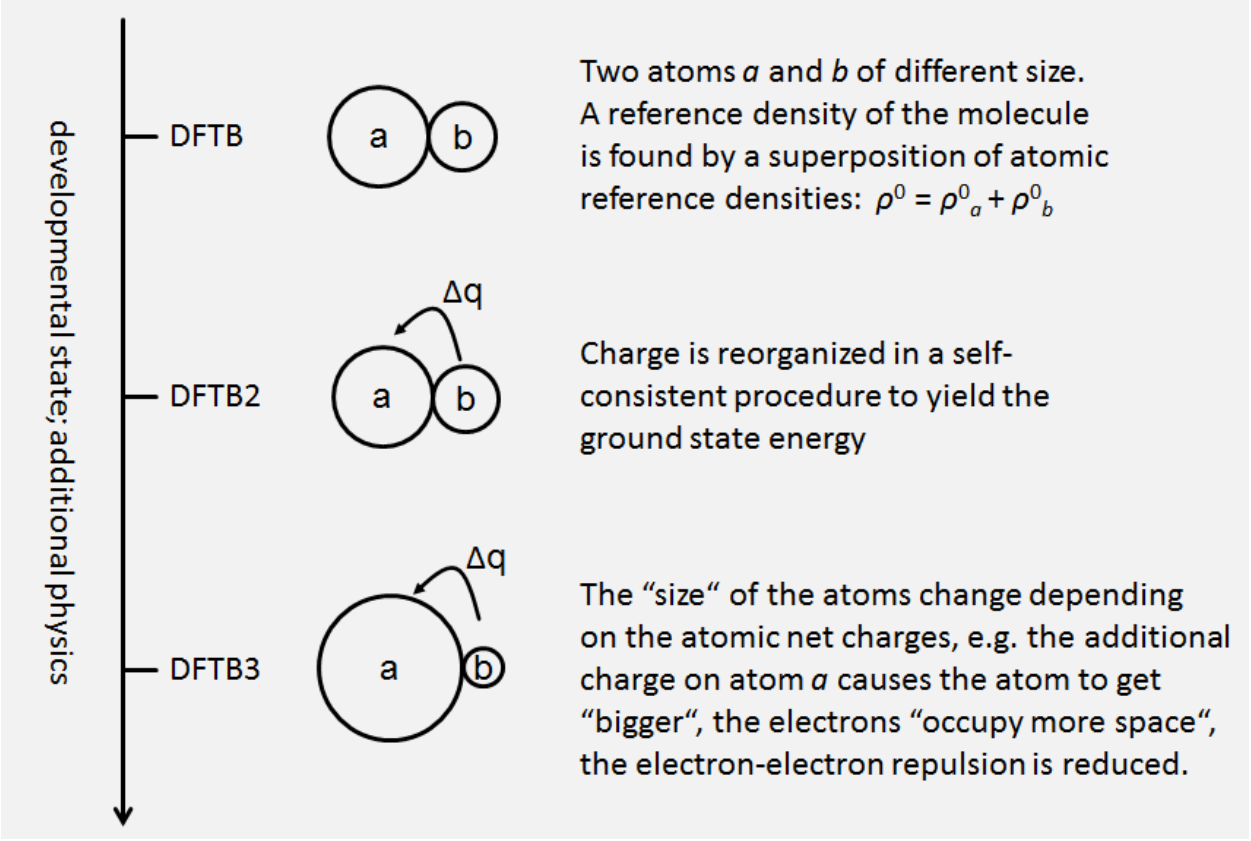


Figure 4.1: Simplified illustration of the development of DFTB.

In the present work the off-diagonal third order contributions are implemented and tested. In combination with the γ^{h}-function and diagonal third order terms, this establishes a third generation of the DFTB methodology which will be called DFTB3. The off-diagonal terms are shown to overall improve the DFTB performance; most importantly, with this new formalism, a single set of parameters is able to reproduce many properties of CHNO and phosphorus containing complexes with good accuracy.

In this chapter the formalism of both extensions is described. It is followed by computational details including an overview of several DFTB variants and the newly introduced parameters for DFTB3. Finally some benchmark tests are shown and concluding remarks are drawn. In Figure 4.1 the development of DFTB is illustrated.

4.1 Third Order Taylor Series Expansion

4.1.1 Total Energy

The third order term as shown in eq 3.5 is given by

$$\begin{aligned} E^{3\mathrm{rd}} &= \frac{1}{6}\int''\int'\int \left[\frac{\delta^3 E^{\mathrm{xc}}[\rho]}{\delta\rho\delta\rho'\delta\rho''}\right]_{\rho^0,\rho^{0\prime},\rho^{0\prime\prime}} \Delta\rho\Delta\rho'\Delta\rho'' \\ &= \frac{1}{6}\int''\int'\int \frac{\delta}{\delta\rho''}\left[\frac{\delta^2 E^{\mathrm{xc}}[\rho]}{\delta\rho\delta\rho'}\right]_{\rho^0,\rho^{0\prime},\rho^{0\prime\prime}} \Delta\rho\Delta\rho'\Delta\rho''. \end{aligned} \tag{4.1}$$

The same approximations as for the second order integrals can be applied here, i.e., the description of the charge density fluctuations in terms of superposition of atomic contributions eq 3.22 and the restriction of the charge density fluctuations to the monopole term in eq 3.23. Inserting these two equations into eq 4.1 yields [67, 93]:

$$E^{\text{3rd}} \approx E^{\Gamma} = \frac{1}{6}\sum_{abc}\Delta q_a\Delta q_b\Delta q_c \left.\frac{\mathrm{d}\gamma_{ab}}{\mathrm{d}q_c}\right|_{q_c^0} \tag{4.2}$$

$$= \frac{1}{6}\sum_{a}\Delta q_a^3 \left.\frac{\partial\gamma_{aa}}{\partial q_a}\right|_{q_a^0} + \frac{1}{6}\sum_{a\neq b}\Delta q_a\Delta q_b\left(\Delta q_a \left.\frac{\partial\gamma_{ab}}{\partial q_a}\right|_{q_a^0} + \Delta q_b \left.\frac{\partial\gamma_{ab}}{\partial q_b}\right|_{q_b^0}\right). \tag{4.3}$$

Therefore, in the third order DFTB formalism the derivative of the γ-function with respect to charge introduces the desired chemical behavior for charged systems. For the diagonal terms (first term in eq 4.3) the derivative of γ implies via eq 3.27 a charge dependent Hubbard parameter (chemical hardness), i.e. the chemical hardness changes with charge state. Since U_a is also used to approximate the size of atom a in the damped Coulomb repulsion term γ, a charge dependent U_a will also make the atomic electron-electron repulsion charge dependent. For the off-diagonal terms (second term in eq 4.3) this effect applies for the electron-electron repulsion between two atoms. Note that γ_{ab} is dependent on the atomic charges only via the Hubbard parameters U_a and U_b. Introducing

$$\begin{aligned}
\Gamma_{ab} &= \left.\frac{\partial\gamma_{ab}}{\partial q_a}\right|_{q_a^0} = \left.\frac{\partial\gamma_{ab}}{\partial U_a}\frac{\partial U_a}{\partial q_a}\right|_{q_a^0} && \text{with } a\neq b,\\
\Gamma_{ba} &= \left.\frac{\partial\gamma_{ab}}{\partial q_b}\right|_{q_b^0} = \left.\frac{\partial\gamma_{ab}}{\partial U_b}\frac{\partial U_b}{\partial q_b}\right|_{q_b^0} && \text{with } a\neq b,\\
\Gamma_{aa} &= \left.\frac{\partial\gamma_{aa}}{\partial q_a}\right|_{q_a^0} = \left.\frac{1}{2}\frac{\partial\gamma_{aa}}{\partial U_a}\frac{\partial U_a}{\partial q_a}\right|_{q_a^0},
\end{aligned} \tag{4.4}$$

where the latter definition is made to ease the summation, the third order energy contribution becomes

$$E^{\Gamma} = \frac{1}{6}\sum_{ab}\Delta q_a\Delta q_b\left(\Delta q_a\Gamma_{ab} + \Delta q_b\Gamma_{ba}\right) = \frac{1}{3}\sum_{ab}\Delta q_a^2\Delta q_b\Gamma_{ab}. \tag{4.5}$$

Details on the derivative $\frac{\partial\gamma_{ab}}{\partial U_a}$ are given in appendix B. The diagonal term $\left.\frac{\partial U_a}{\partial q_a}\right|_{q_a^0}$ can be computed as the third derivative of the total energy of an atom with respect to charge. Practically, chemical hardness values are computed for atoms in different charge states (using Janak's theorem, see above) and these values are used to estimate the third derivative.

Thus, adding the approximated third-order contribution E^{Γ} to E^{DFTB2} (eq 3.32) yields the total energy of the third-order formalism

$$\begin{aligned}
E^{\text{DFTB3}} &= E^{\text{H0}} + E^{\gamma} + E^{\Gamma} + E^{\text{rep}}\\
&= \sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b} n_i c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2}\sum_{ab}\Delta q_a\Delta q_b\gamma_{ab} + \frac{1}{3}\sum_{ab}\Delta q_a^2\Delta q_b\Gamma_{ab} + E^{\text{rep}}
\end{aligned} \tag{4.6}$$

4.1.2 Kohn-Sham Equations

To determine the MO coefficients $c_{\mu i}$ (eq 3.6) approximate Kohn-Sham equations are derived by taking the derivative of the total energy with respect to the constraint[1]

$$\int \psi_i \psi_j \mathrm{d}\boldsymbol{r} = \delta_{ij} \tag{4.7}$$

to get:

$$\frac{\partial}{\partial c_{\delta i}} \left[E^{\mathrm{DFTB3}} - \sum_j n_j \epsilon_j \left(\sum_{ab} \sum_{\mu \in a} \sum_{\nu \in b} c_{\mu j} c_{\nu j} S_{\mu\nu} - 1 \right) \right] \; = \; 0 \qquad \forall d, \delta \in d. \tag{4.8}$$

Again, the Mulliken charge analysis for estimating the charge fluctuations $\Delta q_a = q_a - q_a^0$ is employed,

$$q_a \; = \; \sum_j n_j \sum_{\mu \in a} \sum_b \sum_{\nu \in b} c_{\mu j} c_{\nu j} S_{\mu\nu} \tag{4.9}$$

and as shown in detail in appendix B the Kohn-Sham equations are obtained after combining eqs 4.6–4.9 and $\gamma_{ab} = \gamma_{ba}$:

$$\sum_b \sum_{\nu \in b} c_{\nu i} \left(H_{\mu\nu} - \epsilon_i S_{\mu\nu} \right) \; = \; 0, \qquad \forall a, \mu \in a, i \tag{4.10}$$

$$H_{\mu\nu} \; = \; H^0_{\mu\nu} + S_{\mu\nu} \sum_c \Delta q_c \left(\frac{1}{2}(\gamma_{ac} + \gamma_{bc}) + \frac{1}{3}(\Delta q_a \Gamma_{ac} + \Delta q_b \Gamma_{bc}) + \frac{\Delta q_c}{6}(\Gamma_{ca} + \Gamma_{cb}) \right) \tag{4.11}$$

$$\forall a, b, \mu \in a, \nu \in b.$$

The total energy in eq 4.6 can also be expressed in terms of the ϵ_i (for details see appendix B) as

$$E^{\mathrm{DFTB3}} \; = \; \sum_i n_i \epsilon_i - \frac{1}{2} \sum_{ab} (q_a + q_a^0) \Delta q_b \gamma_{ab} - \frac{1}{3} \sum_{ab} (q_a + q_a^0) \Delta q_a \Delta q_b \Gamma_{ab} - \frac{1}{3} \sum_{ab} q_a \Delta q_b^2 \Gamma_{ba} + E^{\mathrm{rep}}. \tag{4.12}$$

4.1.3 Atomic Forces

An analytical force equation is derived for a Cartesian coordinate system using the derivative of the total energy with respect to the atomic coordinates R_{kx}, while subjecting to the constraints eq 4.7. With k, the respective atom index, and $x \in \{1, 2, 3\}$, the spacial directions (xyz) the force acting on one atom in one direction can be written using eq 3.6:

$$F_{kx} \; = \; -\frac{\partial}{\partial R_{kx}} \left[E^{\mathrm{DFTB3}} - \sum_i n_i \epsilon_i \left(\sum_{ab} \sum_{\mu \in a} \sum_{\nu \in b} c_{\mu i} c_{\nu i} S_{\mu\nu} - 1 \right) \right] \qquad \forall k, x \tag{4.13}$$

The energy depends explicitely on the atomic coordinates via $S_{\mu\nu}$, $H^0_{\mu\nu}$, and γ_{ab} and implicitly via the coefficients $c_{\mu i}$:

$$E^{\mathrm{DFTB3}} \; = \; E^{\mathrm{DFTB3}} \left(c_{\mu i}(R_{kx}), R_{kx} \right) \tag{4.14}$$

$$\frac{\mathrm{d}E^{\mathrm{DFTB3}}}{\mathrm{d}R_{kx}} \; = \; \sum_{\mu i} \frac{\mathrm{d}E^{\mathrm{DFTB3}} \left(c_{\mu i}(R_{kx}), R_{kx} \right)}{\mathrm{d}c_{\mu i}} \frac{\mathrm{d}c_{\mu i}}{\mathrm{d}R_{kx}} + \frac{\partial E^{\mathrm{DFTB3}}}{\partial R_{kx}}. \tag{4.15}$$

[1]Note, in DFT the variational search for the minimum of $E[\rho]$ can be either achieved using eq 2.8 or as presented here which implies $\rho(r) = \sum_i n_i \psi_i^*(r) \psi_i(r)$ (eq 2.4), for details see [29] (p. 146).

Because of the variational principle (see eq 4.8) the implicit dependence via the coefficients $c_{\mu i}$, i.e., the first term of eq 4.14 is equal to zero. The second term, i.e., the explicit dependence of the energy with respect to the coordinates $\frac{\partial E^{\mathrm{DFTB3}}}{\partial R_{kx}}$ can be evaluated using eq 4.6 as shown in appendix B. Taking advantage of the symmetry of $S_{\mu\nu} = S_{\nu\mu}$ and $\gamma_{ab} = \gamma_{ba}$ the equation for the atomic forces is then given by

$$\begin{aligned} F_{kx} \;=\; & -\sum_{a\neq k}\sum_{\mu\in a}\sum_{\nu\in k}\sum_{i} n_i c_{\mu i} c_{\nu i}\Bigg(2\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}} - 2\epsilon_i\frac{\partial S_{\mu\nu}}{\partial R_{kx}} + \frac{\partial S_{\mu\nu}}{\partial R_{kx}}\cdot \\ & \left(\sum_c \Delta q_c\left(\gamma_{ac}+\gamma_{kc}+\frac{1}{3}(2\Delta q_a\Gamma_{ac}+\Delta q_c\Gamma_{ca}+2\Delta q_k\Gamma_{kc}+\Delta q_c\Gamma_{ck})\right)\right)\Bigg) \\ & -\Delta q_k\sum_{a\neq k}\Delta q_a\frac{\partial\gamma_{ak}}{\partial R_{kx}} - \frac{1}{3}\Delta q_k\sum_{a\neq k}\Delta q_a\left(\Delta q_a\frac{\partial\Gamma_{ak}}{\partial R_{kx}}+\Delta q_k\frac{\partial\Gamma_{ka}}{\partial R_{kx}}\right) - \frac{\partial E^{\mathrm{rep}}}{\partial R_{kx}} \qquad \forall k,x. \end{aligned} \tag{4.16}$$

The derivatives $\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}}$ and $\frac{\partial S_{\mu\nu}}{\partial R_{kx}}$ are determined by taking the numerical derivative of the tabulated integrals $H^0_{\mu\nu}$ and $S_{\mu\nu}$ The derivatives $\frac{\partial\gamma_{ak}}{\partial R_{kx}}$, $\frac{\partial\Gamma_{ak}}{\partial R_{kx}}$, and $\frac{\partial\Gamma_{ka}}{\partial R_{kx}}$ are evaluated analytically as shown in appendix B. The force contribution $\frac{\partial E^{\mathrm{rep}}}{\partial R_{kx}}$ is also calculated analytically. Details about the representation of E^{rep} can be found in sections 3.3 and 5.1.

4.2 The γ^{h}-Function

The γ-function represents the Coulomb repulsion between the density fluctuations within the DFTB approximation, i.e., for spherically constrained atomic densities. In ref [17], an analytical function has been derived, which is:

$$\gamma_{ab} \;=\; \frac{1}{r_{ab}} - S(r_{ab}, U_a, U_b), \tag{4.17}$$

where S is a short-range function being responsible for the correct convergence of γ_{ab} at $r_{ab} = 0$. This function imposes a simple rule, which implies that the chemical hardness of an atom is inversely proportional to its size [17]. As has been pointed out earlier, traditional semiempirical methods like MNDO, AM1 or PM3 use a similar approximation for the Coulomb interaction [67]. As discussed above, the Hubbard parameter U_a has a dual role: for the on-site contributions, U_a models the effective Coulomb repulsion at site a, while for the off-diagonal terms, the inverse of U_a models the covalent radius of atom a, i.e., it determines the deviation of γ_{ab} from $1/r_{ab}$. However, this inverse relation of chemical hardness and atomic size is not strictly valid across the periodic table [67], it basically only holds within one period of the system of elements, as can be seen from Figure 4.2, which shows the calculated Hubbard parameters for each element in dependence of the covalent radii.

Therefore, in principle a different γ_{ab} should be applied for different rows of the periodic table. Clearly, the deviation is the largest for hydrogen, therefore we proposed to modify γ_{ab} when hydrogen is involved and introduced a γ^{h}-function as [67, 93]

$$\gamma^{\mathrm{h}}_{ab} \;=\; \frac{1}{r_{ab}} - S(r_{ab}, U_a, U_b)\cdot h(r_{ab}, U_a, U_b), \tag{4.18}$$

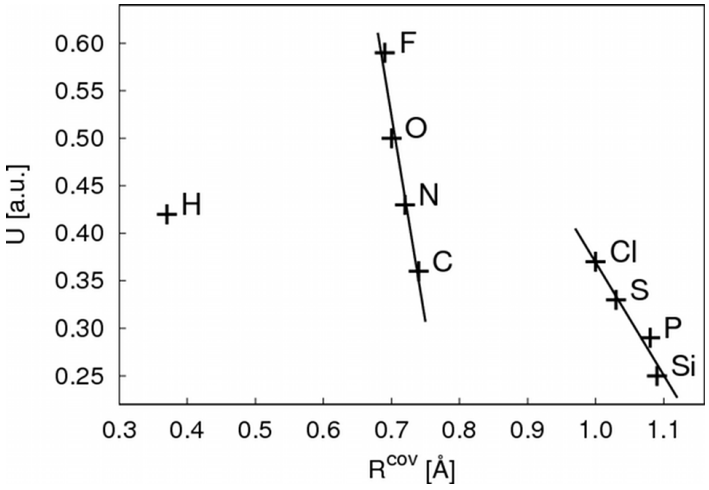


Figure 4.2: Calculated Hubbard parameters U versus covalent radii R^{cov}. The covalent radii are taken from the literature [103]. For C, H, N, O, F values for R^{cov} are plotted that are estimated for bonds to second period elements, for Si, P, S, Cl values for R^{cov} are plotted that are estimated for bonds to third period elements. There is no overall inverse proportional relation as assumed by DFTB2 but only for elements within one period.

where

$$h(r_{ab}, U_a, U_b) = \begin{cases} 1 & \text{if neither atom } a \text{ nor } b \text{ are of type hydrogen} \\ \exp\left[-\left(\frac{U_a+U_b}{2}\right)^{\zeta} r_{ab}^2\right] & \text{if at least one of atoms } a \text{ and } b \text{ is of type hydrogen.} \end{cases} \tag{4.19}$$

In the following it will be referred to this function as the γ^{h}-function in contrast to the γ-function as was used in standard DFTB2 ($h = 1$ for all cases). Note that different than mentioned in ref [93] the γ^{h}-function is also used for the H–H pair. The particular choice of h is to some degree arbitrary. On the other hand, its functional form is quite well physically motivated, correcting the shortcomings of the original function, since the chemical hardness of hydrogen simply cannot be used to represent the hydrogen covalent radius. Clark and coworkers described a similar problem and modification for NDDO-based semiempirical methods [78].

Up to now, the parameters introduced in the second and third order extensions, in principle, can be calculated based on DFT. Unfortunately, the parameter ζ in eq 4.19 cannot be computed from DFT but has to be fitted. However, as shown before [72, 67], by choosing the parameter ζ such that the binding energy of the water dimer is reproduced correctly, γ^{h} becomes more repulsive in the covalent and hydrogen bonding region (see Figure 4.3), and improves hydrogen bonding systematically. The γ^{h}-function is used in combination with the third order terms. Therefore, also the derivative of the γ^{h}-function with respect to charge has to be calculated which is shown in detail in appendix B.

4.3 Computational Details for Benchmarking

For benchmarking the extensions derived in the previous section several DFTB variants are defined in a first part of this section. It is followed by the presentation of parameters newly introcuded within DFTB3. Finally, the way of calculating proton affinities is explained being different than in earlier studies [93, 94].

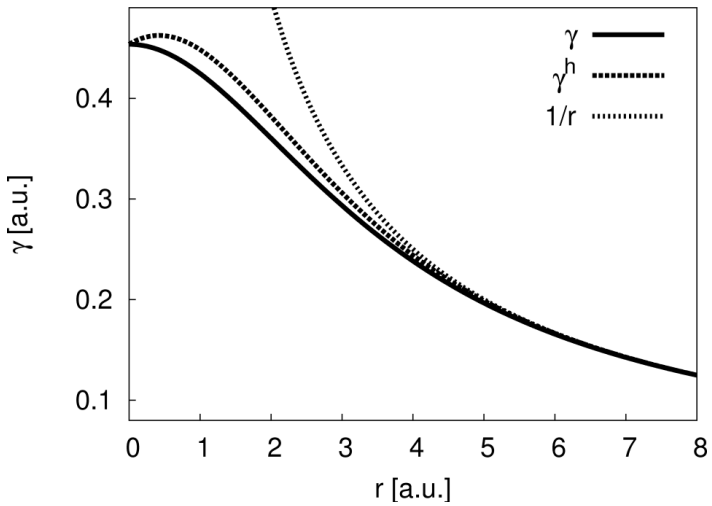


Figure 4.3: The γ^{h}-function plotted for the HO-pair ($U_{\mathrm{H}} = 0.4195$ a.u.,$U_{\mathrm{O}} = 0.4954$ a.u.) is more repulsive than the original γ-function but still yields the same limits at $r = 0$ a.u. and at $r \to \infty$.

4.3.1 DFTB Variants

In DFTB a Taylor series expansion is applied for the DFT exchange-correlation energy. While DFTB, the nonself-consistent variant, includes terms up to first order, SCC-DFTB includes also the second order term and DFTB3 also the third order term. For consistent naming the names DFTB, DFTB2, and DFTB3 are used. Note that DFTB2 corresponds to the formerly called SCC-DFTB. By default the standard γ-function is used for DFTB2 and the γ^{h}-function for DFTB3. In the following benchmark a comparison will be conducted for

- DFTB2: formerly called SCC-DFTB using the standard γ-function (eq 4.17) as derived in ref [17] and
- DFTB3: full third order extension (eq 4.6) including the γ^{h}-function (eq 4.18) as derived in present work.

Starting from DFTB2 three major changes have been made to develop DFTB3. First, the γ^{h}-function substitutes the standard γ-function, second, diagonal third order terms are included, and third, off-diagonal third order terms are taken into account. To illustrate the effects of each of these extensions separately results are presented also for the following intermediate variants:

- DFTB2-γ^{h}: the standard γ-function of DFTB2 is exchanged by the γ^{h}-function
- DFTB3-γ: the standard γ-function is used in connection with DFTB3
- DFTB3-diag: the γ^{h}-function is used and only the diagonal third order terms are included (second term in eq 4.3 is neglected) as suggested in refs [67, 93].

Note that by introducing DFTB3 the intermediate variants become obsolete and are not recommended for practical applications. An overview of all DFTB variants is given in Table 4.1.

The additional computational costs of the γ^{h}-function and the full third order extensions are negligible compared to the diagonalization of the Hamilton matrix, which is the time limiting step in the DFTB methodology. Therefore, the computer time requirements are roughly the same for all variants in Table 4.1.

Table 4.1: DFTB Variants Compared in Present Study

Variant	γ^{h}-function	Diagonal third order terms	Off-diagonal third order terms
DFTB2	no	no	no
DFTB2-γ^{h}	yes	no	no
DFTB3-γ	no	yes	yes
DFTB3-diag	yes	yes	no
DFTB3	yes	yes	yes

4.3.2 Parameters for DFTB3

The definition of E^{rep} (eqs 3.9 and 3.36) requires to newly determine the repulsive potential for a change in the electronic part of the energy. However, the extensions introduced with DFTB3 are mainly relevant for systems with significant amount of localized net charges. Therefore, the MIO parameter set (section 3.3) serves as good starting point. In chapter 6 a new parametrization for DFTB3 is suggested.

New Parameters

The γ^{h}-function for the pairs HX (X$\in$\{C, H, N, O, P, S\}) describes the dependence between the size of the atom and the electron-electron interaction more correctly; one additional, purely empirical parameter ζ is necessary. It can be determined using only one data point, the binding energy of the water dimer, for which the most accurate theoretical value is –5.0 kcal/mol using CCSD(T) [104]. Nevertheless, to stay consistent within the fitting procedure as described below the similar value of –4.9 kcal/mol is chosen, which is the result from the G3B3 [105] method. In Table 4.2 this way of determining ζ is denoted as "calc". Further this parameter will be fitted to reproduce an extended data set, in combination with fitting the Hubbard derivatives, then denoted as "fit".

The third order Taylor series expansion of the exchange correlation energy makes use of the Hubbard derivatives $U_a^{\mathrm{d}} = (\partial U_a/\partial q_a)|_{q_a^0}$ (see eq 4.4). These are determined by taking the numerical derivative of the corresponding Hubbard parameter of a neutral atom with respect to the occupation number of the highest occupied atomic orbital. Thus, there is one additional parameter per element. In Table 4.2 the Hubbard derivatives are summarized and abbreviated as "calc"; they are calculated with the PBE exchange-correlation functional [37] and the in-house program TWOCENT.

Therefore, one parameter set "calc" is provided, where only one parameter (ζ) is fitted to one system (water dimer), and all Hubbard derivatives are calculated. In a different approach, all parameters are fitted for a large set of molecules, resulting in the parameter set "fit". This has been done first for the DFTB3-diag method in ref [93].

It is important to note that fitting of the Hubbard derivatives and ζ basically affects hydrogen bonds and proton affinities, most properties of neutral molecules like equilibrium geometries are not significantly altered. Nevertheless, one has to be careful not to misuse the parameters, i.e. correcting shortcomings that result from the electronic and repulsive parameters. The results indicate that the approximations in the third order terms account very well for the physical effects arising from that level. The remaining errors in the description of H-bonding and proton affinities seem not to result from the third order approximations but from the underlying second order DFTB2. For the moment, DFTB3 parameters are opti-

Table 4.2: Parameter Sets for the Different DFTB Variants

Variant	Set[a]	ζ [b]	$U^{\mathrm{d}}_{\mathrm{C}}$	$U^{\mathrm{d}}_{\mathrm{H}}$	$U^{\mathrm{d}}_{\mathrm{N}}$	$U^{\mathrm{d}}_{\mathrm{O}}$	$U^{\mathrm{d}}_{\mathrm{P}}$	$U^{\mathrm{d}}_{\mathrm{S}}$
DFTB2		—	—	—	—	—	—	—
DFTB2-γ^{h}	calc	3.70[c]	—	—	—	—	—	—
DFTB3-γ	calc	—	–0.1492	–0.1857	–0.1535	–0.1575	–0.0702	–0.0695
DFTB3-diag	calc	4.53	–0.1492	–0.1857	–0.1535	–0.1575	–0.0702	–0.0695
DFTB3-diag	fit	5.0	–0.04	–0.14	–0.11	–0.17	–0.07	—
DFTB3	calc	4.05	–0.1492	–0.1857	–0.1535	–0.1575	–0.0702	–0.0695
DFTB3	fit	4.2	–0.23	–0.16	–0.13	–0.19	–0.14	—

[a] "calc" stands for calculated Hubbard derivatives U^{d} and/or ζ fitted to the water dimer; "fit" for a set of parameters fitted to a large set of binding energies and proton affinities. For details see text. [b] ζ is the unitless parameter as defined in eq 4.19. [c] Note that ζ is fitted to yield a binding energy for the water dimer of –4.9 kcal/mol in contrast to $\zeta = 4.5$ as reported in ref [93] for DFTB2-γ^{h} where ζ was fitted to minimize the error of 22 selected binding energies.

mized to make it applicable to important chemical and biological problems without refining the DFTB2 approximations.

The idea of DFTB2 is to use as many parameters calculated from DFT as possible. By fitting the Hubbard derivative parameters U^{d} this spirit is left behind and it seems natural to challenge the insistence on DFT calculated Hubbard parameters U. Surely, a fitting also of these parameters may lead to an improved chemical accuracy, however, at cost of the following benefits; first, a physically robust and transferable method, second, an easy detection of systematic errors, and third, a small space of parameters allowing an easier fitting of the remaining parameters (especially because the Hubbard parameters affect mainly all chemical properties for systems of biological relevance). Please also note that the third order formalism introduces new physics into the DFTB method that cannot be compensated just by a new set of fitted Hubbard parameters. In that sense future work is focused on avoiding empirical fitting of Hubbard parameters and derivatives by improving the electronic (confining radii) and repulsive parameters, that are till now fitted on DFTB2, at the DFTB3 level.

To optimize the parameters for DFTB3-diag, Yang et al. chose a weighted penalty function where the properties of interest included binding energies, proton affinities as well as the root-mean-square gradient of the included molecules calculated at the reference structure [93]. Finally they minimized the penalty function using a genetic algorithm optimizing the Hubbard derivatives and the ζ parameter. In this work a "brute force" fitting is used. A small set of parameters around the calculated values (ζ fitted to the water dimer) is chosen and the performance is evaluated by calculating the mean unsigned error (MUE) of proton affinities and binding energies using geometry optimized molecules for each parameter set. Whenever the parameter set that performs the best reaches a boundary of the current range of parameters, the range is extended. The latter step is repeated until the best set does not reach any boundaries of the current range.

For the fitting of ζ and the Hubbard derivatives of C, H, N, and O, a set of 22 binding energies and 32 proton affinities as compiled by Yang et al. [93] is used to represent important biological properties. The calculations are carried out in the gas phase at 0 K without including the zero-point corrections for both reference and DFTB. Subsequently, the Hubbard derivative of phosphorus is fitted to a set of 18 proton affinities of phosphorus containing molecules (compilation from ref [94]) in the same manner but keeping all other

parameters fixed. All molecules involved in the fitting procedure are listed in the following subsections.

It is found that the Hubbard derivative of carbon becomes very small during the fitting of DFTB3/fit while all other Hubbard parameters stay close to the calculated values. To avoid getting unphysical values U_C^d is limited to a lower boundary of –0.23 a.u..[2] Similarly, U_C^d becomes quite large during the fit of DFTB3-diag/fit such that it is limited to an upper boundary of –0.04 a.u.. Note that the fitted parameters are different from the ones published by Yang et al. [93] since the way of computing proton affinities is different (details see above).

The additional off-diagonal terms within DFTB3 seem to be more repulsive in comparison with DFTB3-diag, therefore ζ becomes smaller to compensate for that as shown in Table 4.2.

Nitrogen Hybridization: a Problematic Case for a Minimal Basis Set Method

As mentioned in section 3.3 nitrogen hybridization seems to pose a problem for minimal basis set methods like DFTB. This problem is not corrected for by neither γ^h-function nor the third order terms and leads to dramatic errors when computing deprotonation energies of systems containing sp^3-hybridized nitrogen. As pragmatic solution results are presented for a "NHorg" and a "NHmix" parameter set. NHorg denotes the parameters for N–H bonds from the MIO set (H–N potential, see Table 3.2), i.e., in this set no different atom types occur. For the NHmix set the standard MIO potential (H–N potential) is only used for compounds containing sp^2 or sp^1 nitrogen whereas the H–N-mod potential is applied for sp^3-hybridized nitrogen atoms which remedies the problems for the deprotonation energies. For reactions where a nitrogen changes its hybridization state from sp^2 or sp to sp^3 the NHorg parameter set is used in order to have consistent energetic contributions for the N–H atom pairs.[3]

The fitting procedure for ζ and the Hubbard derivatives is applied separately for NHorg and NHmix; however, both optimized parameters turn out to be equal. This extends the transferability of the "fit" parameter sets (see Table 4.2) and implies that besides for "calc" also for "fit" the NHorg and NHmix results differ only for test molecules where a sp^3 nitrogen is bound to hydrogen.

4.3.3 Calculation of Proton Affinities Using DFTB

The proton affinity is defined as the negative of the enthalpy change for the gas phase reaction $A^-(g) + H^+(g) \rightarrow AH(g)$ at a given temperature. To avoid a large number of vibrational calculations only the potential energy change is considered in this work and zero-point correction, thermal contributions, and the PV-term (difference between energy and enthalpy) are not included. This is done consistently for both, reference calculations and DFTB calculations. Due to the neglect of the one-center terms in the repulsive potential eq 3.8, the energy of a proton is not zero in DFTB [106] and can be computed in two ways:

(i) First, it is given by the DFTB2 energy as (see eq 3.32):

$$E^{\mathrm{DFTB2}}(H^+) = \frac{1}{2}\gamma_{HH} = \frac{1}{2}U_H. \tag{4.20}$$

[2]For $U_C^d < -0.40$ a.u. the self-consistent procedure does not converge for several molecules in the training set.

[3]Note, different than in present work Yang et al. [93] defined NHmix that way that H–N-mod is also used for calculating the proton affinity of NH_2^-. Since the orbitals calculated on the NH_2^- molecule are similar to orbitals on sp^2 nitrogen, NHorg is applied for that case.

This is a direct result of neglecting the one center terms in the repulsive potential of eq 3.8, since eq 3.9 is used for all practical implementations and applications [106]. Therefore, the energy of the proton is given by half of the Hubbard parameter of hydrogen, which is 131.62 kcal/mol when computed using the DFT-PBE functional. This value may not be considered an accurate estimate since the Hubbard parameter is computed for the neutral hydrogen atom, however, it is consistent with the DFTB2 formalism.

(ii) Alternatively, the one center contribution to the repulsive potential can also be computed directly [106] as

$$V_H^{\mathrm{rep}}[\rho_{\mathrm{H}}^0] = E^{\mathrm{DFT}} - E^{\mathrm{SCC,el}}. \tag{4.21}$$

With the energy of the hydrogen atom $E^{\mathrm{DFT}} = -0.49772$ a.u. (B3LYP/6-311++G(d,p)) and the electronic part of the DFTB2 energy, $E^{\mathrm{SCC,el}} = E^{\mathrm{H0}} + E^{\gamma} = -0.27164$ a.u. (first and second term in eq 3.32, here $E^{\gamma} = 0$ a.u.) gives a one-center repulsive energy contribution for the hydrogen atom of $V_{\mathrm{H}}^{\mathrm{rep}}[\rho_{\mathrm{H}}^0] = -141.87$ kcal/mol [106].[4] For the proton the energy within DFTB2 is then given by $0.5U_{\mathrm{H}} + V_{\mathrm{H}}^{\mathrm{rep}}[\rho_{\mathrm{H}}^0] = 10.25$ kcal/mol. Clearly, the electronic energy of a proton should be equal to zero, however, U_{H} is calculated as the derivative of the highest occupied atomic orbital with respect to the occupation number for the neutral hydrogen atom and cannot completely compensate for $V_{\mathrm{H}}^{\mathrm{rep}}[\rho_{\mathrm{H}}^0]$ in the case of H^+. For this unique situation where the total charge of the system is removed the perturbative approach of DFTB2 fails. Therefore, the energy of the proton was set to +141.87 kcal/mol in earlier studies [106, 93].

With DFTB3 the direct calculation (i) is slightly modified:

$$E^{\mathrm{DFTB3}}(\mathrm{H}^+) = \frac{1}{2}\Delta q_{\mathrm{H}}^2\gamma_{\mathrm{HH}} + \frac{1}{3}\Delta q_{\mathrm{H}}^3\Gamma_{\mathrm{HH}} = \frac{1}{2}U_{\mathrm{H}} - \frac{1}{6}\left.\frac{\partial U_{\mathrm{H}}}{\partial q_{\mathrm{H}}}\right|_{q_{\mathrm{H}}^0}. \tag{4.22}$$

The Hubbard derivative $\left.\frac{\partial U_{\mathrm{H}}}{\partial q_{\mathrm{H}}}\right|_{q_{\mathrm{H}}^0}$ is –0.1857 a.u. using DFT-PBE and the energy of the proton then equals 151.04 kcal/mol now overestimating the value of $-V_{\mathrm{H}}^{\mathrm{rep}}[\rho_{\mathrm{H}}^0]$.

In previous applications, approach (ii) has been used [23, 93, 94, 106], however, this may not be the best choice for general applications. In principle, the energy of the proton is just a constant and not relevant when relative proton affinities are of interest, as, for example, for proton transfer reactions. However, it becomes important when absolute proton affinities and $\mathrm{pK_a}$ are of interest [107]. In the following, it will be shown that fitting the Hubbard derivatives leads to a drastic improvement of the performance of the method. When Hubbard derivatives are fitted, as in the current work, it is more consistent to use eq 4.22. Computing the PA's with

$$E^{\mathrm{PA}} = E^{\mathrm{A}^-} + E^{\mathrm{H}^+} - E^{\mathrm{AH}} \tag{4.23}$$

and using a fixed value for the DFTB energy of the proton E^{H^+} means that a change in the Hubbard derivative due to fitting affects only the energy of the molecule AH (E^{AH}) such that the reference proton affinity E^{PA} is obtained irrespective of the potential well depth of the A–H bond. This problem is resolved when calculating E^{H^+} with eq 4.22, E^{H^+} being different for different hydrogen Hubbard derivatives.

For this reason the energy of the proton in the present work is consistently determined using eq 4.20 and eq 4.22 depending on the level of theory.

[4] Within the MIO parameter the spin-polarization energies are calculated with LDA, when using PBE values instead the electronic energy contribution for the hydrogen atom is –0.27966 a.u. which gives $V_{\mathrm{H}}^{\mathrm{rep}}[\rho_{\mathrm{H}}^0] = -136.83$ kcal/mol. For details see section 3.3.

Table 4.3: Mean Unsigned and Maximum Absolute Deviation of Geometrical Properties[a] of the G2 Set for 61 CHNO-Containing Closed Shell Molecules for Different DFTB Variants[b]

Property	N[c]	DFTB2		DFTB3					PBE[d]	B3LYP[d]
		γ	γ^{h}	γ	diag		full			
				calc	calc	fit	calc	fit		
r (Å)	223	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.009	0.004
$r^{\max}$ (Å)	223	0.065	0.067	0.061	0.064	0.064	0.062	0.063	0.060	0.041
a (deg)	187	0.9	0.9	0.9	0.9	0.9	0.9	1.0	0.4	0.4
$a^{\max}$ (deg)	187	4.7	6.4	4.9	6.4	6.2	6.6	6.5	1.9	1.9

[a] bond lengths r, bond angles a, max stands for maximum absolute deviation. Geometric data is compared to the MP2/cc-pVTZ calculations. [b] explanations see Table 4.1 and Table 4.2 [c] Number of comparisons [d] Basis set 6-311G(2d,2p)

4.4 Benchmarks and Discussion

In the following subsections benchmark calculations are presented for the different DFTB variants shown in Table 4.1 regarding geometries, binding energies, proton affinities and proton transfer barriers for CHNO containing molecules and also compare the results with commonly used density functionals. Further results are shown for proton affinities and hydrolysis reactions of phosphorus containing molecules.

The parameters used for the γ^{h}-function and third order terms are given in Table 4.2 and if not explicitly stated, the NHorg repulsive potential is used. Binding energies, proton affinities, proton transfer barriers, and reaction energies are computed using the potential energies at 0 K without including any zero-point energy correction. Deviations are given as the difference to high-level ab initio methods ($E^{\mathrm{method}} - E^{\mathrm{high\text{-}level}}$), where the high-level calculations are performed using the Gaussian 03 program [108].

The compilation and notation for binding energies and proton affinities are taken from ref [93], proton affinities and hydrolysis reactions of phosphorus containing molecules from ref [94].

4.4.1 Geometries

The performance for geometrical data of the different DFTB variants is tested for the charge-neutral closed-shell molecules of the G2/97 set [42]. The G2/97 set is a compilation of small molecules containing most relevant binding situations. For each molecule symmetry, bond lengths, angles and dihedrals are collected that provide a complete description of that molecule. The database for all geometrical properties used in this survey can be found in the supporting information of ref [20]. As shown in Table 4.3 the geometries do not change significantly for all tested DFTB variants and parameter sets. Similarly, the different NH repulsive potentials NHorg and NHmix cause only very small differences for geometries.

Significant differences occur for charged molecules, some of them are summarized in Table 4.4. For example, the C–C bond length in the acetate anion is overestimated by DFTB2 in comparison to B3LYP [39, 40, 75]/cc-pVTZ [109]; that error becomes smaller for the DFTB variants including third order terms. Similar findings are obtained for the O–H bond length of the hydroxide anion, even though in this case also the γ^{h}-function has a significant effect. DFTB2 underestimates hydrogen bond length in the water dimer indicating that the Pauli repulsion may be underestimated by DFTB. Inclusion of the γ^{h}-

Table 4.4: Deviation of DFTB in Comparison to B3LYP/cc-pVTZ for Selected Bond Lengths r in Å

Bond type	B3LYP	DFTB2 [a]		DFTB3 [a]				
		γ	γ^{h}	γ	diag		full	
				calc	calc	fit	calc	fit
r_{CC} in CH_3COO^-	1.567	+0.047	+0.051	+0.000	+0.011	–0.003	+0.005	–0.004
r_{HO} in OH^-	0.971	+0.033	+0.011	+0.005	–0.010	–0.009	–0.003	–0.003
r_{OH} in $(H_2O)_2$ [b]	1.945	–0.056	–0.122	–0.060	–0.125	–0.120	–0.116	–0.117

[a] explanations see Table 4.1 and Table 4.2 [b] r_{OH}: hydrogen bond length in water dimer

function even further shortens the hydrogen bond. It is important to note that this is a general trend (also valid for e.g. water clusters), i.e., hydrogen bond lengths are predicted systematically too short by DFTB.

4.4.2 Binding Energies

In a previous study it has been shown that DFTB2 underestimates the strength of hydrogen bonding interactions [93]. The performance for hydrogen bonds is drastically improved using the γ^{h}-function as shown in Table 4.5, while the third order corrections alone does not seem to have a substantial effect on these properties, however, the errors for the negative charged species are now more consistent with the ones of neutral and positive charged systems. The combination of both extensions in DFTB3-diag and DFTB3 adopts both improvements, the mean unsigned error in comparison to G3B3 [105] drops from 8 kcal/mol for DFTB2 to about 3 kcal/mol irrespective of the set of Hubbard derivative parameters (U^{d}) used. In ref [93] the test of DFTB3-diag have been extended to a larger test set, similar results for DFTB3 are expected.

In many biological applications, DFT methods with medium sized basis sets are applied. Calculating the binding energies given in Table 4.5 using PBE and B3LYP with the 6-31+G(d,p) basis set gives a mean unsigned error of 7.0 and 3.7 kcal/mol, respectively. Details are listed in appendix C. These errors are significantly larger when using basis sets without diffuse function. This of course is due to the basis set superposition error (BSSE) which can be remediated when including the counterpoise correction [110, 111], dropping the MUE to 3.7 and 1.3 kcal/mol, respectively. Nevertheless it is important to be aware of these large errors, for example, when studying larger bio-molecular systems where the counterpoise correction is rarely done. Therefore, although it is often claimed that certain DFT functionals perform well for hydrogen bonding [112, 113, 114, 115], this is only true for converged basis sets, which are often not used in practical applications. In such cases the use of a well calibrated approximate method like DFTB can be even a more appropriate choice. For example, the finding that the active site of bacteriorhodopsin is scrambled using QM/MM-CPMD simulation may be related to an imbalanced description of QM, QM/MM and MM interactions, where one factor contributing to the imbalance may be BSSE [116]. The application of empirical dispersion corrections would even worsen the problem, since dispersion further strengthens the interaction, i.e., leads to an even larger overbinding.

Table 4.5: 22 Binding Energies in kcal/mol: Deviation of DFTB in Comparison to G3B3[a]

System	G3B3	DFTB2[b]		DFTB3[b]				
		γ	γ^{h}	γ	diag		full	
				calc	calc	fit	calc	fit
$2H_2O$	–4.9	+1.6	–0.0	+1.5	–0.0	+0.2	–0.0	+0.0
$3H_2O$	–15.1	+5.5	–0.6	+5.4	–0.5	+0.2	–0.3	–0.1
$4H_2O$	–27.4	+9.7	+0.6	+9.4	+0.8	+1.8	+0.8	+1.1
$5H_2O$	–36.3	+13.3	+1.4	+12.5	+1.8	+3.0	+1.3	+1.7
$2H_2O(H^+)$	–33.9	+4.5	–2.0	+5.9	+2.4	+3.4	+0.9	+2.1
$3H_2O(H^+)$	–57.3	+10.4	–0.1	+11.6	+5.3	+6.6	+3.7	+5.3
$4H_2O(H^+)$	–77.2	+13.9	+1.1	+15.0	+6.4	+7.9	+5.0	+6.7
$5H_2O(H^+)$	–91.9	+18.3	+1.8	+19.7	+7.2	+9.1	+6.2	+8.1
$2H_2O(–H^+)$	–27.4	–5.1	–12.8	+1.5	–3.4	–1.5	–5.9	–3.2
$3H_2O(–H^+)$	–48.6	–2.6	–17.0	+5.3	–6.5	–3.8	–8.4	–5.3
$4H_2O(–H^+)$	–66.7	+0.3	–17.5	+9.0	–5.0	–1.8	–7.2	–3.5
$5H_2O(–H^+)$	–86.3	+6.1	–18.2	+14.2	–7.7	–4.1	–7.8	–4.7
$NH_3(H_2O)$ [c]	–6.6	+3.2	+2.1	+3.1	+2.0	+2.2	+2.1	+2.1
$NH_4^+(H_2O)$ [c]	–20.4	+0.6	–3.4	+1.4	–1.1	–0.7	–1.3	–0.9
6H2O_book	–45.8	+16.7	+1.2	+16.5	+1.5	+3.2	+1.7	+2.2
6H2O_cage	–46.6	+17.2	+0.3	+17.6	+0.3	+2.1	+1.5	+1.8
6H2O_prism	–47.2	+17.6	–0.0	+18.0	+0.1	+2.0	+1.3	+1.7
6H2O_ring	–44.7	+16.5	+1.8	+15.3	+2.4	+3.9	+1.5	+2.1
Methylimidazole$(–H^+)(H_2O)$	–15.9	+4.1	+2.0	+3.2	+1.4	+1.5	+1.2	+1.1
Methylimidazole(H_2O)_1	–6.2	+2.4	+1.4	+2.6	+1.8	+1.9	+1.9	+2.0
Methylimidazole(H_2O)_2	–8.2	+3.5	+2.6	+2.8	+2.0	+2.2	+1.9	+1.9
Methylimidazole$H^+(H_2O)$	–16.0	+3.3	+1.2	+3.9	+2.2	+2.3	+2.3	+2.5
MUE		8.0	4.0	8.9	2.8	3.0	2.9	2.7
MSE		+7.3	–2.5	+8.9	+0.6	+1.9	+0.1	+1.1
MAX		18.3	18.2	19.7	7.7	9.1	8.4	8.1

[a] The binding energy is computed as the difference between the complex and the isolated molecules at 0 K, no zero-point energy correction has been included. For the DFTB methods the deviation is given as the difference to the G3B3 method ($E^{\mathrm{method}} - E^{\mathrm{G3B3}}$). Compilation of molecules and notation taken from ref [93], Examples of notation: "$2H_2O$", neutral water dimer; "$2H_2O(H^+)$", protonated water dimer; "$2H_2O(–H^+)$", deprotonated water dimer, "$6H_2O$_book", neutral water hexamer in the book configuration; "Methylimidazole$(–H^+)(H_2O)$", deprotonated methylimidazole comlexed with water; "Methylimidazole(H_2O)_1", neutral methylimidazole complexed with water as hydrogen-bond donor; "Methylimidazole(H_2O)_2", neutral methylimidazole complexed with water as the hydrogen-bond acceptor; "Methylimidazole$H^+(H_2O)$", protonated methylimidazole complexed with water. [b] Explanations see Table 4.1 and Table 4.2. [c] When applying NHmix the results are slightly but not significantly different.

4.4.3 Proton Affinities

As shown in earlier studies [93, 117], DFTB2 overestimates proton affinities (PA) that implicate acidic oxygen. Yang et al. report an improvement with DFTB3-diag for molecules in which charge is strongly localized, a situation where the third-order term contributes accordingly. In these studies the DFTB2 energy of the proton was assumed to be 141.9 kcal/mol; in the present work eq 4.20 for DFTB2 and eq 4.22 for DFTB3 is used. Consequently the

Table 4.6: 23 Proton Affinities with Acidic Oxygen in kcal/mol: Deviation of DFTB in Comparison to G3B3[a]

System	G3B3	DFTB2[b]		DFTB3[b]				
		γ	γ^{h}	γ	diag		full	
				calc	calc	fit	calc	fit
H_2O	398.4	+16.3	+18.5	+8.0	+5.8	–1.6	+7.5	–1.8
$2H_2O$	375.9	+9.6	+5.7	+8.0	+2.3	–3.2	+1.7	–5.1
$3H_2O$	365.0	+8.1	+1.9	+7.7	–0.4	–5.6	–0.7	–7.1
$4H_2O$	359.1	+7.0	+0.3	+7.5	–0.1	–5.2	–0.5	–6.5
$5H_2O$	348.4	+9.2	–1.1	+9.6	–3.7	–8.7	–1.6	–8.3
CH_3OH	392.6	–5.7	–2.6	+3.3	+5.8	–0.7	+5.9	–0.3
CH_3CH_2OH	388.3	–1.5	+1.6	+6.5	+9.3	+2.6	+9.0	+2.2
$CH_3CH_2CH_2OH$	387.6	–2.2	+1.0	+6.0	+8.7	+1.9	+8.6	+2.0
CH_3–CH(OH)–CH_3	385.6	+1.4	+4.7	+8.2	+11.7	+4.6	+10.7	+3.3
HCOOH	351.2	+1.7	+3.4	+8.6	+14.2	+7.1	+10.0	+2.9
CH_3COOH	355.1	+1.1	+3.2	+6.8	+12.7	+5.6	+8.5	+0.6
CH_3CH_2COOH	354.5	+1.0	+3.4	+7.5	+13.1	+6.0	+9.3	+1.5
C_6H_5OH	356.7	–4.7	–2.4	+8.0	+11.0	+5.2	+9.7	+4.0
p-CH_3–C_6H_4OH	357.9	–5.6	–3.1	+7.4	+10.5	+4.5	+9.2	+3.7
p-NO_2–C_6H_4OH	334.6	–9.3	–7.5	+2.2	+5.2	–0.7	+3.5	–1.3
H_3O^+	171.2	–0.4	–4.7	+10.6	+9.0	+5.6	+6.3	+4.3
$2H_2O(H^+)$	200.2	–3.3	–2.7	+6.3	+6.6	+2.3	+5.4	+2.2
$3H_2O(H^+)$	213.4	–5.3	–5.1	+4.4	+3.2	–0.9	+2.3	–1.2
$4H_2O(H^+)$	221.1	–4.7	–5.3	+4.9	+3.3	–0.7	+2.0	–1.4
$5H_2O(H^+)$	226.7	–5.3	–5.1	+3.5	+3.8	–0.4	+1.4	–1.9
$CH_3OH_2^+$	186.8	–8.3	–10.3	+6.5	+6.1	+2.0	+4.6	+2.2
H_2COH^+	177.1	–11.8	–13.8	+4.3	+4.3	+0.5	+2.6	–0.2
CH_3CHOH^+	190.2	–10.1	–10.8	+5.8	+6.5	+2.4	+5.1	+2.0
MUE		5.8	5.1	6.6	6.8	3.4	5.5	2.9
MSE		–1.0	–1.3	+6.6	+6.5	+1.0	+5.2	–0.2
MAX		16.3	18.5	16.6	14.2	8.7	10.7	8.3

[a] The molecules are given in the protonated form. The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction. For the DFTB methods the deviation is given as the difference to the G3B3 method ($E^{\mathrm{method}} - E^{\mathrm{G3B3}}$). The compilation of the molecules is taken from ref [93]. [b] Explanations see Table 4.1 and Table 4.2.

proton affinities as compiled in Table 4.6 are shifted by about 10 kcal/mol for DFTB2 in comparison to the earlier studies.

While the mean signed error (MSE) for DFTB2 in comparison to G3B3 is quite small, the proton affinities of negatively charged molecules are overestimated and the proton affinities for neutral molecules underestimated. This holds true also when including the γ^{h}-function. The situation changes when looking at the third order variants. Even though the MUE is not significantly reduced (or even enlarged) in comparison to DFTB2 the proton affinities for almost all molecules are consistently overestimated and the MSE is (almost) as large as the MUE (+5.2 vs 5.5 kcal/mol in the case of DFTB3/calc). This indicates a consistent overbinding of the O–H bond. This error, however, is not related to the third order formalism but has its roots already in the repulsive potential of DFTB2. As Otte et al [14] mentioned,

Table 4.7: 9 Proton Affinities with Acidic Nitrogen in kcal/mol: Deviation of DFTB and the NHorg Parameter Set in Comparison to G3B3[a]

System	G3B3	DFTB2[b]		DFTB3[b]				
		γ	γ^{h}	γ	diag		full	
				calc	calc	fit	calc	fit
$HCNH^+$	176.0	–12.4	–14.6	+4.5	+4.3	+0.4	+2.8	+0.2
CH_3CNH^+	192.3	–14.3	–15.4	+2.9	+2.6	–1.2	+1.9	–0.9
$C_5H_5NH^+$	229.5	–17.1	–18.3	+1.3	+0.9	–3.5	+0.4	–2.1
MethylimidazoleH$^+$	237.3	–12.7	–13.4	+5.3	+5.1	+0.8	+4.7	+2.1
MethylguanidineH$^+$	249.3	–12.0	–13.4	+0.8	+0.4	–2.2	–0.8	–2.9
NH_3	413.9	+10.4	+10.9	–0.9	–16.8	–0.0	–5.3	–0.2
NH_4^+	212.3	–24.4	–30.5	–9.2	–13.0	–15.0	–14.4	–15.8
$CH_3NH_3^+$	223.3	–26.8	–30.5	–10.2	–11.7	–15.1	–13.3	–15.3
1-AminobutaneH$^+$	228.2	–26.7	–29.9	–9.8	–11.4	–14.9	–12.6	–14.6
MUE		17.4	19.7	5.0	7.4	5.9	6.2	6.0
MSE		–15.1	–17.2	–1.7	–4.4	–5.6	–4.1	–5.5
MAX		26.8	30.5	10.2	16.8	15.1	14.4	15.8

[a] The molecules are given in the protonated form. The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction. For the DFTB methods the deviation is given as the difference to the G3B3 method ($E^{\mathrm{method}} - E^{\mathrm{G3B3}}$). The compilation of the molecules is taken from ref [93]. [b] Explanations see Table 4.1 and Table 4.2.

the O–H bond shows an overbinding of about 6–7 kcal/mol. This overbinding can also be roughly estimated by the half of the atomization energy error of H_2O, which is 5.8 kcal/mol for DFTB3/calc (using PBE spin-polarization energies, details see section 3.3) in comparison to G3B3. This value is very similar to the MSE of DFTB3/calc (5.2 kcal/mol) in Table 4.6, leading to the conclusion that removing this overbinding remedies the error for the proton affinities. Indeed, once fitting the third order and γ^{h}-function parameters (DFTB3/fit) the MSE can be removed to obtain a MUE as small as 2.9 kcal/mol.[5] This would not work for DFTB2, indicating that the third order terms lead to a systematic improvement of DFTB.

Proton affinities with acidic nitrogen are shown in Table 4.7. DFTB2 shows large errors which are systematically improved by all third order variants. Large errors remain for the last three molecules in Table 4.7 with sp^3-nitrogen, which show a systematic error of more than 10 kcal/mol, as discussed in detail already in ref [93]. The use of H–N-mod specifically for sp^3-hybridized nitrogen, although not satisfactory from a theoretical point of view, remedies this problem (Table 4.8). The remaining MUE for DFTB3/calc/NHmix is only 2.5 kcal/mol.

Another encouraging result is the improvement of the proton affinity for NH_2^-. While for DFTB2 the proton affinity is overestimated, it is underestimated for DFTB3-diag. The error is then substantially reduced using the full third order variants, DFTB3-γ and DFTB3, showing the first example where the third order off-diagonal terms seem to be of importance.

[5] For water clusters, the proton affinity can be written as the sum of the proton affinity for a (neutral or protonated) water molecule and the difference in the binding energies of water clusters of different protonation states. Therefore, the errors in the water cluster PAs can be understood in terms of the errors in the PA of a single (neutral or protonated) water and errors in the binding energies of the relevant water clusters. For example, the fairly large error for the PA of a neutral $(H_2O)_5$ is due mainly to the fact that DFTB3/fit overestimates the binding energy of a deprotonated $(H_2O)_5$ (–4.7 kcal/mol, see Table 4.5) but slightly underestimates the binding energy of a neutral $(H_2O)_5$ (+1.7 kcal/mol, see Table 4.5).

Table 4.8: 9 Proton Affinities with Acidic Nitrogen in kcal/mol: Deviation of DFTB and the NHmix Parameter Set in Comparison to G3B3[a]

System	G3B3	DFTB2[b]		DFTB3[b]				
		γ	γ^{h}	γ	diag		full	
				calc	calc	fit	calc	fit
$HCNH^+$	176.0	–12.4	–14.6	+4.5	+4.3	+0.4	+2.8	+0.2
CH_3CNH^+	192.3	–14.3	–15.4	+2.9	+2.6	–1.2	+1.9	–0.9
$C_5H_5NH^+$	229.5	–17.1	–18.3	+1.3	+0.9	–3.5	+0.4	–2.1
MethylimidazoleH$^+$	237.3	–12.7	–13.4	+5.3	+5.1	+0.8	+4.7	+2.1
MethylguanidineH$^+$	249.3	–12.0	–13.4	+0.8	+0.4	–2.2	–0.8	–2.9
NH_3	413.9	+10.4	+10.9	–0.9	–16.8	–0.0	–5.3	–0.2
NH_4^+	212.3	–13.1	–19.5	+2.0	–2.0	–3.8	–3.4	–4.8
$CH_3NH_3^+$	223.3	–15.4	–19.2	+1.2	–0.4	–3.8	–2.0	–4.0
1-AminobutaneH$^+$	228.2	–15.3	–18.6	+1.6	–0.1	–3.5	–1.2	–3.3
MUE		13.6	15.9	2.3	3.6	2.1	2.5	2.3
MSE		–11.3	–13.5	+2.1	–0.7	–1.9	–0.3	–1.8
MAX		17.1	19.5	5.3	16.8	3.8	5.3	4.8

[a] The molecules are given in the protonated form. The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction. For the DFTB methods the deviation is given as the difference to the G3B3 method ($E^{\mathrm{method}} - E^{\mathrm{G3B3}}$). The compilation of the molecules is taken from ref [93]. [b] Explanations see Table 4.1 and Table 4.2.

Due to the hybridization problem the error analysis for H–N bond is more involved. Nevertheless, the overbinding of the H–N bond calculated as a third of the error in the atomization energy of NH_3 for DFTB3/calc as compared to G3B3 is 2.9 kcal/mol (using PBE spin-polarization energies, details see section 3.3), which is comparably small. With the H–O overbinding of 5.8 kcal/mol, the error for the relative proton affinity between oxygen and nitrogen containing molecules can be estimated to be roughly (5.8–2.9) kcal/mol = 2.9 kcal/mol, which is an important measure for the accuracy of proton transfer energetics between different donor and acceptor species.

Also DFT methods with medium sized basis sets for proton affinities are benchmarked. The mean unsigned error of PBE/6-31+G(d,p) and B3LYP/6-31+G(d,p) in comparison to G3B3 is 4.7 kcal/mol and 2.5 kcal/mol, which is comparable to the performance of DFTB3. Note that the use of diffuse functions is essential here, and errors for calculations without diffuse functions are much larger (for details see appendix C). For example, the use of HF/4-31G for the description of a proton transfer reaction may not yield a correct description of the dynamics due to errors in the PAs of donor and acceptor [118].

Overall one can find a clear difference in the performance of DFTB2 and DFTB3 due to the inclusion of the third order terms, whereas DFTB3-diag and DFTB3 perform very similar on proton affinities. As for the binding energies, Yang et al. [93] compiled larger test sets and showed that DFTB3-diag overall improves the description of proton affinities. This is true for both, using calculated Hubbard derivatives or fitted Hubbard derivatives. With these findings a similar behavior for DFTB3 is expected. An improved performance for both, hydrogen binding energies and proton affinities of DFTB2 is only found when including both extensions, γ^{h}-function and third order terms. Therefore, further benchmark tests are shown in the following for the combination of these extensions and also the improvement of DFTB3 over DFTB3-diag will be discussed.

4.4.4 Proton Transfer Barriers

For testing proton transfer barriers, several simple models are considered. For the O–H–O models, a proton is placed between two water molecules and between two hydroxide anions. The barriers are calculated for a fixed oxygen-oxygen distance with the shared proton at half the distance between both oxygens. All other hydrogen atoms are geometry optimized. For the relaxed structure, the shared proton is allowed to relax. While for the cationic complex the barriers calculated with MP2 [119]/G3large [120] are already well reproduced with DFTB2, large errors occur for the anionic model for large O–O distance. These errors are completely removed for both DFTB3-diag and DFTB3. Table 4.9 summarizes the results, DFTB3 notably improves over popular DFT methods with an intermediate basis set.

Similarly, proton transfer barriers for nitrogen species are tested. DFTB2 underestimates the barriers severely while DFTB3/calc reduces this error and even slightly overestimates the barrier for the negatively charged complex. The DFTB models with fitted parameters show further improved results.

The proton transfer barriers for the models containing one oxygen and one nitrogen atom are computed keeping both heavy atoms fixed and translating the shared proton along the straight line between oxygen and nitrogen. The barrier is then given by the highest energy surrounded by two minima. For the relaxed structure the shared proton is again geometry optimized together with all other hydrogen atoms. Rather large deviations are found for DFTB2 which are reduced with DFTB3-diag/calc and DFTB3/calc. Again, an overall good performance is found for the DFTB3-diag/fit and DFTB3/fit versions; the largest errors appear for $[NH_3–H–OH_2]^+$, where surprisingly DFT-GGA methods also reveal comparably large errors (see Table 4.9) in comparison to MP2/G3large.

The use of the H–N-mod potential has the following consequences on barriers. The H–N bond is energetically shifted by about 10 kcal/mol being more attractive in the binding region. The strength of the bond decreases with larger H–N distances. As a consequence no barrier can be found for the models containing one oxygen and one nitrogen with small N–O distances. Here it becomes obvious that H–N-mod is not parametrized and not applicable to proton transfer barriers. Nevertheless, H–N-mod is a practical solution for correcting errors for proton affinities as has been shown in several applications, e.g. ref [99]. For models with two nitrogen atoms very similar results for original H–N and the H–N-mod potential are found. Future work will have to concentrate on solving the hybridization problem and balancing H–N and H–O repulsive potentials such that proton transfer barriers with oxygen and nitrogen participation are described correctly.

To point out this more clearly, $[H_3N–H–OH_2]^+$ is a model for a proton transfer between an amino acid with an acidic nitrogen in the side chain (lysine, histidine, arginine) and an oxygen. In the model the nitrogen is sp^3-hybridized as would be the case for lysine and an error is found of about 10 kcal/mol; therefore, H–N-mod should be used. However, when doing so the barrier vanishes, i.e., the energy monotonically rises as the hydrogen moves towards oxygen. Thus, proton affinities and proton transfer barriers can be well described for systems including histidine or arginine (applying the original H–N potential) but special care must be taken for lysine. Using the H–N-mod potential the proton affinity is described well, but not the barrier height of a proton transfer. The same problem arises for DNA proton transfer reactions, where the MUE is of about 5 kcal/mol for proton affinities of DNA bases as found for DFTB3-diag in ref [93].

Table 4.9: Proton Transfer Barrier in kcal/mol for a Fixed Distance (rXY) between the Heavy Atoms (X,Y∈{O,N}): Deviation of DFTB and DFT in Comparison to MP2/G3large[a]

				DFTB3-diag[b]		DFTB3[b]			
Barrier	rXY	MP2	DFTB2	calc	fit	calc	fit	PBE[c]	B3LYP[c]
$[H_2O–H–OH_2]^+$	2.5	0.6	–0.6	–0.6	–0.6	–0.6	–0.6	–0.5	–0.4
	2.6	2.4	–1.4	–0.2	+0.0	–1.1	–0.8	–1.9	–1.0
	2.7	5.2	–1.1	+0.5	+0.8	–0.7	–0.3	–3.2	–1.5
	2.8	8.9	–1.3	+0.5	+0.9	–0.9	–0.4	–4.4	–1.9
$[HO–H–OH]^-$	2.5	0.5	–0.5	–0.5	–0.5	–0.5	–0.5	–0.5	–0.2
	2.6	2.3	–2.3	–0.6	–0.2	–0.7	+0.0	–1.6	–0.6
	2.7	5.2	–4.6	–0.0	+0.5	–0.2	+0.9	–2.7	–0.9
	2.8	8.8	–6.7	–0.0	+0.7	–0.3	+1.2	–3.7	–1.2
$[H_3N–H–NH_3]^+$	2.6	0.4	–0.4	–0.4	–0.4	–0.4	–0.4	–0.4	–0.3
	2.7	1.9	–1.8	–1.0	–1.4	–1.7	–1.7	–1.9	–1.0
	2.8	4.4	–2.4	–0.5	–1.4	–2.0	–2.2	–3.0	–1.6
	2.9	7.7	–2.5	–0.1	–1.2	–2.1	–2.2	–4.1	–2.2
$[H_2N–H–NH_2]^-$	2.5	0.1	–0.0	+0.3	–0.1	–0.1	–0.1	+0.0	+0.0
	2.6	1.4	–1.4	+5.2	–1.1	–0.4	–1.0	–1.2	–0.5
	2.7	3.5	–3.5	+5.5	–0.6	+1.5	–0.5	–2.1	–0.8
	2.8	6.3	–4.9	+6.9	+0.6	+2.9	+0.9	–3.0	–1.1
$[H_3N–H–OH_2]^+$ [d]	2.9	25.3	–8.2	–5.2	–4.2	–5.8	–5.3	—[e]	–2.7
	3.0	30.0	–9.5	–6.3	–5.4	–7.3	–6.7	–6.8	–3.3
	3.1	35.1	–11.0	–7.7	–6.8	–8.9	–8.2	–7.8	–3.8
	3.2	40.5	–12.5	–9.1	–8.2	–10.5	–9.8	–8.8	–4.2
$[H_2O–H–NH_3]^+$ [d]	2.9	0.8	+0.7	+2.2	+1.1	+0.2	+0.1	—[e]	–0.8
	3.0	3.3	+0.9	+2.9	+1.5	+0.2	+0.1	–3.0	–1.6
	3.1	6.7	+0.3	+2.3	+0.9	–0.6	–0.7	–4.5	–2.3
	3.2	10.7	–0.8	+1.3	–0.2	–1.9	–2.0	–5.9	–2.9
$[H_2N–H–OH]^-$ [d]	2.8	10.1	–6.3	–4.6	+2.4	–2.5	+2.9	–3.3	–0.8
	2.9	14.2	–8.6	–5.2	+2.1	–3.1	+2.7	–4.2	–1.1
	3.0	18.6	–11.1	–6.0	+1.4	–3.9	+2.1	–5.1	–1.4
	3.1	23.3	–13.4	–6.7	+0.7	–4.7	+1.5	–5.9	–1.6
$[HO–H–NH_2]^-$ [d]	2.8	4.4	–4.1	+14.4	+0.3	+6.6	+0.7	–2.8	–1.2
	2.9	7.8	–5.2	+15.8	+1.1	+7.9	+1.6	–4.0	–1.7
	3.0	11.6	–6.6	+16.5	+1.2	+8.4	+1.7	–4.9	–2.0
	3.1	15.8	–8.8	+16.2	+0.1	+7.8	+0.7	–5.8	–2.3

[a] Barriers are computed as described in the text at 0 K and no zero-point energy correction has been included. For the DFT and DFTB methods the deviation is given as the difference to the MP2 method ($E^{\text{method}} - E^{\text{MP2}}$). For all models the NHorg parameter set is applied. [b] Explanations see Table 4.1 and Table 4.2. [c] Basis set 6-31+G(d,p). [d] Barrier in comparison to the relaxed structure with the proton binding to the heavy atom that is written on the left hand side of that proton. [e] A barrier does not exist.

Table 4.10: 18 Proton Affinities for Phosphorus Containing Molecules in kcal/mol: Deviation of DFTB in Comparison to G3B3[a]

			DFTB3-diag[c]		DFTB3[c]	
Molecule[b]	G3B3	DFTB2	calc	fit	calc	fit
H_3PO_4	334.0	+17.1	+23.9	+18.5	+18.3	+5.5
$H_2PO_4^-$	464.5	+26.2	+26.6	+20.0	+17.2	–4.3
DMPH[d]	336.3	+9.6	+20.3	+14.9	+15.9	+4.8
MMP[d]	336.7	+12.0	+20.9	+15.4	+15.8	+3.8
MMP^- [d]	460.5	+21.5	+26.3	+19.9	+18.3	–1.2
PH_3OH^+	201.6	–8.6	+7.1	+2.7	+4.8	–0.0
PH_2OHOH^+	201.6	–2.8	+10.8	+6.3	+8.2	+2.1
$PHOHOHOH^+$	200.8	+4.5	+16.3	+11.7	+13.6	+6.2
$PH_2(OH)=O$	336.6	+3.0	+15.7	+10.4	+12.2	+3.3
PH(OH)(OH)=O	334.7	+10.7	+20.4	+15.0	+16.0	+5.3
P(O)(OH)(–O–CH_2CH_2–O–)	336.3	+7.2	+17.7	+12.3	+13.4	+2.4
P(OH)(OH)(–O–CH_2CH_2–O–)(OH^*)	359.0	–3.3	+18.6	+12.8	+12.9	+0.3
P(OH^*)(OH)(–O–CH_2CH_2–O–)(OH)	350.4	+6.7	+15.9	+10.6	+11.0	–0.4
P(OH^*)(OH)(–O–CH_2CH_2–O–)(OCH_3)	351.2	+1.8	+12.8	+7.4	+8.9	–1.4
P(OH)(OCH_3)(–O–CH_2CH_2–O–)(OH^*)	359.6	–8.3	+7.9	+2.1	+3.3	–7.2
P(OH^*)(OCH_3)(–O–CH_2CH_2–O–)(OH)	352.9	+3.6	+14.5	+9.2	+10.1	–0.5
P(OH)(OH)(OH)(OH^*)(OH)_ax	357.3	+4.0	+21.8	+15.9	+14.2	–1.2
P(OH)(OH)(OH)(OH^*)(OH)_eq [e]	347.0	+14.0	—	—	—	–0.0
MUE		9.2	17.5	12.1	12.6	2.8
MSE		+6.6	+17.5	+12.1	+12.6	+1.0
MAX		26.2	26.6	20.0	18.3	7.2

[a] The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction. For the DFTB methods the deviation is given as the difference to the G3B3 method ($E^{\text{method}} - E^{\text{G3B3}}$). The compilation of the molecules is taken from ref [94]. [b] The molecules are given in the protonated form. [c] Explanations see Table 4.1 and Table 4.2. [d] "DMPH" refers to dimethyl hydrogen phospate, "MMP" to P(O)(OH)(OH)(OCH_3), and "MMP^-" to P(O)(O)(OH)$(OCH_3)^-$ [e] The molecule P(OH)(OH)(OH)(OH^*)(OH)_eq dissociates forming H_2O for DFTB3-diag and DFTB3/calc. Depending on the basis set this dissociation also occurs for the DFT functionals PBE and B3LYP, e.g. dissociation for basis set 6-311G(2d,2p), no dissociation for basis set cc-pVTZ.

4.4.5 Phosphorus Containing Molecules

Proton Affinity and Hydrolysis Energetics

For phosphorus containing molecules, two tests are carried out. Table 4.10 shows 18 proton affinities of biological relevance, in Table 4.11 elementary steps for a representative set of phosphate hydrolysis reactions are listed, which include the hydrolysis of monophosphate ester (MMP) and dimethyl monophosphate ester (DMP) with different protonation states, number of water involved, and dissociative/associative mechanisms. In Figure 4.4 selected structures are illustrated. The complete set of geometrical properties can be found in the supporting information of ref [20].

Proton affinities are generally overestimated for DFTB2 and DFTB3/calc. The MUE shows even worse results for DFTB3/calc than for DFTB2; however, similar as for the proton affinities with acidic oxygen, the deviation in comparison to the MP2 results are much

Table 4.11: Deviation of Exothermicity and Barrier Height from the DFTB Variants in Comparison to MP2/G3Large Single Point Calculations at B3LYP/6-31+G(d,p) Structures for 37 Elementary Steps in the Hydrolysis of MMP and DMP[a]

			DFTB3-diag [c]		DFTB3 [c]	
Process[b]	MP2	DFTB2	calc	fit	calc	fit
com1 → ts1 (MMP,B)	31.0	–1.0	–3.9	–4.2	–6.7	–7.2
com1 → int1 (MMP,E)	30.6	–2.2	–2.4	–2.6	–6.1	–7.0
com1 → ts1_2 (MMP,B)	41.5	+0.8	–1.7	–1.9	–3.9	–3.3
com1 → int1_2 (MMP,E)	31.0	–4.4	+0.5	+0.3	–4.6	–5.9
int1_2 → ts2_0 (MMP,B)	11.9	–3.0	–3.0	–3.0	–0.2	+2.7
int1_2 → ts2 (MMP,B)	3.6	–5.5	–4.9	–4.9	–2.8	–1.1
int1_2 → com2 (MMP,E)	–28.8	+2.1	+0.0	+0.1	+3.3	+4.4
com1 → diss_tsa (MMP,B)	36.8	+4.6	+2.2	+2.4	+3.3	+4.9
com1 → diss_int (MMP,E)	19.6	–7.3	–7.3	–6.7	–3.2	+0.3
com1_w2 → ts1_2_w2 (MMP,B)	39.9	–8.3	–12.3	–12.1	–13.6	–12.7
com1_w2 → int1_2a_w2 (MMP,E)	28.0	–5.0	–1.5	–1.8	–6.6	–7.9
int1_2a_w2 → int1_2_w2 (MMP,E)	0.4	+0.3	+1.7	+1.5	+2.5	+2.5
int1_2_w2 → ts2_0_w2 (MMP,B)	11.4	–4.1	–10.8	–10.5	–7.7	–5.4
com1_da → ts1_da (MMP,B)	55.0	–22.4	–8.8	–9.1	–7.1	–0.2
com1_da → int_da (MMP,E)	4.5	–3.3	+0.6	+0.5	–1.6	–2.0
com1 → ts1 (DMP,B)	38.6	–1.7	–5.3	–5.8	–7.1	–7.3
com1 → int1 (DMP,E)	35.4	–5.7	–2.2	–2.4	–6.4	–7.6
int1 → int1_2 (DMP,E)	1.3	–3.2	–0.3	–0.3	+0.4	+1.3
int1_2 → ts2 (DMP,B)	0.6	+0.2	–0.5	–0.6	+0.4	+1.2
int1_2 → com2 (DMP,E)	–35.2	+7.0	+4.0	+4.1	+5.9	+6.0
n_com1 → n_ts3 (DMP,B)	33.6	+4.9	+3.0	+2.7	+0.7	+0.2
n_com1 → n_int1 (DMP,E)	13.2	–3.7	–0.0	–0.4	–3.8	–4.9
n_int1 → n_ts4 (DMP,B)	22.9	+6.4	+4.1	+4.0	+4.8	+5.1
n_int1 → n_com2 (DMP,E)	–15.8	+2.2	–0.1	+0.0	+3.2	+3.7
DMP_P → diss_ts (DMP,B)	40.9	+11.6	+8.3	+8.1	+8.8	+8.8
DMP_P → diss_prod (DMP,E)	28.2	+0.6	–1.7	–1.8	–0.0	–0.9
diss_prod2 → diss_ts2 (DMP,B)	13.5	+13.2	+11.3	+11.2	+9.8	+10.4
diss_prod2 → MMP_P (DMP,E)	–29.8	+0.8	+2.9	+2.8	+0.3	+0.6
diss_w_reac → diss_w_ts (DMP,B)	20.9	+5.9	+2.2	+2.2	+2.3	+0.7
diss_w_reac → diss_w_prod (DMP,E)	18.4	+4.3	+0.2	+0.1	+1.4	+0.3
diss_w_prod2 → diss_w_ts2 (DMP,B)	1.9	+2.7	+2.2	+2.2	+0.9	+0.3
diss_w_prod2 → diss_w_reac2 (DMP,E)	–21.0	–2.5	+0.4	+0.4	–1.4	–0.7
n_w_com1 → n_w_ts3 (DMP,B)	28.2	–2.7	–4.4	–4.4	–7.6	–8.9
n_w_com1 → n_w_int1 (DMP,E)	13.1	–4.1	–0.7	–1.0	–4.5	–5.8
n_w_int1 → n_w_int2 (DMP,E)	–0.5	+0.1	+0.7	+0.7	+0.7	+0.6
n_w_int2 → n_w_ts4 (DMP,B)	15.1	+2.0	–2.2	–2.0	–2.3	–3.1
n_w_int2 → n_w_com2 (DMP,E)	–13.0	+1.4	–0.7	–0.7	+3.0	+4.1
MUE		4.4	3.2	3.2	4.0	4.1
MSE		–0.5	–0.8	–0.9	–1.2	–0.9
MAX		22.4	12.3	12.1	13.6	12.7

[a] Compilation from ref [94]; no zero-point corrections are included in either exothermicity or barrier heights. All quantities are given in kcal/mol. [b] The processes are labeled as in ref [94], "E" stands for "Exothermicity", "B" for "Barrier", all structures are listed in detail in the supporting information of ref [20]. [c] Explanations see Table 4.2.

more consistent with DFTB3. For DFTB3 all proton affinities are overestimated, the MSE being as large as the MUE (12.6 kcal/mol); additionally, the maximal absolute deviation is smaller than that for DFTB2. As discussed above, about 6 kcal/mol of this error is due to the overbinding of the H–O bond, which can be removed by refitting the H–O repulsive potential. The remaining error may be reduced by fitting the P Hubbard derivative. For now, the method of choice is DFTB3/fit, for which the MUE is only 2.8 kcal/mol. Important is that the Hubbard derivatives for C, H, N, and O are taken from the fit on non-phosphate molecules and only $U_{\mathrm{P}}^{\mathrm{d}}$ is fitted to the 18 listed proton affinities; this is in contrast to earlier work, where the best performance for DFTB3-diag could only be achieved by fitting *all* parameters at once.

Table 4.10 also shows the results for DFTB3-diag/calc which looks similar to DFTB3/calc. However, when using DFTB3-diag/fit the error cannot be reduced as much as is the case for DFTB3/fit. The parameter $U_{\mathrm{P}}^{\mathrm{d}}$ has very small influence on the proton affinities. The MUE ranges from 10.7–12.5 kcal/mol when choosing $U_{\mathrm{P}}^{\mathrm{d}}$ in the range of –0.40 to –0.04 atomic units, therefore, the calculated parameter $U_{\mathrm{P}}^{\mathrm{d}} = -0.07$ H is kept. This observation highlights that DFTB3-diag/fit does not properly account for some part of the interactions within these molecules, i.e., the flexibility of the model is not sufficient to yield good results for non-phosphate and phosphate molecules at the same time.

For the hydrolysis reactions the MUE for DFTB2 is 4.4 kcal/mol and is only slightly reduced for DFTB3/fit (Table 4.11). Note that for the latter the parameter $U_{\mathrm{P}}^{\mathrm{d}}$ is fitted to the proton affinities only. A special fit also for these reactions does not improve this situation significantly. Surprisingly, DFTB3-diag performs somehow superior with a MUE of 3.2 kcal/mol.

Additional Discussion of Transferability of Parameters

Earlier extensions of DFTB2 have suggested a lack of general transferability, for example, the two phosphorus related parameter sets (see additional discussions in the next subsection), SCC-DFTBPA and SCC-DFTBPR [94], need to be developed for different properties. Both

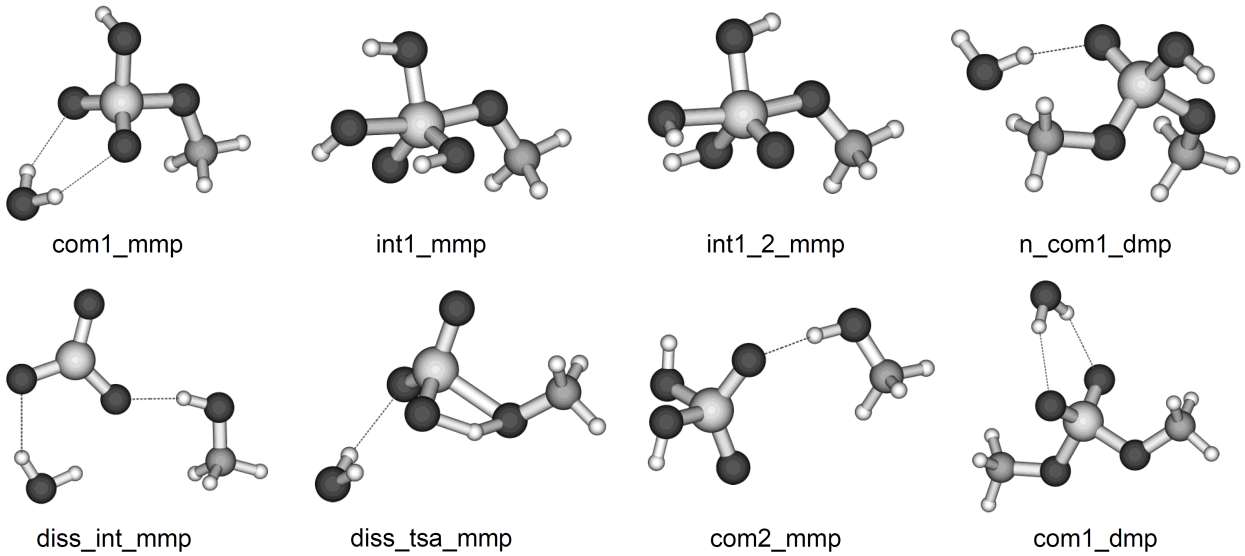


Figure 4.4: Selected structures for the phosphate hydrolysis reactions.

sets are based on DFTB3-diag (without the γ^{h}-function) with fitted Hubbard derivatives and an additional empirical Gaussian term (with three additional parameters) to adjust the Hubbard derivatives within the SCC procedure. SCC-DFTBPA is specifically designed for proton affinities of phosphorus containing molecules and yields a MUE for the 18 proton affinities of Table 4.10 of only 2.6 kcal/mol but performs inferior for proton affinities of non-phosphate molecules. SCC-DFTBPR, on the other hand, is designed for the hydrolysis reactions of Table 4.11 and shows a MUE for these reactions of only 2.4 kcal/mol but is less accurate for the proton affinities (in particular for non-phosphate molecules).

DFTB3 is a consistent extension of the DFTB2 methodology and transferable to a wide range of chemical properties. Instead of different methods with a different number of parameters (six or nine parameters additionally to the ones from DFTB2) DFTB3 is a method that shows an overall good performance for binding energies and proton affinities of non-phosphate and phosphate molecules using only six additional parameters in comparison to DFTB2 (ζ, U_{C}^{d}, U_{H}^{d}, U_{N}^{d}, U_{O}^{d}, U_{P}^{d}). A limitation is found, however, for the hydrolysis reactions, for which only a slight improvement is achieved in comparison to DFTB2. DFTB3 is not performing as well as SCC-DFTBPR with that respect, which suggests that further improvements are necessary for the phosphorus parameters and/or for the DFTB formalism.

4.5 Conclusion

DFTB3 is a new method that extends the standard second order DFTB2 (formerly SCC-DFTB) by two conceptually independent improvements. DFTB3 maintains the strengths of DFTB2, such as rapid computation of large scale molecular systems with reliable geometry, but improves transferability and overall accuracy for several properties.

The first concept is the γ^{h}-function ameliorating the electron-electron interaction of charge fluctuation. The γ^{h}-function corrects the original function, which incorrectly imposes a linear relation between the chemical hardness and the atomic size. This relationship is only valid within one row of the periodic table, and particularly fails when interactions of first row atoms with hydrogen are involved. Therefore one additional, purely empirical parameter (ζ) is introduced which can be adjusted to a single reference system like the water dimer. Previous tests have shown that this improves the performance of DFTB2 for hydrogen bonding systematically. Therefore, this correction does not introduce additional terms to total energy in an *ad hoc* fashion, but establishes a consistent improvement of the electron-electron interaction in the second (and third) order terms of DFTB2 (DFTB3). As a result, the mean unsigned error for hydrogen binding energies drops from 8.0 kcal/mol for DFTB2 to 4.0 kcal/mol for DFTB2-γ^{h} for a fairly broad set of test systems. A drawback is found for the hydrogen bond lengths which turn out to be too short.[6]

The second improvement concerns the extension of DFTB2 to include third order terms of the Taylor series expansion of the DFT exchange-correlation energy. The third order terms cause the chemical hardness (Hubbard parameter) of an atom to be dependent on its charge, which becomes particularly important for the description of systems with localized charges. One additional parameter is introduced for each element, the Hubbard derivative with respect to charge, which can be either computed from DFT for atoms or can be fitted. With the first approach, the DFTB3-γ method does not involve any new empirical parameters. Geometries

[6] In principle, these repulsive energy potentials are intimately coupled to the electronic DFTB terms with which they have been determined. Therefore, they have to be refitted when the DFTB Hamiltonian is modified.

for charged molecules are slightly improved. Regarding proton affinities the errors become consistently overestimated in contrast to an underestimation for negatively charged systems and an overestimation for positively charged systems with DFTB2.

The combination of both improvements into DFTB3 also combines the effects. The accuracy of DFTB2 for geometries of C, H, N, O containing molecules is maintained. For charged molecules a slight geometrical improvement is found whereas hydrogen bonds are consistently too short. The mean unsigned error for a set of 22 hydrogen binding energies drops below 3.0 kcal/mol. It should be noted that this improved DFTB model outperforms standard DFT functionals using medium sized basis sets without correction for BSSE, a methodology typically used in, for example, QM/MM applications to biological systems.[7] For proton affinities the mean unsigned error is not significantly reduced when using calculated Hubbard derivatives. However, the remaining errors arise not due to third order approximations (and neither the γ^{h}-function) but result from the repulsive potential terms of second order DFTB2; i.e., the errors could be removed in principle by re-optimizing the DFTB repulsive potentials, with which an empirical fitting of the Hubbard derivative parameters is likely no longer necessary. For the time being, empirically fitted parameters are presented (Hubbard derivatives and ζ), which result in a significant improvement over DFTB2. The mean unsigned deviation for oxygen containing test systems in comparison to G3B3 results are 5.8, 2.9, and 2.5 kcal/mol for DFTB2, DFTB3 with fitted ζ and Hubbard derivatives, and B3LYP/6-31+G(d,p), respectively.

Further it is found in present work that the energy of a proton is a constant and not equal to zero for DFTB2 (and DFTB3) due to the neglect of atomic contributions within the repulsive energy contribution. There are different eligible ways of how to compute this constant leading to different constants in order to obtain an absolute proton affinity. For reasons of consistency the constant as calculated directly from the respective level of theory (DFTB2 or DFTB3) is used. It should be noted that for most applications only relative proton affinities are important, i.e. the value of this constant does not matter at all. Only for specific applications where the absolute proton affinity is needed the value of that constant becomes important, e.g. determining the $\mathrm{pK_a}$ of a molecule. An empirical but helpful choice different than fitting parameters would then be to use a constant which compensates the consitent over- or underestimation of the respective DFTB variant.

In earlier work, the diagonal part of the third order corrections has been implemented and tested and several different parametrizations are provided (Hubbard derivatives, ζ and in some cases also additional parameters) [93, 94]. In a comparison between DFTB3-diag and DFTB3, the newly implemented off-diagonal terms do not seem to lead to a large improvement for molecules consisting of O, N, C and H, except for the $\mathrm{NH_2}$ molecule, since the diagonal part is already quite accurate. The most significant advantage of DFTB3 over DFTB3-diag and earlier published extensions of DFTB2 is its consistent performance for hydrogen bonding energies and proton affinities including atoms of type C, H, N, O, and P. While all earlier extensions needed different parameterizations for different properties, DFTB3 with fitted ζ and Hubbard derivatives is more transferable and covers all properties with a single parametrization. One persistent limitation is found for phosphate hydrolysis reactions, where a model based on DFTB3-diag with an empirical Gaussian term and "reaction specific" parametrization of the Hubbard derivatives (SCC-DFTBPR) [94] is still needed for better accuracy.

[7] Note that adding empirical dispersion corrections to DFT-GGA would even worsen the situation, since DFT over-binds the H-bonded complexes already due to BSSE.

Despite all progress major limitations for DFTB3 remain. First, the error of proton affinities of nitrogen containing molecules seems to correlate with the hybridization state of nitrogen. The use of different repulsive potentials, NHorg and NHmix, is discussed which provides a pragmatic way for calculating accurate proton affinities but is unreliable for studying reactions and proton transfer barriers. Moreover, the scheme is conceptually unsatisfactory. Second, the hydrogen bond lengths are generally too short, and third, large errors are found for non-isodesmic reactions of phosphorus containing species. Addressing these limitations new set of electronic and repulsive parameters for DFTB3 is developed in chapter 6.

Chapter 5

Partially Automatized Scheme for Parametrizing DFTB

During the MIO parametrization of DFTB2 the fitting of the repulsive potentials turned out to be the most time consuming part. This is due to the fact that only one property and one potential at a time could be constructed (section 3.3). For several properties, e.g. single, double, triple bond energetics in the C–C potential case, the potential had to be constructed by piecewisely merging the resulting potentials together. First and second derivatives of a potential, being important for atomic forces and vibrational frequencies, are only considered indirectly due to the stretching of bonds in a dense mesh and using the total energy of each structure. Further energy shifts of the potentials had to be introduced in order to yield proper reaction energetics and also assure adequate cut-off radii. On the technical side very often numerical noise appeared in the second derivate.

Thus, a first requirement to the new parametrization scheme is to include the possibility to fit to several properties (energies, forces, vibrational frequencies) of a set of molecules (e.g. ethane, ethene, and ethyne in the case of the C–C potential) without the cumbersome manual work. Moreover, the repulsive potentials should be constructed in a numerically stable way.

Earlier studies have shown that the DFTB2 performance for distinct properties can be partially tweaked by more or less elaborate fitting procedures. For example Witek and coworkers could significantly improve vibrational frequencies [88]. A question arises whether DFTB2 performance can be systematically improved for all the considered properties using better fitting strategies for the repulsive potentials, or, whether there are optimization conflicts, in which one property is improved on the cost of others. In general, such a behavior is expected since the electronic part constitutes an approximation to full DFT, and its limited flexibility (minimal basis, charge self-consistency, fixed reference density, etc.) may lead to a limited transferability and conflicts in parameter optimization. In this chapter, the aim is to explore this point in detail, using a new representation of repulsive potentials and an improved fitting strategy. The proposed methodology is applied to construct a set of repulsive potentials for C and H, which are subsequently tested in a systematic fashion using large sets of molecules.

The following section describes the details of the algorithm for semi-automatized construction of the repulsive DFTB2 potentials V^{rep}. The procedure uses a set of molecular equilibrium geometries together with corresponding atomization energies as input. Because both, the energies and forces associated with the employed molecular structures, can be represented as linear functions of repulsive potentials, a solution of an inverse problem is

capable—in principle—to yield a set of V^{rep} that reproduces the input quantities. In a second section the fitting procedure is applied to hydrocarbons for finding repulsive potentials and also electronic parameters for the DFTB2 formalism. Finally, the experience on the appropriate degree of automatization is summarized and conclusions are drawn on the ability of the fitting procedure for systematically improving the DFTB2 performance for hydrocarbons.

5.1 Formalism

In this section an alternative derivation of the formalism for constructing repulsive potentials is provided than published in [121]. First, a detailed description of the respresentation of repulsive potentials is given. Second, a linear equation system is defined including so called continuity equations, constraints to be exactly fulfilled, and other equations to be fitted to. In a third subsection it is shown how to solve this equation system fulfilling the constraints and solving the other equations in a least square sense. For an easier understanding a simple example illustrates the equations system in a fourth subsection followed by some additional notes.

5.1.1 Analytical Representation of Repulsive Potentials

In the MIO parametrization as described in section 3.3, the repulsive potentials V_{ab}^{rep} between atoms a and b are parametrized for pairs of atom types $T_a = A$ and $T_b = B$ (i.e. $V_{ab}^{\text{rep}} = V_{T_aT_b}^{\text{rep}^*} \quad \forall a, b$) and are represented by cubic splines. Here, the repulsive potentials are represented as fourth-order splines. This choice ensures that the first- and second-order derivatives of the repulsive potentials—necessary for geometry optimization and for computing harmonic vibrational frequencies—will be given as sufficiently smooth functions of interatomic distance. The previous choice of cubic splines yielded piecewisely linear functions for the second-order derivatives that had not enough flexibility to reproduce adequately the Hessian of the repulsive potentials. As shown below, where the exact construction of spline functions is given in more detail, using fourth-order splines does not introduce any additional complexity in comparison to the traditional third-order approach.[1]

The definition of the repulsive potential $V_{AB}^{\text{rep}^*}$ between atom types A and B starts with dividing the range of possible atomic separations $[0, \infty)$ into a set of intervals $I_{AB0} = [0, r_{AB1})$, $I_{AB1} = [r_{AB1}, r_{AB2})$, ..., $I_{ABn} = [r_{ABn}, r_{AB(n+1)})$, and $I_{AB(n+1)} = [r_{AB(n+1)}, \infty)$ using a set of division points $r_{AB1}, r_{AB2}, \ldots, r_{AB(n+1)}$. Subsequently, for each of the intervals I_{AB1}, I_{AB2}, ..., I_{ABn}, a fourth-order polynomial is defined by

$$S_{ABi}(r) = \sum_{k=0}^{4} s_{ABik}(r - r_{ABi})^k, \tag{5.1}$$

where r denotes the interatomic distance and the unknowns s_{ABik} are to be determined. The total number of unknowns for one repulsive potential $V_{AB}^{\text{rep}^*}$ is thus $5n_{AB}$ (n can be different for different pairs AB, however, $V_{AB}^{\text{rep}^*} = V_{BA}^{\text{rep}^*}$!). $S_{AB(n+1)}$ as well as its first three derivatives are required to be identically equal to zero over the whole interval $I_{AB(n+1)}$. Therefore, the last division point, $r_{AB(n+1)}$, can be interpreted as a cut-off, beyond which the repulsive

[1] In principle, higher order splines could be used, however preliminary tests have found that a higher order leads to technical disadvantages concerning the flexibility of the spline function.

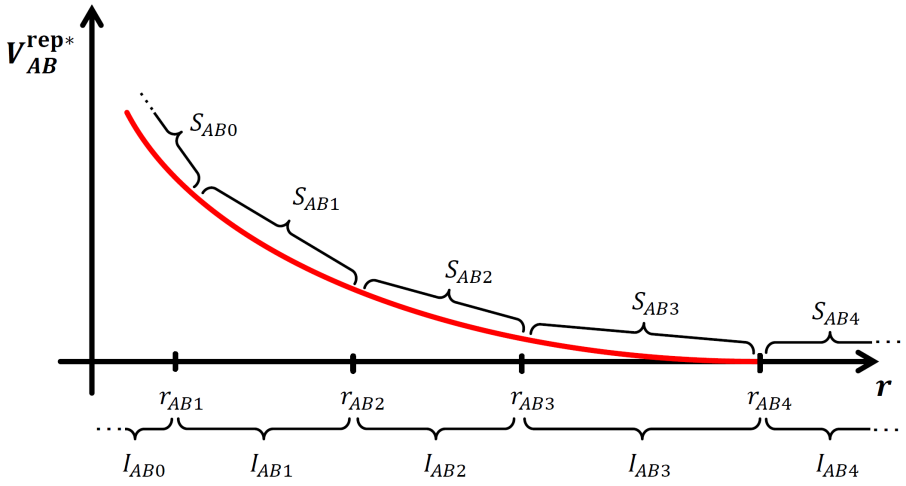


Figure 5.1: Scheme of a repulsive potential for the atom type pair AB with $n = 3$ including an exponential part S_{AB0}, a spline function with three polynomial sections S_{AB1}, S_{AB2}, and S_{AB3}, and a cut-off r_{AB4} beyond which the repulsive potential is defined to be $S_{AB4} = 0$.

potential vanishes. To conclude this part of this presentation, it should be mentioned that in interval I_{AB0} the function S_{AB0} is defined as

$$S_{AB0}(r) = \alpha_{AB} \exp(\beta_{AB} r + \gamma_{AB}), \tag{5.2}$$

where the three parameters α_{AB}, β_{AB}, and γ_{AB} are chosen to match the corresponding values of S_{AB1} and its two lowest derivatives at $r = r_{AB1}$. Figure 5.1 summarizes the representation of the repulsive potential for a simple example.

5.1.2 Defining a Linear Equation System

For finding the unknowns there are several objectives. At first, the definition of the spline function has to be fullfilled, i.e. the continuity equations, further the fitting is carried out to reproduce atomization energies, atomic forces and vibrational frequencies. All these objectives can be expressed in linear dependence of the unkown spline coefficients s_{ABik} forming a linear equation system which can be written in matrix form as

$$M \cdot \boldsymbol{s} = \boldsymbol{b}, \tag{5.3}$$

where vector $\boldsymbol{s}$ contains the unknowns s_{ABik}, each line of matrix M together with the corresponding entry of the reference vector $\boldsymbol{b}$ represents one objective to the problem. The number of columns of matrix M is defined by the number of unknowns s_{ABik}. For each intervall I_{ABi} $(i = 1, 2, ..., n_{AB})$ of each atom type pair AB there are 5 columns $(k = 0, 1, 2, 3, 4$ for the fourth order splines as defined in eq 5.1).

Continuity Equations

The definition of the splines is that the spline function is continously differentiable up to the second derivative in the interval $[r_{AB1}, r_{AB(n+1)})$. Thus one can write

$$\begin{aligned}
S_{ABi}(r_{AB(i+1)}) - S_{AB(i+1)}(r_{AB(i+1)}) &= 0, & (5.4)\\
S'_{ABi}(r_{AB(i+1)}) - S'_{AB(i+1)}(r_{AB(i+1)}) &= 0, & (5.5)\\
S''_{ABi}(r_{AB(i+1)}) - S''_{AB(i+1)}(r_{AB(i+1)}) &= 0, & (5.6)\\
S'''_{ABi}(r_{AB(i+1)}) - S'''_{AB(i+1)}(r_{AB(i+1)}) &= 0, & (5.7)
\end{aligned}$$

where the primes denote subsequent derivatives and $i = 1, 2, ..., n_{AB}$. With the definition of the spline function eq 5.1 these conditions can be easily written as $4n_{AB}$ equations for each pair-potential $V_{AB}^{\mathrm{rep}^*}$ into the equation system. Note that $S_{AB(n+1)}(r_{AB(n+1)})$ and derivatives are equal zero. An example is given in subsection 5.1.4.

Energy Equations

The DFTB2 total energy of a molecule M can be divided into two parts, the electronic part E^{el} given by the first two terms of eq 3.32 and the repulsive part E^{rep} given by the third term of eq 3.32

$$E^{\mathrm{DFTB2}} = E^{\mathrm{el}} + E^{\mathrm{rep}}. \tag{5.8}$$

The atomization energy E^{at} of the molecule M is given by (see also eq 3.34)

$$E^{\mathrm{at}} = E^{\mathrm{DFTB2}} - \sum_a E_a^{\mathrm{el}} \tag{5.9}$$

where the summation runs over all atoms constituting the molecule M. A simple rearrangement of these equations gives

$$E^{\mathrm{rep}} = \sum_{a,b>a} V_{ab}^{\mathrm{rep}}(r_{ab}) = E^{\mathrm{at}} - E^{\mathrm{el}} + \sum_a E_a^{\mathrm{el}}. \tag{5.10}$$

The right hand side of eq (5.10) can be treated as a constant because it does not depend on the sought spline coefficients. The atomization energy E^{at} needs to be an appropriate reference value, E^{el} can be calculated from eq 3.32, and the values of atomic electronic energies E_a^{el} on the right hand side of eq (5.10) are usually computed as (see also eq 3.35)

$$E_a^{\mathrm{el}} = \sum_i n_i \epsilon_i^a + E_a^{\mathrm{spin}}, \tag{5.11}$$

where E_a^{spin} denotes the spin polarization energy of atom a, n_i the orbital occupation number and ϵ_i^a the atomic Kohn-Sham eigenvalues.

To derive the working energy equation for molecule M, the left hand side of eq 5.10 is expressed in terms of the unknowns. The first division points $r_{AB1} (= r_{T_a T_b 1})$ are chosen to be smaller than the smallest distance r_{ab} between atoms a and b, respectively. Thus, using eq 5.1 the left hand side of eq 5.10 can be written as

$$E^{\mathrm{rep}} = \sum_{a,b>a} V_{ab}^{\mathrm{rep}}(r_{ab}) = \sum_{a,b>a} \sum_{k=0}^{4} s_{T_a T_b pk} (r_{ab} - r_{T_a T_b p})^k \tag{5.12}$$

where $T_a = A$ and $T_b = B$ denote atom types and each r_{ab} lies in some interval $I_{ABp} = [r_{ABp}, r_{AB(p+1)})$. Note that for interatomic distances that are larger than the cut-off radius ($r_{ab} > r_{AB(n+1)}$) the repulsive energy contribution equals zero. An illustrative example of an energy equation is shown in subsection 5.1.4.

It is possible to replace E_a^{el} with empirical atomic electronic energies, i.e. one parameter $E_A^{\mathrm{el}^*} = E_{T_a}^{\mathrm{el}^*}$ for each atom type. The actual values will be determined via the fitting process described below. With this ansatz, eq (5.10) can be written as

$$\sum_{a,b>a} V_{ab}^{\mathrm{rep}}(r_{ab}) - \sum_a E_{T_a}^{\mathrm{el}^*} = E^{\mathrm{at}} - E^{\mathrm{el}}, \tag{5.13}$$

where the quantities on the left hand side depend on the unknown parameters and the quantities on the right hand side are constants. Both fitting strategies, i.e., with using the calculated atomic electronic energies given by eqs 5.10 and 5.11 and optimizing atomic electronic energies (eq 5.13), will be employed in the following.

Force Equations

The force vector can be decomposed in a similar way as the DFTB2 total energy 5.8 as

$$\boldsymbol{F}^{\mathrm{DFTB2}} = \boldsymbol{F}^{\mathrm{el}} + \boldsymbol{F}^{\mathrm{rep}}. \tag{5.14}$$

At the equilibrium geometry this vector is identically equal to zero and can be written as

$$\boldsymbol{F}^{\mathrm{rep}} = -\boldsymbol{F}^{\mathrm{el}}. \tag{5.15}$$

While the right hand side can be calculated and enters the reference vector $\boldsymbol{b}$ of the total equation system (eq 5.3), the repulsive part of the force vector on atom a can be expressed through the first derivatives of the repulsive potentials as

$$\boldsymbol{F}_a^{\mathrm{rep}} = -\sum_{b\neq a} \frac{\boldsymbol{r}_a - \boldsymbol{r}_b}{r_{ab}} \left.\frac{\mathrm{d}V_{ab}^{\mathrm{rep}}}{\mathrm{d}r}\right|_{r=r_{ab}}. \tag{5.16}$$

For r_{ab} lying inside the interval I_p ($1 \leq p \leq n$) the derivative in eq 5.16 is given by (compare also eq 5.12)

$$\left.\frac{\mathrm{d}V_{ab}^{\mathrm{rep}}}{\mathrm{d}r}\right|_{r=r_{ab}} = \sum_{k=1}^{4} s_{T_aT_bpk} \cdot k \cdot (r_{ab} - r_{T_aT_bp})^{k-1}, \tag{5.17}$$

Note, the first division points $r_{AB1} = r_{T_aT_b1}$ are chosen to be smaller than the smallest distances r_{ab} and for $r_{ab} > r_{AB(n+1)}$ the derivative equals zero.

The total number of force equations (eq 5.15) for the molecule M is $3n$, where n is the total number of atoms constituting the molecule M and 3 a factor corresponding to each cartesian coordinate. An example is discussed in subsection 5.1.4.

In principle it is also possible to employ non-equilibrium structures in the presented fitting procedure. The only complication is that for such a non-equilibrium structure, the total force vector in eq 5.14 is not equal to zero but equal to a reference total force vector which must be calculated by some quantum chemical method.

Additional Equations

It is sometimes convenient to impose some special conditions on the fitted repulsive potentials. A requirement that a repulsive potential $V_{AB}^{\mathrm{rep}^*}$ has a particular value (say V_z) at some particular point (say r_z) can be easily expressed in a similar fashion to the previously discussed energy and force equations. For the point r_z inside the interval I_{ABp} with r_{ABp} being the corresponding division point one can write

$$V_{AB}^{\mathrm{rep}^*}(r_z) = V_z = S_{ABp}(r_z) = \sum_{k=0}^{4} s_{ABpk}(r_z - r_{ABp})^k. \tag{5.18}$$

It is easy to see that similar additional conditions can be imposed not only on repulsive potentials, but also on their derivatives. Note that particularly useful in practice are the

additional conditions imposed on the second order derivatives of the repulsive potentials at some particular points because they help to adjust the values of harmonic vibrational frequencies of a given type of vibration. An example for an additional equation adjusting the H–O stretching fequencies is given in Figure 5.3.

Reaction Equations

A simple modification of the presented energy equations enables fitting the repulsive potentials also to energies of chemical reactions. For example a simple reaction between molecules M_1 and M_2 is given as

$$n_1 M_1 + n_2 M_2 \rightarrow n_3 M_3 + n_4 M_4. \tag{5.19}$$

The aim is to determine the shape of the repulsive DFTB2 potentials that yields the correct value of reaction energy E^{rea} for this reaction. For this purpose, four energy equations for molecules M_1, M_2, M_3, and M_4 can be constructed as described in eq (5.13). Adding these equations with the appropriate reaction coefficients yields a working equation for reaction energies:

$$\sum_m \nu_m \sum_{a,b>a} V_{ab}^{\mathrm{rep}}(r_{ab}) = \sum_m \nu_m \left(E_m^{\mathrm{at}} - E_m^{\mathrm{el}}\right) \tag{5.20}$$

where ν_m is the appropriate reaction coefficient for molecule m (i.e. $\nu_1 = n_1$, $\nu_2 = n_2$, and $\nu_3 = -n_3$, $\nu_4 = -n_4$), and a and b are atomic indices referring to molecules M_m. Similar as in eqs 5.10 and 5.12, the left hand side of eq 5.20 can be expressed in terms of the coefficients s_{ABik}. Note that the right hand side of eq 5.20 does not depend explicitly on the atomic electronic energies E_a^{el} since they cancel out upon addition.

5.1.3 Solving the Linear Equation System

As described in the previous subsection (eq 5.3) an equation system including the unknowns s_{ABik} is defined as

$$M \cdot \boldsymbol{s} = \boldsymbol{b}. \tag{5.21}$$

The equations can be divided into 1) conditions that are necessary to be fulfilled exactly, those are the continuity equations, and 2) conditions designed to fit the curve to, those are energy, force, additional, and reaction equations. Among 2) the conditions may contradict to each other due to a) the limits of the electronic part of DFTB and approximations of the two-center approach for the repulsive energy, b) limited flexibility of the repulsive potential, and c) inaccurate reference values. Thus, when fitting to these objectives it is desired to yield a curve with lowest deviation in the least square sense. In the following, several steps are performed to exactly fulfill the continuity equations by reducing the number of unknowns and to fit the remaining unknowns to all other conditions in the least square sense.

First, the equation system is seperated into the continuity equation part and the other equation part as

$$\begin{bmatrix} M^{\mathrm{cont}} \\ M^{\mathrm{fit}} \end{bmatrix} \cdot \boldsymbol{s} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{b}^{\mathrm{fit}} \end{bmatrix} \tag{5.22}$$

Second, the vector of unknowns $\boldsymbol{s}$ is separated into two vectors, one containing the unknowns which are implicitly determined by the continuity equations ($\boldsymbol{s}^{\mathrm{impl}}$) and the other containing the unknowns which have to be fitted explicitly to the other conditions ($\boldsymbol{s}^{\mathrm{expl}}$). When inspecting eq 5.1 one finds that for one spline function of fourth order there are four continuity equations (5.4–5.7) and five unknowns per interval. Thus, $\boldsymbol{s}^{\mathrm{expl}}$ consists of one

unknown for each interval of a spline segment[2], $\boldsymbol{s}^{\text{impl}}$ includes the remaining unknowns. Separation of the matrices M^{cont} and M^{fit} in the same way leads to

$$M^{\text{cont,impl}}\boldsymbol{s}^{\text{impl}} + M^{\text{cont,expl}}\boldsymbol{s}^{\text{expl}} = \mathbf{0} \tag{5.23}$$

$$M^{\text{fit,impl}}\boldsymbol{s}^{\text{impl}} + M^{\text{fit,expl}}\boldsymbol{s}^{\text{expl}} = \boldsymbol{b}^{\text{fit}}. \tag{5.24}$$

Third, eq 5.23 can be rearranged to

$$\boldsymbol{s}^{\text{impl}} = -\left(M^{\text{cont,impl}}\right)^{-1} M^{\text{cont,expl}}\boldsymbol{s}^{\text{expl}} \tag{5.25}$$

and inserted into eq 5.24

$$\left(M^{\text{fit,impl}}\left(M^{\text{cont,impl}}\right)^{-1} M^{\text{cont,expl}} + M^{\text{fit,expl}}\right) \cdot \boldsymbol{s}^{\text{expl}} = \boldsymbol{b}^{\text{fit}} \tag{5.26}$$

giving a final equation system which can be rewritten using the abbreviation M^{final} for the terms in brackets as

$$M^{\text{final}} \cdot \boldsymbol{s}^{\text{expl}} = \boldsymbol{b}^{\text{fit}}. \tag{5.27}$$

Fourth, the set of linear equations in eq 5.27 is usually either over- or underdetermined (compare also subsection 5.2.1). A solution can be found in the most numerically stable fashion via singular value decomposition (SVD) [122] of matrix M^{final}. Thus, eq 5.27 can be rewritten as

$$M^{\text{final}} \cdot \boldsymbol{s}^{\text{expl}} = U\Sigma V^{\text{T}}\boldsymbol{s}^{\text{expl}} = \boldsymbol{b}^{\text{fit}} \tag{5.28}$$

where U and V are orthogonal square matrices and Σ is a diagonal rectangular matrix containing the singular values sorted from the largest to the smallest one. Assuming that all singular values are different than zero, the solution $\boldsymbol{s}^{\text{expl}}$ can be expressed as

$$\boldsymbol{s}^{\text{expl}} = V\Sigma^{-1}U^{\text{T}}\boldsymbol{b}^{\text{fit}}. \tag{5.29}$$

It might happen that some singular values are zero or close to zero. In these cases, the singular values are truncated to the first t nonzero numbers. This procedure can be considered as using the concept of a generalized inverse to solve eq 5.27. The solution $\boldsymbol{s}^{\text{expl}}$ is given by the same formula as that in eq 5.29, with V and U truncated to their first t columns and Σ^{-1} being the inverse of the upperleft $t \times t$ segment of Σ.

For an underdetermined equation system singular values will be obtained being equal zero. The SVD-technique then solves for vector $\boldsymbol{s}^{\text{expl}}$ to be of minimum length, i.e. no vector lying in the nullspace is contributing to the obtained solution [122]. For an overdetermined equation system there is no solution but it can be shown that eq 5.29 constructs a "solution" vector $\boldsymbol{s}^{\text{expl}}$ being closest to the set of equations in the least square sense [122]. However, some singular values might be very small. A small singular value means that some equations are not quite but almost linear dependent and cause only a small contribution to the least square error (residual)

$$r^{\text{res}} = \left|M^{\text{final}} \cdot \boldsymbol{s}^{\text{expl}} - \boldsymbol{b}^{\text{fit}}\right| \tag{5.30}$$

but add a linear combination of variables to the solution vector with possibly large values. That might deteriorate the performance of the unknowns for objectives that are not directly included in the fit. To prevent this, small singular values and their inverse are set to zero.

[2]Technically in this work the coefficient of the highest power for each interval of all spline functions is written to the vector $\boldsymbol{s}^{\text{expl}}$ which would be the unknowns s_{abi4} in eq 5.1. However, this choice is arbitrary, also the coefficients of any other power could be used.

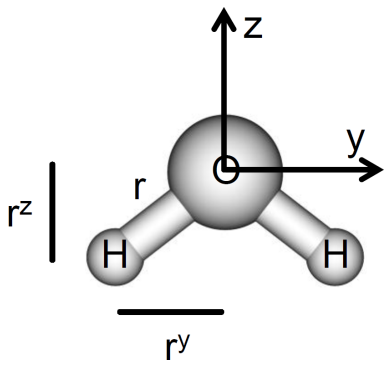


Figure 5.2: H_2O molecule of C_{2v} symmetry lying in the yz-plane.

For the applications shown in this work singular values being smaller than a threshold of 10^{-4} are set equal zero and also their inverse are set equal zero.

Finally, while eq 5.29 determines all the unknowns $\boldsymbol{s}^{\text{expl}}$, the vector $\boldsymbol{s}^{\text{impl}}$ is computed from eq 5.25.

5.1.4 Illustrative Example of Fitting $V_{\mathbf{HO}}^{\mathbf{rep}}$

In this subsection an example of an equation system as derived in the previous subsections is given. For keeping it simple an equation system for only one repulsive potential is constructed, the H–O potential. In addition, only three objectives are considered in order to properly describe a reference H_2O molecule at equilibrium (Figure 5.2) with respect to a reference atomization energy, H–O bond length and symmetric H–O stretching frequency.[3]

Three intervals are chosen denoted by I_1, I_2, and I_3 where the H–O bond length of the H_2O molecule r lies within the first interval I_1. With that choice a solution can be found that exactly fulfills all three objectives. Using eqs 5.1 and 5.4–5.7 then form twelve continuity equations, four for each interval. In Figure 5.3 the definition $\Delta_i = r_{i+1} - r_i$ $(i = 1, 2, 3)$ is used representing the difference between consecutive spline division points. For example, the first line in Figure 5.3 can be derived using eqs 5.4 and 5.1 as (since only one potential is considered the indices A and B are dropped)

$$\begin{aligned} & S_1(r_2) - S_2(r_2) = 0 \\ =& \sum_{k=0}^{4} s_{1k}(r_2 - r_1)^k - \sum_{k=0}^{4} s_{2k}(r_2 - r_2)^k = 0 \\ =& \sum_{k=0}^{4} s_{1k}\Delta_1^k - s_{2k} = 0 \\ =& s_{10} + s_{11}\Delta_1 + s_{12}\Delta_1^2 + s_{13}\Delta_1^3 + s_{14}\Delta_1^4 - s_{20} = 0 \end{aligned} \tag{5.31}$$

The energy equation in Figure 5.3 can then be derived starting from eq 5.10 as

$$E^{\text{rep}} = E^{\text{at}} - E^{\text{el}} + \sum_a E_a^{\text{el}} = E^{\text{ref}}, \tag{5.32}$$

[3] In the following either experimental data is used as reference or atomization energies are calculated with G3B3 [105] and equilibrium geometries are determined from B3LYP/cc-pVTZ. Even though this example seems very simplified, parts of the more involved examples as shown in subsection 5.2.1 are very similar to this simple problem.

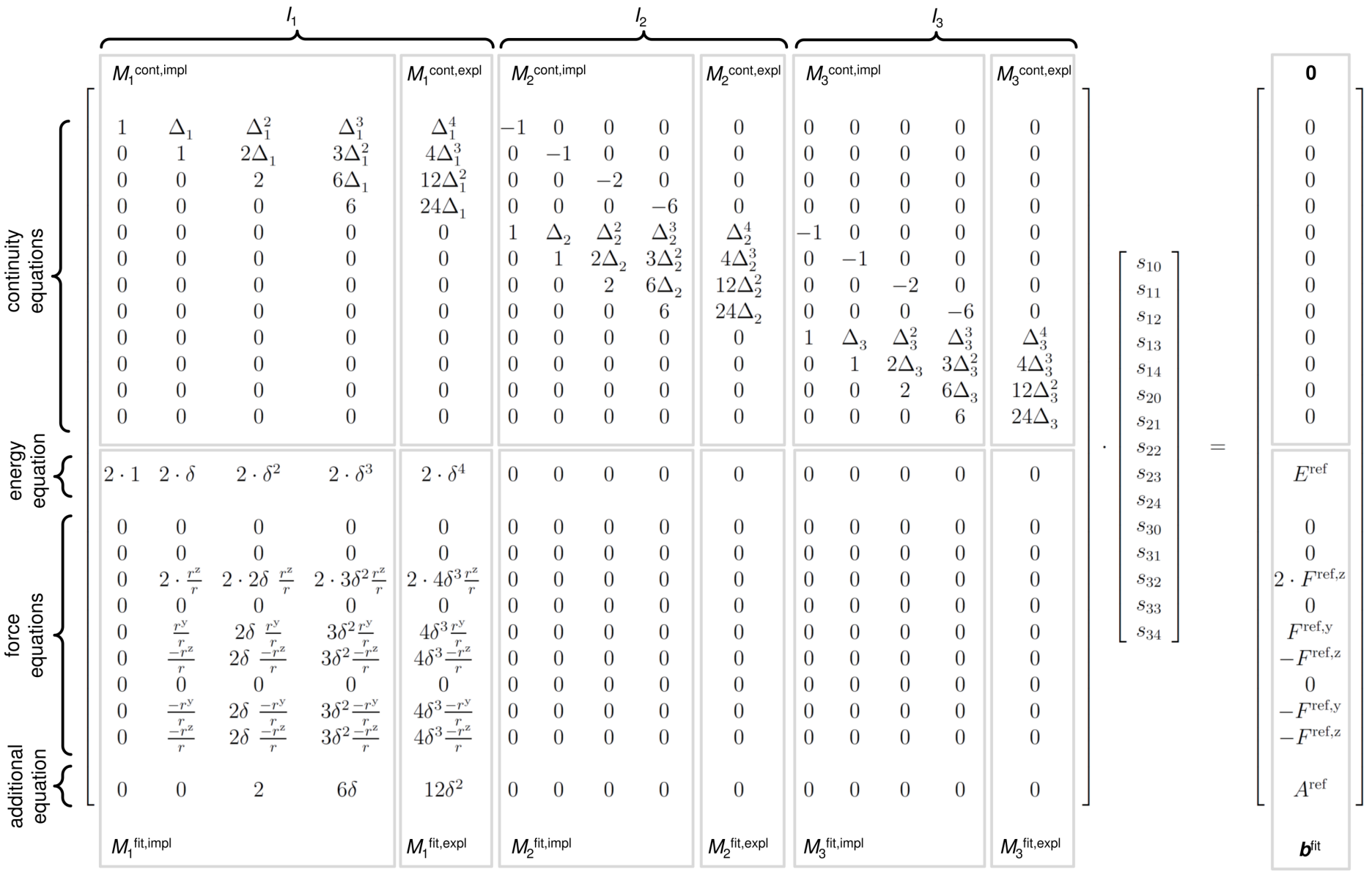


Figure 5.3: Example of an equation system for determining a single repulsive potential.

E^{ref} being an abbreviation for the right hand side. Following eq 5.12 the repulsive energy can be written as

$$E^{\mathrm{rep}} = \sum_{a,b>a} V^{\mathrm{rep}}(r_{\mathrm{HO}}) = 2V^{\mathrm{rep}}(r_{\mathrm{HO}}) = E^{\mathrm{ref}}, \tag{5.33}$$

where the factor two enters because two H–O interactions of the water molecule with exactly the same interatomic distance $r = r_{\mathrm{HO}}$ contribute to the repulsive energy. Note that the H–O potential gives the only contribution to the repulsive energy (and its derivatives) of the H_2O molecule. For the H–H potential a cut-off point is defined to be at a distance smaller than the H–H distance in the H_2O molecule (second neighbor distance). That is not only done to construct an easy example but is used as general strategy as will be further discussed in subsection 5.2.1. Remembering that r_{HO} lies in the first interval I_1, eq 5.33 can be recasted to (eq 5.12)

$$E^{\mathrm{rep}} = 2\sum_{k=0}^{4} s_{1k}(r_{\mathrm{HO}} - r_1)^k = E^{\mathrm{ref}}, \tag{5.34}$$

and with $\delta = r_{\mathrm{HO}} - r_1$ being the difference between the H–O bond length r_{HO} and the first spline division point, the energy equation in Figure 5.3 is given by

$$2s_{10} + 2\delta s_{11} + 2\delta^2 s_{12} + 2\delta^3 s_{13} + 2\delta^4 s_{14} = E^{\mathrm{ref}}. \tag{5.35}$$

Nine force equations are listed in Figure 5.3, one for each atom and one cartesian coordinate as follows from eqs 5.15–5.17. Because an equilibrium structure is taken as reference, the right hand side of the equation shown in Figure 5.3 (arbitrarily named $F^{\mathrm{ref,y}}$ and $F^{\mathrm{ref,z}}$) corresponds to the negative electronic force of eq 5.15. For example, the first three force equations stem from the forces acting on the oxygen atom. While the forces in x- and y-directions are equal zero (first and second line of the force equations in Figure 5.3) as can be easily derived from inspecting Figure 5.2, the force in z-direction can be calculated as

$$F_{\mathrm{O}}^{\mathrm{rep,z}} = -2\frac{r_{\mathrm{HO}}^{\mathrm{z}}}{r_{\mathrm{HO}}} \left.\frac{\mathrm{d}V^{\mathrm{rep}}}{\mathrm{d}r}\right|_{r=r_{\mathrm{HO}}} = -2F^{\mathrm{ref,z}} \tag{5.36}$$

Again, the factor two stems from the two H–O bonds involved, and the indices A and B are dropped in comparison to eq 5.16. For the sake of simplicity eq 5.36 is multiplied by (-1). Using $\delta = r_{\mathrm{HO}} - r_1$ and writing eq 5.36 in terms of the unknowns (eq 5.17) yields

$$2\frac{r_{\mathrm{HO}}^{\mathrm{z}}}{r_{\mathrm{HO}}} \sum_{k=1}^{4} s_{1k} k \left(r_{\mathrm{HO}} - r_1\right)^{k-1} = 2F^{\mathrm{ref,z}} \tag{5.37}$$

$$= 2\frac{r_{\mathrm{HO}}^{\mathrm{z}}}{r_{\mathrm{HO}}} s_{11} + 2\frac{r_{\mathrm{HO}}^{\mathrm{z}}}{r_{\mathrm{HO}}} 2\delta s_{12} + 2\frac{r_{\mathrm{HO}}^{\mathrm{z}}}{r_{\mathrm{HO}}} 3\delta^2 s_{13} + 2\frac{r_{\mathrm{HO}}^{\mathrm{z}}}{r_{\mathrm{HO}}} 4\delta^3 s_{14} = 2F^{\mathrm{ref,z}}. \tag{5.38}$$

Dropping the indices HO, eq 5.38 is exactly the one shown in the third line of the force equations in Figure 5.3.

The force equations for the hydrogen atoms can be derived analogously. Note that for all nine force equations there is only one linear independent equation which represents the condition to properly describe the H–O bond length.

Further, one additional equation for a second order derivative in Figure 5.3 serves for yielding a proper H–O symmetric stretching frequency (compare eq 5.18),

$$\left.\frac{\mathrm{d}^2 V^{\mathrm{rep}}(r)}{\mathrm{d}r^2}\right|_{r=r_{\mathrm{HO}}} = 2s_{12} + 6\delta s_{13} + 12\delta^2 s_{14} = A^{\mathrm{ref}}, \tag{5.39}$$

where again $\delta = r_{\mathrm{HO}} - r_1$ is used and A^{ref} is a reference value.[4] Note that the choice of $r = r_{\mathrm{HO}}$ for this additional equation has been made for the sake of simplicity, any other choice is possible, too.

To solve the equation system it is splitted as shown in eqs 5.23 and 5.24. For the example in Figure 5.3 these equations can be constructed by introducing

$$M^{\mathrm{cont,impl}} = \left[M_1^{\mathrm{cont,impl}} \; M_2^{\mathrm{cont,impl}} \; M_3^{\mathrm{cont,impl}} \right] \tag{5.40}$$

$$M^{\mathrm{cont,expl}} = \left[M_1^{\mathrm{cont,expl}} \; M_2^{\mathrm{cont,expl}} \; M_3^{\mathrm{cont,expl}} \right] \tag{5.41}$$

$$M^{\mathrm{fit,impl}} = \left[M_1^{\mathrm{fit,impl}} \; M_2^{\mathrm{fit,impl}} \; M_3^{\mathrm{fit,impl}} \right] \tag{5.42}$$

$$M^{\mathrm{fit,expl}} = \left[M_1^{\mathrm{fit,expl}} \; M_2^{\mathrm{fit,expl}} \; M_3^{\mathrm{fit,expl}} \right] \tag{5.43}$$

$$\left(s^{\mathrm{impl}}\right)^{\mathrm{T}} = \left[s_{10} \; s_{11} \; s_{12} \; s_{13} \; s_{20} \; s_{21} \; s_{22} \; s_{23} \; s_{30} \; s_{31} \; s_{32} \; s_{33} \right] \tag{5.44}$$

$$\left(s^{\mathrm{expl}}\right)^{\mathrm{T}} = \left[s_{14} \; s_{24} \; s_{34} \right]. \tag{5.45}$$

Using

$$M^{\mathrm{final}} = \left(M^{\mathrm{fit,impl}} \left(M^{\mathrm{cont,impl}} \right)^{-1} M^{\mathrm{cont,expl}} + M^{\mathrm{fit,expl}} \right) \tag{5.46}$$

and $\boldsymbol{b}^{\mathrm{fit}}$ as defined in Figure 5.3, the equation system can thus be solved by applying eqs 5.28, 5.29, and 5.25.

5.1.5 Further Notes

Fitting New Pair-Potentials on Top of Existing Potentials

Very often a set of pair-potentials is already parametrized and additional pair-potentials are desired. An example is the MIO parameter set. While all pairs between the elements C, H, N, O, S, and P are parametrized, DFTB2 is not applicable for molecular systems containing e.g. halogen atoms. With the presented formalism it is easily possible to fit only these new pair-potentials to atomization energies, forces and vibrational frequencies of molecules that consist of the already parametrized as well as the additional elements. The only difference to the formalism as presented here is that the contribution of the repulsive energy or its first or second derivative stemming from the existing potentials are constants now and no variables anymore. Thus, in case of the energy equation 5.10 those pair-potentials of V_{ab}^{rep} that are already parametrized need to be written on the right hand side (entering the vector $\boldsymbol{b}$ of eq 5.3) and only the unknown pair-potentials remain on the left hand side. Similar changes are also required for force, additional, and reaction equations.

Weighting

It was discussed that for an overdetermined equation system the vector of unknowns $\boldsymbol{s}$ is constructed in the least square sense. This of course allows a weighting for the different objectives. Within the formalism the equation system 5.3 can then be rewritten as

$$M^{\mathrm{w}} \cdot \boldsymbol{s} = \boldsymbol{b}^{\mathrm{w}}, \tag{5.47}$$

[4] The determination of A^{ref} is discussed in section 5.2.1.

where the weighted matrix elements M^{w}_{ji} and weighted reference vector elements b^{w}_j are given by

$$\begin{aligned} M^{\mathrm{w}}_{ji} &= w_j M_{ji} \qquad \forall i \\ b^{\mathrm{w}}_j &= w_j b_j. \end{aligned} \tag{5.48}$$

The index j stands for a specific energy equation, set of force equations, additional equation, or reaction equation.

5.2 Application to Hydrocarbons and DFTB2

In this section the partially automatized machinery is employed to find the optimal shape of the H–H, C–H, and C–C repulsive potentials using DFTB2.

5.2.1 Computational Details

The Electronic Parameters

For constructing repulsive potentials it is necessary to define electronic parameters. The physically legitimated electronic parameters (as described in section 3.3) therefore serve as a good starting point. However, there are two parameters per atom type A, namely the compression (or confining) radii r^{wf}_A and r^{dens}_A (see eq 3.17).

The following initial compression radii are chosen: $r^{\mathrm{wf}}_{\mathrm{H}}$=3.0 a.u., $r^{\mathrm{dens}}_{\mathrm{H}}$=2.5 a.u., $r^{\mathrm{wf}}_{\mathrm{C}}$=2.7 a.u., and $r^{\mathrm{dens}}_{\mathrm{C}}$=7.0 a.u. being the ones from the MIO parameters. These values of compression radii for hydrogen and carbon will be further optimized in the present study. As mentioned in eq 5.11 the electronic energy of an atom a is defined as

$$E^{\mathrm{el}}_a = \sum_i n_i \epsilon^a_i + E^{\mathrm{spin}}_a. \tag{5.49}$$

While the ϵ^a_i are used as defined in the MIO set (section 3.3) for the spin polarization energies E^{spin}_a the PBE values are used instead of the LDA ones. The resulting values E^{el}_a are shown in Table 5.1.

In this work, the fitting procedure is applied using the following two approaches. In approach 1, the atomic electronic energies E^{el}_a are treated as free parameters (one for each element, see eq 5.13) and are optimized together with the parameters of the repulsive potentials. In approach 2 the calculated atomic electronic energies from Table 5.1 are used and only the parameters of the repulsive potentials are optimized. Therefore, in approach 2 the atomic energies are not optimized.

In total, six different sets of repulsive potentials are constructed, three using approach 1 and three using approach 2. Within each approach, one set of repulsive potentials is determined for the original compression radii, one for optimized compression radii, and one

Table 5.1: Spin Polarization Energies E^{spin}, Atomic Orbital Energies ϵ_{p} and ϵ_{s}, and Total Atomic Electronic Energies E^{el} (in Hartree) Obtained from Atomic PBE Calculations

Element	E^{spin}	ϵ_{p}	ϵ_{s}	E^{el}
H	–0.0411	—	–0.2386	–0.2797
C	–0.0455	–0.1944	–0.5049	–1.4440

Table 5.2: A Summary of Parameter Sets Developed in this Study

	Original compression radii	Optimized compression radii	Frequency optimized compression radii
Approach 1	opt1	opt3	opt5
Approach 2	opt2	opt4	opt6

Table 5.3: Compression Radii (in atomic units)

	opt1	opt2	opt3	opt4	opt5	opt6
$r_{\mathrm{H}}^{\mathrm{wf}}$	3.0	3.0	3.3	2.5	2.7	2.5
$r_{\mathrm{H}}^{\mathrm{dens}}$	2.5	2.5	3.0	2.5	2.5	2.5
$r_{\mathrm{C}}^{\mathrm{wf}}$	2.7	2.7	2.7	3.0	3.0	3.0
$r_{\mathrm{C}}^{\mathrm{dens}}$	7.0	7.0	5.0	7.0	7.0	7.0

for compression radii optimized to give optimal performance for the vibrational frequencies. The considered parameter sets are listed in Tables 5.2 and 5.3. The optimal values of the atomic electronic energies for the three sets of repulsive potentials, opt1, opt3, and opt5 are given in Table 5.4.

Training Sets

The actual fitting procedure of E^{rep} is done in two consecutive steps. In the first step, a training set 1 is used to find the parameters defining each E^{rep}. The training set 1 contains experimental structures of hydrogen [123], methane [123], ethyne [123], ethene [124], and ethane [124]. These molecules are chosen because they represent the most important bonding situations occuring in molecules consisting of carbon and hydrogen. Unfortunately, the repulsive potentials obtained using such a training set produce large errors for the equilibrium geometries of cyclopropene and 2-butyne in case of approach 1. To remedy this problem, the experimental equlibrium geometries of these two molecules have been included in the training set 1 for opt1, opt3, and opt5. The reference experimental atomization energies are taken from the CCCBDB database [123]. The vibrational zero-point energies have to be excluded from these values because DFTB2 is not parametrized to enthalpies at 0 K or to heats of formation at 298.15 K but to energies excluding the vibrational and thermal contribution. The zero-point energy contribution has been approximated as half of the sum of the experimental fundamental frequencies. For hydrogen, methane, ethyne, ethene, and ethane these energies are 109.6, 419.5, 405.1, 562.7, and 711.4 kcal/mol, respectively. For cyclopropene and 2-butyne, only the force equations are used in the parametrization process.

In the second step, the performance of the fit is evaluated using a training set 2, which contains atomization energies, equilibrium geometries, and vibrational frequencies of 15 molecules together with selected 32 reaction energies. Details are given in appendix C. The reason for employing a second, larger training set is the following. The division points

Table 5.4: Optimized Atomic Electronic Energies (in Hartree)

	opt1	opt3	opt5
$E_{\mathrm{H}}^{\mathrm{el}}$	–0.2555	–0.2514	–0.2524
$E_{\mathrm{C}}^{\mathrm{el}}$	–1.4527	–1.4176	–1.4472

and additional equations for each repulsive potential have to be specified as an initial step of the fitting procedure. This is a non-trivial problem and will be discussed in more detail below. Dependent on the performance of the fit for the training set 2, the division points are changed, either moved, deleted, or added. Then, the repulsive potentials are refitted according to the first step, and again tested using the training set 2. This procedure is iterated until a satisfactory performance is found.

Parameters Defining the Repulsive Potentials

In principle the division points r_p and the cut-off radii r_{n+1} defining the spline functions of the repulsive H–H, C–H, and C–C pair potentials can be freely chosen. Clearly, the shortest bond length between atom types X and Y should lie in the first interval I_{XY1}, i.e., it should be larger than the first division point r_{XY1}.

As discussed in detail in ref [15], a cut-off radius is introduced for the repulsive potential beyond which the potential and its derivatives are zero. The cut-off radius in the standard parametrization of DFTB2 was chosen to be smaller than second neighbor distances, resulting in cut-off radii of 4.3, 3.5, and 2.64 atomic units (a.u.) for C–C, C–H, and H–H, respectively. Here, the same cut-off radii are used except for C–C, where a cut-off of 4.8 a.u. is applied. This is larger than second neighbor distances in some molecules (e.g. benzene, propene, propane, cyclohexane), however, the potential decays rapidly being negligible at the second neighbor distance.

The number of intervals is another free parameter which is related to the number of fitting objectives. For example, in the parametrization of the H–H potential only three objectives are considered, the atomization energy, the equilibrium bond length, and the stretching frequency of the hydrogen molecule, which can directly be connected to the repulsive potential and its first and second derivative. As described in the previous section, for each interval there is only one free parameter to be determined, all others are defined by the continuity equations. Thus, three intervals for the H–H potential are necessary to fulfill the three objectives. The same holds for the C–H potential. The different C–H bonds of the molecules in training set 1 are very similar such that there are again only three objectives as above and only three intervals are necessary for the repulsive potential. For C–C the situation is different, since the single, double, and triple bonds have substantially different characteristics. In principle, one could fit the zeroth, first, and second derivatives to reference data for ethane, ethene, and ethyne using nine intervals (three intervals for each type of bond as above). However, for systems with intermediate bond lengths (e.g. benzene) such a fit results in large errors. This is due to the fact that the slope and curvature of the intervals of the single, double, and triple bonds do not tend to match at the division points. A pragmatic solution to this problem is to use intervals covering more than one bond type, which leads to an interpolation between these two regions. This of course means that choosing the optimal interval division for the C–C potential is an empirical procedure of trial and error. As a result, an extensive scan for different numbers and position of intervals was performed, because the properties of training set 2 depend sensitively on the choice of the intervals. The division points found this way define the intervals and are listed for each parameter set in Table 5.5.

For the fitting approach 1 (i.e., including the fit of atomic electronic energies) the situation becomes even more complicated, since two more parameters, $E_{\mathrm{C}}^{\mathrm{el}}$ and $E_{\mathrm{H}}^{\mathrm{el}}$, have to be determined. Both parameters appear in all energy equations (except for hydrogen, where only $E_{\mathrm{H}}^{\mathrm{el}}$ appears) and thus affect all repulsive potentials. That means, for example, that

Table 5.5: Division Points for the Repulsive Potentials

Division points	opt1	opt3,opt5	opt2,opt4,opt6
H–H			
r_1	1.40	1.40	1.40
r_2	1.80	1.80	1.80
r_3	2.40	2.64	2.60
r_4	2.64	—	2.64
C–H			
r_1	2.00	2.00	2.00
r_2	2.50	2.50	2.50
r_3	3.50	3.50	3.40
r_4	—	—	3.50
C–C			
r_1	2.00	2.00	2.00
r_2	2.80	2.80	2.80
r_3	3.22	3.22	3.22
r_4	4.80	4.80	4.00
r_5	—	—	4.80

for opt1 only two intervals are used for the C–H potential in order to leave one degree of freedom for the determination of the atomic electronic energy. several numbers and positions of the intervals have been tested in a "brute force" manner. Note, that the search for the best possible intervals depends on the choice of additional equations (see below) if good performance of vibrational frequencies is also desired.

So far, only fitting schemes using equilibrium geometries and atomization energies have been discussed. In order to achieve a good performance also for vibrational frequencies, additional equations have to be included in the fitting process (see eq 5.18). The form $V_{XY}''^{\mathrm{rep}^*}(r_z) = V$ is chosen, where the second derivative of a pair-potential X–Y at the distance r_z is set to a certain value V. For H_2, r_z is set to the equilibrium distance and V is chosen appropriately in order to reproduce the experimental fundamental frequency of H_2 with good accuracy. Note that in general the additional predefined conditions are not exactly fulfilled in the optimization scheme due to the limitations resulting from the applied least square fit.

As an initial guess for V one can take the value of the second derivative of the H–H potential at a given point, which is then varied by hand. Similarily, for C–H V is chosen to minimize the error of the C–H stretching frequencies in the training set 2. In the more complicated case of the C–C potential for simplicity only the C–C stretching modes of ethane, ethene, and ethyne are considered. Unfortunately, inclusion of additional equations improves the C–C stretching frequencies at the cost of geometries and atomization energies, which shows the limits of optimization within the DFTB2 framework. For a further improvement, modifications of the electronic part seem to be necessary. By choosing a set of additional equations, the ratio of errors for frequencies and for geometries or atomization energies can be controlled. One could of course come up with an appropriately weighted objective function combined with an optimization algorithm to automatically find a desired ratio. Algorithms for similar problems have recently been reported for example in refs [6, 125]. However, during the application to CHNO (next chapter) the empirical procedure is further developed. The advantage is that the fitting process—made as easy as possible—allows an easy detection of intrinsic errors of the DFTB formalism beyond the E^{rep} scope.

Table 5.6: Additional Equations of the Form $V_{XY}''^{\text{rep}*}(r_z/\text{Å}) = V$ in a.u.

X–Y	r_z	V(opt1)	V(opt2)	V(opt3)	V(opt4)	V(opt5)	V(opt6)
H–H	0.743	0.413	0.420	0.416	0.415	0.4075	0.415
C–H	1.081	0.400	0.400	0.400	0.380	0.400	0.380
C–C	1.203	—	—	—	—	—	1.300
C–C	1.300	0.950	—	0.980	—	0.950	—
C–C	1.330	—	—	—	—	—	0.800
C–C	1.505	0.340	—	0.320	—	0.310	0.370

Table 5.7: Experimental Fundamental C–C Stretching Frequencies and Deviation of Different Parameter Sets (in cm^{-1})

Molecule	exp[a]	opt1	opt2	opt3	opt4	opt5	opt6	MIO[b]	M2005[c]
Ethyne	1974	+13	+50	–30	+33	+62	–35	+146	+3
Ethene	1623	+138	+191	+132	+142	+129	+32	+200	+67
Ethane	995	+62	+43	+16	+30	+15	+29	+137	–26

[a] Experimental values from ref [88]. [b] Parameter set from ref [17]. [c] Parameter set from ref [88].

Neglecting the additional equations in the fitting scheme lead to substantial errors in the vibrational frequencies as large as 2000 cm^{-1} (H_2) and 500 cm^{-1} (C–H stretching in hydrocarbons). For all the parameter sets shown in Table 5.2 one additional equation for the H–H potential and one for the C–H potential is needed. Due to limits of optimization, additional sets of repulsive potentials, opt5 and opt6, are created which show improved performance for vibrational frequencies. All additional equations are listed in Table 5.6.

Determination of E^{rep} Using the Original Compression Radii

In the first step the original electronic compression radii from the MIO parameter set is used and only the repulsive potentials are optimized. Set opt1 is created using fitting approach 1, i.e. fitting also atomic electronic energies. For set opt2, fitting approach 2 is applied using the calculated atomic electronic energies from Table 5.1. Interestingly, set opt1 shows large errors in the C–C stretching frequencies if constructed without additional equations for C–C. Including two additional equations for C–C reduces these errors significantly. For set opt2, the errors in the frequencies are reasonably small already without additional equations and even lower in comparison to the original MIO parameter set (see Table 5.7). Including additional equations for the C–C potential reduces errors for frequencies but worsens geometries and atomization energies, therefore, they were not applied.

Table 5.8 summarizes the performance of opt1 and opt2 in comparison to the original MIO parameter set and with the parameter set M2005 optimized by Małolepsza et al. for vibrational frequencies [88]. The main advantage of the new sets of repulsive potentials is a significant improvement in the computed atomization energies; the mean absolute error for 14 atomization energies is reduced from 36.5 kcal/mol (MIO) to 4.1 and 4.0 kcal/mol (opt1 and opt2, respectively). It is worth mentioning here, that in the original MIO parameter set the atomization energies were consistently overestimated by roughly 5% and that the error of 36.5 kcal/mol could be reduced to only 8.8 kcal/mol just by employing the spin polarization energies derived from PBE (instead of LDA). Since the LDA values are consistently larger than the PBE ones, it can be concluded that the MIO parameters lead to a consistent overbinding for all molecules. It is interesting to mention that this error can be further

Table 5.8: Mean Unsigned Error and Maximum Absolute Deviation of Several Properties of Training Set 2

Property[a]	N[b]	opt1	opt2	opt3	opt4	opt5	opt6	MIO[c]	M2005[c,d]
E^{at} (kcal/mol)	14	4.1	4.0	1.4	4.0	8.3	5.0	36.5	40.6
$E^{\mathrm{at}}_{\mathrm{max}}$ (kcal/mol)		14.6	19.1	4.0	16.2	27.8	23.3	87.1	88.3
r (Å)	41	0.007	0.006	0.008	0.006	0.007	0.014	0.010	0.016
r_{max} (Å)		0.023	0.030	0.025	0.029	0.022	0.041	0.026	0.040
a (deg)	25	0.6	0.6	0.6	0.5	0.5	0.7	0.5	1.2
a_{max}		1.8	1.7	1.8	1.9	2.0	2.1	1.7	2.7
ν (cm^{-1})	305	42	46	48	39	38	32	61	31
ν_{max} (cm^{-1})		161	229	250	185	175	163	352	123
E^{rea} (kcal/mol)	32	5.1	4.2	2.4	4.7	6.6	7.2	5.2	8.3
$E^{\mathrm{rea}}_{\mathrm{max}}$ (kcal/mol)		14.8	14.3	6.9	15.5	32.4	23.9	14.9	46.4

[a] Atomization energies E^{at}, bond lengths r, bond angles a, harmonic vibrational wavenumbers ν, reaction energies E^{rea}, max stands for maximum absolute deviation. The reaction energy is compared to G3B3 [105] results at 0K. The zero point energies are calculated using harmonic vibrational frequencies calculated analytically with DFTB2 [126]. All other properties are compared to expermiental values. For details see appendix C. [b] Number of comparisons. [c] For the calculation of E^{at} the originally used LDA spin polarization energies were taken, for PBE spin polarization energies the mean unsigned error for MIO is 8.8 and the maximal absolute deviation is 30.5 kcal/mol, for M2005 72.4 and 155.0 kcal/mol, respectively. [d] Hydrogen was excluded, since these parameters were not optimized for this molecule.

reduced to 3.8 kcal/mol if the original electronic atomic energies defined by eq 5.11 are replaced by empirical values obtained by fitting the total DFTB2 atomization energies to the experimental atomization energies of the training set 2. Geometries are well reproduced by all sets, while vibrational frequencies are more accurate with opt1 and opt2 in comparison to the original MIO set but do not achieve the accuracy of M2005. For reaction energies no major improvement can be achieved, showing again, that the original MIO fit leads to a set of quite consistent repulsive potentials, however, including a large overbinding per bond.

Figure 5.4 shows the shape of repulsive potentials for MIO, opt1 and opt2. Set opt1 is more repulsive than opt2 for the C–H and H–H potentials whereas for the C–C potential this tendency is reversed. One main difference between opt2 and MIO origins from the different spin polarization energies used in the fitting (PBE calculated values versus LDA ones). This is reflected in the relatively small shift of the curves between MIO and opt2. The eminent shift of the repulsive potentials C–H and H–H of opt1 shows that the optimization of the atomic parameters $E^{\mathrm{el}}_{\mathrm{C}}$ and $E^{\mathrm{el}}_{\mathrm{H}}$ leads to a difference in the potential depth. This difference is effectively incorporated into the optimized atomic electronic energies. Tests indicate, that this shift can be different for different atom pairs. This can become problematic, since the effective binding energy is no more represented by the "depth" of the total energy (electronic plus repulsive), but is transferred to the atomic values. When comparing the opt1 binding energy curves with those from some high-level reference quantum calculation, the opt1 potentials can be more shallow, the difference is captured by the atomic values. This may lead to problems, when new atom types are included without refitting the parameters of the old atom types, i.e. consistent fits can only be achieved when all desired atom types are fitted at once.

This fact can be illustrated with the example reaction $C_2H_6 + H_2O \longrightarrow CH_3OH + CH_4$. The reaction enthalpy at 0 K of the MIO set is 12.1 kcal/mol and reproduces the G3B3 [105]

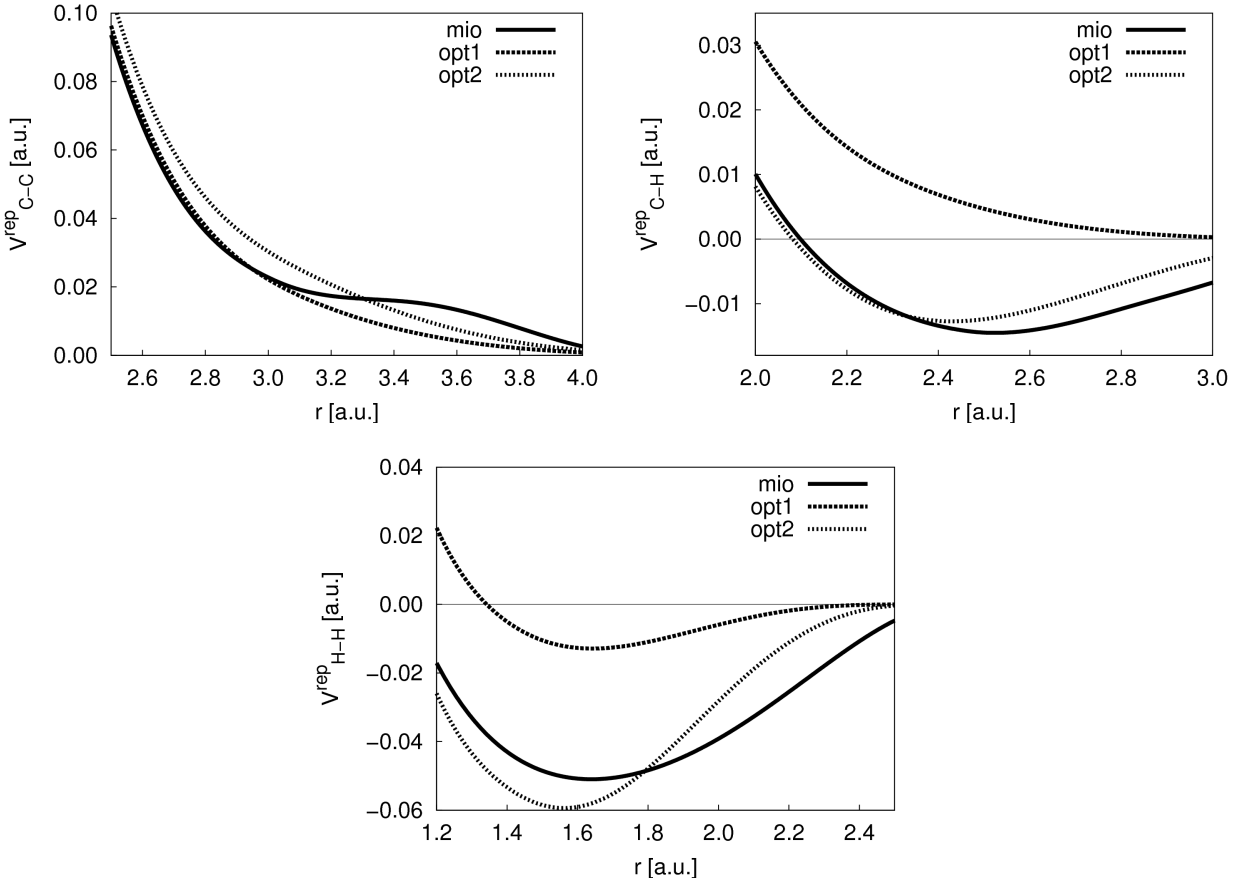


Figure 5.4: Original and optimized repulsive potentials for element pairings C–C, C–H, and H–H.

result (11.7 kcal/mol). Replacement of the C–C, C–H, and H–H repulsive potentials with the opt1 potentials gives 24.7 kcal/mol, and 5.4 kcal/mol for opt2. Set opt1 yields a large error enforcing the assumption that the opt1 potentials C–C and C–H are not transferable to the MIO potentials C–O and O–H. In opt2 the difference to the MIO result is much smaller and mainly due to the shift of the C–C repulsive potential (see Figure 5.4). While the MIO parameter set overbinds C–C and C–O, this overbinding is reduced for C–C in opt2. In chapter 6 it will be shown that when also reducing the C–O overbinding the reaction enthalpies are improved. Note that for reactions atomic electronic energies cancel out.[5]

The MIO repulsive potential corresponding to the C–C pair displays a characteristic hump around 3.6 a.u. (compare section 3.3). The origin of this hump is closely related to the previous paradigm of determination of the repulsive potentials. The short range part of the MIO C–C potential was obtained from auxiliary DFT calculations for a set of small molecules (C2H2, C2H4, and C2H6). This short-range potential was subsequently down-shifted to assure correct energetics of hydrocarbons. On the other hand, the long-range part of the potential was required to vanish beyond 4.2 a.u. Combining the two segments of the MIO C–C potential into a single curve resulted in producing the unphysical hump mentioned earlier. Clearly, this superfluous behavior of the potential is one of the drawbacks of constructing the repulsive potentials in a semimanual manner. As can be seen from

[5]Another example reaction is $C_2H_6 + NH_3 \longrightarrow CH_3NH_2 + CH_4$, where the same arguments hold. The MIO parameter set gives a reaction enthalpy at 0 K of 8.1 kcal/mol, being very similar to the G3B3 result of 7.9 kcal/mol. Using the C–C, C–H, and H–H potentials of opt1 and all further pair potentials from the MIO set yield 20.7 kcal/mol, for opt2 1.4 kcal/mol.

the curves presented in Figure 5.4, this behavior is eliminated when using the presented semiautomatized approach.

The fact that the H–H repulsive potential becomes negative (i.e. attractive) may lead in practice to problems in molecular dynamics simulations when the interatomic distance between two non-bonded hydrogen atoms becomes too small. This potential difficulty can be alleviated using modified electronic compression radii, as discussed below.

Optimization of E^{rep} and the Compression Radii

As the next step, different values of compression radii used for obtaining the atomic wave functions (2.4, 2.7, and 3.0 for carbon, and 2.5, 2.7, 3.0, and 3.3 for hydrogen) and the atomic input densities (5.0, 7.0, and 9.0 for carbon, and 2.5, 3.0, 3.5, and 5.0 for hydrogen) are tested. The repulsive potentials $V_{AB}^{\mathrm{rep}^*}$ have been determined for every combination of these compression radii. Again the two approaches mentioned above are used: including (opt3 and opt5) and excluding (opt4 and opt6) the fit of the electronic atomic energies $E_A^{\mathrm{el}^*}$. The parameter sets opt3 and opt4 are optimized to give accurate atomization energies and geometries, while opt5 and opt6 are designed to improve vibrational frequencies. The additional conditions are found to be similar for different compression radii. They are further optimized once the "best" compression radii have been found and are shown in Table 5.6. The compression radii are determined by testing all parameter sets on training set 2. Table 5.8 gives an overview over all constructed repulsive potentials indicating that various properties may require different electronic parameters, i.e. this demonstrates the limits of the transferability of the electronic part of DFTB.

The repulsive potential sets opt3 and opt4 are fitted to give the most accurate atomization and reaction energies with acceptable errors for equilibrium geometries and frequencies. While geometries are described satisfactorily, opt3 shows much smaller errors for atomization and reaction energies. In parameter set opt4 (as well as in opt1 and opt2) large atomization energy errors are found for small cyclic structures such as cyclopropane and cyclopropene. The superb performance of opt3 for the atomization and reaction energies is further confirmed on a larger testing set in the following section. A slight shadow on the future performance of opt3 is cast by the fact that adding additional elements to the fitting scheme may severely perturb the delicate balance between $E_{\mathrm{C}}^{\mathrm{el}}$ and $E_{\mathrm{H}}^{\mathrm{el}}$ necessary for obtaining the reported very good agreement between the DFTB2 and experimental atomization and reaction energies. The perturbation is inevitably associated with the fitting procedure since the parameters for all elements are linked via the atomic electronic energies as discussed above.

The repulsive potential sets opt5 and opt6 are constructed to obtain small errors for vibrational frequencies. Geometries are still accurate, however, the errors of atomization and reaction energies are much larger now. The results for the training set 2 (see Table 5.8) show improved vibrational frequencies with errors comparable to those of the M2005 parameter set. In comparison to the MIO parameter set one remarkable improvement is the C–C stretching frequencies as shown in Table 5.7.

The optimization of the compression radii significantly improves the atomization and reaction energies for opt3. For the other parameter sets, only small improvement is achieved. This shows that there is only a limited influence of the compression radii on the performance of DFTB2. The difference between opt5 and opt3 stems from a different choice of the compression radii and the values for the additional equations. The input for the fitting procedure of opt6 differs from opt4 only by three additional equations for the C–C potential

Table 5.9: Mean Unsigned Error and Maximum Absolute Deviation of Several Molecular Properties of the G3/99 Test Set

Property[a]	N[b]	opt1	opt2	opt3	opt4	MIO	PBE[c]	B3LYP[c]
ΔH_f (kcal/mol)	39	4.8	5.8	2.2	7.1	55.4[d]	26.0	7.4
$\Delta H_\mathrm{f}^\mathrm{max}$ (kcal/mol)		19.2	29.0	16.3	25.2	114.8[d]	81.3	17.3
ΔH_f (kcal/mol)[e]	39	4.3	4.0	2.1	3.7	4.1	1.8	3.4
$\Delta H_\mathrm{f}^\mathrm{max}$ (kcal/mol)[e]		19.7	25.1	16.2	20.5	21.7	7.4	9.6
r (Å)	196	0.008	0.008	0.008	0.007	0.011	0.008	0.003
r_max (Å)		0.155	0.267	0.038	0.142	0.035	0.020	0.016
a (deg)	177	0.6	0.7	0.6	0.6	0.5	0.4	0.4
a_max (deg)		14.4	19.9	8.2	13.4	4.8	1.9	1.9
d (deg)	5	1.5	1.5	1.3	2.1	1.5	0.7	1.0
d_max (deg)		6.3	6.3	5.5	8.8	6.7	2.0	2.6

[a] Heats of Formation ΔH_f at 298.15 K, bond lengths r, bond angles a, dihedral angles d, max stands for maximum absolute deviation. Geometric data is compared to the MP2/cc-pVTZ calculations. [b] Number of comparisons. [c] Basis set 6-311G(2d,2p). [d] ΔH_f calculated using LDA spin polarization energies, taking PBE values instead gives MUE of 13.2 kcal/mol and a maximal absolute deviation of 34.2 kcal/mol. [e] Calculated using atomic electronic energies fitted to 14 experimental atomization energies of training set 2.

as shown in Table 5.6. It is clear from the presented data that the quality of the C–C stretching frequencies can only be improved at the cost of deteriorating the atomization and reaction energies. As discussed above, this seems to be an effect of the approximations inherently present in the current DFTB2 framework.

5.2.2 Benchmarks

The optimized repulsive potentials opt1–opt6 are now evaluated in several benchmark studies.

Benchmarks for the G3/99 Molecule Set

In this subsection a test on the G3/99 [43] molecule set containing H_2 and 38 hydrocarbons is performed. The tested properties include heats of formation at 298.15 K and equilibrium geometries (see Table 5.9). The heats of formation are calculated as described in ref [42]. The enthalpies of formation for gaseous atoms at 0 K and the $(H^{298}-H^0)$ values for hydrogen and carbon in their standard states are taken from experiment (ΔH_f^0(0 K) for hydrogen and carbon is 51.63 and 169.98 kcal/mol, respectively, and the corresponding values of $(H^{298}-H^0)$ are 1.01 and 0.25 kcal/mol) [42]. The heat capacity corrections for molecular vibrations are estimated as

$$E^\mathrm{vib} = R\sum_k \theta_k \left(\frac{1}{2} + \frac{1}{e^{\theta_k/T}-1}\right) \text{ where } \theta_k = \frac{h\nu_k}{k_B}, \tag{5.50}$$

using harmonic vibrational frequencies ν calculated analytically with DFTB2 [126]. R is the molar gas constant, θ the vibrational temperature, T the temperature, h and k_B the Planck and the Boltzmann constants, respectively. In addition, thermal corrections are included within the classical approximation for translations ($\frac{3}{2}RT$), rotations ($\frac{3}{2}RT$ for nonlinear and RT for linear molecules), and the PV term (RT).

Further two approaches for the calculation of atomization energies have been chosen. In the first approach the atomic electronic energies are used as discussed in the theory section.

Table 5.10: Atomic Electronic Energies Fitted to 14 Experimental Atomization Energies of Training Set 2

Element	opt1	opt2	opt3	opt4	opt5
H	–0.2564	–0.2802	–0.2514	–0.2808	–0.2544
C	–1.4516	–1.4416	–1.4176	–1.4404	–1.4392
	opt6	MIO	M2005	PBE[a]	B3LYP[a]
H	–0.2819	–0.2762	–0.2603	–0.4965	–0.5026
C	–1.4422	–1.4545	–1.4409	–37.8066	–37.8538

[a] Basis set 6-311G(2d,2p).

The results are displayed in the first two lines of Table 5.9. In the second approach, the atomic electronic energies are fitted to yield least square errors for the 14 experimental atomization energies of training set 2. These results are displayed in the third and fourth lines of Table 5.9. It is not surprising that the latter approach reveals smaller errors for the heats of formation. The fitted atomic electronic energies are listed for each parameter set in Table 5.10. Heats of formation are also calculated for the widely used density functionals PBE and B3LYP. For PBE calculations the heats of formation are largely underestimated whereas for B3LYP overestimated. The errors are readily reduced when fitting the atomic energies. The corresponding values can be found in Table 5.10. Substantial improvements using optimized atomic energies have been noted in other studies [127, 11, 128, 129, 130].

Set opt3 shows an excellent performance with an MUE of 2.2 kcal/mol for heats of formation; the largest deviations are found for azulene and methylene (respectively –7.3 and –16.3 kcal/mol). The DFTB2 bond lengths and angles compared to MP2/cc-pVTZ geometries give an overall MUE of 0.008/0.007 Å and 0.6/0.6 degrees for the opt3/opt4 parameter sets, showing good agreement with the reference data. Similar accuracy is yielded for the original MIO, opt1, and opt2 parameter sets, and both tested DFT methods. The largest discrepancies in the opt3/opt4 equilibrium geometries are observed for bicyclobutane: 0.038/0.142 Å for bond distances and 8.2/13.4 degree for bond angles. With respect to bond types the largest MUE for opt3 is observed for the C=C and C−H bonds (0.008 Å), and the largest MUE for opt4, for the C−C bond (0.013 Å). Keeping in mind that for bicyclobutane an exceptional large deviation is found, all other bond lengths are reproduced with good accuracy as shown in Table 5.11. The mean signed error (MSE) shows that the C≡C bonds tend to be shorter whereas all other bond lengths tend to be larger than the MP2/cc-pVTZ reference.

Benchmarks for the Jorgensen Molecule Sets

Recently, Jorgensen and coworkers published a collection of experimental heats of formation, isomerization enthalpies, conformational energetics, and MP2/6-31g(d) geometries [131] which are very suitable to benchmark approximate methods.[6] This set is used to test the new parametrization, as shown in Tables 5.12, 5.13, and 5.14.

Heats of Formation. The best performance is found for the parameter set opt3, with an MUE of 2.6 kcal/mol it is comparable to the results of PDDG-PM3. The largest deviations

[6] While the comparison on the G3/99 test set were made with MP2/cc-pVTZ geometries, MP2/6-31G(d) geometries for the Jorgensen test set were used. This facilitates comparison to the data compiled in ref [131, 11]. The mean unsigned error (MUE) of MP2/6-31G(d) compared to MP2/cc-pVTZ bond lengths for the G3/99 test set is found to be only 0.004 Å [121].

Table 5.11: Deviation of Different Bond Types of the G3/99 Test Set Compared to MP2/cc-pVTZ Calculations in Angstrom

		opt1			opt2			MIO		
Type[a]	N[b]	MAX	MSE	MUE	MAX	MSE	MUE	MAX	MSE	MUE
rC–C	53	0.155	–0.010	0.011	0.267	+0.014	0.014	0.035	–0.007	0.010
rC=C	14	0.017	–0.002	0.003	0.024	+0.008	0.008	0.020	+0.004	0.005
rC≡C	3	0.024	–0.024	0.024	0.008	–0.007	0.007	0.008	–0.008	0.008
rC%C	14	0.013	+0.008	0.008	0.015	+0.009	0.009	0.014	+0.003	0.005
rC–H	111	0.038	+0.006	0.007	0.072	+0.004	0.006	0.023	+0.013	0.013
rH–H	1	0.004	+0.004	0.004	0.004	+0.004	0.004	0.006	+0.006	0.006
all	196	0.155	+0.006	0.008	0.267	+0.007	0.008	0.035	+0.006	0.011
		opt3			opt4			MIO		
Type[a]	N[b]	MAX	MSE	MUE	MAX	MSE	MUE	MAX	MSE	MUE
rC–C	53	0.038	+0.002	0.007	0.142	+0.013	0.013	0.035	–0.007	0.010
rC=C	14	0.019	+0.008	0.008	0.023	+0.007	0.007	0.020	+0.004	0.005
rC≡C	3	0.006	–0.005	0.005	0.009	–0.009	0.009	0.008	–0.008	0.008
rC%C	14	0.014	+0.007	0.008	0.017	+0.011	0.011	0.014	+0.003	0.005
rC–H	111	0.020	+0.008	0.009	0.039	+0.003	0.004	0.023	+0.013	0.013
rH–H	1	0.004	+0.004	0.004	0.004	+0.004	0.004	0.006	+0.006	0.006
all	196	0.038	+0.006	0.008	0.142	+0.006	0.007	0.035	+0.006	0.011

[a] The abbreviations stand for "–" single bond, "=" double bond, "≡" triple bond, "%" aromatic bond. [b] Number of comparisons.

Table 5.12: Mean Unsigned Error and Maximum Absolute Deviation of Several Molecular Properties of the Jorgensen Test Set

Property[a]	N[b]	opt1	opt2	opt3	opt4	opt5	opt6	MIO	M2005	PDDG[c]
ΔH_f	254	6.4	6.1	2.6	9.2	24.3	15.7	87.9[d]	159.0	2.6
$\Delta H_\mathrm{f}^\mathrm{max}$		32.8	29.1	35.5	29.4	76.3	66.9	184.6[d]	337.5	39.1
ΔH_f [e]	254	5.1	3.2	2.4	3.8	6.8	7.7	3.9	7.2	—
$\Delta H_\mathrm{f}^\mathrm{max}$ [e]		31.6	25.2	35.2	20.5	27.8	56.6	21.7	30.5	—
r	111	0.007	0.008	0.005	0.007	0.008	0.018	0.009	0.020	0.011
r_max		0.154	0.266	0.037	0.141	0.196	0.178	0.034	0.266	0.057
a	57	0.8	1.0	0.7	0.7	0.8	0.9	0.6	1.3	1.3
a_max		16.4	21.9	10.2	15.4	18.6	17.5	6.8	22.1	11.8
d	20	1.6	1.9	1.0	1.7	1.9	1.7	1.3	2.7	2.9
d_max		8.2	9.6	5.9	9.3	8.9	7.8	9.4	13.4	17.1

[a] Heats of formation ΔH_f at 298.15 K in kcal/mol, bond lengths r in Å, bond angles a and dihedral angles d in degree, max stands for maximum absolut deviation. Geometric data is compared to MP2/6-31g(d) calculations. For details see ref [131]. [b] Number of comparisons. [c] PDDG-PM3 values from ref [11]. [d] ΔH_f calculated using LDA spin polarization energies, taking PBE values instead gives MUE of 19.9 kcal/mol and a maximal absolute deviation of 88.9 kcal/mol. [e] Calculated using atomic electronic energies fitted to 14 experimental atomization energies of training set 2.

Table 5.13: Deviation from Experiment[a] for Selected Isomerization Enthalpies (kcal/mol) at 298.15 K

Isomerization			Exp	MIO	opt3	opt4	PDDG
Propyne	→	Allene	1.2	+4.0	+2.7	+4.8	+4.63
Propyne	→	Cyclopropene	21.8	+16.5	+6.5	+18.0	–0.23
Propene	→	Cyclopropane	7.9	+6.3	+0.5	+6.8	+0.29
trans-2-Butene	→	*cis*-2-Butene	1.1	–0.0	+0.1	–0.5	+0.53
2-Methylpropene	→	*trans*-2-Butene	1.3	–0.1	+0.6	–1.1	–1.91
trans-2-Butene	→	1-Butene	2.8	+1.3	+0.8	+0.6	+2.18
1,3-Butadiene	→	Cyclobutene	11.3	+2.2	–3.8	+1.7	–3.51
Cyclopentene	→	Vinylcyclopropane	22.2	+11.5	+3.2	+9.1	+2.73
1-*trans*-3-Pentadiene	→	1,4-Pentadiene	7.1	+1.0	+0.6	+0.2	–1.77
2,2–Dimethylpropane	→	*n*–Pentane	5.1	–2.4	–0.1	–5.3	+2.09
2,2,3,3-Tetramethylbutane	→	*n*-Octane	4.1	–3.4	+0.8	–7.4	+3.14
Toluene	→	Norbornadiene	46.7	+2.6	–4.2	+4.9	–2.72
Styrene	→	Cyclooctatetraene	35.8	+7.7	+5.3	+4.8	–4.92
MUE				4.5	2.2	5.0	2.4

[a] Experimental and PDDG-PM3 values from ref [11].

are found for cubane (–35.5 kcal/mol) and diamantane (–13.0 kcal/mol). In general, large deviations are found for large aromatic systems such as anthracene, azulene, and biphenylene, bicyclic structures, and highly substituted cyclic compounds. Similar problems have been detected for opt4. As already mentioned for the G3/99 test set a refitting of the atomic electronic energies on the 14 experimental atomization energies of training set 2 generally reduces this error (third and fourth line of Table 5.12). This effect is significant e.g. for the MIO set where the mean unsigned error is reduced to only 3.9 kcal/mol. Note that this refitting of atomic electronic energies is only done for hydrocarbons. When including further elements a refit also influences hydrogen and carbon, thus also the heats of formation for hydrocarbons are expected to be less accurate.

Geometries. Generally DFTB2 describes geometries very well, being slightly more accurate than PDDG-PM3 (with exception of the frequency optimized sets opt6 and M2005). Again, the best performance is observed for the opt3 and MIO sets with MUE for bond lengths smaller than 0.04 Å. The largest bond length error is consistently found for the bridging C–C bond in bicyclobutane (up to 0.27 Å with opt2 and M2005), the second largest deviation is already much smaller: The error for the C−C single bond in 1,3-butadiyne with opt2, opt3, and opt4 is smaller than 0.024 Å and the error for the C≡C triple bond in acetylene with opt1 and opt5 is smaller than 0.031 Å. Only for the frequency optimized set opt6 the second largest error is 0.046 Å for the C−C bond of 1,3-butadiyne. Similar observations are found for bond angles and dihedrals. While bicyclobutane gives exceptionally large errors, deviations for all other bond angles are within 3.4° and for dihedrals, within 8°.

Isomerization Enthalpies. Table 5.13 compares calculated enthalpies (including corrections for zero-point energy and temperature as described above) for a selected set of isomerization reactions [11]. The MUE for the MIO parameter set is 4.5 kcal/mol. The performance of DFTB2 can be improved with the parameter set opt3 giving MUE of only 2.2 kcal/mol, which is of comparable accuracy as for PDDG-PM3 (2.4 kcal/mol). The main reason for the better performance of the opt3 set over the MIO set is a better description of small cyclic hydrocarbons (cyclopropene and cyclopropane) by the former set. The

Table 5.14: Deviation for Conformational Energetics (kcal/mol)

Conformers		ΔE^{ref} [a]	MIO	opt3	opt4	PDDG[a]
Butane	*anti* vs *gauche*	0.7	–0.2	–0.2	–0.4	–0.4
Ethane	*anti* vs *eclipsed*	2.8	–0.5	–0.7	–0.8	–1.7
Methylcyclohexane	*eq* vs *ax*	1.8	–0.9	–0.8	–1.3	–0.9
cis-1,3-Dimethylcyclohexane	*eq,eq* vs *ax,ax*	5.5	–2.1	–1.9	–3.3	–2.3
Propene	*eclipsed* vs *anti*	2.0	–0.9	–0.9	–1.0	–1.3
1,3-Butadiene	*s-trans* vs *s-cis*	2.5	–1.4	–0.9	–1.2	–1.8
MUE			1.0	0.9	1.3	1.4

[a] Reference and PDDG-PM3 values from ref [11].

parameter set opt4 on the other hand does not show any improvement over the original MIO set. Probably the most serious problem of opt4 is the wrong sign of the isomerization enthalpies for the isomerization of sterically crowded alkanes to the corresponding linear isomers (2,2-Dimethylpropane → n-Pentane and 2,2,3,3-Tetramethylbutane → n-Octane), leading to qualitatively wrong information about the relative stability of linear alkanes. A dispersion correction as proposed in ref [64] improves these results only marginally, yielding isomerization enthalpies of 0.0 and –2.0 kcal/mol. For the sets opt1, opt2, opt5, opt6, and M2005 the mean unsigned error are 4.6, 4.9, 5.9, 5.7, and 8.1 kcal/mol, respectively.

Conformational Energetics. The comparison of results shown in Table 5.14 is based on a small compilation of conformational energies given in ref [11]. All tested parameter sets give performance similar or slightly better than PDDG-PM3. The largest error is found for *cis*-1,3-dimethylcyclohexane. Qualitative errors are found only with the M2005 parameter set for butane, methylcyclohexane, and *cis*-1,3-dimethylcyclohexane, for which the stability pattern is reversed with respect to the experiment. This behavior is probably related to significantly larger cut-off radii used to construct the M2005 repulsive potential. For M2005, the cut-off radii are set to 7 atomic units, whereas for MIO, the cut-off radii are chosen in a way ensuring that the second neighbor interaction vanishes. This results in cut-off radii of 2.64 a.u. for the H–H, 3.5 a.u. for the C–H, and 4.3 a.u. for the C–C repulsive pair potential. A geometrical analysis of the structures of *anti*- and *gauche*-butane shows that the distance between the first and third carbon atoms is 4.916 and 4.939 a.u., respectively. Since for shorter distance the repulsive energy is higher, the *anti* conformer is artificially destablized by the M2005 parameter set. Clearly, for the MIO set, the analogous repulsive energy contribution is zero for both conformers.

Benchmarks for Vibrational Frequencies

The implementation of analytical second derivatives [126] greatly simplified the calculation of harmonic frequencies within the DFTB2 formalism. This functionality is used to determine the DFTB2 frequencies for a group of 14 hydrocarbons representing a variety of typical bonding situations. Mean unsigned errors and maximal absolute deviations with respect to experimental fundamental vibrational frequencies are given in Tables 5.15 and 5.16, respectively. The contributions for doubly degenerate vibrations have been included twice in the averaging and those for triply degenerate vibrations, thrice. The total number of computed vibrational frequencies is 349. The number of available experimental modes [123, 132] is only 346. Therefore, for two modes of 1,3,5-hexatriene and one mode of bicyclo[2,1,0]pentane, the DFT harmonic vibrational frequencies (B3LYP[39, 40, 75]/cc-pVTZ[109] with scaling factor

Table 5.15: Mean Unsigned Error of Calculated Harmonic Vibrational Frequencies from Experimental Fundamental Frequencies for a Group of 14 Hydrocarbons

Molecule	BLYP[a]	opt1	opt2	opt3	opt4	opt5	opt6	MIO	M2005
Methane	23	55	30	47	34	41	30	65	44
Ethyne	32	30	83	101	57	47	54	70	28
Ethene	19	40	60	50	39	36	36	49	34
Ethane	20	40	34	28	29	26	24	52	28
Allene	13	53	74	73	54	52	45	58	39
Cyclopropane	17	43	56	52	42	35	37	55	33
Propene	21	27	32	29	26	21	19	42	20
1,3-Butadiene	18	35	48	49	33	31	27	46	21
Bicyclo[2,1,0]pentane	50	40	46	45	40	41	34	52	42
Spiropentane	32	52	58	59	47	41	40	68	52
Cyclohexane	27	31	25	23	28	21	24	49	23
Benzene	17	53	67	70	46	42	34	58	34
1,3,5-Hexatriene	19	39	49	51	37	36	33	46	24
Cubane	22	52	47	63	52	64	43	96	35
MUE (total)[b]	25	42	48	50	40	38	33	58	32

[a] Calculated with cc-pVTZ basis set. [b] Mean unsigned error of all 349 vibrational frequencies of the 14 molecules shown.

Table 5.16: Maximal Unsigned Error of Calculated Harmonic Vibrational Frequencies Versus Experimental Fundamental Frequencies for a Group of 14 Hydrocarbons

Molecule	BLYP[a]	opt1	opt2	opt3	opt4	opt5	opt6	MIO	M2005
Methane	48	113	79	102	50	82	37	137	51
Ethyne	70	69	159	250	107	81	103	140	72
Ethene	64	145	189	158	152	143	102	201	82
Ethane	58	64	63	54	53	45	45	136	48
Allene	36	159	229	181	185	175	110	245	89
Cyclopropane	40	81	133	111	77	78	87	149	75
Propene	70	150	192	138	156	140	83	219	54
1,3-Butadiene	72	149	191	142	153	143	94	214	64
Bicyclo[2,1,0]pentane	158	125	136	137	145	163	127	154	140
Spiropentane	110	283	274	306	259	194	180	435	137
Cyclohexane	117	64	63	103	72	74	95	116	103
Benzene	52	161	199	144	156	127	99	230	100
1,3,5-Hexatriene	88	147	189	169	152	140	122	203	89
Cubane	51	112	115	188	102	135	143	352	83
MAX (total)[b]	158	283	274	306	259	194	180	435	140

[a] Calculated with cc-pVTZ basis set. [b] Mean unsigned error of all 349 vibrational frequencies of the 14 molecules shown.

of 0.965 [123]) have been used instead. The unscaled DFTB2 frequencies are compared with the DFT results (BLYP[39, 40]/cc-pVTZ, unscaled). In order to compare the intrisic accuracy unscaled DFTB2 frequencies were used. Further it has been shown that for the MIO set a uniform scaling factor is very close to one [84]. All the DFT calculations are performed using the Gaussian 03 program [108].

The MUEs of the analyzed 349 vibrational modes are noticeably reduced for all the optimized parameter sets in comparison to the original MIO parameter set. The results are most significantly improved for the frequency-optimized sets opt5 and opt6 with deviations of 38 and 33 cm^{-1}, respectively. The latter set of repulsive potentials gives similar accuracy as the M2005 parameter set (32 cm^{-1}), but both of them give approximately 15% larger errors than the BLYP/cc-pVTZ computational scheme (25 cm^{-1}). The largest deviations to experiment are observed for the DFTB2 frequencies of allene, bicyclo[2,1,0]pentane, and spiropentane.

Benchmarks for Linear Alkanes

Recent comparisons of heats of formation for linear alkanes using G3 [133] and B3LYP [127] showed that both computational schemes display a cummulative error, i.e., that the total error in the computed heats of formation grows with the length of the chain. This is a systematic error which is considered worth to reduce or even eliminate. Using a dispersion correction [63] and fitting the atomic energies for DFT improves the situation but does not solve this problem completely [127]. Similar results are found for DFTB2 as shown in Figure 5.5. For the MIO set, the overbinding is obvious. For opt4, a small underbinding is detected; it can be partially eliminated including a dispersion correction [64]. No cumulative error is found for any of these sets, if the atomic energies are fitted to minimize the errors of heats of formation of this linear alkanes. With this result, it is not too much of a surprise that for opt3 no error accumulation is found. As discussed above, the atomic energies of opt3 are already fitted to training set 1, which partially accounts for the accumulation problem.

5.3 Conclusion

A new methodology is presented which greatly simplifies the generation of the DFTB2 repulsive potentials and considerably improves the accuracy, in particular for heats of formation and vibrational frequencies. The method is based on a solution of a linear inverse problem for a set of repulsive potentials for a given group of elements. Since in most cases, the effective linear problem is not directly invertible (being either over- or underdetermined), a singular value decomposition approach is used to extract the meaningful portion of information from the effective linear problem. The application of fourth order splines results in smooth second derivatives, leading to an improved description of vibrational frequencies.

In this work, the formalism of the electronic part of DFTB2 has not been modified, however, the values of atomic electronic energies and compression radii for atomic orbitals and densities—entering the Hamilton matrix elements—are adjusted to improve the overall performance. The present work clearly shows that the approximations inherent in DFTB2 limit its overall performance. Heats of formation and vibrational frequencies can not be simultaneously optimized to an accuracy comparable with full DFT methods. Therefore, a special parametrization for vibrational frequencies is suggested for use when needed (sets opt5 and opt6).

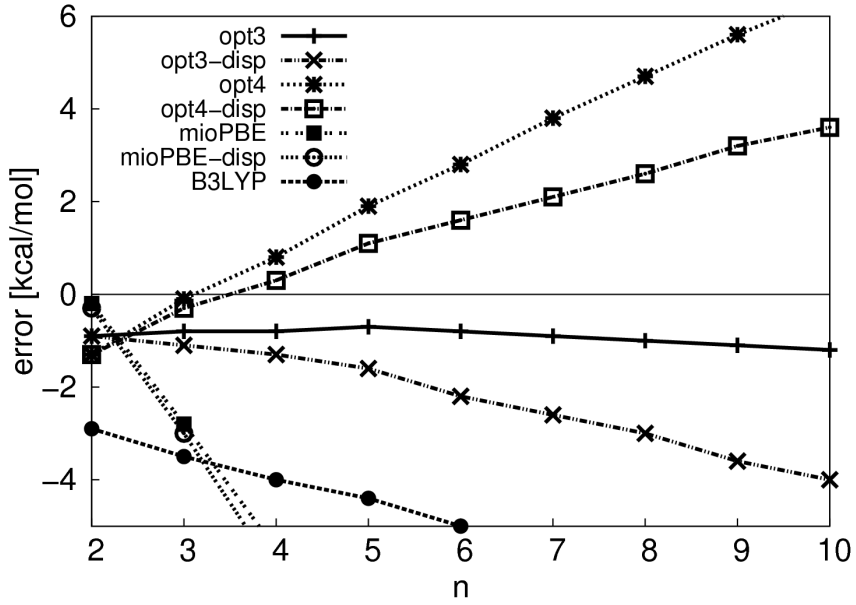


Figure 5.5: Errors in computed heats of formation for linear alkanes C_nH_{2n+2} for opt3, opt4, and the MIO(PBE) parameter sets and also including disperion corrections. The results are compared with B3LYP/6-31G(d) results including a dispersion correction from ref [127]. The line connecting the points is plot to facilitate the comparisons.

The DFTB2 repulsive potentials have up to now been derived for a variety of elements, including C, H, N, O, Mg, P, S, etc.. The repulsive potentials depend on the electronic parameters they have been fitted for, in particular the compression radii. In a first step, optimized repulsive parameters using the MIO compression radii are derived. Therefore, the sets opt1 and opt2, which show improved properties for the repulsive potentials for C and H, can be used in conjunction with the parameters with the other elements parametrized so far.

On the other hand, also the compression radii have been optimized (opt3, opt4, opt5, opt6). These sets are no more consistent with the parameters of the other elements, i.e., in the following chapter parameters of other elements are also reparametrized. The variation of electronic parameters improves all properties, most significantly the heats of formation, however, within a limited range. This means, that the electronic parameters can not be used for a significant improvement of performance e.g. within a "brute force" fitting approach. On the other hand, this shows the robustness of DFTB2, i.e. a variation of electronic parameters will also not lead to significant failures, of course within the limits of DFT-GGA. With opt3, DFTB2 can successfully compete with the special parametrization PDDG-PM3 and also with B3LYP, being much better than PBE at least for hydrocarbons. This shows, that with a suitable choice of parameters systematic deficiencies of the underlying PBE functional can be improved.

A fundamental decision in the parametrization procedure is, whether the atomic parameters (atomic electronic energies) and repulsive potentials are optimized in one step. This is done for the sets opt1, opt3 and opt5 and is a general strategy for the parametrization of NDDO type semiempirical methods. This however, influences the repulsive potential itself, i.e. the repulsive potentials are arbitrarily shifted to optimize the performance. This means, that part of the energy of a bond, which is described by parameters depending on two centers (pairwise potentials) are shifted to atomic parameters, in the case of DFTB2 to

the atomic electronic energies. So far, this seems to pose no problem as long as parameters for all atom types are determined in one optimization step. This means, however, that the parametrization for new atom types will lead to an adjustment of the already parametrized ones, which will lead to an increasing complexity of the problem, when more elements are to be included. Therefore, approach 2, where only repulsive potentials are optimized (opt2, opt4, opt6), is considered to be more practical in this respect. The atomic values can still be reoptimized after determining E^{rep}, leading still to a slight improvement, although the performance of approach 1 can not be completely matched. In the next chapter the approach 2 is applied giving an improvement also for the CHNO parameter set.

Chapter 6

Parametrization of DFTB3 for Elements C, H, N, and O

In the previous chapters it has been shown that the MIO parametrization is excellent for describing geometries and vibrational frequencies. Nevertheless, a new parametrization was motivated because of two reasons. First, the hydrogen bond lengths (e.g. within the water and ammonia dimers) are described too short. During the MIO parametrization that was result of a compromise with a too low hydrogen binding energy. The methodological extensions within DFTB3 show more flexibility in that respect. Therefore, the idea is to change the electronic parameters in order to enlarge the hydrogen bond length while maintaining the good description of the binding energy due to the modified Coulomb interaction within DFTB3. The second reason is the overbinding tendency causing an overestimation of atomization energies. Within the last chapter it has already been shown for DFTB2 that this can be eliminated for hydrocarbons by a proper choice of electronic and repulsive parameters.

In this chapter a new parametrization for hydrogen, carbon, nitrogen, and oxygen will be presented using the semi-automatized procedure for determining the repulsive potentials as derived in the last chapter. The new parameters are then tested for geometries, vibrational frequencies and energetic properties on large sets of organic molecules.

The parametrization presented was carried out in collaboration with Albrecht Goez. During work on his Master's thesis [134], his focus was to find parameters related with nitrogen. In order to give a detailed and comprehensive description of the fitting procedure, also parts of his results will be discussed in the following.

6.1 Parametrization Scheme

A parametrization scheme is suggested containing four steps which allow to find well performing parameters in a reasonable amount of time. It further helps to detect systematic deficiencies of the DFTB3 methodology being motivation for further conceptual improvements. These steps will first be presented in a general sense. The following subsections describe the application of these steps for parametrizing the elements C, H, N, and O concluded by a summary of the final parameter set.

1. *Initial guess for electronic and repulsive parameters*: Most electronic parameters are derived from DFT and can be calculated. For the remaining parameters, the com-

pression radii[1], one may either start with values from a yet existing parameter set or choosing r^{wf} as twice the covalent radius and $r^{\mathrm{dens}} = 10$ a.u. for elements of the second and third period of the systems of elements. For elements of higher periods r^{dens} may be increased. For the repulsive parameters a fit-input is prepared containing the most simple reference molecules (tentatively including different bond types such as single, double, and triple bonds). The cut-off radii should lie between first and second neighbor distance such that the repulsive parameters affect only covalent bonds (a list of selected distances is provided in appendix C). The only importance of the resulting parameters is that they should give reasonable covalent bond lengths of the reference systems that are necessary for the next step. Note that in this approach calculated atomic electronic energies will be used (approach 2 described in the last chapter) rather than fitting them together with the repulsive potentials.

2. *Tuning the compression radii*: A set of criteria needs to be compiled that depend only on the compression radii. That includes bond angles, dihedral angles, vibrational bending modes (not stretching modes because they are influenced by the repulsive parameters), and hydrogen bonds (which should not be influenced by the repulsive parameters) of selected reference systems. The compression radii should then be adjusted to yield sufficient accuracy. There might be several degrees of freedom within the paramter space, these will be reduced in the third parametrization step. Note that for every set of compression radii the repulsive parameters need to be refitted using the input as described in the first step. This is usually a fast procedure because the input does not need to be re-adjusted for a large range of compression radii and the runtime of the fit itself takes only a few seconds.

3. *Tuning the repulsive parameters*: Obviously, properties are of interest that are affected directly by the choice of repulsive parameters. Those are atomization energies and reaction energies, bond lengths, vibrational stretching frequencies of equilibrium structures, and also the energetic behavior for molecules with stretched bonds. A selection needs to be made for a training set 1 (using the nomenclature of the last chapter) which contains reference molecules and properties that enter the fit and a slightly larger training set 2 which is necessary for quickly estimating the performance of the resulting parameters. Thus, in training set 2 (if not already in training set 1) most important chemical bonding situations are included. For tuning the parameters reference systems are included in or eliminated from training set 1, additional equations for describing vibrational frequencies are taken into account, and division points for the spline function representing the repulsive potentials are varied. This iterative way of finding an optimal input for creating the parameters is guided by several concepts as shown below but also contains an empirical portion.

   During this process usually conflicts of objectives are detected. In order to reduce them as much as possible the compression radii need to be readjusted within the reasonable range as found in the second step. This of course may be repeated many times, however, in case of the C, H, N, and O parametrization the answer quickly converges to a final parameter set.

4. *Tuning the Hubbard derivatives (and the ζ parameter)*: In chapter 4 it has been shown that the calculated Hubbard derivative values might not be the best choice for re-

[1]The parameter ζ is a specialty necessary only for fits that include hydrogen and will be discussed in the following subsections.

producing proton affinities and hydrogen binding energies. Of course, fitting of such parameters—which could principally be determined from DFT based calculations—is methodologically not desired. However, as has been demonstrated in that chapter the only effect of changing Hubbard derivatives in a physically reasonable range is seen for species containing highly localized net charges. For neutral molecules and properties such as geometries, atomization energies, and vibrational frequencies, this effect is negligible. Further in chapter 4 it has been found that only for two atom types, namely carbon and phosphorus, a smaller Hubbard derivative is needed which indicates that for these elements a stronger net charge dependence is necessary. For all other elements the optimized value is very similar to the calculated one.

In that sense a fitting or rather a tuning of the Hubbard derivatives should be carried out with caution. Along this line the fit is accomplished as a last step in the fitting procedure keeping electronic and repulsive parameters fixed.

6.1.1 Initial Guess

As initial guess for the electronic parameters the ones of the MIO set are used. The only exceptions are that the atomic spin polarization energies are calculated using the PBE functional instead of LDA and ϵ_s for nitrogen is set to the calculated value rather than a manipulated one (compare section 3.3). The Hubbard derivatives and the ζ parameter are determined as described in subsection 4.3.2 (the "calc" parameter set).

The repulsive parameters are fitted to the B3LYP/cc-pVTZ equilibrium geometries and atomization energies calculated with G3B3 of the molecules hydrogen, methane, ethane, ethene, ethyne, ammonia, nitrogen, methylamine, methanimine, water, methanol, formaldehyde, oxygen (singlet), and nitromethane (training set 1). The search for division points of the spline intervals and additional equations is carried out in the way described in the previous chapter, however, focusing only on qualitative trends. It is carefully checked that the covalent bond lengths for the reference systems in the second step are reasonably well described.

6.1.2 Compression Radii

The list of criteria necessary for this parametrization step is defined as bond angles of ammonia and water, vibrational bending modes for ammonia, water, methane, heavy atom distances for water dimer, water-ammonia dimer, and ammonia dimer, and the binding energy of the water dimer. Further dihedral angles of the lowest lying conformer (C^{7eq}) of N-acetylalanyl-N'-methylamide (Ac-Ala-NHMe) [92] are taken into account resembling the important features of peptide bonds (Figure 6.1). The reference geometries are calculated from B3LYP/cc-pVTZ[2], for vibrational frequencies experimental values are considered [123]. The reference value for the binding energy of the water dimer is 5.0 kcal/mol which stems from a very accurate calculation from ref [104].

Several sets of compression radii are created. For each set the repulsive parameters are fitted using the input of the initial guess, taking care that the covalent bond lengths of the above systems are reasonably well described. Starting with the compression radii of the initial guess one compression radius is changed at a time in order to observe the effect

[2]Note that B3LYP/cc-pVTZ gives accurate results, e.g. the oxygen-oxygen distance in the water dimer is 2.910 Å being in accord with the high-level calculation of Klopper et al. [104] (2.912 Å).

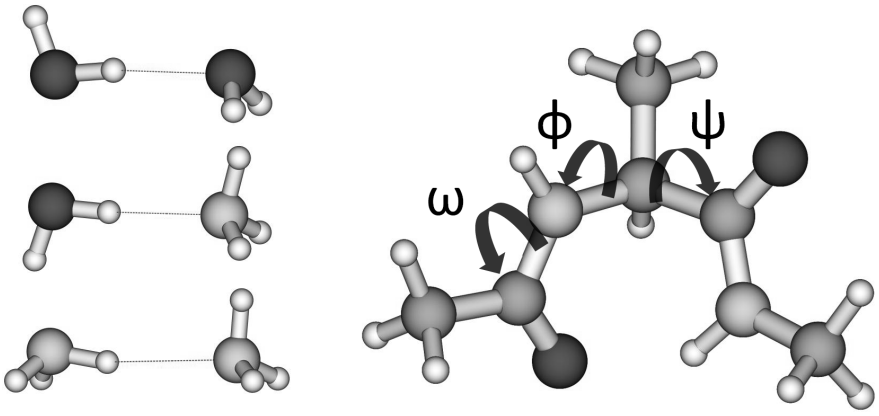


Figure 6.1: The heavy atom distances of the water dimer, water-ammonia, and ammonia dimer, as well as the dihedral angles in Ac-Ala-NHMe enter the criteria list defined for choosing the compression radii.

of each compression radius. Two further sets are generated starting from the initial guess where ζ is changed. The ζ parameter enters the γ^{h} function (eq 4.18) and is responsible for the charge-charge interaction of X–H atom pairs (X=C,N,O). The results are shown in Table 6.1.

Using DFTB2 in combination with the MIO parameter set (called MIO2) all properties are very well described, the only drawback is the nitrogen-nitrogen distance within the ammonia dimer. MIO3 (MIO applied in combination with DFTB3) slightly worsens the situation. The bond angles become larger, the vibrational bending mode of water is underestimated, and also the oxygen-oxygen distance in the water dimer is shortened.

Parameter set a01 shown in Table 6.1 is the initial guess using the same compression radii as MIO. The difference to MIO3 is the different on-site s-orbital energy of nitrogen $\epsilon_{\mathrm{s}}^{\mathrm{N}}$. Most prominent is the effect on the dihedral angles of Ac-Ala-NHMe where a01 reveals large errors. These errors were actually the reason in MIO to change the $\epsilon_{\mathrm{s}}^{\mathrm{N}}$ (see discussion below). Nevertheless, for the moment the empirical choice of the $\epsilon_{\mathrm{s}}^{\mathrm{N}}$ is avoided.

The effect of changing a single compression radius or ζ is investigated for the parameter sets b02–b19 in Table 6.1. Combining these effects—even though they of course do not add linearly—lead to the final parameter set of this parametrization step (b20). A few comments will illucidate the choice of compression radii and ζ in b20:

- The choice of $r_{\mathrm{C}}^{\mathrm{wf}}$ is most sensitive for the dihedral angle of Ac-Ala-NHMe in comparison to all other comression radii, a large value is necessary.

- For changing $r_{\mathrm{C}}^{\mathrm{dens}}$ only minor changes are detected, thus, the original MIO value of 7.0 a.u. is kept. It is a degree of freedom which can principally be changed in the next parametrization step without destroying the performance on the current step.

- The values for $r_{\mathrm{H}}^{\mathrm{wf}}$ and $r_{\mathrm{H}}^{\mathrm{dens}}$ affect all properties. Extensive tests have shown that the MIO values for these two parameters still seem to provide the best balance. An important issue is that $r_{\mathrm{H}}^{\mathrm{wf}}$ should not be larger because then the bond angles of ammonia and water are overestimated, and should not be to small because that would shorten the dimer distances.

- A large $r_{\mathrm{N}}^{\mathrm{wf}}$ lengthens the dimer distances. Care is taken not to overestimate the N–O distance of the water-ammonia dimer.

Table 6.1: Errors of Several Parameter Sets for the Criteria Defined in Order to Choose the Compression Radii[a]

Name[b]	Change[c]	a^{NH_3}	a^{H_2O}	$\nu_1^{NH_3}$	$\nu_2^{NH_3}$	ν^{H_2O}	$\nu_1^{CH_4}$	$\nu_2^{CH_4}$	$r_{OO}^{(H_2O)_2}$	$r_{NO}^{H_2O-NH_3}$	$r_{NN}^{(NH_3)_2}$	ω	ϕ	ψ	$\Delta E^{(H_2O)_2}$
REF		106.5	104.5	950	1627	1595	1306	1534	2.910	2.950	3.288	2.5	–83.7	73.1	–5.0
MIO2	—	+3.4	+2.7	+40	–4	–70	+31	–19	–0.05	+0.04	–0.22	–1	+2	–3	+1.7
MIO3	—	+4.7	+5.6	–54	–55	–180	+19	–31	–0.10	+0.03	–0.23	+1	+2	–4	+0.1
a01	—	+1.7	+6.0	+117	+9	–164	+35	–13	–0.10	–0.13	–0.33	–16	+13	+14	+0.3
b02	r_C^{wf}=2.4	+1.7	+6.0	+118	+10	–164	+57	+11	–0.10	–0.13	–0.33	–19	+15	+20	+0.3
b03	r_C^{wf}=3.0	+1.6	+6.0	+117	+8	–165	+17	–32	–0.10	–0.13	–0.33	+11	–1	–1	+0.3
b04	r_C^{dens}=5.0	+1.7	+6.0	+117	+9	–164	–6	–51	–0.10	–0.13	–0.33	–16	+12	+14	+0.3
b05	r_C^{dens}=9.0	+1.7	+6.0	+117	+9	–164	+55	+7	–0.10	–0.13	–0.33	–16	+14	+14	+0.3
b06	r_H^{wf}=2.5	+0.2	+4.4	+210	+34	–111	+36	–14	–0.12	–0.14	–0.35	–18	+16	+15	+0.4
b07	r_H^{wf}=3.5	+2.9	+7.4	+38	–13	–208	+34	–9	–0.09	–0.12	–0.31	–14	+11	+13	+0.3
b08	r_H^{dens}=2.0	+1.5	+6.0	+90	–35	–212	–4	–55	–0.11	–0.13	–0.33	–16	+14	+14	+0.1
b09	r_H^{dens}=3.0	+1.8	+6.0	+138	+40	–128	+62	+17	–0.09	–0.12	–0.32	–16	+13	+14	+0.5
b10	r_N^{wf}=1.9	+1.8	+6.0	+138	+44	–164	+34	–13	–0.10	–0.20	–0.45	–15	+13	+13	+0.3
b11	r_N^{wf}=2.5	+1.6	+6.0	+100	–19	–164	+35	–13	–0.10	–0.04	–0.20	–17	+14	+15	+0.3
b12	r_N^{dens}=9.0	+1.5	+6.0	+123	+1	–164	+35	–13	–0.10	–0.12	–0.33	–17	+14	+15	+0.3
b13	r_N^{dens}=13.0	+1.8	+6.0	+115	+14	–164	+34	–13	–0.10	–0.13	–0.33	–15	+13	+13	+0.3
b14	r_O^{wf}=2.0,	+1.7	+5.9	+117	+7	–131	+34	–13	–0.18	–0.14	–0.33	–15	+15	+11	–0.7
b15	r_O^{wf}=2.6,	+1.7	+6.2	+117	+7	–191	+34	–13	+0.00	–0.11	–0.33	–17	+12	+17	+1.0
b16	r_O^{dens}=7.0	+1.7	+5.9	+117	+9	–176	+34	–13	–0.10	–0.14	–0.33	–16	+13	+14	+0.2
b17	r_O^{dens}=11.0	+1.7	+6.1	+117	+9	–158	+35	–13	–0.10	–0.12	–0.33	–16	+13	+14	+0.4
b18	ζ=3.50	+2.5	+7.3	+61	–24	–211	+24	–25	–0.13	–0.14	–0.36	–15	+13	+13	–0.6
b19	ζ=4.50	+1.3	+5.4	+141	23	–141	+39	–8	–0.08	–0.12	–0.31	–16	+13	+14	+0.8
b20	several [d]	+1.7	+6.0	+81	–47	–212	–6	–57	–0.04	+0.03	–0.09	+1	+2	–2	+0.5
c21	several [e]	+3.2	+6.0	–3	–76	–215	–10	–60	–0.04	+0.04	–0.08	+1	+2	–3	+0.4
d22	several [e]	+3.2	+6.0	+3	–74	–218	–15	–66	–0.04	+0.03	–0.08	+1	+1	–3	+0.3

[a] The criteria are bond angles a in °, vibrational bending modes ν in cm^{-1}, heavy atom distances r in Å, dihedral angles of Ac-Ala-NHMe ω, ϕ, and ψ in °, and the binding energy of the water dimer $\Delta E^{(H_2O)_2}$ in kcal/mol. [b] References REF are given by B3LYP/cc-pVTZ geometries, experimental fundamental frequencies [123] and $\Delta E^{(H_2O)_2}$ by a high-level calculation [104], all other numbers are deviations. MIO2 are calculations using DFTB2. MIO3 and all other sets are calculated using DFTB3; set a01 is the initial guess, b02–b20, c21, and d22 denote sets of the second, third and fourth step of the fitting procedure, respectively. [c] The compression radii for MIO2 and MIO3 set are r_C^{wf}=2.7, r_C^{dens}=7.0, r_H^{wf}=3.0, r_H^{dens}=2.5, r_N^{wf}=2.2, r_N^{dens}=11.0, r_O^{wf}=2.3, r_O^{dens}=9.0. For MIO3 ζ=4.05 is applied. [d] r_C^{wf}=3.5, r_C^{dens}=7.0, r_H^{wf}=3.0, r_H^{dens}=2.5, r_N^{wf}=2.8, r_N^{dens}=11.0, r_O^{wf}=2.5, r_O^{dens}=6.0, ζ=4.00. [e] r_C^{wf}=3.3, r_C^{dens}=6.5, r_H^{wf}=3.0, r_H^{dens}=2.5, r_N^{wf}=2.8, r_N^{dens}=10.0, r_O^{wf}=2.5, r_O^{dens}=6.0, ζ=4.00 (c21), ζ=3.90 (d22). Note, for c21 and d22 an empirical value for $\epsilon_s = -0.66$ is applied.

- The compression $r_{\mathrm{N}}^{\mathrm{dens}}$ of 11.0 a.u. is maintained because otherwise it is difficult to keep the good description of the Ac-Ala-NHMe dihedral angles. Further tests have shown that for even larger $r_{\mathrm{N}}^{\mathrm{dens}}$ problems arise for fitting the repulsive potentials (within the next parametrization step).

- Similar as for nitrogen the value for $r_{\mathrm{O}}^{\mathrm{wf}}$ is risen to increase the water dimer distance. However, a balance with the decreasing water dimer binding energy is necessary.

- Lowering $r_{\mathrm{O}}^{\mathrm{dens}}$ helps to strengthen the water dimer bond and for a smaller water monomer angle but also slightly lowers the bending mode (increasing the error). When fitting the repulsive potentials in the next parametrization step it is found that a small $r_{\mathrm{O}}^{\mathrm{dens}}$ is helpful.

- A lower value of ζ means that the repulsion between a partially positive charged hydrogen and other partially positive charged atoms becomes stronger and the attraction to partially negative charged atoms is smaller (compare section 4.2). This causes the bond angles of ammonia and water to become larger and on the other hand the dimer distances to become smaller. Additionally the water dimer binding energy decreases. For the b20 parameter set the ζ parameter remains almost unchanged.

To summarize, the b20 set considerably improves the dimer distances, but the drawbacks of a too large water monomer angle and a too low water vibrational bending mode persist comparing to MIO3. When comparing b20 to MIO2 it is worth noting that the hydrogen binding energy of the water dimer is decreased almost matching the reference value even though the distance between the heavy atoms has become slightly larger, an effect which was not possible to achieve during the MIO parametrization for DFTB2 [18].

It is remarkable that the empirical shift of the parameter $\epsilon_{\mathrm{s}}^{\mathrm{N}}$ (s-orbital energy of the neutral nitrogen atom) as used in MIO is not necessary for properly describing the dihedral angles of Ac-Ala-NHMe. This led to some further investigations as summarized in Table 6.2 for parameter sets b20–b23. While b20 uses the calculated $\epsilon_{\mathrm{s}}^{\mathrm{N}} = -0.6819693$, that value is changed for parameter sets b21–b23. The compression radii remain constant but the repulsive potentials are refitted for every set. (For completeness also the final parameter set d22 is listed in Table 6.2.) Elstner already described [18] that a shift of $\epsilon_{\mathrm{s}}^{\mathrm{N}}$ affects the hybridization of nitrogen. For a larger energy difference between s and p orbital sp^3-hybridized systems are favoured. A simple test is the inversion barrier of ammonia $\Delta E_{\mathrm{NH_3}}^{\mathrm{bar}}$ (Figure 6.2). In the ground state ammonia is sp^3-hybridized, at the barrier s and p orbitals hybridize to sp^2. Thus, when shifting $\epsilon_{\mathrm{s}}^{\mathrm{N}}$ towards $\epsilon_{\mathrm{p}} \approx -0.26$ a.u. the barrier becomes smaller. A G3B3 reference calculation gives a value of 5.0 kcal/mol and can be only matched for even smaller $\epsilon_{\mathrm{s}}^{\mathrm{N}}$ than the calculated one. On the other hand Table 6.2 compares the OCNH dihedral angle of formamide d^{form} (compare Figure 6.2). While formamide should be planar with an sp^2-hybridized nitrogen it is calculated with an dihedral of 14.8° using parameter set b20. Consequently, rising $\epsilon_{\mathrm{s}}^{\mathrm{N}}$ lowers the error. Similar case can be found for the dihedral of acetamide d^{acet}. The reported experimental value is 7.6° [123]. Note that also high-level calculations show a considerable discrepancy; B3LYP/cc-pVTZ calculation gives an angle of 2.1° while MP2/cc-pVTZ yields 9.1°.

In formamide a partial double bond between carbon and nitrogen is formed ($r_{\mathrm{CN}}^{\mathrm{form}}$). While a typical C=N double bond lies at 1.26 Å (methanimine) and a C–N single bond at 1.46 Å (aminomethane) it is 1.36 Å for formamide. The bond length should clearly depend on the hybridization state. As shown in Table 6.2 this, however, is only a subtle effect for parameter sets b20–b23.

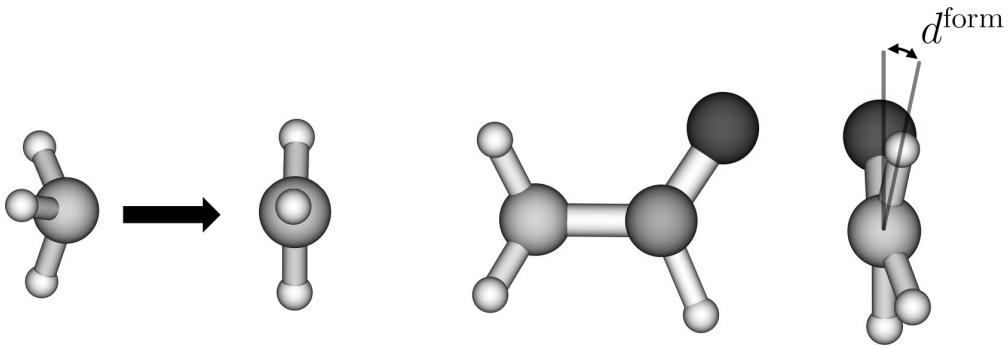


Figure 6.2: Inversion barrier of ammonia and dihedral angle d^{form} of formamide are two criteria to choose the parameter $\epsilon_{\text{s}}^{\text{N}}$.

Table 6.2: Dependence of $\epsilon_{\text{s}}^{\text{N}}$ on the Performance for Nitrogen Hybridization Related Properties[a]

Parameter set	$\epsilon_{\text{s}}^{\text{N}}$	$\Delta E_{\text{NH}_3}^{\text{bar}}$ in kcal/mol	d^{form} in °	d^{acet} in °	$r_{\text{CN}}^{\text{form}}$ in Å
Reference	—	5.0	0.0	7.6	1.357
MIO2	–0.640	3.4	6.4	10.0	1.375
MIO3	–0.640	2.3	0.1	4.4	1.375
b20	–0.6819693[b]	4.3	14.8	15.9	1.391
b21	–0.670	3.6	10.7	12.7	1.388
b22	–0.665	3.3	8.2	10.8	1.387
b23	–0.660	3.0	4.0	8.5	1.386
d22	–0.660	3.1	8.7	10.5	1.385

[a] $\Delta E_{\text{NH}_3}^{\text{bar}}$—inversion barrier of ammonia; d^{form} and d^{acet}—OCNH dihedral angle of formamide and acetamide; $r_{\text{CN}}^{\text{form}}$—C–N bond length in formamide. [b] This is the calculated value.

Extensive tests were carried out whether one can match the dihedrals and the barrier just by changing the compression radii without introducing an empirical shift of $\epsilon_{\text{s}}^{\text{N}}$, however, no set could be found that at the same time shows a reasonable performance for all the criteria shown in Table 6.1. Therefore, it was decided to use parameter set b23 with $\epsilon_{\text{s}}^{\text{N}} = -0.660$ for the following parametrization steps. Set b23 reveals an acceptable balance between a too small ammonia inversion barrier and too large dihedral angles.

6.1.3 Repulsive Parameters

Technical details and the general procedure to find repulsive potentials has been described in chapter 5 and will only be repeated for new aspects appearing within the parametrization of C, H, N, and O.

A training set 1 was already defined in the initial guess and will now be modified. As training set 2 the G2/97 molecule set is used [42] extended by the molecules oxygen (singlet), methanimine, *trans*-diazene (HN=NH), nitrous acid, and nitric acid in order to include further relevant bonding situations. The reference atomization energies are calculated using G3B3 at 0 K excluding zero point and thermal corrections, reference geometries are calculated using B3LYP/cc-pVTZ. Further experimental fundamental vibrational frequencies were used as listed in the CCCBDB database [123].

Training set 1 is used for simultaneously fitting all repulsive potentials. The performance of new potentials is tested on that set, i.e. equilibrium geometries, atomization energies and vibrational frequencies are determined. Depending on the performance division points and additional equations defining the potentials are added, changed, or removed.[3] When a reasonable performance is found the potentials are tested on training set 2. This is important because one needs to check whether certain bonding situations which are not included in training set 1 are reasonably well described and should possibly be included to training set 1. Note that training set 1 is kept as small as possible in order to find systematic deficiencies within the potentials. For a large training set 1 a lot of information for a few bond length ranges (e.g. single bond, double bond, triple bond) would enter the fit and it would be more difficult to determine the division points and additional equations for smooth potentials, first and second derivatives.

A general trend is found concerning the dependence of the compression radii on the repulsive potentials. For large values of r^{dens} the fitted potential is strongly repulsive. The same effect is found for the wave function compression r^{wf}, but in opposite manner. For a larger r^{wf} the potential becomes less repulsive. These effects are used to find optimal shapes for the repulsive potentials, i.e. curves that smoothly yield zero at the cut-off distance. The compression radii are slightly changed in this parametrization step keeping in mind the findings of the second parametrization step (last subsection). In the following, comments for each pair potential are given. A detailed list of all parameters defining the repulsive potentials is shown in subsection 6.1.5. Graphs for all potentials are illustrated in appendix A. The resulting parameter set of this parametrization step is called c21.

H–H: Three intervals and one additional equation are needed to match the reference atomization energy, bond length, and vibrational frequency of the H_2 molecule (compare also to subsection 5.2.1). The potential is negative in the binding region mainly due to the very small density compression of hydrogen. As mentioned in section 3.3 this might cause problems in molecular dynamics simulations in the case of two hydrogens coming too close to each other and forming an unphysical bond due to the repulsive potential minima. However, in comparison to the MIO potential this problem is mildened because the gradient is very small in the cut-off region.

H–C: The H–C bond lengths for all molecules within training set 2 are almost alike. Therefore, similar as for the H–H potential one additional equation and three intervals are used.

H–N: Besides the covalent H–N bond as apparent in ammonia, proton transfer barriers of the two model systems $[H_3N–H–NH_3]^+$ and $[H_2N–H–NH_2]^-$ are included in the fit as reaction equations. The reference energies are calculated using MP2/cc-pVTZ for several fixed nitrogen-nitrogen distances (compare Table 6.5). Both, the relaxed geometry where the shared proton is bound to one of the ammonia molecules and the barrier structure where the proton is at half the distance between the nitrogen atoms include bond lengths in the range of 1.1–1.8 Å. Therefore, besides three intervals and one additional equation for the covalent H–N bond one further interval is introduced giving the spline function more flexibility. The reaction energies, however, do not include information about the gradient of the potential. Thus, while the fit gives exact results for the barriers, a geometry optimization of the model system (with the constraint of two fixed nitrogens) using the fitted potential relaxes to some unspecific structure. To correct the gradient two additional equations are necessary. This

[3] This is a small difference compared to the approach described in section 5.2 where always the performance of training set 2 was evaluated. However, for CHNO the number of molecules has become too large making this procedure too cumbersome.

causes a small "hump" in the potential at around 1.6 Å, i.e. at 1.3 Å the potential is almost zero and rises to about 0.5 kcal/mol at 1.6 Å and then decreases again towards the cut-off.

H–O: Similar as for the H–N potential, the water molecule and two proton transfer barrier models are included, $[H_2O–H–OH_2]^+$ and $[HO–H–OH]^-$. In this case it turned out that one additional division point is necessary, but no additional equations.

C–C: As has been shown in section 5.2 ethane, ethene and ethyne resembling single, double and triple bond seem to be sufficient for describing the bond lengths of all hydrocarbons for a proper choice of intervals. For further tuning the performance the compression radii are slightly changed (r^{wf} from 3.5 to 3.3 a.u. and r^{dens} from 7.0 to 6.5 a.u.), and also benzene and some weighting factors different from 1.0 have been applied (for details see Table 6.5). This, of course, is an empirical procedure of trial and error, however, it does slightly reduce the errors for atomization energies and bond lengths. For the vibrational frequencies one additional equation is added, however, the C=C double bond stretching frequency of ethene seems not to be adjustable without deteriorating the geometrical properties. The error remains at about 150 cm^{-1}.

C–N: Single, double, and triple C–N bond are included in the fit via the molecules hydrogen cyanide, methanimine, and methylamine. Energetics and bond lengths for single and double bonds can be well reproduced and are even improved by the change of the carbon compression radii as mentioned above and also by lowering the nitrogen density compression to 10.0 a.u.. For a slight improvement, trimethylamine is included in training set 1. One drawback is found, the C=N double bond stretching frequency of methanimine is overestimated by roughly 230 cm^{-1}. Reducing this error cannot be enforced by an additional equation without loosing accuracy for energetics and geometries. Further, as will be shown below, the atomization energy of hydrogen cyanide has an error of more than 10 kcal/mol, additional tests indicate that it might be possible to reduce this error in a future parameter set by a different choice of division points.

C–O: Single and double bond lengths for methanol and formaldehyd lie at 1.42 and 1.20 Å. For formic acid the C–O single bond length lies at 1.34 Å and has to be included into the fit because otherwise the C–O bond length becomes too large. For a more subtle balance between the performance of these different bonding situations also methyl formate and acetic acid have been included into training set 1. It was found that carbon monoxide and carbon dioxide are not well described, a further tuning of division points might help to alleviate these errors.

N–N: With dinitrogen and hydrazine the most common N–N bonds are covered. It is also found that the N=N double bond length in *trans*-diazene is well described even though is not contained in training set 1.

N–O: There are three substantially different N–O bond lengths for all molecules of training set 2 which can be characterized as single, double, and partial double bond. The two molecules nitrous acid and nitromethane resemble these bond lengths (1.18 Å, 1.39 Å, 1.22 Å) and are included in the fit. However, potential and gradients cannot be smoothly interpolated between these different bond types, introducing "humps" in the intermediate regions (see appendix A for the graph of that potential). Another issue is that nitrous acid contains an H–O bond. During the fit, this drastically affects the H–O potential. Therefore, a small weigthing factor of 0.1 for this molecule (energy and force equations) is introduced.

O–O: Due to the fact that spin symmetry is not considered, the singlet oxygen molecule is applied to the fit. Another relevant O–O bond length appears for peroxides. Therefore, the H_2O_2 molecule is included in training set 1. Together with one additional equation, also bond lengths and the vibrational stretching modes of both molecules can be well described.

Table 6.3: Calculated and Fitted Hubbard Derivatives for Parameter Sets c21 and d22

Parameter set	ζ	$U_{\mathrm{C}}^{\mathrm{d}}$	$U_{\mathrm{H}}^{\mathrm{d}}$	$U_{\mathrm{N}}^{\mathrm{d}}$	$U_{\mathrm{O}}^{\mathrm{d}}$
c21	4.05	–0.1492	–0.1857	–0.1535	–0.1575
d22	3.90	–0.24	–0.16	–0.12	–0.15

Table 6.4: Overview of the New Electronic Parameters (in Atomic Units if not Unitless), $\zeta = 3.9$

Parameter	H	C	N	O
l_{max}	0	1	1	1
n_{max}	2	2	2	2
α_0	0.50	0.50	0.50	0.50
α_1	1.00	1.14	1.21	1.26
α_2	2.00	2.62	2.90	3.17
α_3	—	6.00	7.00	8.00
p	2	2	2	2
r^{wf}(s,p)	3.0	3.3	2.8	2.5
r^{dens}	2.5	6.5	10.0	6.0
ϵ_{s}	–0.2386004	–0.5048917	–0.66	–0.8788325
ϵ_{p}	—	–0.1943551	–0.2607280	–0.3321317
E^{spin}	–0.0410614	–0.0454791	–0.1147656	–0.0557761
U	0.4195	0.3647	0.4309	0.4954
U^{d}	–0.16	–0.24	–0.12	–0.15

6.1.4 Tuning the Hubbard Derivatives

Several indications were found in section 4.3 that tuning the Hubbard derivatives might not be necessary for new repulsive potentials, i.e. the H–O overbinding within the MIO parameters caused systematically too large proton affinities for 23 oxygen containing systems tested (Table 4.6). Indeed it is found that the mean signed error of +5.2 kcal/mol is reduced for the new parameter set c21 to +1.3 kcal/mol. The mean unsigned error drops from 5.5 kcal/mol for MIO3 down to 4.0 kcal/mol. Using MIO3 and fitted Hubbard derivatives (MIO3/fit) the error is at 2.9 kcal/mol. Therefore, also for the new parameter set the Hubbard derivatives were tuned.

Starting from parameter set c21 the Hubbard derivatives U^{d} and the ζ parameter are fitted in the way described in section 4.3.2 using 22 hydrogen binding energies and 32 proton affinities. It is found that the fitted values are almost the same as the calculated ones with one exception for carbon. Similar as found before (section 4.3.2) this value becomes smaller indicating that a stronger charge dependency of carbon improves the description of the proton affinities. In order not to use unphysically small values, U_C^{d} is restricted to a lower boundary of –0.24 a.u.. The fitted values used for the final parameter set d22 are listed in Table 6.3. Results for binding energies and proton affinities are shown in appendix C.

6.1.5 Summary of New Parameters

This subsection provides a compact presentation of the new parameter set (d22). The new electronic parameters are listed in Table 6.4. Details of each parameter entering the DFTB method are shown in section 3.3. The element independent parameter for the γ^{h} function

Table 6.5: Overview of Parameters Defining the Repulsive Potentials

Molecules of training set 1[a] (w^{eeq}, w^{feq} if not equal 1.0) and E^{at}					
Dihydrogen	109.8	Hydrogen cyanide	313.3	Methyl formate	785.9
Methane	419.7	Methanimine	438.5	Dinitrogen	227.8
Ammonia	296.8	Methylamine	580.5	Hydrazine	436.1
Water	231.9	Trimethylamine	1159.2	Nitrous acid (0.1, 0.1)	311.3
Ethyne (2.0, 1.0)	404.9	Formaldehyde	374.4	Nitromethane	601.1
Ethene (2.0, 2.0)	563.9	Methanol	512.1	Dioxygen (singlet)	91.4
Ethane (2.0, 4.0)	711.4	Formic acid	500.9	Hydrogen peroxide	267.5
Benzene (1.5, 1.0)	1365.9	Acetic acid	802.2		

Proton transfer reactions [b]	r^{XX}	E^{bar}
$[H_2O–H–OH_2]^+$	2.5, 2.6, 2.7, 2.8	0.61, 2.36, 5.20, 8.85
$[HO–H–OH]^-$	2.5, 2.6, 2.7, 2.8	0.50, 2.30, 5.16, 8.81
$[H_3N–H–NH_3]^+$	2.6, 2.7, 2.8, 2.9	0.40, 1.94, 4.44, 7.70
$[H_2N–H–NH_2]^-$	2.5, 2.6, 2.7, 2.8	0.05, 1.44, 3.49, 6.26

Potential	Division points (a.u.)	Additional equations (a.u.)
H–H	1.4, 1.8, 2.5, 2.6	$V''(1.404) = 0.423$
H–C	2.0, 2.5, 3.4, 3.5	$V''(2.043) = 0.375$
H–N	1.8, 2.5, 3.0, 3.5, 4.4	$V''(1.916) = 0.500, V''(2.362) = 0.150,$ $V''(2.835) = -0.020$
H–O	1.7, 2.2, 2.7, 3.4, 3.7	$V''(1.816) = 0.620$
C–C	2.0, 2.8, 3.0, 4.0, 4.8	$V''(2.513) = 0.920$
C–N	2.0, 2.4, 3.1, 3.8, 4.2	$V''(2.767) = 0.350, V''(2.166) = 1.700$
C–O	2.0, 2.9, 3.5, 4.6	$V''(2.683) = 0.370$
N–N	2.0, 3.3, 3.4, 4.8	$V''(2.062) = 2.400, V''(2.714) = 0.350$
N–O	1.9, 2.4, 2.6, 3.5, 3.6, 3.9	$V''(2.268) = 1.300, V''(2.589) = 0.900$
O–O	2.0, 2.8, 3.4, 4.4, 4.5	$V''(2.279) = 0.750$

[a] E^{at} is the reference atomization energy in kcal/mol, w^{eeq} and w^{feq} are the weighting factors for energy and force equations in 1/a.u.; reaction equations and additional equations are weighted with 1/a.u. [b] E^{bar} is the reference proton transfer barrier in kcal/mol for a heavy atom distance r^{XX} in Å.

is $\zeta = 3.9$.[4] The parameters defining the repulsive potentials are shown in Table 6.5. The reference atomization energies are calculated at 0 K using the G3B3 method excluding zero point vibrational energy and thermal corrections. As reference geometries B3LYP/cc-pVTZ equilibrium structures are applied. For the proton transfer barriers, energetics and geometries have been determined using MP2/G3Large. All reference calculations have been carried out using the Gaussian 03 software package [108].

6.2 Performance of the New Parameters

In this section the performance of the new parameter set d22 is compared to the MIO parameters using DFTB3 (called MIO3). For completeness, also results of MIO in combination with DFTB2 are presented (MIO2).

[4] The hamilton and overlap matrix elements defined in eqs 3.20 and 3.12 are tabulated for 600 equidistant interatomic distances starting from 0 a.u. with an increment of 0.02 a.u.

Table 6.6: Mean Unsigned Errors of Atomization Energies of the G2/97 Set for 61 CHNO-Containing Closed Shell Molecules

Set	N[a]	MIO2[b]	MIO3[b]	d22
Calculated atomic energies				
All	61	44.1	45.4	5.5
Hydrocarbons	22	40.0	41.2	2.2
CHN-compounds	10	51.0	52.1	6.9
CHO-compounds	15	42.5	44.3	4.4
XNO-compounds	3	94.6	93.1	12.0
Fitted atomic energies				
All	61	6.4	6.1	4.7
Hydrocarbons	22	6.4	5.9	2.3
CHN-compounds	10	2.9	2.9	5.2
CHO-compounds	15	5.1	5.3	3.3
XNO-compounds	3	26.1	22.9	12.0

[a] Number of comparisons. [b] These values are for the original MIO parametrization. The errors reduce already by taking PBE spin-polarization values instead of the LDA ones. The overall error are then 14.8 and 15.4 kcal/mol for MIO2 and MIO3.

6.2.1 Atomization Energies, Geometries, Vibrational Frequencies

A test for atomization energies for the G2/97 molecule set (only CHNO containing molecules) is shown in Table 6.6. As reference atomization energies G3B3 values excluding zero-point vibrational and thermal corrections are used. The d22 set yields a remarkable low mean unsigned error of 5.5 kcal/mol. Breaking down the statistics to different compound types, it is seen that, even though improved, large errors still remain for compounds containing N–O bonds. Also for systems containing C≡N triple bonds, in the G2 set those are HCN, H_3C–C≡N, and H_3C=CH–C≡N, large errors of up to 15 kcal/mol are found. Even worse, for N≡C–C≡N the largest deviation for the G2 set of 38 kcal/mol is found. Further, carbon monoxide and carbon dioxide reveal errors of about 20 kcal/mol.

Due to the strong overbinding MIO2 and MIO3 reveal extremely large errors. Nevertheless, these errors are consistent and can be substantially reduced by refitting the atomic energies. The fit has been carried out to training set 1 (Table 6.5), a detailed description for this fitting is given in subsection 5.2.1, the atomic energies are listed in Table 6.7. The mean unsigned error for all 61 molecules is found to be 6.4 kcal/mol. A fit to the complete G2/97 molecule set does not significantly further reduce this error (6.3 kcal/mol) and also when removing the problematic N–O bond containing systems from the fit procedure an overall error of 6.1 kcal/mol remains. Such a fit has also been carried out for the d22 set as shown in Table 6.6, the overall error dropping down almost 1 kcal/mol to 4.7 kcal/mol. However, the above mentioned large errors persist and a further improvement for these cases is subject of ongoing work.

DFTB results for geometrical properties of the G2/97 set are compared to MP2/cc-pVTZ calculations. The definition of 223 bond lengths and 187 bond angles that are compared are taken from ref [20]. In appendix C (Table C.5) errors for specific bonding situations are listed. The overall good performance of the MIO parameter set is maintained for the d22 set. The mean unsigned error for all bond lengths drops from 0.014 down to 0.009 Å, however, this is mainly due to a statistic effect because the C–H bond lengths—dominating

Table 6.7: Atomic Electronic Energies Fitted to 23 Atomization Energies of Training Set 1

Parameter set	H	C	N	O
MIO2	–0.2727	–1.4585	–2.2116	–3.1674
MIO3	–0.2738	–1.4571	–2.2114	–3.1685
d22	–0.2789	–1.4454	–2.2208	–3.1436

Table 6.8: C–C, C–N, and C–O Vibrational Stretching Frequencies in cm^{-1}

Molecule	Experiment	BLYP/cc-pVTZ	MIO2	MIO3	d22
$H_3C–CH_3$	995	958	1131	1125	1014
$H_2C=CH_2$	1623	1639	1824	1819	1774
HC≡CH	1974	2009	2114	2114	2018
$H_3C–OH$	1033	1051	1046	1046	1029
$H_2C=O$	1746	1748	1876	1869	1705
C≡O	2170	2114	2394	2395	2027
$H_3C–NH_2$	1130	1144	1204	1192	1069
$H_2C=NH$	1638	1646	1904	1902	1865
HC≡N	2089	2113	2253	2254	2058

the statistics with 102 bonds compared—are slightly improved. The largest error for MIO2 and MIO3 is found for the N–O bond within methylnitrite deviating 0.065 and 0.062 Å while for d22 the carbon monoxide bond length is about 0.066 Å too large in comparison to MP2. Surprisingly, the large error of bicyclobutane as found in section 5.2 is removed, the reason is not understood. Another finding is that the C–H bond length of aldehyde groups is overestimated for both parameter sets (independent of whether DFTB2 or DFTB3 is used). For example, the C–H bond length in formaldehyde is 1.142 Å for MIO2 and MIO3, 1.133 Å for d22 while the MP2/cc-pVTZ value lies at 1.101 Å. Bond angles are described excellently with a mean unsigned error of 0.9° for all sets, MIO2, MIO3, and d22. As has already been discussed large deviations for MIO3 and d22 are found for the water monomer. The HOH angle is calculated as 110.1° and 110.5° for MIO3 and d22 and thus overestimates the MP2/cc-pVTZ reference which gives an angle of 103.5°. This angle is the largest error for MIO3 and d22, all further deviations are below 4°.

Information of vibrational frequencies have been included during the parametrization process. Those are the bending modes of ammonia and water for adjusting the compression radii and vibrational stretching frequencies during the fit of the repulsive potentials. Table 6.8 shows unscaled stretching frequencies for a few characteristic bonding situations. All stretching modes besides the C–O mode of methanol are overestimated by MIO2 and MIO3. While the errors are significantly reduced for the d22 parameters set large errors remain for C=C and C=N double bond stretching modes. A comprehensive benchmark test for a large number of vibrational frequencies still remains to be carefully examined.

6.2.2 Proton Affinities, Hydrogen Binding Energies, and Proton Transfer Barriers

The tests on 23 proton affinities with acidic oxygen, 9 proton affinities with acidic nitrogen, and 22 hydrogen binding energies as have been carried out in section 4.3 are repeated for the new d22 parameter set. The details are listed in appendix C. The results are very

similar to the MIO parameter set used in combination with DFTB3 and fitted Hubbard derivatives (MIO3/fit). For the proton affinities with acidic oxygen the mean unsigned error is at 3.0 kcal/mol in comparison to 2.9 kcal/mol for MIO3. For proton affinities with acidic nitrogen the same systematic error is detected as has been described for the MIO parameters (section 4.4); systems containing sp^3-hybridized nitrogen underestimate the proton affinity by more than 10 kcal/mol. Similar to the MIO parametrization a pragmatic solution is to fit a modified H–N repulsive potential being energetically shifted by that amount.

For the hydrogen binding energy the mean unsigned error from MIO2 dropped from 8.0 kcal/mol to 2.7 kcal/mol for MIO3/fit. Even though the binding energies greatly improves going from DFTB2 to DFTB3, for the latter still a systematic trend is visible; the binding energy for protonated water clusters is overestimated, while it is underestimated for the deprotonated water clusters. This trend is somewhat more pronounced for d22 leading to a mean unsigned error of 3.4 kcal/mol.

The proton transfer barriers for positive and negative charged water and ammonia model systems have been included in the fit for the repulsive potential for both, the MIO and the d22 parametrization. Therefore, the barriers are described very well and are comparable to the DFT calculations. Details are shown in appendix C.

6.2.3 Reaction Energies and Isomerization Enthalpies

The repulsive potentials of the MIO parametrization were shifted in order to yield consistent reaction energies, i.e. all bonds are too strong (details see section 3.3). However, a very nice balance between the overbinding of different bonds could be established finding a reasonably well description of reaction energies as shown in Table 6.9.

Within the new parametrization the repulsive potentials were fitted directly to atomization energies.[5] Thus, if only small errors appear for the atomization energies also reaction energies are well described. As consequence, errors of reaction energies for d22 can directly be detected from the errors of atomization energies.

The compilation in Table 6.9 are mainly hydrogenation reactions which help to directly compare different bond types, e.g. for the first reaction in that table the C–C bond of ethane (and the H–H bond of hydrogen) breaks and a C–H bond is formed. This of course requires that hydrogen is well described, as is the case for the d22 parameter set; in case of the MIO parameters that means that the overbinding of hydrogen should be properly balanced with other bond types. This of course is difficult to judge because it is very dependent on the selection of reactions. Therefore, further tests are carried out on isomerization enthalpies as shown below.

For d22 the mean unsigned error of 49 reaction energies is 5.7 kcal/mol. MIO2 and MIO3 show relatively large mean unsigned errors of 8.2 and 8.3 kcal/mol. They mainly stem from the wrong description of N–O bonds. For the new parametrization these large errors are reduced but still deviations up to 16 kcal/mol are found. Furthermore, for d22 large errors are found for carbon monoxide, carbon dioxide, hydrazine, and *trans*-diazene. Leaving all these problematic cases out of the statistics a mean unsigned error of only 4.3, 6.4, and 4.1 kcal/mol is found for MIO2, MIO3, and d22. In comparison, B3LYP with the quite large basis set cc-pVTZ yields 2.6 kcal/mol for this statistic.

[5] This was not achieved during the MIO parametrization. It is possible in the new parametrization due to a different choice of compression radii and the usage of atomic spin-polarization energies calculated with PBE rather than LDA. For details see section 3.3.

Table 6.9: Deviation of DFTB for 49 Reaction Energies Compared to G3B3[a]

Reaction			G3B3	B3LYP[b]	MIO2	MIO3	d22
$H_3C–CH_3 + H_2$	→	$2\ CH_4$	–18.2	–1.9	+1.8	+1.0	+1.9
$H_2C{=}CH_2 + 2\ H_2$	→	$2\ CH_4$	–56.8	–1.8	–5.7	–6.6	–1.1
$HC{\equiv}CH + 3\ H_2$	→	$2\ CH_4$	–105.1	–3.4	–4.7	–5.7	+3.5
$C_6H_6 + 9\ H_2$	→	$6\ CH_4$	–163.9	–9.3	–7.0	–12.5	–0.1
$H_2N–NH_2 + H_2$	→	$2\ NH_3$	–47.7	+1.5	+4.0	–0.4	–11.8
$HN{=}NH + 2\ H_2$	→	$2\ NH_3$	–78.6	+0.8	–2.6	–10.1	–13.5
$N_2 + 3\ H_2$	→	$2\ NH_3$	–36.3	–2.0	+1.1	–7.5	–0.2
$HO–OH + H_2$	→	$2\ H_2O$	–86.5	+6.1	+0.6	–8.6	–4.5
$^1O_2 + 2\ H_2$	→	$2\ H_2O$	–152.8	+1.9	+7.4	–6.7	+9.0
$H_3C–NH_2 + H_2$	→	$CH_4 + NH_3$	–26.2	–0.8	+1.1	–1.5	+1.0
$H_2C{=}NH + 2\ H_2$	→	$CH_4 + NH_3$	–58.4	–0.3	–0.1	–4.2	+3.2
$HCN + 3\ H_2$	→	$CH_4 + NH_3$	–73.7	–2.6	+5.6	+0.4	+10.4
Pyridine $+ 9\ H_2$	→	$5\ CH_4 + NH_3$	–170.2	–7.2	–3.8	–12.6	+3.1
$H_3C–OH + H_2$	→	$CH_4 + H_2O$	–29.7	+1.0	+0.4	–4.3	+1.7
$H_2C{=}O + 2\ H_2$	→	$CH_4 + H_2O$	–57.6	+2.0	+0.3	–7.5	+6.9
$C{\equiv}O + 3\ H_2$	→	$CH_4 + H_2O$	–62.3	–1.0	+10.3	+2.3	+22.9
$CO_2 + 4\ H_2$	→	$CH_4 + 2\ H_2O$	–53.6	+5.1	+3.4	–12.8	+16.9
$HC(=O)OH + 3\ H_2$	→	$CH_4 + 2\ H_2O$	–53.1	+4.7	–7.1	–19.4	+2.3
$N_2O + 4\ H_2$	→	$2\ NH_3 + H_2O$	–115.5	+6.2	+45.9	+28.3	+15.9
$HNO_2 + 3\ H_2$	→	$NH_3 + 2\ H_2O$	–119.9	+7.8	+32.7	+16.5	–15.3
$HNO_3 + 4\ H_2$	→	$NH_3 + 3\ H_2O$	–165.7	+12.9	+79.4	+49.3	–1.3
$H_2C{=}CH_2 + H_2$	→	$H_3C–CH_3$	–38.5	+0.1	–7.5	–7.7	–3.0
$HC{\equiv}CH + H_2$	→	$H_2C{=}CH_2$	–48.3	–1.7	+1.0	+0.9	+4.6
$C_6H_6 + 6\ H_2$	→	$3\ H_3C–CH_3$	–109.2	–3.5	–12.5	–15.5	–5.6
$HN{=}NH + H_2$	→	$H_2N–NH_2$	–30.9	–0.8	–6.6	–9.7	–1.7
$N_2 + H_2$	→	$HN{=}NH$	42.3	–2.8	+3.7	+2.6	+13.3
$^1O_2 + H_2$	→	$HO–OH$	–66.3	–4.2	+6.8	+2.0	+13.5
$H_2C{=}NH + H_2$	→	$H_3C\text{-}NH_2$	–32.1	+0.4	–1.2	–2.7	+2.2
$HCN + H_2$	→	$H_2C{=}NH$	–15.3	–2.3	+5.7	+4.6	+7.2
$H_2C{=}O + H_2$	→	$H_3C–OH$	–27.9	+1.0	–0.1	–3.3	+5.2
$HC(=O)OH + H_2$	→	$H_2C{=}O + H_2O$	4.5	+2.7	–7.4	–11.9	–4.6
$HC(=O)OH + 2\ H_2$	→	$H_2C–OH + H_2O$	–23.4	+3.7	–7.5	–15.1	+0.5
$HNO_3 + H_2$	→	$HNO_2 + H_2O$	–45.8	+5.1	+46.6	+32.8	+13.9
$HN(CH_3)_2 + H_2$	→	$H_3C–NH_2 + CH_4$	–22.1	–1.9	–1.5	–3.7	–2.1
$N(CH_3)_3 + H_2$	→	$HN(CH_3)_2 + CH_4$	–19.0	–3.6	–3.2	–5.0	–4.5
$O(CH_3)_2 + H_2$	→	$H_3C–OH + CH_4$	–24.6	–0.9	–1.4	–5.2	–1.2
$HC(=O)OCH_3 + H_2$	→	$HC(=O)OH + CH_4$	–24.9	–1.9	–2.4	–5.2	–0.7
$HC(=O)OCH_3 + H_2$	→	$H_2C{=}O + H_3C–OH$	9.3	–0.2	–10.2	–12.8	–7.0
$H_2C{=}CH_2 + CH_4$	→	$CH_3–CH_2–CH_3$	–22.6	+3.2	–7.7	–7.0	–2.3
$HC{\equiv}CH + CH_4$	→	$CH_2{=}CH–CH_3$	–35.0	+0.4	–0.9	–0.1	+3.7
$H_3C–NH_2 + CH_4$	→	$H_3C–CH_3 + NH_3$	–8.0	+1.2	–0.7	–2.6	–0.9
$H_2C{=}NH + CH_4$	→	$H_2C{=}CH_2 + NH_3$	–1.6	+1.4	+5.6	+2.4	+4.3
$HCN + CH_4$	→	$HC{\equiv}CH + NH_3$	31.4	+0.8	+10.3	+6.1	+7.0
$H_3C–OH + CH_4$	→	$H_3C–CH_3 + H_2O$	–11.5	+2.9	–1.4	–5.3	–0.1
$H_2C{=}O + CH_4$	→	$H_2C{=}CH_2 + H_2O$	–0.8	+3.7	+6.0	–0.9	+8.0
$HC(=O)OH + CH_4$	→	$H_2C{=}O + H_2O$	13.2	+4.0	–5.8	–9.8	–2.8
$H_3C–NH_2 + H_2O$	→	$H_3C–OH + NH_3$	3.5	–1.7	+0.7	+2.7	–0.8
$H_2C{=}NH + H_2O$	→	$H_2C{=}O + NH_3$	–0.8	–2.3	–0.4	+3.3	–3.7
$HCN + 2\ H_2O$	→	$HC(=O)OH + NH_3$	–20.6	–7.3	+12.7	+19.8	+8.2
MUE				3.0	8.2	8.3	5.7
MAX				12.9	79.4	49.5	22.9

[a] Energies are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] Basis set cc-pVTZ.

Table 6.10: Deviation from Experiment for Selected Isomerization Enthalpies of Hydrocarbons

Isomerization			Exp[a]	B3LYP[a]	MIO2	MIO3	d22
Propyne	→	Allene	1.2	–4.0	+4.9	+4.7	+3.2
Propyne	→	Cyclopropene	21.8	–0.3	+16.9	+16.9	+9.3
Propene	→	Cyclopropane	7.9	–0.1	+4.3	+4.3	–3.0
trans-2-Butene	→	*cis*-2-Butene	1.1	+0.3	–0.2	–0.1	–0.2
2-Methylpropene	→	*trans*-2-Butene	1.3	–1.1	–2.5	–2.1	–2.8
trans-2-Butene	→	1-Butene	2.8	+0.9	+1.3	+1.2	+0.9
1,3-Butadiene	→	Cyclobutene	11.3	+0.6	+1.5	+1.7	–3.6
Cyclopentene	→	Vinylcyclopropane	22.2	+0.0	+12.8	+12.1	+1.1
1-*trans*-3-Pentadiene	→	1,4-Pentadiene	7.1	+2.7	+1.3	+1.1	+0.7
2,2-Dimethylpropane	→	*n*-Pentane	5.1	–4.3	–3.2	–2.7	–4.9
2,2,3,3-Tetramethylbutane	→	*n*-Octane	4.1	–11.6	–5.2	–4.1	–9.0
Toluene	→	Norbornadiene	46.7	+9.4	+3.0	+2.8	+4.6
Styrene	→	Cyclooctatetraene	35.8	+4.8	+8.6	+8.9	+6.8
MUE				3.1	5.0	4.8	3.9

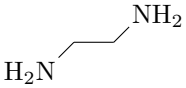
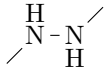
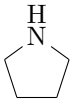
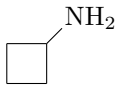
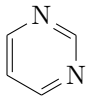
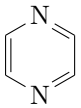
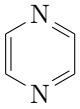
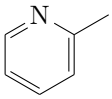
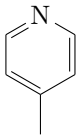
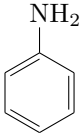
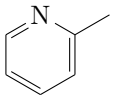
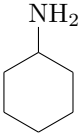
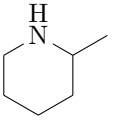
[a] Experimental and B3LYP/6–31G(d) values taken from ref [11]. All values are given in kcal/mol.

A good test for the treatment of variations in bonding, steric effects, conjugation, and ring strain are isomerization enthalpies of hydrocarbons, nitrogen and oxygen containing molecules as compiled by Sattelmeyer et al. [11] and shown in Tables 6.10, 6.11, and 6.12. DFTB and B3LYP results are compared to experimental enthalpies at 25 °C. It was shown that zero point and thermal corrections affect the results insignificantly [11], therefore these corrections were not made. The new parametrization lowers the mean unsigned error by about 1 kcal/mol for hydrocarbons and oxygen containing systems in comparison to MIO3, but does not quite reach the performance of B3LYP/6-31G(d). For nitrogen containing systems, the MUE is slightly larger. This again stems mainly from the wrong energetic description of the N–N single bond—a refined N–N potential needs to be developed in order to remediate this problem.

6.3 Conclusion

In this chapter a procedure has been presented to find electronic and repulsive parameters for the elements carbon, hydrogen, nitrogen, and oxygen designed to use in combination with DFTB3. In comparison to MIO, the new parametrization improves the description of hydrogen bonded complexes. For example, the too short hydrogen bond length within the water dimer is increased while at the same time the binding energy is correctly reproduced. Moreover, the overbinding for all bond types produced with the MIO parameter set is eliminated. This substantially reduces errors for atomization energies. One exception is found for the C≡N triple bond, the atomization energies of molecules containing such a bond are systematically overestimated. Further improvements are found for reaction energies and vibrational frequencies. However, the stretching frequencies for C=C and C=N double bonds still show errors of more than 150 cm^{-1}. This error could not be further reduced without spoiling geometric and energetic properties revealing limitations inherent in the DFTB3 formalism. The excellent description of the MIO parametrization with respect to geometries is

Table 6.11: Deviation from Experiment for Selected Isomerization Enthalpies of Nitrogen Containing Molecules

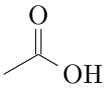
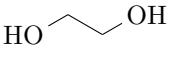
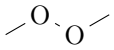
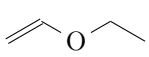
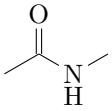
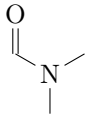
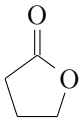
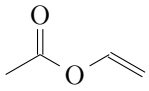
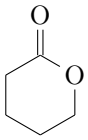
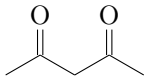
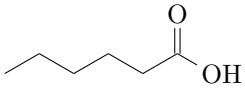
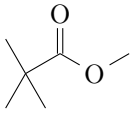
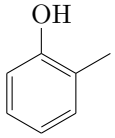
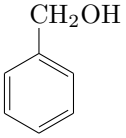
Isomerization			Exp[a]	B3LYP[a]	MIO2	MIO3	d22
CH_3CN	→	CH_3NC	21.4	+2.7	–10.9	–10.8	–11.9
$CH_3CH_2NH_2$	→	$HN(CH_3)_2$	6.9	–2.4	+2.1	+3.6	+1.8
	→		11.7	+2.7	+16.3	+15.1	+5.9
	→		26.1	–5.8	–0.6	+2.4	+15.1
	→		10.7	+3.2	+2.1	+0.2	–2.1
	→		0.0	+4.0	+3.9	+4.2	+4.3
	→		19.7	–1.1	–8.1	–8.0	+8.7
	→		1.1	–0.2	+0.2	+0.3	–0.1
	→		2.9	–5.4	+0.6	+1.8	+0.3
	→		4.9	–3.6	+2.1	+3.5	+2.0
MUE				3.1	4.7	5.0	5.2

[a] Experimental and B3LYP/6-31G(d) values taken from ref [11]. All values are given in kcal/mol.

maintained and even slightly improved. Despite this good performance also some drawbacks have been detected for specific cases. Those are energetic and geometric properties of some small molecules such as carbon monoxide and carbon dioxide.

For the tests carried out so far, the new parametrization leads to a higher flexibility of DFTB3 to treat different chemical environments more efficiently. In addition, the presented scheme may lead to an easy access for a future parametrization of further elements.

Table 6.12: Deviation from Experiment for Selected Isomerization Enthalpies of Oxygen Containing Molecules

Isomerization			Exp[a]	B3LYP[a]	MIO2	MIO3	d22
CH_3CH_2OH	→	$(CH_3)_2O$	12.3	–6.8	+0.0	+3.2	–0.6
	→		27.2	+0.4	+9.2	+9.3	+3.5
	→		18.3	–6.5	+1.3	+4.2	–0.1
	→		62.6	–12.5	–0.3	+6.2	+5.3
	→		32.7	–0.9	–4.2	–3.5	–4.8
	→		10.4	+1.9	+9.7	+9.3	+4.1
	→		13.5	–6.4	–1.5	–0.4	–2.5
	→		12.3	+0.3	+9.5	+9.5	+5.6
	→		0.3	+5.3	+5.3[b]	+5.4	+4.4
	→		4.4	+0.4	+6.1	+8.4	+6.7
	→		8.1	–0.0	+3.4	+2.8	+3.1
MUE				3.4	4.6	5.7	3.7

[a] Experimental and B3LYP/6-31G(d) values taken from ref [11]. [b] This number was reported to be +10.9 kcal/mol in ref [11] but could not be reproduced.

Chapter 7

Summary and Outlook

In this work important extensions and an efficient parametrization of the approximate quantum chemical calculation method DFTB is presented. The developments are based on the self-consistent charge density functional tight binding method (DFTB2) that has proven to yield rapid, robust, and accurate results for a broad field of applications. However, DFTB2 is not flexible enough for several chemical situations that are of utmost importance for biological applications, i.e. hydrogen binding energies, proton affinities, and proton transfer reactions for differently charged systems.

Therefore, an extension is derived in the spirit of DFTB2 (chapter 4). It is deduced from density functional theory by an expansion of the total energy up to third order terms around a given reference density. The consequences of the assumptions made in DFTB2 are analyzed and additional improvements direct to a higher transferability of this method for a larger variety of chemical environments. The new method is called DFTB3 which—besides improving the deficiencies mentioned above—also allows to properly treat phosphate chemistry. It is noteworthy that all these extensions do not significantly increase computation time.

Important for the accuracy of DFTB are the repulsive potentials. Until now, tedious manual work was necessary to fit them to reference calculations. In chapter 5, a new procedure is presented that partially automatizes the search for these parameters. It is applied to hydrocarbons, and it is found that the existent parameters are of remarkable quality but can still be slightly improved for geometries, atomization and reaction energies. In addition, conflicts of objectives were detected leading to the suggestion to use different parameter sets if one specific property is of ultimate importance for a certain application. In that sense a parametrization is provided specifically designed for vibrational frequencies which comes close to the quality of high-level DFT calculations.

This scheme is adapted in chapter 6, further refined, and applied to a parametrization of the elements carbon, hydrogen, nitrogen, and oxygen. The parameters are constructed for use in combination with DFTB3. Extensive tests for a large set of organic molecules and several properties show that deficiencies of the old parametrization are alleviated, in specific, atomization and reaction energies are improved as well as vibrational stretching frequencies.

The success of DFTB for applications in biology induced a strong demand for parameters of further elements. With the new scheme provided in this work they will now be much easier accessible. A parametrization of the halogens is almost completed. Moreover, the experience made on phosphorus (chapter 4) will help to release improved parameters for that element soon.

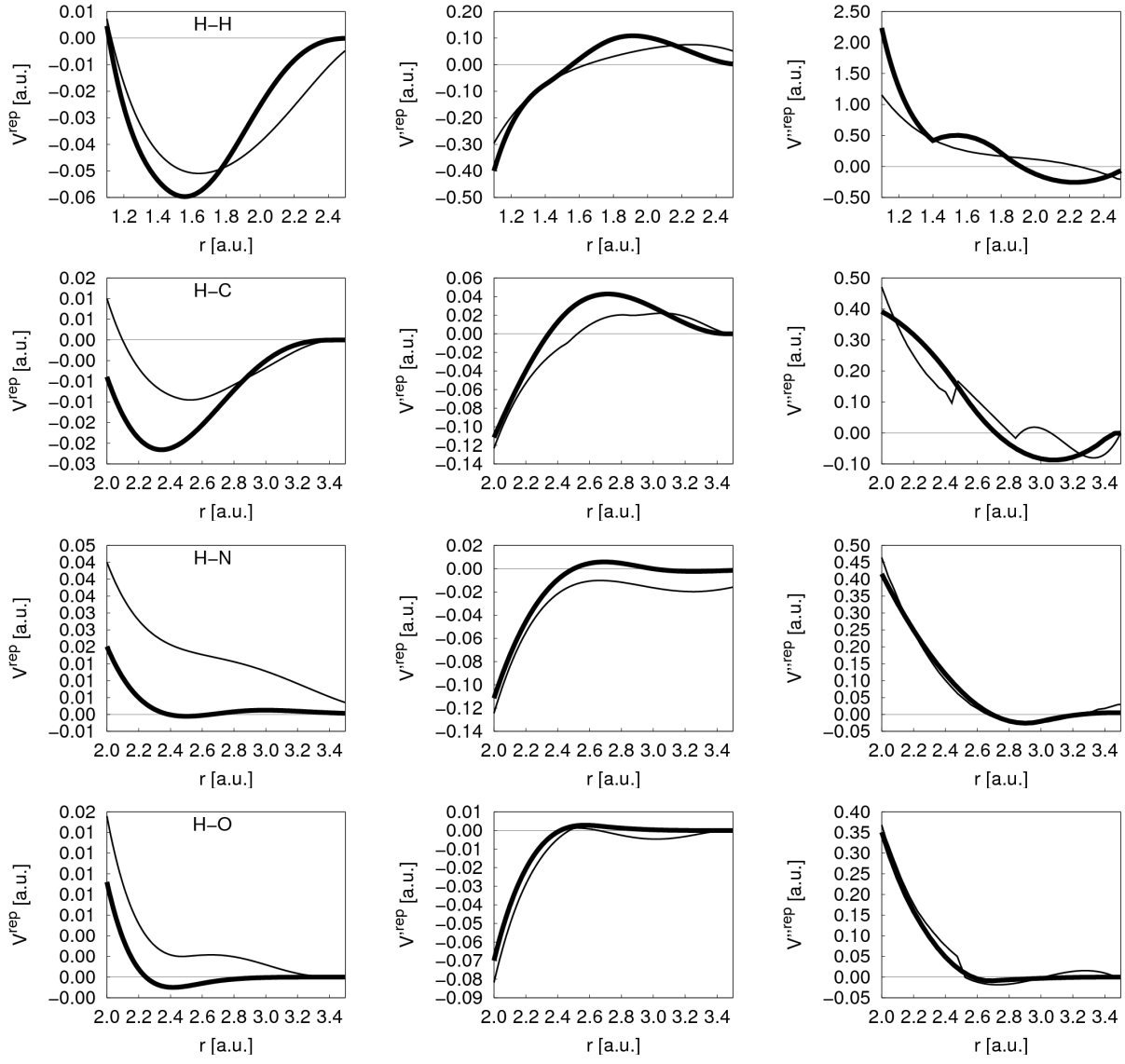
On the other hand, future methodological improvements will further augment efficiency and transferability of the DFTB3 method. A self-interaction correction, possibly crucial for describing charge tranport, has already been implemented (see appendix D), first tests are currently carried out. Further ideas concern the consideration of three-center terms in an approximate fashion. Another path that is currently followed is the correction of molecular polarizabilities for acquiring insights into internal structure and behavior of large molecules and also for reliable computation of Raman spectra.

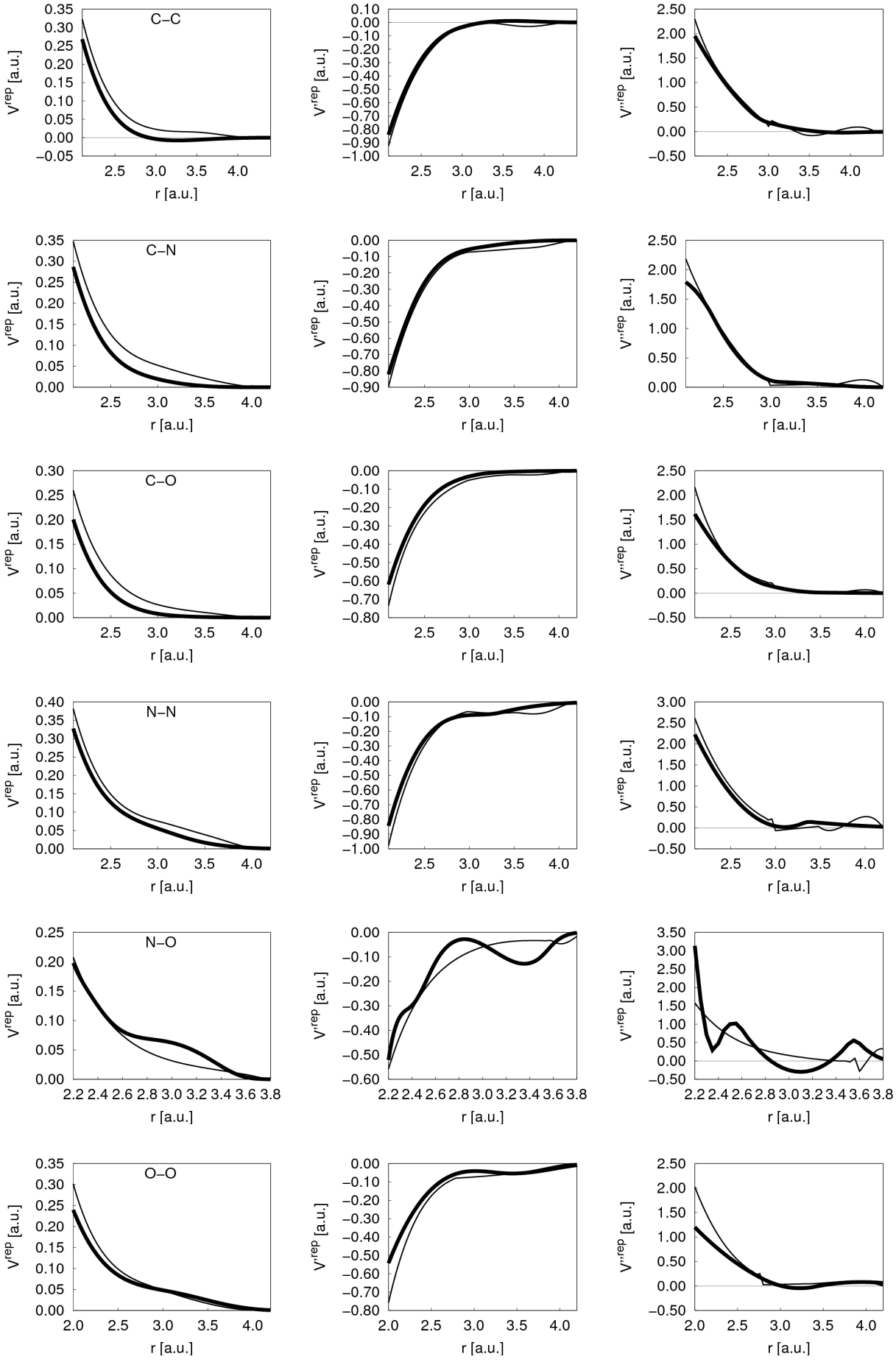
Prospectively, the methodological progress presented in this work will enhance the ability to investigate structural and dynamic properties of large biological molecules and will allow to move on to theoretical experiments that have not yet been accessible on a quantum mechanical basis.

Appendix A

Illustration of Repulsive Potentials

For each element pair the repulsive potential V^{rep}, its first (V'^{rep}), and its second derivative (V''^{rep}) is shown. The thick line represents the new parameter set d22 produced in this work (chapter 6), the thin line the MIO parameters. For H–H, the H–H-mod potential has been shown instead of the original MIO H–H potential. For details see section 3.3.2.




C–C
C–N
C–O
N–N
N–O
O–O
V^rep [a.u.]
V'^rep [a.u.]
V''^rep [a.u.]
r [a.u.]

Appendix B

Derivation of DFTB3 Expressions

B.1 Functional Form of γ and its Derivative $\frac{\partial \gamma_{ab}}{\partial U_a}$

The functional form of γ as introduced in eq 3.26—using the abbreviations $\alpha = \frac{16}{5}U_a$, $\beta = \frac{16}{5}U_b$, $r = |\boldsymbol{R}_b - \boldsymbol{R}_a|$— is given by (for the γ^{h}-function see below)

$$\gamma_{ab} = \begin{cases} \frac{1}{r} - S^{\mathrm{f}} & r \neq 0, \alpha \neq \beta \\ \frac{1}{r} - S^{\mathrm{g}} & r \neq 0, \alpha = \beta \\ \frac{5}{16}\alpha & r = 0 \end{cases} \tag{B.1}$$

$$S^{\mathrm{f}} = \mathrm{e}^{-\alpha r} f(\alpha, \beta, r) + \mathrm{e}^{-\beta r} f(\beta, \alpha, r) \tag{B.2}$$

$$S^{\mathrm{g}} = \mathrm{e}^{-\alpha r} g(\alpha, r) \tag{B.3}$$

$$f(\alpha, \beta, r) = \frac{\alpha\beta^4}{2(\alpha^2 - \beta^2)^2} - \frac{\beta^6 - 3\alpha^2\beta^4}{(\alpha^2 - \beta^2)^3 r} \tag{B.4}$$

$$g(\alpha, r) = \frac{1}{48r}\left(48 + 33\alpha r + 9\alpha^2 r^2 + \alpha^3 r^3\right). \tag{B.5}$$

Its derivative $\frac{\partial \gamma_{ab}}{\partial U_a} = \frac{\partial \gamma_{ba}}{\partial U_a}$ as introduced in eq 4.4 can then be written as (for the derivative of the γ^{h}-function see below)

$$\frac{\partial \gamma_{ab}}{\partial U_a} = \frac{16}{5}\frac{\partial \gamma_{ab}}{\partial \alpha} = \frac{16}{5}\begin{cases} -\frac{\partial S^{\mathrm{f}}}{\partial \alpha} & r \neq 0, \alpha \neq \beta \\ -\frac{\partial S^{\mathrm{g}}}{\partial \alpha} & r \neq 0, \alpha = \beta \\ \frac{5}{16} & r = 0 \end{cases} \tag{B.6}$$

$$\frac{\partial S^{\mathrm{f}}}{\partial \alpha} = \mathrm{e}^{-\alpha r}\frac{\partial}{\partial \alpha} f(\alpha, \beta, r) - r\mathrm{e}^{-\alpha r} f(\alpha, \beta, r) + \mathrm{e}^{-\beta r}\frac{\partial}{\partial \alpha} f(\beta, \alpha, r) \tag{B.7}$$

$$\frac{\partial S^{\mathrm{g}}}{\partial \alpha} = \mathrm{e}^{-\alpha r}\frac{\partial}{\partial \alpha} g(\alpha, r) - r\mathrm{e}^{-\alpha r} g(\alpha, r) \tag{B.8}$$

$$\tag{B.9}$$

$$\frac{\partial}{\partial\alpha}f(\alpha,\beta,r) = -\frac{\beta^6+3\alpha^2\beta^4}{2(\alpha^2-\beta^2)^3} - \frac{12\alpha^3\beta^4}{(\alpha^2-\beta^2)^4 r} \tag{B.10}$$

$$\frac{\partial}{\partial\alpha}f(\beta,\alpha,r) = \frac{2\beta^3\alpha^3}{(\beta^2-\alpha^2)^3} + \frac{12\beta^4\alpha^3}{(\beta^2-\alpha^2)^4 r} \tag{B.11}$$

$$\frac{\partial}{\partial\alpha}g(\alpha,r) = \frac{1}{48}\left(33+18\alpha r+3\alpha^2 r^2\right). \tag{B.12}$$

Note the symmetry of γ: $\gamma_{ab}=\gamma_{ba}$ and $\frac{\partial\gamma_{ab}}{\partial U_a}=\frac{\partial\gamma_{ba}}{\partial U_a}$.

B.2 Derivation of the Kohn-Sham Equations

Combining eqs 4.6 and 4.8 and taking advantage of the symmetry of $H_{\mu\nu}=H_{\nu\mu}$ and $S_{\mu\nu}=S_{\nu\mu}$ gives:

$$2\sum_b\sum_{\nu\in b} n_i c_{\nu i} H^0_{\delta\nu} + \frac{\partial E^\gamma}{\partial c_{\delta i}} + \frac{\partial E^\Gamma}{\partial c_{\delta i}} - 2n_i\epsilon_i\sum_b\sum_{\nu\in b} c_{\nu i}S_{\delta\nu} = 0, \qquad \delta\in d \tag{B.13}$$

$$\frac{\partial E^\gamma}{\partial c_{\delta i}} = \frac{\partial}{\partial c_{\delta i}}\frac{1}{2}\sum_{ab}\Delta q_a\Delta q_b\gamma_{ab} = \frac{1}{2}\sum_{ab}\frac{\partial q_a}{\partial c_{\delta i}}\Delta q_b\left(\gamma_{ab}+\gamma_{ba}\right) \tag{B.14}$$

$$\frac{\partial E^\Gamma}{\partial c_{\delta i}} = \frac{\partial}{\partial c_{\delta i}}\frac{1}{3}\sum_{ab}\Delta q_a^2\Delta q_b\Gamma_{ab} = \frac{1}{3}\sum_{ab}\frac{\partial q_a}{\partial c_{\delta i}}\Delta q_b\left(2\Delta q_a\Gamma_{ab}+\Delta q_b\Gamma_{ba}\right). \tag{B.15}$$

The derivation of eq 4.9 with respect to $c_{\delta i}$, where $\delta\in d$ can be written as

$$\frac{\partial q_a}{\partial c_{\delta i}} = \delta_{ad}\sum_c\sum_{\nu\in c} n_i c_{\nu i}S_{\delta\nu} + \sum_{\mu\in a} n_i c_{\mu i}S_{\delta\mu}. \tag{B.16}$$

Thus, eqs B.14 and B.15 expand to

$$\frac{\partial E^\gamma}{\partial c_{\delta i}} = \frac{1}{2}\sum_{bc}\Delta q_b(\gamma_{db}+\gamma_{bd}+\gamma_{cb}+\gamma_{bc})\sum_{\nu\in c} n_i c_{\nu i}S_{\delta\nu} \tag{B.17}$$

$$\frac{\partial E^\Gamma}{\partial c_{\delta i}} = \frac{1}{3}\sum_{bc}\Delta q_b\left(2\Delta q_d\Gamma_{db}+\Delta q_b\Gamma_{bd}+2\Delta q_c\Gamma_{cb}+\Delta q_b\Gamma_{bc}\right)\sum_{\nu\in c} n_i c_{\nu i}S_{\delta\nu}. \tag{B.18}$$

Dividing eq B.13 by $(2n_i)$ and combining with eqs B.17 and B.18 gives

$$\sum_b\sum_{\nu\in b} c_{\nu i}\left(H_{\mu\nu}-\epsilon_i S_{\mu\nu}\right) = 0, \qquad \forall a,\mu\in a,i \tag{B.19}$$

$$\begin{aligned} H_{\mu\nu} = {} & H^0_{\mu\nu} + S_{\mu\nu}\sum_c\Delta q_c\left(\frac{1}{4}(\gamma_{ac}+\gamma_{ca}+\gamma_{bc}+\gamma_{cb})+\right.\\ & \left.+\frac{1}{3}(\Delta q_a\Gamma_{ac}+\Delta q_b\Gamma_{bc})+\frac{\Delta q_c}{6}(\Gamma_{ca}+\Gamma_{cb})\right)\\ & \forall a,b,\mu\in a,\nu\in b. \end{aligned} \tag{B.20}$$

Taking advantage of the symmetry $\gamma_{ab}=\gamma_{ba}$ (note, that $\Gamma_{ab}\neq\Gamma_{ba}$ is not symmetric!), and renaming the indices results in eqs 4.10 and 4.11.

B.3 Total Energy

The total energy is given by eq 4.6. Another possibility is to express the energy in terms of ϵ_i using eqs B.19, B.20, the symmetry $S_{\mu\nu} = S_{\nu\mu}$, and the definition of q_a (eq 4.9):

$$\begin{aligned}
\sum_i n_i\epsilon_i &= \sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b} n_i c_{\mu i} c_{\nu i} H_{\mu\nu} \\
&= \sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b} n_i c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b} n_i c_{\mu i} c_{\nu i} \\
&\quad S_{\mu\nu}\sum_c \Delta q_c \left(\frac{1}{4}(\gamma_{ac}+\gamma_{ca}+\gamma_{bc}+\gamma_{cb}) + \frac{1}{3}(\Delta q_a\Gamma_{ac} + \Delta q_b\Gamma_{bc}) + \frac{\Delta q_c}{6}(\Gamma_{ca}+\Gamma_{cb})\right) \\
&= \sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b} n_i c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \sum_{ac} q_a\Delta q_c\left(\frac{1}{2}(\gamma_{ac}+\gamma_{ca}) + \frac{2}{3}\Delta q_a\Gamma_{ac} + \frac{1}{3}\Delta q_c\Gamma_{ca}\right).
\end{aligned} \tag{B.21}$$

Thus, inserting eq B.21 into eq 4.6 and using $\Delta q_a = q_a - q_a^0$, the total energy can be written as

$$\begin{aligned}
E^{\mathrm{DFTB3}} &= \sum_i n_i\epsilon_i - \sum_{ac} q_a\Delta q_c\left(\frac{1}{2}(\gamma_{ac}+\gamma_{ca}) + \frac{2}{3}\Delta q_a\Gamma_{ac} + \frac{1}{3}\Delta q_c\Gamma_{ca}\right) + \frac{1}{2}\sum_{ab}\Delta q_a\Delta q_b\gamma_{ab} \\
&\quad + \frac{1}{3}\sum_{ab}\Delta q_a^2\Delta q_b\Gamma_{ab} + E^{\mathrm{rep}} \\
&= \sum_i n_i\epsilon_i - \frac{1}{2}\sum_{ab}\Delta q_b(q_a\gamma_{ba} + q_a^0\gamma_{ab}) - \frac{1}{3}\sum_{ab}\Delta q_a\Delta q_b\Gamma_{ab}(q_a + q_a^0) \\
&\quad - \frac{1}{3}\sum_{ab} q_a\Delta q_b^2\Gamma_{ba} + E^{\mathrm{rep}}.
\end{aligned} \tag{B.22}$$

Considering the symmetry $\gamma_{ab} = \gamma_{ba}$ the total energy can be simplified to eq 4.12.

B.4 Atomic Forces

The atomic forces are derived by combining eqs 4.13 and 4.6 to

$$\begin{aligned}
F_{kx} &= -\frac{\partial}{\partial R_{kx}}\left[E^{\mathrm{H0}} + E^{\gamma} + E^{\Gamma} + E^{\mathrm{rep}} - \sum_i n_i\epsilon_i\left(\sum_{ab}\sum_{\mu\in a}\sum_{\nu\in b} c_{\mu i}c_{\nu i}S_{\mu\nu} - 1\right)\right] \\
&= F^{\mathrm{H0}}_{kx} + F^{\gamma}_{kx} + F^{\Gamma}_{kx} + F^{\mathrm{rep}}_{kx} + F^{\mathrm{norm}}_{kx} \qquad \forall k, x,
\end{aligned} \tag{B.23}$$

where as shown above only the explicit dependence of the energy with respect to the atomic coordinate has to be considered:

$$F^{\mathrm{H0}}_{kx} = -\frac{\partial}{\partial R_{kx}}\sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b} n_i c_{\mu i}c_{\nu i}H^0_{\mu\nu} = -\sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b} n_i c_{\mu i}c_{\nu i}\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}} \tag{B.24}$$

$$\begin{aligned}
F^{\gamma}_{kx} &= -\frac{\partial}{\partial R_{kx}}\frac{1}{2}\sum_{ab}\Delta q_a\Delta q_b\gamma_{ab} \\
&= -\frac{1}{2}\sum_{ab}\gamma_{ab}\left(\frac{\partial\Delta q_a}{\partial R_{kx}}\Delta q_b + \Delta q_a\frac{\partial\Delta q_b}{\partial R_{kx}}\right) - \frac{1}{2}\sum_{ab}\Delta q_a\Delta q_b\frac{\partial\gamma_{ab}}{\partial R_{kx}} \\
&= F^{\gamma 1} + F^{\gamma 2}
\end{aligned} \tag{B.25}$$

$$
\begin{aligned}
F_{kx}^{\Gamma} &= -\frac{\partial}{\partial R_{kx}}\frac{1}{3}\sum_{ab}\Delta q_a^2\Delta q_b\Gamma_{ab} \\
&= -\frac{1}{3}\sum_{ab}\Delta q_a\Gamma_{ab}\left(2\frac{\partial\Delta q_a}{\partial R_{kx}}\Delta q_b+\Delta q_a\frac{\partial\Delta q_b}{\partial R_{kx}}\right)-\frac{1}{3}\sum_{ab}\Delta q_a^2\Delta q_b\frac{\partial\Gamma_{ab}}{\partial R_{kx}} \\
&= -\frac{1}{3}\sum_{ab}\frac{\partial q_a}{\partial R_{kx}}\Delta q_b\left(2\Delta q_a\Gamma_{ab}+\Delta q_b\Gamma_{ba}\right)-\frac{1}{3}\sum_{ab}\Delta q_a^2\Delta q_b\frac{\partial\Gamma_{ab}}{\partial R_{kx}} \\
&= F^{\Gamma 1} + F^{\Gamma 2} \qquad \text{(B.26)}
\end{aligned}
$$

$$
F_{kx}^{\mathrm{rep}} = -\frac{\partial E^{\mathrm{rep}}}{\partial R_{kx}} = -\sum_{ab}\frac{\partial V_{ab}^{\mathrm{rep}}}{\partial R_{kx}} \qquad \text{(B.27)}
$$

$$
F_{kx}^{\mathrm{norm}} = -\frac{\partial}{\partial R_{kx}}\sum_i\epsilon_i\left(\sum_{ab}\sum_{\mu\in a}\sum_{\nu\in b}c_{\mu i}c_{\nu i}S_{\mu\nu}-1\right) = \sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b}\epsilon_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}}. \qquad \text{(B.28)}
$$

Because $H^0_{\mu\nu} = H^0_{\nu\mu}$, $\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}} = \frac{\partial H^0_{\nu\mu}}{\partial R_{kx}}$, and $H^0_{\mu\nu}$ with $\mu \in a$ and $\nu \in b$ is only dependent on the coordinates R_a and R_b [17], and for $a = b$ it is not dependent on any coordinate [17] F_{kx}^{H0} can be simplified to

$$
\begin{aligned}
F_{kx}^{\mathrm{H0}} &= -\sum_{ab}\sum_{\mu\in a}\sum_{\nu\in b}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}} \\
&= -\sum_{b\neq k}\sum_{\mu\in k}\sum_{\nu\in b}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}}-\sum_{a\neq k}\sum_{\mu\in a}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}} \\
&\quad -\sum_{\mu\in k}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}}-\sum_{a\neq k}\sum_{b\neq k}\sum_{\mu\in a}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}} \\
&= -2\sum_{a\neq k}\sum_{\mu\in a}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}}. \qquad \text{(B.29)}
\end{aligned}
$$

The same applies for F_{kx}^{norm} which can be simplified to

$$
F_{kx}^{\mathrm{norm}} = \sum_{iab}\sum_{\mu\in a}\sum_{\nu\in b}n_i\epsilon_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}} = 2\sum_{a\neq k}\sum_{\mu\in a}\sum_{\nu\in k}n_i\epsilon_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}}. \qquad \text{(B.30)}
$$

The derivative of the atomic charge with respect to an atomic coordinate is given by

$$
\frac{\partial q_{a\neq k}}{\partial R_{kx}} = \sum_{\mu\in a}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}} \qquad \text{(B.31)}
$$

$$
\begin{aligned}
\frac{\partial q_k}{\partial R_{kx}} &= \sum_{\mu\in k}\sum_{b\neq k}\sum_{\nu\in b}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}}+\sum_{\mu\in k}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}} \\
&= \sum_{\mu\in k}\sum_{b\neq k}\sum_{\nu\in b}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}}, \qquad \text{(B.32)}
\end{aligned}
$$

where the last recasting in latter equation is due to the fact that the overlap of two basis functions centered at one atom does not change with a change in the coordinate of that atom. Further the basis functions centered on one atom are orthogonal to each other.

Taking advantage of the symmetry $S_{\mu\nu} = S_{\nu\mu}$ and $\frac{\partial S_{\mu\nu}}{\partial R_{kx}} = \frac{\partial S_{\nu\mu}}{\partial R_{kx}}$ one can rewrite $F^{\gamma 1}$ and $F^{\Gamma 1}$ as (see eqs B.25 and B.26)

$$\begin{aligned}
F_{kx}^{\gamma 1} &= -\frac{1}{2}\sum_{a\neq k}\frac{\partial q_a}{\partial R_{kx}}\sum_b \Delta q_b(\gamma_{ab}+\gamma_{ba}) - \frac{1}{2}\frac{\partial q_k}{\partial R_{kx}}\sum_b \Delta q_b(\gamma_{kb}+\gamma_{bk}) \\
&= -\frac{1}{2}\sum_{a\neq k}\sum_b \Delta q_b(\gamma_{ab}+\gamma_{ba})\sum_{\mu\in a}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}} \\
&\quad -\frac{1}{2}\sum_b \Delta q_b(\gamma_{kb}+\gamma_{bk})\sum_{\mu\in k}\sum_{c\neq k}\sum_{\nu\in c}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}} \\
&= -\frac{1}{2}\sum_{a\neq k}\sum_b \Delta q_b(\gamma_{ab}+\gamma_{ba}+\gamma_{kb}+\gamma_{bk})\sum_{\mu\in a}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}} \qquad \text{(B.33)} \\
F_{kx}^{\Gamma 1} &= -\frac{1}{3}\sum_{a\neq k}\frac{\partial q_a}{\partial R_{kx}}\sum_b \Delta q_b\left(2\Delta q_a\Gamma_{ab}+\Delta q_b\Gamma_{ba}\right) - \frac{1}{3}\frac{\partial q_k}{\partial R_{kx}}\sum_b \Delta q_b\left(2\Delta q_k\Gamma_{kb}+\Delta q_b\Gamma_{bk}\right) \\
&= -\frac{1}{3}\sum_{a\neq k}\sum_{\mu\in a}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}}\sum_b \Delta q_b\left(2\Delta q_a\Gamma_{ab}+\Delta q_b\Gamma_{ba}\right) \\
&\quad -\frac{1}{3}\sum_{\mu\in k}\sum_{b\neq k}\sum_{\nu\in b}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}}\sum_c \Delta q_c\left(2\Delta q_k\Gamma_{kc}+\Delta q_c\Gamma_{ck}\right) \\
&= -\frac{1}{3}\sum_{a\neq k}\sum_b \Delta q_b\left(2\Delta q_a\Gamma_{ab}+\Delta q_b\Gamma_{ba}+2\Delta q_k\Gamma_{kb}+\Delta q_b\Gamma_{bk}\right)\sum_{\mu\in a}\sum_{\nu\in k}\sum_i n_i c_{\mu i}c_{\nu i}\frac{\partial S_{\mu\nu}}{\partial R_{kx}}. \qquad \text{(B.34)}
\end{aligned}$$

As mentioned above the derivatives $\frac{\partial S_{\mu\nu}}{\partial R_{kx}}$ are evaluated numerically using the tabulated values for $S_{\mu_a\nu_b}(r_{ab})$ with $r_{ab} = |R_b - R_a|$. Taking advantage of the derivatives $\frac{\partial \gamma_{kk}}{R_{kx}} = 0$ and $\frac{\partial \Gamma_{kk}}{R_{kx}} = 0$ one can also expand $F_{kx}^{\gamma 2}$ (eq B.25) and $F_{kx}^{\Gamma 2}$ (eq B.26) to

$$\begin{aligned}
F_{kx}^{\gamma 2} &= -\frac{1}{2}\sum_a \Delta q_a\Delta q_k\frac{\partial \gamma_{ak}}{\partial R_{kx}} - \frac{1}{2}\sum_b \Delta q_k\Delta q_b\frac{\partial \gamma_{kb}}{\partial R_{kx}} \\
&= -\frac{1}{2}\Delta q_k\sum_{a\neq k}\Delta q_a\left(\frac{\partial \gamma_{ak}}{\partial R_{kx}}+\frac{\partial \gamma_{ka}}{\partial R_{kx}}\right) \qquad \text{(B.35)} \\
F_{kx}^{\Gamma 2} &= -\frac{1}{3}\left(\sum_{a\neq k}\Delta q_a^2\Delta q_k\frac{\partial \Gamma_{ak}}{\partial R_{kx}} + \sum_{b\neq k}\Delta q_k^2\Delta q_b\frac{\partial \Gamma_{kb}}{\partial R_{kx}}\right) \\
&= -\frac{1}{3}\Delta q_k\sum_{a\neq k}\Delta q_a\left(\Delta q_a\frac{\partial \Gamma_{ak}}{\partial R_{kx}}+\Delta q_k\frac{\partial \Gamma_{ka}}{\partial R_{kx}}\right). \qquad \text{(B.36)}
\end{aligned}$$

Thus, the final expression for the force can be shortly written by inserting eqs B.25, B.26, B.29, B.30, and B.33–B.36 into eq B.23, using $\gamma_{ab} = \gamma_{ba}$ and reordering terms as (eq 4.16)

$$\begin{aligned} F_{kx} \;=\; & -\sum_{a\neq k}\sum_{\mu\in a}\sum_{\nu\in k}\sum_{i} n_i c_{\mu i} c_{\nu i}\Bigg(2\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}} - 2\epsilon_i\frac{\partial S_{\mu\nu}}{\partial R_{kx}} + \frac{\partial S_{\mu\nu}}{\partial R_{kx}}\cdot \\ & \left(\sum_c \Delta q_c\left(\gamma_{ac}+\gamma_{kc}+\frac{1}{3}(2\Delta q_a\Gamma_{ac}+\Delta q_c\Gamma_{ca}+2\Delta q_k\Gamma_{kc}+\Delta q_c\Gamma_{ck})\right)\right)\Bigg) \\ & -\Delta q_k\sum_{a\neq k}\Delta q_a\frac{\partial\gamma_{ak}}{\partial R_{kx}} - \frac{1}{3}\Delta q_k\sum_{a\neq k}\Delta q_a\left(\Delta q_a\frac{\partial\Gamma_{ak}}{\partial R_{kx}}+\Delta q_k\frac{\partial\Gamma_{ka}}{\partial R_{kx}}\right) - \frac{\partial E^{\mathrm{rep}}}{\partial R_{kx}} \qquad \forall k,x. \end{aligned} \tag{B.37}$$

The derivatives $\frac{\partial H^0_{\mu\nu}}{\partial R_{kx}}$ and $\frac{\partial S_{\mu\nu}}{\partial R_{kx}}$ are determined by taking the numerical derivative of the tabulated integrals $H^0_{\mu\nu}$ and $S_{\mu\nu}$. The force contribution $\frac{\partial E^{\mathrm{rep}}}{\partial R_{kx}}$ is also calculated analytically. Details about the representation of E^{rep} can be found in ref [121].

In the remainder of this subsection the analytical expressions for $\frac{\partial\gamma_{kb}}{\partial R_{kx}}$ and $\frac{\partial\Gamma_{kb}}{\partial R_{kx}}$ are derived. With eqs B.1, 4.4, B.6, and

$$\begin{aligned} r_{kb} &= |\boldsymbol{R}_b - \boldsymbol{R}_k| & r_{ak} &= |\boldsymbol{R}_k - \boldsymbol{R}_a| \\ r_{kb} &= \sqrt{\textstyle\sum_{x=1}^{3}(R_{bx}-R_{kx})^2} & r_{ak} &= \sqrt{\textstyle\sum_{x=1}^{3}(R_{kx}-R_{ax})^2} \\ \frac{\partial r_{kb}}{\partial R_{kx}} &= \frac{-(R_{bx}-R_{kx})}{r_{kb}} & \frac{\partial r_{ak}}{\partial R_{kx}} &= \frac{(R_{kx}-R_{ax})}{r_{ak}} \end{aligned} \tag{B.38}$$

$\forall a,b,k,x$, one can write

$$\frac{\partial\gamma_{kb}}{\partial R_{kx}} = \frac{\partial\gamma_{kb}}{\partial r_{kb}}\frac{\partial r_{kb}}{\partial R_{kx}} \qquad \frac{\partial\gamma_{ak}}{\partial R_{kx}} = \frac{\partial\gamma_{ak}}{\partial r_{ak}}\frac{\partial r_{ak}}{\partial R_{kx}} \tag{B.39}$$

$$\frac{\partial\Gamma_{kb}}{\partial R_{kx}} = \frac{\partial\Gamma_{kb}}{\partial r_{kb}}\frac{\partial r_{kb}}{\partial R_{kx}} \qquad \frac{\partial\Gamma_{ak}}{\partial R_{kx}} = \frac{\partial\Gamma_{ak}}{\partial r_{ak}}\frac{\partial r_{ak}}{\partial R_{kx}}, \tag{B.40}$$

and for easier writing $\alpha = \frac{16}{5}U_a$, $\beta = \frac{16}{5}U_b$, $r = r_{ab}$ is used to yield (for the derivative of the γ^{h}-function see below)

$$\frac{\partial\gamma_{ab}}{\partial r_{ab}} = \frac{\partial\gamma_{ab}}{\partial r} = \begin{cases} -\frac{1}{r^2} - \frac{\partial S^{\mathrm{f}}}{\partial r} & r\neq 0, \alpha\neq\beta \\ -\frac{1}{r^2} - \frac{\partial S^{\mathrm{g}}}{\partial r} & r\neq 0, \alpha=\beta \\ 0 & r=0 \end{cases} \tag{B.41}$$

$$\frac{\partial S^{\mathrm{f}}}{\partial r} = \mathrm{e}^{-\alpha r}\frac{\partial f(\alpha,\beta,r)}{\partial r} - \alpha\mathrm{e}^{-\alpha r}f(\alpha,\beta,r) + \mathrm{e}^{-\beta r}\frac{\partial f(\beta,\alpha,r)}{\partial r} - \beta\mathrm{e}^{-\beta r}f(\beta,\alpha,r) \tag{B.42}$$

$$\frac{\partial S^{\mathrm{g}}}{\partial r} = \mathrm{e}^{-\alpha r}\frac{\partial g(\alpha,r)}{\partial r} - \alpha\mathrm{e}^{-\alpha r}g(\alpha,r) \tag{B.43}$$

$$\frac{\partial f(\alpha,\beta,r)}{\partial r} = \frac{\beta^6 - 3\alpha^2\beta^4}{(\alpha^2-\beta^2)^3 r^2} \tag{B.44}$$

$$\frac{\partial g(\alpha,r)}{\partial r} = -\frac{1}{r^2} + \frac{3\alpha^2}{16} + \frac{\alpha^3}{24}\cdot r. \tag{B.45}$$

Functions $f(\alpha,\beta,r)$ and $g(\alpha,r)$ are defined in eqs B.4 and B.5, respectively. Similarly,

$$\begin{aligned}\frac{\partial\Gamma_{ab}}{\partial r_{ab}} &= \frac{\partial\Gamma_{ab}}{\partial r} = \frac{\partial}{\partial r}\left(\frac{\partial\gamma_{ab}}{\partial U_a}\frac{\partial U_a}{\partial q_a}\right) = \frac{\partial^2\gamma_{ab}}{\partial U_a\partial r}\frac{\partial U_a}{\partial q_a}+\frac{\partial\gamma_{ab}}{\partial U_a}\frac{\partial^2 U_a}{\partial q_a\partial r} = \\ &= \frac{\partial^2\gamma_{ab}}{\partial U_a\partial r}\frac{\partial U_a}{\partial q_a}. \end{aligned} \tag{B.46}$$

The latter term on right hand side of the first line is equal to zero, because the Hubbard derivative is not dependent on any atomic coordinate. Thus, eqs B.6, B.41 are needed and (for the derivative of the γ^{h}-function see below):

$$\frac{\partial^2\gamma_{ab}}{\partial U_a\partial r} = \begin{cases} -\dfrac{\partial^2 S^{\mathrm{f}}}{\partial U_a\partial r} & r\neq 0, \alpha\neq\beta \\ -\dfrac{\partial^2 S^{\mathrm{g}}}{\partial U_a\partial r} & r\neq 0, \alpha=\beta \\ 0 & r=0 \end{cases} \tag{B.47}$$

$$\begin{aligned}\frac{\partial^2 S^{\mathrm{f}}}{\partial U_a\partial r} &= \frac{\partial}{\partial r}\left(\frac{16}{5}\frac{\partial S^{\mathrm{f}}}{\partial\alpha}\right) = \\ &= \frac{16}{5}\left[\mathrm{e}^{-\alpha r}\left(f(\alpha,\beta,r)(\alpha r-1)-\alpha\frac{\partial f(\alpha,\beta,r)}{\partial\alpha}+\frac{\partial^2 f(\alpha,\beta,r)}{\partial\alpha\partial r}-r\frac{\partial f(\alpha,\beta,r)}{\partial r}\right)\right. \\ &\quad \left.+\mathrm{e}^{-\beta r}\left(\frac{\partial^2 f(\beta,\alpha,r)}{\partial\alpha\partial r}-\beta\frac{\partial f(\beta,\alpha,r)}{\partial\alpha}\right)\right] \end{aligned} \tag{B.48}$$

$$\begin{aligned}\frac{\partial^2 S^{\mathrm{g}}}{\partial U_a\partial r} &= \frac{\partial}{\partial r}\left(\frac{16}{5}\frac{\partial S^{\mathrm{g}}}{\partial\alpha}\right) = \\ &= \frac{16}{5}\mathrm{e}^{-\alpha r}\left((\alpha r-1)g(\alpha,r)-\alpha\frac{\partial g(\alpha,r)}{\partial\alpha}+\frac{\partial^2 g(\alpha,r)}{\partial\alpha\partial r}-r\frac{\partial g(\alpha,r)}{\partial r}\right) \end{aligned} \tag{B.49}$$

$$\frac{\partial^2 f(\alpha,\beta,r)}{\partial\alpha\partial r} = \frac{12\alpha^3\beta^4}{(\alpha^2-\beta^2)^4 r^2} \tag{B.50}$$

$$\frac{\partial^2 f(\beta,\alpha,r)}{\partial\alpha\partial r} = -\frac{12\alpha^3\beta^4}{(\beta^2-\alpha^2)^4 r^2} \tag{B.51}$$

$$\frac{\partial^2 g(\alpha,r)}{\partial\alpha\partial r} = \frac{3}{8}\alpha+\frac{1}{8}\alpha^2 r. \tag{B.52}$$

B.5 The γ^{h}-Function

The $\gamma^{\mathrm{h}}_{ab}=\gamma^{\mathrm{h}}_{ba}$-function as described in eq 4.18 differs from the original γ-function specifically for HX atom pairs (X$\in${C, H, N, O, P, S}):

$$\gamma_{ab} = \frac{1}{r}-S\cdot h \tag{B.53}$$

$$h = \exp\left(-\left(\frac{U_a+U_b}{2}\right)^{\zeta} r^2\right), \tag{B.54}$$

where $S = S^{\mathrm{f}}$ when $\alpha \neq \beta$ and $S = S^{\mathrm{g}}$ when $\alpha = \beta$. S^{f} and S^{g} as well as α, β, and r are defined in eqs B.2 and B.3. Eqs B.6, B.41, and B.47 then change to:

$$\frac{\partial \gamma_{ab}}{\partial U_a} = -\left(\frac{\partial S}{\partial U_a} h + S \frac{\partial h}{\partial U_a}\right) \tag{B.55}$$

$$\frac{\partial \gamma_{ab}}{\partial r} = -\frac{1}{r^2} - \left(\frac{\partial S}{\partial r} h + S \frac{\partial h}{\partial r}\right) \tag{B.56}$$

$$\frac{\partial^2 \gamma_{ab}}{\partial U_a \partial r} = -\left(\frac{\partial^2 S}{\partial U_a \partial r} h + \frac{\partial S}{\partial U_a}\frac{\partial h}{\partial r} + \frac{\partial S}{\partial r}\frac{\partial h}{\partial U_a} + S \frac{\partial^2 h}{\partial U_a \partial r}\right) \tag{B.57}$$

with

$$\frac{\partial h}{\partial U_a} = -\frac{\zeta r^2}{2}\left(\frac{U_a + U_b}{2}\right)^{\zeta - 1} \cdot h \tag{B.58}$$

$$\frac{\partial h}{\partial r} = -2r\left(\frac{U_a + U_b}{2}\right)^{\zeta} \cdot h \tag{B.59}$$

$$\frac{\partial^2 h}{\partial U_a \partial r} = \zeta r \left(\frac{U_a + U_b}{2}\right)^{\zeta - 1} \left(r^2 \left(\frac{U_a + U_b}{2}\right)^{\zeta} - 1\right) \cdot h. \tag{B.60}$$

Appendix C

Molecule Sets and Benchmarks

Table C.1: 32 Reaction Energies (kcal/mol) of Training Set 2 as Defined in Section 5.2 Calculated Using G3B3 at 0 K Excluding Zero-Point and Thermal Corrections

Reaction					G3B3
Isodesmic reactions					
Butadiyne	+	Ethane	→	2 Propyne	–1.5
Biphenyl	+	Ethane	→	2 Toluene	2.3
1,3-Butadiene	+	Butane	→	2 E-2-Butene	–1.8
Propane	+	Methane	→	2 Ethane	3.0
Isobutane	+	2 Methane	→	3 Ethane	8.1
2,2-Dimethylpropane	+	3 Methane	→	4 Ethane	14.2
Propene	+	Methane	→	Ethane + Ethene	5.6
1,3-Butadiene	+	2 Methane	→	2 Ethene + Ethane	14.3
E-2-Butene			→	1-Butene	2.9
1,3-Pentadiene			→	1,4-Pentadiene	6.5
Non-isodesmic reactions					
Ethene	+	2 Methane	→	2 Ethane	–15.9
Ethyne	+	4 Methane	→	3 Ethane	–39.4
3 Ethyne			→	Benzene	–138.7
Ethyne	+	Ethane	→	2 Ethene	–9.4
Allene	+	Methane	→	2 Ethene	–2.1
Isobutane			→	Ethene + Ethane	23.1
Naphthalene	+	Butadiyne	→	Anthracene	–89.3
Cyclohexane	+	Ethane	→	Hexane + Ethene	20.0
1,3-Butadiene	+	Ethene	→	Cyclohexene	–37.5
Propene	+	Ethene	→	Pentene	–19.8
Propyne			→	Allene	1.0
Hydrogenations					
Ethane	+	Hydrogen	→	2 Methane	–15.2
Ethene	+	2 Hydrogen	→	2 Methane	–45.4
Ethyne	+	3 Hydrogen	→	2 Methane	–85.0
Ethene	+	Hydrogen	→	Ethane	–30.2
Ethyne	+	Hydrogen	→	Ethene	–39.6
Ethyne	+	2 Hydrogen	→	Ethane	–69.8
Reactions with small cyclic structures					
Benzene	+	Cyclopropane	→	Mesitylene	–38.1
Cyclopropane	+	3 Methane	→	3 Ethane	–18.5
Cyclopropene	+	3 Methane	→	2 Ethane + Ethene	–40.7
Propene			→	Cyclopropane	9.1
1,3-Butadiene			→	Cyclobutene	13.3

Table C.2: Experimental Bond Lengths (Å), Bond Angles (°), and Fundamental Vibrational Frequencies (cm^{-1}) of Training Set 2 as Defined in Section 5.2

Molecule	Variable	Value	Symmetry	Frequency	Reference
Dihydrogen	rHH	0.741	Σ_g	4401	[123]
$D_{\infty h}$					
Methane	rCH	1.087	A_1	2917	[123]
T_d			E	1534	
			T_2	3019, 1306	
Ethyne	rC≡C	1.203	Σ_g	3374, 1974	[123]
$D_{\infty h}$	rCH	1.063	Σ_u	3289	
			Π_g	612	
			Π_u	730	
Ethene	rC=C	1.330	A_g	3026, 1623, 1342	[123, 124]
D_{2h}	rCH	1.080	A_u	1023	
	aHCH	117.1	B_{1g}	3086, 1222	
			B_{1u}	949	
			B_{2g}	940	
			B_{2u}	3105, 826	
			B_{3u}	2989, 1444	
Ethane	rCC	1.522	A_{1g}	2896, 1388, 995	[123, 124]
D_{3d}	rCH	1.089	A_{1u}	289	
	aCCH	111.2	A_{2u}	2915, 1370	
			E_g	2969, 1468, 1190	
			E_u	2974, 1460, 822	
Cyclopropene	rCC	1.505	A_1	3158, 2909, 1653	[123, 124]
C_{2v}	rC=C	1.293	A_1	1483, 1110, 905	
	rCH=	1.072	A_2	996, 820	
	rCH–	1.085	B_1	2995, 1088, 569	
	aC=CH	150.0	B_2	3124, 1043, 1011, 769	
	aHCH	114.3			
Propyne	rCC	1.460	A_1	3334, 2918, 2124	[123]
C_{3v}	rC≡C	1.207	A_1	1382, 931	
	rCH1	1.096	E	3008, 1452, 1053, 633	
	rCH2	1.060	E	328	
	aHCC	110.6			
Cyclopropane	rCC	1.501	A_1'	3038, 1479, 1188	[123]
D_{3h}	rCH	1.083	A_1''	1126	
	aHCH	114.5	A_2'	1070	
			A_2''	3103, 854	
			E'	3025, 1438, 1029, 866	
			E''	3082, 1188, 739	
Propene	rC=C	1.333	A'	3090, 3013, 2991	[123, 135]
C_s	rCC	1.496	A'	2954, 2871, 1650	
	rCH11	1.094	A'	1470, 1420, 1378	
	rCH12	1.082	A'	1297, 1171, 963, 920	
	rCH23	1.091	A'	428	
	rCHs	1.084	A''	2954, 1443, 1045, 991	
	rCHa	1.108	A''	912, 578, 174	
	aH2C=C	121.6			
	aCCC	124.3			
	aH1C=C	120.4			
	aH3C–C	117.4			
	aHsCC	111.3			
	aHaCC	109.9			
	aHaCHa	106.7			

Table C.3: Experimental Bond Lengths (Å), Bond Angles (°), and Fundamental Vibrational Frequencies (cm^{-1}) of Training Set 2 as Defined in Section 5.2

Molecule	Variable	Value	Symmetry	Frequency	Reference
Allene	rC=C	1.308	A_1	3015, 1443, 1073	[123, 124]
D_{2d}	rCH	1.076	B_1	865	
	aHCH	118.2	B_2	3007, 1957, 1398	
			E	3086, 999, 841, 355	
Propane	rCC	1.521	A_1	2977, 2962, 2887	[123, 135]
C_{2v}	rCH	1.093	A_1	1476, 1462, 1392	
	rCHs	1.088	A_1	1158, 869, 369	
	rCHa	1.091	A_2	2967, 1451, 1278, 940	
	aCCC	112.4	A_2	216	
	aHCH	106.1	B_1	2973, 2968, 1472	
	aHaCHs	107.0	B_1	1192, 748, 268	
	aCCHs	111.6	B_2	2968, 2887, 1464	
	aCCHa	110.6	B_2	1378, 1338, 1054, 922	
2-Butyne	rCC	1.468	A_1'	2916, 2240, 1380, 725	[123]
D_{3h}	rC≡C	1.214	A_1''	28[a]	
	rCH	1.116	A_2''	2938, 1382, 1152	
	aHCC	110.7	E′	2973, 1456, 1054, 213	
			E″	2966, 1448, 1029, 371	
Cyclohexane	rCC	1.530	A_{1g}	2930, 2852, 1465	[123]
D_{3d}	rCHax	1.101	A_{1g}	1157, 802, 383	
	rCHeq	1.093	A_{1u}	1383, 1157, 1057	
	aCCC	111.3	A_{2g}	1437, 1090	
	aHCCax	108.8	A_{2u}	2915, 2860, 1437	
	aHCCeq	110.6	A_{2u}	1030, 523	
	aHCH	106.7	E_g	2930, 2897, 1443	
	dCCCC	55.3	E_g	1347, 1266, 1027, 785	
			E_g	426	
			E_u	2933, 2863, 1457	
			E_u	1355, 1261, 907, 863	
			E_u	248	
Benzene	rC%C	1.390	A_{1g}	3062, 992	[123, 124]
D_{6h}	rCH	1.086	A_{2g}	1326	
			A_{2u}	673	
			B_{1u}	3068, 1010	
			B_{2g}	995, 703	
			B_{2u}	1310, 1150	
			E_{1g}	849	
			E_{1u}	3063, 1486, 1038	
			E_{2g}	3047, 1596, 1178, 606	
			E_{2u}	975, 410	
Cubane	rCC	1.562	A_{1g}	2995, 1002	[123, 132]
O_h	rCH	1.096	A_{2u}	2978, 839	
			E_g	1083, 912	
			E_u	1151, 617	
			T_{1g}	1130	
			T_{1u}	2978, 1230, 853	
			T_{2g}	2970, 1182, 821, 665	
			T_{2u}	1036, 829	

[a] Calculated with BLYP/cc-pVTZ.

Table C.4: Experimental Atomization Energies at 0 K (E^{at}) and Zero-Point Energies (ZPE) from ref [123] in kcal/mol of Training Set 2 as Defined in Section 5.2

Molecule	E^{at}	ZPE	E^{at}+ZPE
Dihydrogen	103.3	6.3	109.6
Methane	392.4	27.1	419.5
Ethyne	388.7	16.2	404.9
Ethene	531.9	30.8	562.7
Ethane	666.1	45.2	711.3
Allene	669.1	33.3	702.4
Cyclopropene	648.2	33.8	682.0
Propyne	670.7	33.8	704.5
Cyclopropane	802.9	49.5	852.4
Propene	811.4	48.2	859.6
Propane	942.7	63.0	1005.7
2-Butyne	951.7	51.1	1002.8
Cyclohexane	1659.5	103.6	1763.1
Benzene	1305.7	61.2	1366.9

Table C.5: Deviation of Different Bond Types of 61 Molecules in the G2/97 Molecule Set Compared to MP2/cc-pVTZ Calculations in Angstrom

		MIO2			MIO3			d22		
Type[a]	N[b]	MAX	MSE	MUE	MAX	MSE	MUE	MAX	MSE	MUE
rC–C	39	0.034	–0.001	0.009	0.033	–0.001	0.008	0.034	+0.003	0.010
rC=C	11	0.020	+0.007	0.008	0.020	+0.007	0.008	0.016	+0.004	0.006
rC≡C	3	0.008	–0.008	0.008	0.008	–0.007	0.007	0.007	–0.006	0.006
rC%C	3	0.008	+0.005	0.005	0.008	+0.005	0.005	0.007	+0.004	0.004
rC–H	102	0.047	+0.016	0.016	0.045	+0.014	0.014	0.034	+0.006	0.006
rH–H	1	0.006	+0.006	0.006	0.006	+0.006	0.006	0.006	+0.006	0.006
rC–O	13	0.048	+0.024	0.024	0.041	+0.026	0.026	0.028	+0.013	0.015
rC=O	9	0.028	–0.009	0.010	0.027	–0.009	0.010	0.034	–0.014	0.015
rC≡O	1	0.038	–0.038	0.038	0.038	–0.038	0.038	0.066	–0.066	0.066
rH–O	7	0.027	+0.017	0.017	0.023	+0.015	0.015	0.016	+0.005	0.006
rO–O	1	0.003	+0.003	0.003	0.002	+0.002	0.002	0.021	+0.021	0.021
rO=O	1	0.035	–0.035	0.035	0.035	–0.035	0.035	0.039	–0.039	0.039
rC–N	7	0.026	–0.012	0.012	0.022	–0.010	0.011	0.033	+0.006	0.012
rC≡N	4	0.027	–0.025	0.025	0.027	–0.025	0.025	0.020	–0.018	0.018
rC%N	1	0.001	–0.001	0.001	0.001	–0.001	0.001	0.005	+0.005	0.005
rC@N	1	0.016	+0.016	0.016	0.014	+0.014	0.014	0.025	+0.025	0.025
rH–N	10	0.027	+0.020	0.020	0.025	+0.019	0.019	0.012	+0.007	0.008
rN–N	1	0.033	–0.033	0.033	0.029	–0.029	0.029	0.023	+0.023	0.023
rN≡N	1	0.006	–0.006	0.006	0.006	–0.006	0.006	0.022	–0.022	0.022
rN–O	1	0.065	+0.065	0.065	0.062	+0.062	0.062	0.004	–0.004	0.004
rN@O	3	0.008	–0.005	0.005	0.007	–0.002	0.002	0.011	–0.009	0.009
all	223	0.065	+0.009	0.014	0.062	+0.008	0.014	0.066	+0.004	0.009

[a] The abbreviations stand for "–" single bond, "=" double bond, "≡" triple bond, "%" aromatic bond, and "rC@N" partial double bond within acetamide, "rN@O" partial double bonds within nitrous acid and nitromethane. [b] Number of comparisons.

Table C.6: 22 Binding Energies in kcal/mol: Deviation of DFTB in Comparison to G3B3[a]

							Without counterpoise correction				With counterpoise correction			
							PBE		B3LYP		PBE		B3LYP	
System	G3B3	MIO2	MIO3/calc	MIO3/fit	c21	d22	Set1[b]	Set2[c]	Set1[b]	Set2[c]	Set1[b]	Set2[c]	Set1[b]	Set2[c]
$2H_2O$	–4.9	+1.6	–0.0	+0.0	+0.3	+0.2	–1.8	–3.3	–1.1	–2.4	–0.9	+0.1	–0.3	+0.5
$3H_2O$	–15.1	+5.5	–0.3	–0.1	+0.7	+0.2	–5.0	–11.6	–2.9	–8.1	–2.9	–1.7	–1.0	+0.3
$4H_2O$	–27.4	+9.7	+0.8	+1.1	+2.7	+1.9	–8.7	–16.3	–4.8	–11.2	–4.4	–2.2	–1.1	+1.1
$5H_2O$	–36.3	+13.3	+1.3	+1.7	+3.6	+2.6	–11.3	–19.8	–6.2	–13.3	–5.5	–2.6	–1.1	+1.6
$2H_2O(H^+)$	–33.9	+4.5	+0.9	+2.1	+2.5	+2.0	–5.4	–7.6	–3.6	–5.3	–4.0	–4.1	–2.3	–2.1
$3H_2O(H^+)$	–57.3	+10.4	+3.7	+5.3	+4.7	+3.9	–7.7	–11.7	–5.2	–8.3	–5.0	–5.5	–2.7	–2.9
$4H_2O(H^+)$	–77.2	+13.9	+5.0	+6.7	+6.1	+5.1	–8.7	–14.8	–5.7	–10.7	–4.8	–5.8	–2.2	–2.8
$5H_2O(H^+)$	–91.9	+18.3	+6.2	+8.1	+8.3	+7.0	–10.6	–20.3	–6.1	–14.3	–5.3	–5.7	–1.2	–1.4
$2H_2O(-H^+)$	–27.4	–5.1	–5.9	–3.2	–4.6	–5.4	–5.3	–22.2	–2.6	–18.0	–3.9	–5.0	–1.5	–2.4
$3H_2O(-H^+)$	–48.6	–2.6	–8.4	–5.3	–6.4	–7.9	–8.6	–35.5	–5.1	–28.9	–6.0	–10.8	–2.9	–7.7
$4H_2O(-H^+)$	–66.7	+0.3	–7.2	–3.5	–5.5	–7.1	–11.5	–43.0	–6.2	–34.2	–7.0	–5.1	–2.5	–0.7
$5H_2O(-H^+)$	–86.3	+6.1	–7.8	–4.7	–5.8	–8.1	–10.4	–51.2	–3.8	–40.7	–4.5	–1.3	+1.0	+4.6
$NH_3(H_2O)$	–6.6	+3.2	+2.1	+2.1	+3.1	+2.9	–2.3	–2.3	–1.2	–1.4	–1.2	+0.5	–0.3	+1.1
$NH_4^+(H_2O)$	–20.4	+0.6	–1.3	–0.9	+0.2	–0.1	–2.7	–4.8	–1.7	–3.5	–1.7	–2.2	–0.8	–1.2
6H2O_book	–45.8	+16.7	+1.7	+2.2	+4.8	+3.4	–13.1	–28.8	–6.5	–17.9	–5.9	–2.7	–0.2	+2.9
6H2O_cage	–46.6	+17.2	+1.5	+1.8	+4.6	+2.9	–11.5	–28.1	–4.7	–19.1	–4.7	–1.6	+1.1	+4.2
6H2O_prism	–47.2	+17.6	+1.3	+1.7	+4.4	+2.7	–10.8	–28.4	–4.0	–19.5	–4.0	–1.1	+1.9	+4.8
6H2O_ring	–44.7	+16.5	+1.5	+2.1	+4.2	+3.0	–14.1	–23.0	–8.0	–15.5	–6.9	–3.0	–1.7	+1.9
Methylimidazole$(-H^+)(H_2O)$	–15.9	+4.1	+1.2	+1.1	+0.7	+0.6	–2.6	–4.6	–1.1	–2.7	–1.9	+0.5	–0.5	+1.6
Methylimidazole(H_2O)_1	–6.2	+2.4	+1.9	+2.0	+2.3	+2.2	–0.6	–2.5	–0.0	–1.4	+0.4	+0.7	+0.8	+1.1
Methylimidazole(H_2O)_2	–8.2	+3.5	+1.9	+1.9	+3.1	+3.0	–1.0	–2.1	–0.0	–1.0	–0.3	+1.2	+0.5	+1.8
Methylimidazole$H^+(H_2O)$	–16.0	+3.3	+2.3	+2.5	+3.6	+3.4	–1.0	–2.9	–0.2	–1.8	+0.2	+0.0	+1.0	+0.8
MUE		8.0	2.9	2.7	3.7	3.4	7.0	17.5	3.7	12.7	3.7	2.9	1.3	2.2
MSE		+7.3	+0.1	+1.1	+1.7	+0.8	–7.0	–17.5	–3.7	–12.7	–3.6	–2.6	–0.7	+0.3
MAX		18.3	8.4	8.1	8.3	8.1	14.1	51.2	8.0	40.7	7.0	10.8	2.9	7.7

[a] The notation is defined in Table 4.5. G3B3, PBE, and B3LYP calculations were carried out using the Gaussian 03 software package [108]. [b] Basis set 6-31+G(d,p). [c] Basis set 6-311G(2d,2p).

Table C.7: 23 Proton Affinities with Acidic Oxygen in kcal/mol: Deviation of DFTB in Comparison to G3B3[a]

							PBE			B3LYP		
System	G3B3	MIO2	MIO3/calc	MIO3/fit	c21	d22	Set1[b]	Set2[c]	Set3[d]	Set1[b]	Set2[c]	Set3[d]
H_2O	398.4	+16.3	+7.5	–1.8	+1.8	+1.0	–4.5	+31.4	–3.2	–2.7	+31.1	–1.3
$2H_2O$	375.9	+9.6	+1.7	–5.1	–3.1	–4.6	–8.0	+12.6	–6.9	–4.2	+15.5	–2.8
$3H_2O$	365.0	+8.1	–0.7	–7.1	–5.4	–7.2	–8.3	+7.4	–7.2	–5.0	+10.3	–3.7
$4H_2O$	359.1	+7.0	–0.5	–6.5	–6.3	–8.1	–7.4	+4.7	–6.9	–4.1	+8.1	–3.2
$5H_2O$	348.4	+9.2	–1.6	–8.3	–7.5	–9.7	–3.6	+0.1	–3.8	–0.3	+3.7	+0.0
CH_3OH	392.6	–5.7	+5.9	–0.3	+3.0	+0.4	–7.6	+4.2	–8.3	–4.0	+8.2	–4.2
CH_3CH_2OH	388.3	–1.5	+9.0	+2.2	+5.4	+2.3	–7.7	+3.3	–8.5	–3.2	+7.7	–3.4
$CH_3CH_2CH_2OH$	387.6	–2.2	+8.6	+2.0	+5.0	+2.0	–7.5	+3.0	–7.8	–3.0	+7.6	–2.9
CH_3–CH(OH)–CH_3	385.6	+1.4	+10.7	+3.3	+6.7	+2.7	–7.3	+3.1	–7.7	–2.6	+7.5	–2.5
HCOOH	351.2	+1.7	+10.0	+2.9	+6.6	+2.3	–5.0	+9.5	–4.0	–3.1	+10.7	–1.9
CH_3COOH	355.1	+1.1	+8.5	+0.6	+3.9	–1.1	–4.3	+8.2	–3.5	–2.2	+9.5	–1.4
CH_3CH_2COOH	354.5	+1.0	+9.3	+1.5	+4.7	–0.1	–4.2	+6.9	–3.4	–1.9	+8.6	–1.1
C_6H_5OH	356.7	–4.7	+9.7	+4.0	+7.5	+3.4	–6.2	+2.5	–5.0	–2.3	+6.2	–1.3
p–CH_3–C_6H_4OH	357.9	–5.6	+9.2	+3.7	+6.9	+3.0	–6.2	+1.8	–5.1	–2.1	+5.6	–1.2
p–NO_2–C_6H_4OH	334.6	–9.3	+3.5	–1.3	–1.0	–3.9	–9.1	+0.5	–7.2	–5.7	+3.4	–4.0
H_3O^+	171.2	–0.4	+6.3	+4.3	+1.6	–0.3	+0.4	+5.4	–0.7	+0.6	+5.0	–0.5
$2H_2O(H^+)$	200.2	–3.3	+5.4	+2.2	–0.6	–2.1	+4.0	+9.6	+2.1	+3.0	+7.9	+1.0
$3H_2O(H^+)$	213.4	–5.3	+2.3	–1.2	–2.4	–4.0	+3.0	+5.5	+1.2	+2.8	+5.2	+1.0
$4H_2O(H^+)$	221.1	–4.7	+2.0	–1.4	–1.8	–3.5	+0.3	+3.7	–0.3	+1.4	+4.4	+0.6
$5H_2O(H^+)$	226.7	–5.3	+1.4	–1.9	–3.0	–4.6	–0.2	+5.9	–0.8	+0.5	+6.1	–0.2
$CH_3OH_2^+$	186.8	–8.3	+4.6	+2.2	+2.2	–0.0	–1.0	+2.2	–1.8	+0.5	+3.4	–0.3
H_2COH^+	177.1	–11.8	+2.6	–0.2	+1.1	–2.0	–2.2	+1.0	–1.8	–0.6	+2.4	–0.3
CH_3CHOH^+	190.2	–10.1	+5.1	+2.0	+3.9	+0.7	+0.5	+3.7	+1.1	+1.0	+3.9	+1.5
MUE		5.8	5.5	2.9	4.0	3.0	4.7	5.9	4.3	2.5	7.9	1.8
MSE		–1.0	+5.2	–0.2	+1.3	–1.5	–4.0	5.9	–3.9	–1.6	7.9	–1.4
MAX		16.3	10.7	8.3	7.5	9.7	9.1	31.4	8.5	5.7	31.1	4.2

[a] The notation is defined in Table 4.6. G3B3, PBE, and B3LYP calculations were carried out using the Gaussian 03 software package [108]. [b] Basis set 6-31+G(d,p). [c] Basis set 6-311G(2d,2p). [d] Basis set G3large [120].

Table C.8: 9 Proton Affinities with Acidic Nitrogen in kcal/mol: Deviation of DFTB and the NHorg Parameter Set in Comparison to G3B3[a]

System	G3B3	MIO2	MIO3/calc	MIO3/fit	c21	d22
$HCNH^+$	176.0	–12.4	+2.8	+0.2	+3.4	+0.5
CH_3CNH^+	192.3	–14.3	+1.9	–0.9	+2.9	–0.1
$C_5H_5NH^+$	229.5	–17.1	+0.4	–2.1	+3.9	+1.3
MethylimidazoleH^+	237.3	–12.7	+4.7	+2.1	+7.2	+4.6
MethylguanidineH^+	249.3	–12.0	–0.8	–2.9	–0.6	–2.5
NH_3	413.9	+10.4	–5.3	–0.2	–8.4	–0.6
NH_4^+	212.3	–24.4	–14.4	–15.8	–14.3	–16.5
$CH_3NH_3^+$	223.3	–26.8	–13.3	–15.3	–11.0	–13.5
1-AminobutaneH^+	228.2	–26.7	–12.6	–14.6	–10.0	–12.4
MUE		17.4	6.2	6.0	6.9	5.8
MSE		–15.1	–4.1	–5.5	–3.0	–4.4
MAX		26.8	14.4	15.8	14.3	16.5

[a] The notation is defined in Table 4.7.

Table C.9: Proton Transfer Barrier in kcal/mol for a Fixed Distance (rXY) between the Heavy Atoms (X,Y∈{O,N}): Deviation of DFTB and DFT in Comparison to MP2/G3large[a]

Barrier	rXY	MP2	MIO2	MIO3/calc	MIO3/fit	d22	PBE[b]	B3LYP[b]
$[H_2O–H–OH_2]^+$	2.5	0.6	–0.6	–0.6	–0.6	–0.5	–0.5	–0.4
	2.6	2.4	–1.4	–1.1	–0.8	–0.6	–1.9	–1.0
	2.7	5.2	–1.1	–0.7	–0.3	–0.9	–3.2	–1.5
	2.8	8.9	–1.3	–0.9	–0.4	–1.5	–4.4	–1.9
$[HO–H–OH]^-$	2.5	0.5	–0.5	–0.5	–0.5	–0.2	–0.5	–0.2
	2.6	2.3	–2.3	–0.7	+0.0	–0.1	–1.6	–0.6
	2.7	5.2	–4.6	–0.2	+0.9	–0.3	–2.7	–0.9
	2.8	8.8	–6.7	–0.3	+1.2	–0.9	–3.7	–1.2
$[H_3N–H–NH_3]^+$	2.6	0.4	–0.4	–0.4	–0.4	–0.4	–0.4	–0.3
	2.7	1.9	–1.8	–1.7	–1.7	–1.9	–1.9	–1.0
	2.8	4.4	–2.4	–2.0	–2.2	–2.6	–3.0	–1.6
	2.9	7.7	–2.5	–2.1	–2.2	–3.3	–4.1	–2.2
$[H_2N–H–NH_2]^-$	2.5	0.1	–0.0	–0.1	–0.1	–0.1	+0.0	+0.0
	2.6	1.4	–1.4	–0.4	–1.0	–1.4	–1.2	–0.5
	2.7	3.5	–3.5	+1.5	–0.5	–1.6	–2.1	–0.8
	2.8	6.3	–4.9	+2.9	+0.9	–0.9	–3.0	–1.1
$[H_3N–H–OH_2]^+$	2.9	25.3	–8.2	–5.8	–5.3	–3.9	—[c]	–2.7
	3.0	30.0	–9.5	–7.3	–6.7	–5.2	–6.8	–3.3
	3.1	35.1	–11.0	–8.9	–8.2	–6.5	–7.8	–3.8
	3.2	40.5	–12.5	–10.5	–9.8	–7.8	–8.8	–4.2
$[H_2O–H–NH_3]^+$	2.9	0.8	+0.7	+0.2	+0.1	–0.7	—[c]	–0.8
	3.0	3.3	+0.9	+0.2	+0.1	–1.7	–3.0	–1.6
	3.1	6.7	+0.3	–0.6	–0.7	–2.5	–4.5	–2.3
	3.2	10.7	–0.8	–1.9	–2.0	–3.2	–5.9	–2.9
$[H_2N–H–OH]^-$	2.8	10.1	–6.3	–2.5	+2.9	–0.6	–3.3	–0.8
	2.9	14.2	–8.6	–3.1	+2.7	–1.0	–4.2	–1.1
	3.0	18.6	–11.1	–3.9	+2.1	–1.7	–5.1	–1.4
	3.1	23.3	–13.4	–4.7	+1.5	–2.3	–5.9	–1.6
$[HO–H–NH_2]^-$	2.8	4.4	–4.1	+6.6	+0.7	+0.4	–2.8	–1.2
	2.9	7.8	–5.2	+7.9	+1.6	+0.3	–4.0	–1.7
	3.0	11.6	–6.6	+8.4	+1.7	–0.2	–4.9	–2.0
	3.1	15.8	–8.8	+7.8	+0.7	–0.9	–5.8	–2.3

[a] The notation is defined in Table 4.9. [b] Basis set 6-31+G(d,p). [c] A barrier does not exist.

Table C.10: Selected First and Second Neighbor Distances r for B3LYP/cc-pVTZ Equilibrium Geometries

Pairing	r in a.u.	r in Å	Molecule	Pairing	r in a.u.	r in Å	Molecule
CC	2.260	1.196	Ethyne	HN	1.911	1.011	Hydrazine
	2.502	1.324	Ethene		1.914	1.013	Methylamine
	2.628	1.391	Benzene		1.916	1.014	Ammonia
	2.886	1.527	Ethane		1.918	1.015	Hydrazine
	2.960	1.567	Acetate-Anion		1.930	1.021	Methanimine
	3.960	2.095	Cyclobutene		3.757	1.988	Hydrazine
	4.552	2.409	Benzene		3.833	2.029	Methanimine
	4.742	2.509	Propene		3.865	2.045	Hydrazine
CH	2.006	1.062	Ethyne		3.957	2.094	Methylamine
	2.045	1.082	Benzene	HO	1.815	0.961	Methanol
	2.046	1.083	Ethene		1.817	0.961	Water
	2.057	1.088	Methane		1.825	0.966	Hydrogen peroxide
	2.057	1.089	Methanol		1.834	0.970	Formic acid
	2.062	1.091	Ethane		3.560	1.884	Hydrogen peroxide
	2.070	1.095	Methanol		3.777	1.999	Formic acid
	3.687	1.951	Methanol		3.815	2.019	Formaldehyde
	3.979	2.106	Ethene		3.826	2.025	Methanol
	4.057	2.147	Benzene		3.964	2.098	Methanol
	4.113	2.176	Ethane	NN	2.062	1.091	Nitrogen
	4.266	2.258	Ethyne		2.713	1.436	Hydrazine
CN	2.166	1.146	Hydrogen cyanide		4.055	2.146	MethylimidazoleH+[a]
	2.387	1.263	Methanimine		4.196	2.220	Methylimidazole[a]
	2.767	1.464	Methylamine	NO	2.225	1.178	Nitrous acid
	4.186	2.215	Pyrrole		2.302	1.218	Nitromethane
	4.797	2.539	Ethylamine		2.628	1.391	Nitrous acid
CO	2.266	1.199	Formaldehyde		4.304	2.278	Formamide
	2.541	1.345	Formic acid	OO	2.280	1.206	Dioxygen (singlet)
	2.684	1.420	Methanol		2.744	1.452	Hydrogen peroxide
	4.216	2.231	Furan		4.070	2.154	Nitrous acid
	4.497	2.380	Ethanol		4.100	2.169	Nitromethane
HH	1.404	0.743	Dihydrogen		4.266	2.258	Formic acid
	2.873	1.521	Water				
	3.070	1.625	Ammonia				
	3.358	1.777	Methane				

[a] B3LYP/6-31+G(d,p).

Appendix D

Self-Interaction Correction for DFTB3

In chapter 2, the self-interaction error (SIE) which is rooted in all standard exchange-correlation functionals of the density functional theory has been discussed. While the nonself-consistent DFTB variant does not lack this error, it emerges in DFTB2 (and DFTB3) due to the approximations made for the second order (and third order) terms. It has been discussed in previous publications, that this causes an artificial stabilization of delocalized states [136, 137, 138]. In this section a self-interaction correction (SIC) is suggested for DFTB3 (SIC-DFTB3) by assimilating the ideas of Perdew and Zunger [53] to eliminate the SIE orbital by orbital. The SIC has been implemented in the DFTB3 code, first tests are currently carried out.

In analogy to eq 4.6 the SIC-DFTB3 total energy is written as

$$E^{\text{SIC-DFTB3}} = E^{\text{H0}} + E^{\gamma} + E^{\Gamma} + E^{\text{rep}} + E^{\text{SIC}\gamma} + E^{\text{SIC}\Gamma}, \tag{D.1}$$

where the SIC terms remove the charge-charge interaction within each molecular orbital i as[1]

$$E^{\gamma} = \frac{1}{2}\sum_{ab}\Delta q_a \Delta q_b \gamma_{ab} \tag{D.2}$$

$$E^{\Gamma} = \frac{1}{3}\sum_{ab}\Delta q_a^2 \Delta q_b \Gamma_{ab} \tag{D.3}$$

$$E^{\text{SIC}\gamma} = -\frac{1}{2}\sum_{iab}\Delta q_{ai} \Delta q_{bi} \gamma_{ab} \tag{D.4}$$

$$E^{\text{SIC}\Gamma} = -\frac{1}{3}\sum_{iab}\Delta q_{ai}^2 \Delta q_{bi} \Gamma_{ab}. \tag{D.5}$$

Similar to the definition of the Mulliken charges and charge fluctuations, orbital charges are introduced:

$$q_a = \sum_i n_i \sum_{\mu\in a}\sum_b \sum_{\nu\in b} c_{\mu i} c_{\nu i} S_{\mu\nu} \tag{D.6}$$

$$\Delta q_a = q_a - q_a^0 \tag{D.7}$$

[1] In the following a, b, c, and d denote indices for atoms in a molecular system.

$$q_{ai} = n_i \sum_{\mu\in a}\sum_{b}\sum_{\nu\in b} c_{\mu i}c_{\nu i}S_{\mu\nu} \tag{D.8}$$

$$\Delta q_{ai} = q_{ai} - n_{ai}^{\mathrm{SIC}} q_a^0, \tag{D.9}$$

where q_a^0 is the charge of the valence electrons of the neutral atom a. Because an analogon for the charge of a neutral atom within a molecular orbital i does not exist, the term $n_{ai}^{\mathrm{SIC}} q_a^0$ is suggested. The value of n_{ai}^{SIC} can be determined using the charges of a preceding self-consistent charge (SCC) energy calculation (regular DFTB3 calculation) as

$$n_{ai}^{\mathrm{SIC}} = \frac{q_{ai}^{\mathrm{SCC}}}{q_a^{\mathrm{SCC}}}. \tag{D.10}$$

Note that $\sum_i n_{ai}^{\mathrm{SIC}} = 1$ and $\sum_i \Delta q_{ai} = \Delta q_a$.

Due to the orbital dependence, the Kohn-Sham scheme cannot be applied. The energy is minimized as suggested in ref [54] by orbital variations utilizing the Broyden-Fletcher-Goldfarb-Shanno (BFGS) variant of a quasi-Newton method [122]. The orbitals are re-orthogonalized after each minimization step using Löwdin's procedure (for details see [54]). The gradients, necessary for the minimization are given by

$$\frac{\partial E^{\mathrm{SIC}\gamma}}{\partial c_{\delta i}} = -\sum_{bc} \Delta q_{bi}(\gamma_{db} + \gamma_{cb}) \sum_{\nu\in c} n_i c_{\nu i} S_{\delta\nu} \qquad \delta \in d \tag{D.11}$$

$$\frac{\partial E^{\mathrm{SIC}\Gamma}}{\partial c_{\delta i}} = -\frac{1}{3}\sum_{bc} \Delta q_{bi}\left(2\Delta q_{di}\Gamma_{db} + \Delta q_{bi}\Gamma_{bd} + 2\Delta q_{ci}\Gamma_{cb} + \Delta q_{bi}\Gamma_{bc}\right) \sum_{\nu\in c} n_i c_{\nu i} S_{\delta\nu}, \tag{D.12}$$

where use was made of

$$\frac{\partial q_{ai}}{\partial c_{\delta i}} = \delta_{ad}\delta_{\mu\delta}\sum_{b}\sum_{\nu\in b} n_i c_{\nu i} S_{\delta\nu} + n_i \sum_{\mu\in a} c_{\mu i} S_{\mu\delta} \qquad \delta \in d \tag{D.13}$$

(δ_{ad} and $\delta_{\mu\delta}$ are Kronecker delta).

At the energy minimum the following expression is valid:

$$\sum_{b}\sum_{\nu\in b} c_{\nu i}\left(H_{\mu\nu i} - \epsilon_i S_{\mu\nu}\right) = 0, \qquad \forall a, \mu \in a, i, \tag{D.14}$$

where the orbital dependent Hamiltonian is given by (compare eq 4.11)

$$\begin{aligned} H_{\mu\nu i} = H^0_{\mu\nu} &+ \frac{1}{2}S_{\mu\nu}\sum_c \Delta q_c\left(\gamma_{ac} + \gamma_{bc}\right) \\ &+ \frac{1}{3}S_{\mu\nu}\sum_c \Delta q_c\left(\Delta q_a\Gamma_{ac} + \Delta q_b\Gamma_{bc}\right) + \frac{1}{6}S_{\mu\nu}\sum_c \Delta q_c^2\left(\Gamma_{ca} + \Gamma_{cb}\right) \\ &- \frac{1}{2}S_{\mu\nu}\sum_c \Delta q_{ci}\left(\gamma_{ac} + \gamma_{bc}\right) \\ &- \frac{1}{3}S_{\mu\nu}\sum_c \Delta q_{ci}\left(\Delta q_{ai}\Gamma_{ac} + \Delta q_{bi}\Gamma_{bc}\right) - \frac{1}{6}S_{\mu\nu}\sum_c \Delta q_{ci}^2\left(\Gamma_{ca} + \Gamma_{cb}\right). \end{aligned} \tag{D.15}$$

The orbital energies can then be expressed as

$$\epsilon_i = \sum_{ab}\sum_{\mu\in a}\sum_{\nu\in b} c_{\mu i}c_{\nu i}H_{\mu\nu i}. \tag{D.16}$$

Bibliography

[1] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory;* MacMillan Publishing Co.: New York, 1982.

[2] Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory;* John Wiley & Sons, Inc.: Weinheim, 2nd ed.; 2002.

[3] TURBOMOLE V5-10 2008, a development of Universität Karlsruhe (TH) and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

[4] Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007,** *111,* 5678.

[5] Zerner, M. C. Semiempirical Molecular Orbital Methods. In *Reviews in Computational Chemistry, Vol 2*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: New York, 1991.

[6] Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002,** *23,* 1601.

[7] Stewart, J. J. P. *J. Mol. Model.* **2007,** *13,* 1173.

[8] Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993,** *14,* 775.

[9] Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000,** *103,* 495.

[10] Foulkes, W. M. C.; Haydock, R. *Phys. Rev. B* **1989,** *39,* 12520.

[11] Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. A* **2006,** *110,* 13551.

[12] Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2000,** *256,* 15.

[13] Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001,** *263,* 203.

[14] Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007,** *111,* 5751.

[15] Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995,** *51,* 12947.

[16] Hazebroucq, S.; Picard, G. S.; Adamo, C.; Heine, T.; Gemming, S.; Seifert, G. *J. Chem. Phys.* **2005,** *123,* 1.

[17] Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998,** *58,* 7260.

[18] Elstner, M. *Weiterentwicklung quantenmechanischer Rechenverfahren für organische Moleküle und Polymere,* Thesis, Universität-Gesamthochschule Paderborn, 1998.

[19] Density Functional Tight Binding: Contributions from the American Chemical Society Symposium, J. Phys. Chem. A, 2007, 111, 5607.

[20] Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011,** *7,* 931.

[21] Niehaus, T. A.; Elstner, M.; Frauenheim, T.; Suhai, S. *J. Mol. Struct.: THEOCHEM* **2001,** *541,* 185.

[22] Cai, Z.; Lopez, P.; Reimers, J. R.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007,** *111,* 5743.

[23] Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003,** *24,* 565.

[24] Moreira, N. H.; Dolgonos, G.; Aradi, B.; Da Rosa, A. L.; Frauenheim, T. *J. Chem. Theory Comput.* **2009,** *5,* 605.

[25] Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964,** *136,* 864–871.

[26] Born, M.; Oppenheimer, R. *Annalen der Physik* **1927,** *84,* 457–484.

[27] Thomas, L. H. *Math. Proc. Camb. Phil. Soc.* **1927,** *23,* 542.

[28] Fermi, E. *Rend. Accad. Naz. Lincei* **1927,** *6,* 602.

[29] Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules;* Oxford University Press: New York, 1989.

[30] Jensen, F. *Introduction to Computational Chemistry;* John Wiley & Sons, Inc.: Chichester, 1999.

[31] Kohn, W.; Sham, L. J. *Phys. Rev.* **1965,** *140,* A1133.

[32] Dirac, P. A. M. *Math. Proc. Camb. Phil. Soc.* **1930,** *26,* 376.

[33] Ceperley, D. M.; Alder, B. J. *Phys. Rev. Lett.* **1980,** *45,* 566.

[34] Vosko, S. H.; Wilk, L.; Nusair, M. *Canadian J. Phys.* **1980,** *58,* 1200.

[35] Perdew, J. P.; McMullen, E. R.; Zunger, A. *Phys. Rev. A* **1981,** *23,* 2785.

[36] Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992,** *45,* 13244.

[37] Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996,** *77,* 3865.

[38] Perdew, J. P.; Schmidt, K. *AIP Conf. Proc.* **2001,** *577,* 1.

[39] Becke, A. D. *Phys. Rev. A* **1988,** *38,* 3098.

[40] Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988,** *37,* 785.

[41] Mattsson, A. E. *Science* **2002,** *298,* 759.

[42] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997,** *106,* 1063.

[43] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000,** *112,* 7374.

[44] Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005,** *123,* 1.

[45] Cramer, C. J. *Essentials of Computational Chemistry – Theories and Models;* John Wiley & Sons, Inc.: Chichester, 2nd ed.; 2004.

[46] Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A. *J. Chem. Theory Comput.* **2010,** *6,* 3688.

[47] Riley, K. E.; Op't Holt, B. T.; Merz Jr., K. M. *J. Chem. Theory Comput.* **2007,** *3,* 407.

[48] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Science* **2008,** *321,* 792.

[49] Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz Jr., J. L. *Phys. Rev. Lett.* **1982,** *49,* 1691.

[50] Zhang, Y.; Yang, W. *Theor. Chem. Acc.* **2000,** *103,* 346.

[51] Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *Phys. Rev. Lett.* **2008,** *100,* 146401.

[52] Janak, J. F. *Phys. Rev. B* **1978,** *18,* 7165.

[53] Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981,** *23,* 5048.

[54] Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2004,** *121,* 8187.

[55] Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2005,** *122,* 1.

[56] Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I. *J. Chem. Phys.* **2006,** *124,* 094108.

[57] Mantz, Y. A.; Gervasio, F. L.; Laino, T.; Parrinello, M. *J. Phys. Chem. A* **2007,** *111,* 105.

[58] Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005,** *123,* 1.

[59] Perdew, J. P.; Ruzsinszky, A.; Constantin, L. A.; Sun, J.; Csonka, G. I. *J. Chem. Theory Comput.* **2009,** *5,* 902.

[60] Baerends, E. J. *Theor. Chem. Acc.* **2000,** *103,* 265.

[61] Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *Phys. Rev. Lett.* **2009,** *102,* 066403.

[62] Perdew, J. P.; Ruzsinszky, A. *Int. J. Quantum Chem.* **2010,** *110,* 2801.

[63] Grimme, S. *J. Comput. Chem.* **2006,** *27,* 1787.

[64] Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001,** *114,* 5149.

[65] Grimme, S. *J. Comput. Chem.* **2004,** *25,* 1463.

[66] Slater, J. C.; Koster, G. F. *Phys. Rev.* **1954,** *94,* 1498.

[67] Elstner, M. *J. Phys. Chem. A* **2007,** *111,* 5614.

[68] Seifert, G. *J. Phys. Chem. A* **2007,** *111,* 5609.

[69] Eschrig, H.; Bergert, I. *Phys. Stat. Sol. B* **1978,** *90,* 621.

[70] Eschrig, H. *Optimized LCAO Method and Electronic Structure of Extended Systems;* Springer: Berlin, 1989.

[71] Seifert, G.; Eschrig, H. *Phys. Stat. Sol. B* **1985,** *127,* 573.

[72] Elstner, M. *Theor. Chem. Acc.* **2006,** *116,* 316.

[73] Elsner, J. *Surfaces and extended defects in Wurtzite GaN,* Thesis, Universität-Gesamthochschule Paderborn, 1998.

[74] Köhler, C.; Seifert, G.; Gerstmann, U.; Elstner, M.; Overhof, H.; Frauenheim, T. *Phys. Chem. Chem. Phys.* **2001,** *3,* 5109.

[75] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. *J. Phys. Chem.* **1994,** *98,* 11623.

[76] Hariharan, P. C.; Pople, J. A. *Theor. Chem. Acc.* **1973,** *28,* 213.

[77] Gaus, M. "Weiterentwicklung und Parametrisierung einer genäherten Dichtefunktionaltheorie unter Verwendung moderner Optimierungsmethoden", Diploma thesis, Technische Universität Braunschweig, 2007.

[78] Winget, P.; Selçuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. *Theor. Chem. Acc.* **2003,** *110,* 254.

[79] Bondar, A.-N.; Fischer, S.; Smith, J. C.; Elstner, M.; Suhai, S. *J. Am. Chem. Soc.* **2004,** *126,* 14668.

[80] Wanko, M.; Hoffmann, M.; Frauenheim, T.; Elstner, M. *J. Comput.-Aided Mol. Des.* **2006,** *20,* 511.

[81] Wanko, M.; Hoffmann, M.; Strodel, P.; Koslowski, A.; Thiel, W.; Neese, F.; Frauenheim, T.; Elstner, M. *J. Phys. Chem. B* **2005,** *109,* 3606.

[82] Liu, H.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Hermans, J.; Yang, W. *Proteins: Structure, Function and Genetics* **2001,** *44,* 484.

[83] Krüger, T.; Elstner, M.; Schiffels, P.; Frauenheim, T. *J. Chem. Phys.* **2005,** *122,* 1.

[84] Witek, H. A.; Morokuma, K. *J. Comput. Chem.* **2004,** *25,* 1858.

[85] Witek, H. A.; Morokuma, K.; Stradomska, A. *J. Chem. Phys.* **2004,** *121,* 5171.

[86] Witek, H. A.; Morokuma, K.; Stradomska, A. *J. Theor. Comput. Chem.* **2005,** *4,* 639.

[87] Witek, H. A.; Irle, S.; Zheng, G.; De Jong, W. A.; Morokuma, K. *J. Chem. Phys.* **2006,** *125,* 214706.

[88] Małolepsza, E.; Witek, H. A.; Morokuma, K. *Chem. Phys. Lett.* **2005,** *412,* 237.

[89] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985,** *107,* 3902.

[90] Stewart, J. J. P. *J. Comput. Chem.* **1989,** *10,* 209.

[91] Seabra, G. D. M.; Walker, R. C.; Roitberg, A. E. *J. Phys. Chem. A* **2009,** *113,* 11938.

[92] Möhle, K.; Hofmann, H.; Thiel, W. *J. Comput. Chem.* **2001,** *22,* 509.

[93] Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007,** *111,* 10861.

[94] Yang, Y.; Yu, H.; York, D.; Elstner, M.; Cui, Q. *J. Chem. Theory Comput.* **2008,** *4,* 2067.

[95] Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006,** *110,* 6458.

[96] Riccardi, D.; Yang, S.; Cui, Q. *Biochim. Biophys. Acta* **2010,** *1804,* 342.

[97] Yang, Y.; Cui, Q. *J. Phys. Chem. A* **2009,** *113,* 12439.

[98] Phatak, P.; Frähmcke, J. S.; Wanko, M.; Hoffmann, M.; Strodel, P.; Smith, J. C.; Suhai, S.; Bondar, A.-N.; Elstner, M. *J. Am. Chem. Soc.* **2009,** *131,* 7064.

[99] Phatak, P.; Ghosh, N.; Yu, H.; Cui, Q.; Elstner, M. *Proc. Natl. Acad. Sci. USA* **2008,** *105,* 19672.

[100] Yang, Y.; Yu, H.; Cui, Q. *J. Mol. Biol.* **2008,** *381,* 1407.

[101] Riccardi, D.; König, P.; Guo, H.; Cui, Q. *Biochemistry* **2008,** *47,* 2369.

[102] Yang, Y.; Cui, Q. *J. Phys. Chem. B* **2009,** *113,* 4930.

[103] Politzer, P.; Murray, J. S.; Lane, P. *J. Comput. Chem.* **2003,** *24,* 505.

[104] Klopper, W.; Van Duijneveldt-van De Rijdt, J. G. C. M.; Van Duijneveldt, F. B. *Phys. Chem. Chem. Phys.* **2000,** *2,* 2227.

[105] Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999,** *110,* 7650.

[106] Zhou, H.; Tajkhorshid, E.; Frauenheim, T.; Suhai, S.; Elstner, M. *Chem. Phys.* **2002,** *277,* 91.

[107] Ghosh, N.; Xavier, P.-R.; Gunner, M. R.; Cui, Q. *Biochemistry* **2009,** *48,* 2468.

[108] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. "Gaussian 03, Revision C.02", Gaussian, Inc., Wallingford, CT, 2004.

[109] Dunning Jr., T. H. *J. Chem. Phys.* **1989,** *90,* 1007.

[110] Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970,** *19,* 553.

[111] Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996,** *105,* 11024.

[112] Ireta, J.; Neugebauer, J.; Scheffler, M. *J. Phys. Chem. A* **2004,** *108,* 5692.

[113] Zhao, Y.; Truhlar, D. *J. Chem. Theory Comput.* **2005,** *1,* 415.

[114] Santra, B.; Michaelides, A.; Scheffler, M. *J. Chem. Phys.* **2007,** *127,* 184104.

[115] Rao, L.; Ke, H.; Fu, G.; Xu, X.; Yan, Y. *J. Chem. Theory Comput.* **2009,** *5,* 86.

[116] Baer, M.; Mathias, G.; Kuo, I.-W.; Tobias, D. J.; Mundy, C. J.; Marx, D. *ChemPhysChem* **2008,** *9,* 2703.

[117] Range, K.; Riccardi, D.; Cui, Q.; Elstner, M.; York, D. M. *Phys. Chem. Chem. Phys.* **2005,** *7,* 3070.

[118] Lee, Y.-S.; Krauss, M. *J. Am. Chem. Soc.* **2004,** *126,* 2225.

[119] Møller, C.; Plesset, M. S. *Phys. Rev.* **1934,** *46,* 618.

[120] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998,** *109,* 7764.

[121] Gaus, M.; Chou, C.-P.; Witek, H.; Elstner, M. *J. Phys. Chem. A* **2009,** *113,* 11866.

[122] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes – The Art of Scientific Computing;* Cambridge University Press: New York, 3rd ed.; 2007.

[123] NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101 Release 14, Sept 2006, Editor: Russell D. Johnson III http://srdata.nist.gov/cccbdb.

[124] Kuchitsu, K. *Structure of Free Polyatomic Molecules – Basic Data;* Springer: Berlin, 1998.

[125] Knaup, J. M.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007,** *111,* 5637.

[126] Witek, H. A.; Irle, S.; Morokuma, K. *J. Chem. Phys.* **2004,** *121,* 5163.

[127] Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2008,** *4,* 297.

[128] Stewart, J. J. P. *J. Mol. Model.* **2004,** *10,* 6.

[129] Csonka, G. I.; Ruzsinszky, A.; Tao, J.; Perdew, J. P. *J. Phys. Chem.* **2005,** *101,* 505.

[130] Brothers, E. N.; Scuseria, G. E. *J. Chem. Theory Comput.* **2006,** *2,* 1045.

[131] Sattelmeyer, K. W.; Tubert-Brohman, I.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006,** *2,* 413.

[132] Miaskiewicz, K.; Smith, D. A. *Chem. Phys. Lett.* **1997,** *270,* 376.

[133] Redfern, P. C.; Zapol, P.; Curtiss, L. A.; Raghavachari, K. *J. Phys. Chem. A* **2000,** *104,* 5850.

[134] Goez, A. "Anpassen neuer Repulsivpotentiale im Rahmen einer genäherten Dichtefunktionaltheorie", Master's thesis, Technische Universität Braunschweig, *in preparation.*

[135] Tam, H. S.; Choe, J.-I.; Harmony, M. D. *J. Phys. Chem.* **1991,** *95,* 9267.

[136] Hourahine, B.; Sanna, S.; Aradi, B.; Köhler, C.; Niehaus, T.; Frauenheim, T. *J. Phys. Chem. A* **2007,** *111,* 5671 5671–5677.

[137] Kubař, T.; Elstner, M. *J. Phys. Chem. B* **2010,** *114,* 11221.

[138] Rapacioli, M.; Spiegelman, F.; Scemama, A.; Mirtschink, A. *J. Chem. Theory Comput.* **2011,** *7,* 44.

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die zum Gelingen meiner Doktorarbeit beigetragen haben.

Herrn Prof. Dr. Marcus Elstner danke ich für die ausgezeichnete Betreuung und die Schaffung exzellenter Arbeitsbedingungen in Braunschweig und Karlsruhe.

I would like to thank Prof. Dr. Henryk Witek for his support, inspiring discussions, and the very successful collaboration during my visit in Taiwan.

Special thanks to Prof. Dr. Qiang Cui for fruitful discussions on DFTB3.

I thank Dr. Bálint Aradi and Guanhua Hou for implementing DFTB3 into DFTB+ and CHARMM and valuable comments, Dr. Yang Yang and Puja Goyal for providing compilations of molecule sets and helpful discussions.

Thanks to Chien-Pin Chou for sharing his great programming skills.

Vielen Dank an Albrecht Goez für die produktive Zusammenarbeit während unseres Parametrisierungprojektes.

Der gesamten Arbeitsgruppe danke ich für die vielen hilfreichen Diskussionen und die angenehme Arbeitsatmosphäre.

Mein besonderer Dank gilt Benjamin Woiczikowski für das unkomplizierte Zuhause auf Zeit in Karlsruhe.

Den Korrekturlesern Jan Frähmcke, Albrecht Goez, Susanne Haeder, Katharina und Günter Gaus herzlichen Dank.

Meinen Eltern danke ich ganz herzlich für die stete Unterstützung während des Studiums und der Promotion.

Bei den Referenten bedanke ich mich für die Übernahme der Gutachten.

Lebenslauf

Michael Gaus, deutsch, ledig

30.10.1983	geboren in Worms
1989–1993	Grundschule Biblis-Nordheim
1994–2002	Altes Kurfürstliches Gymnasium Bensheim
1999–2000	Salisbury High School, Salisbury, NC, USA
04.06.2002	Abitur
2002–2003	Wehrdienst
2003–2007	Studium der Chemie, Technische Universität Braunschweig
2005–2006	Auslandsstudium in der Chemie, Universidad de Zaragoza, Spanien
01.10.2007	Hochschulabschluss zum Diplom-Chemiker, Technische Universität Braunschweig
2007–2010	Wissenschaftlicher Mitarbeiter, Technische Universität Braunschweig
02–05.2008	Wissenschaftlicher Aufenthalt an der National Chiao Tung University, Hsinchu, Taiwan
2010–2011	Wissenschaftlicher Mitarbeiter, Karlsruher Institut für Technologie

Karlsruhe, 31. Mai 2011

Veröffentlichungen

1. Gaus, M.; Chou, C.-P.; Witek, H.; Elstner, M. Automatized Parametrization of SCC-DFTB Repulsive Potentials: Application to Hydrocarbons *J. Phys. Chem. A* **2009**, *113*, 11866–11881.

2. Kaminski, S.; Gaus, M.; Phatak, P.; Von Stetten, D.; Elstner, M.; Mroginski, M. A. Vibrational Raman Spectra from the Self-Consistent Charge Density Functional Tight Binding Method via Classical Time-Correlation Functions *J. Chem. Theory Comput.* **2010**, *6*, 1240–1255.

3. Elstner, M.; Gaus, M. The Self-Consistent-Charge Density-Functional Tight-Binding (SCC-DFTB) Method: An Efficient Approximation to Density Functional Theory. In *Computational Methods for Large Systems*; Reimers, J. R.; John Wiley & Sons Inc., **2011**; Chapter 9; ISBN: 978-0-470-48788-4.

4. Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB) *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

5. Hujo, W.; Gaus, M.; Schultze, M.; Kubař, T.; Grunenberg, J.; Elstner, M.; Bauerecker, S. Effect of Nitrogen Adsorption on the Mid-Infrared Spectrum of Water Clusters *J. Phys. Chem. A* **2011**, *115*, 6218–6225.

6. Goyal, P.; Ghosh, N.; Phatak, P.; Clemens, M.; Gaus, M.; Elstner, M.; Cui, Q. Proton Storage Site in Bacteriorhodopsin: New Insights from QM/MM Simulations of Infrared Spectra and Microscopic pKa, *in revision.*

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel und Literatur angefertigt habe. Die Satzung der Universität Karlsruhe (TH) zur Sicherung guter wissenschaftlicher Praxis habe ich beachtet.

Karlsruhe, 31. Mai 2011