Expanding the Pareto Frontier of Electronic Structure Methods with Advanced Basis Sets

Marcel Müller

Dissertation

Expanding the Pareto Frontier of Electronic Structure Methods with Advanced Basis Sets

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Marcel Müller

aus
Waldbröl

Bonn 2025

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter/Betreuer:	Prof. Dr. Stefan Grimme
Gutachter:	Prof. Dr. Thomas Bredow
Tag der Promotion:	2. Juli 2025
Erscheinungsjahr:	2025

Publications

Parts of this thesis have been published in peer-reviewed journals.

1. M. Müller, A. Hansen, and S. Grimme, *ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set*, The Journal of Chemical Physics **158** *1* (2023) 014103, DOI: 10.1063/5.0133026.

2. S. Grimme, M. Müller, and A. Hansen, *A non-self-consistent tight-binding electronic structure potential in a polarized double-ζ basis set for all spd-block elements up to Z = 86*, The Journal of Chemical Physics **158** *12* (2023) 124111, DOI: 10.1063/5.0137838

3. M. Müller, A. Hansen, and S. Grimme, *An atom-in-molecule adaptive polarized valence single-ζ atomic orbital basis for electronic structure calculations*, The Journal of Chemical Physics **159** *16* (2023) 164108, DOI: 10.1063/5.0172373

4. M. Müller, T. Froitzheim, A. Hansen, and S. Grimme, *Advanced Charge Extended Hückel (CEH) Model and a Consistent Adaptive Minimal Basis Set for the Elements Z = 1–103*, The Journal of Physical Chemistry A **128** *49* (2024), Publisher: American Chemical Society 10723, DOI: 10.1021/acs.jpca.4c06989

For the following articles significant contributions have been made.

5. K. Škoch, C. G. Daniliuc, G. Kehr, S. Ehlert, M. Müller, S. Grimme, and G. Erker, *Frustrated Lewis-Pair Neighbors at the Xanthene Framework: Epimerization at Phosphorus and Cooperative Formation of Macrocyclic Adduct Structures*, Chemistry - A European Journal **27** *47* (2021), Publisher: John Wiley & Sons, Ltd 12104, DOI: 10.1002/chem.202100835

6. N. Fleck, R. M. Thomas, M. Müller, S. Grimme, and B. H. Lipshutz, *An environmentally responsible route to tezacaftor, a drug for treatment of cystic fibrosis prepared in water via ppm Au catalysis as entry to 2-substituted indoles*, Green Chemistry **24** *17* (2022), Publisher: The Royal Society of Chemistry 6517, DOI: 10.1039/D2GC01828D

7. K. Škoch, C. G. Daniliuc, M. Müller, S. Grimme, G. Kehr, and G. Erker, *Stereochemical Behavior of Pairs of P-stereogenic Phosphanyl Groups at the Dimethylxanthene Backbone*, Chemistry - A European Journal **28** *20* (2022), Publisher: John Wiley & Sons, Ltd e202200248, DOI: 10.1002/chem.202200248

8. P. Pracht, S. Grimme, C. Bannwarth, F. Bohle, S. Ehlert, G. Feldmann, J. Gorges, M. Müller, T. Neudecker, C. Plett, S. Spicher, P. Steinbach, P. A. Wesołowski, and F. Zeller, *CREST – A program for the exploration of low-energy molecular chemical space*, The Journal of Chemical Physics **160** *11* (2024) 114110, DOI: 10.1063/5.0197592

The following scientific contributions have been presented at conferences, workshops, and summer schools:

1. 60th Symposium on Theoretical Chemistry, Braunschweig, **2024**: Poster on the Charge Extended Hückel (CEH) model and a consistent adaptive minimal basis set (q-vSZP).

2. Chemical Compound Space Conference, Heidelberg, **2024**: Poster on BayBE – A Bayesian Back End for Design of Experiments.

3. Leopoldina Symposium on Molecular Machine Learning "#LeopoldinAIchem", Leopoldina, Halle, **2023**: Poster on the new dxtb and PTB methods and applications in Bayesian optimization and molecular machine learning.

4. Meeting of the "SPP2363: Molecular Machine Learning", Jena, **2023**: Poster on the new PTB method and potential applications in molecular machine learning.

5. 17th International Congress of Quantum Chemistry, Bratislava, **2023**: Poster on the new PTB-RPBE-D4 composite electronic structure method.

6. CECAM Psi-k Research Conference 2023, Berlin, **2023**: Poster on the new PTB method and potential applications in molecular machine learning.

7. Bunsen-Tagung 2023, Berlin, **2023**: Talk on efficient composite DFT approaches for large molecular systems and screening applications.

8. Kick-off Meeting of the "SPP2363: Molecular Machine Learning", Leopoldina, Halle, **2022**: Poster on the new PTB method and potential applications in molecular machine learning.

9. European Summerschool in Quantum Chemistry, Palermo, **2022**: Poster on the new composite range-separated hybrid DFT method "ωB97X-3c".

10. 19th International Conference on Density Functional Theory and its Applications, Brussels, **2022**: Poster on the new composite range-separated hybrid DFT method "ωB97X-3c".

11. 12th Triennial Congress of the World Association of Theoretical and Computational Chemists (WATOC), Vancouver, **2022**: Talk on simulation of intensities in vibrational spectroscopy by (semi-)empirical methods.

12. 57th Symposium on Theoretical Chemistry, Würzburg (virtual), **2021**: Poster on improved simulation of IR spectra at the force field level.

Abstract

PAUL Dirac famously stated in 1929:[9]

> *The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known* […]. *It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.*

Nearly a century later, Dirac's perspective remains as relevant as ever. The rational design of new drugs and advanced materials is increasingly driven by high-throughput virtual screenings and machine learning techniques, fueled by vast molecular databases. This accelerating demand for physically grounded data necessitates electronic structure methods that are not only accurate but also computationally efficient. Yet, the complexity of the electronic Schrödinger equation makes exact solutions impractical for most systems, forcing electronic structure theory to rely on a hierarchy of approximations that balance cost and accuracy. Two approaches stand out in this trade-off: density functional theory (DFT), the workhorse of computational chemistry for the past 30 years, and semiempirical quantum mechanical (SQM) methods. DFT reformulates the electronic structure problem in terms of the electron density, deriving interactions from first principles, whereas SQM methods replace some of these interactions with semiempirical expressions to reduce the computational complexity. While SQM methods achieve speed-ups of three to four orders of magnitude, their accuracy remains limited in many applications. A key factor in balancing computational efficiency and accuracy in both DFT and SQM methods is the Gaussian atomic orbital (AO) basis set, which dictates how the electronic wavefunction is represented. This thesis explores several approaches to improve this crucial factor, aiming to advance efficient electronic structure methods to meet future demands.

The first two chapters introduce common approaches in electronic structure theory and different ways of approximating those. Chapter 3 deals with an approach to lower the computational cost of DFT by developing a small basis set composite method. It combines a newly developed polarized valence double-ζ basis set, vDZP, with a state-of-the-art range-separated hybrid density functional. Unlike standard basis sets optimized for atoms, vDZP is variationally optimized for molecular environments and features a deep contraction to ensure well-suited AOs, particularly relevant for SQM applications. Together with large-core effective potentials and the D4 dispersion correction, the resulting ωB97X-3c method extends the "3c" composite DFT series, outperforming conventional triple-ζ approaches in cost-to-accuracy ratio. It is particularly well-suited for large chemical systems, including biomolecules and supramolecular complexes, where non-local Fock exchange plays a crucial role.

The vDZP basis set is employed in Chapter 4 to develop the so-called density matrix tight-binding (PTB) method, a new SQM ansatz incorporating two key advancements: *(i)* a Hamiltonian expanded in a double-ζ

basis set and *(ii)* a parameterization scheme aimed at reproducing the density matrx (i.e., the molecular orbitals) of a converged ωB97X-3c DFT calculation. PTB predicts infrared and Raman spectral intensities with near-DFT accuracy and reliably computes atomic charges and bond orders with minimal deviations from the reference DFT method. As with all projects presented in this thesis, consistent accuracy over different chemical spaces and across the periodic table is a core focus of PTB.

However, the development of a robust energy expression within this framework proved infeasible. To overcome this limitation, Chapter 5 introduces q-vSZP, a new environment-adaptive single-ζ basis set. It dynamically contracts or expands based on atomic charge and the number of coordinating atoms, capturing effects that typically require multiple-ζ basis sets. q-vSZP includes all features of vDZP, such as the molecular optimization and deep contration of the AOs. While q-vSZP is formally a single-ζ basis set, it can keep up with conventional double-ζ basis sets in standard DFT thermochemistry benchmarks.

The atomic charges required for q-vSZP are computed using the Charge Extended Hückel (CEH) method, which is further refined and extended up to actinide elements ($Z < 103$) in Chapter 6. By designing an Extended Hückel Theory approach specifically to reproduce atomic charges from a consistent reference DFT calculation, CEH achieves accuracy comparable to more advanced SQM methods while being 10–20 times faster due to the absence of self-consistent field iterations and the employed minimal basis set.

This work systematically examines the challenges of efficient electronic structure methods and proposes, develops, and tests solutions that expand the *Pareto front* of methods with optimal cost-to-accuracy balances, as summed up in Chapter 7. All advancements presented in this thesis are implemented in accessible computational frameworks, either as open-source software or integrated into established closed-source platforms. The q-vSZP basis set together with the findings from the PTB and CEH development represent a cornerstone for a next-generation SQM method, addressing the limitations outlined above to the fullest extent imaginable today.

Contents

CHAPTER 1

From Observation to Simulation

Science, in its most fundamental form, seeks to describe, understand, and predict the natural world. The methodology of natural sciences is rooted in observation, experimentation, and theorization: observations inspire hypotheses, experiments test these hypotheses, and theories are formulated to explain and generalize the observed phenomena. This cycle of empirical investigation and theoretical refinement has driven scientific progress for centuries, providing insights into the fundamental principles governing nature. In chemistry, empirical observations initially dictated how substances interact, react, and transform, with theories emerging only retrospectively to explain experimental findings. While this approach led to remarkable progress, it was inherently limited in its ability to reveal the underlying causes of chemical behavior at the atomic level.

Alchemy, though based on mystical notions, laid the groundwork for systematic chemical experimentation. The discovery of mass conservation in the 18$^{\text{th}}$ century[10] enabled the first structured understanding of chemical elements. The idea that matter consists of fundamental building blocks, atoms, was formalized in the early 19$^{\text{th}}$ century by John Dalton,[11] marking a shift from qualitative to quantitative chemistry. Yet, the precise arrangement of atoms in molecules remained elusive. This changed with Kekulé's valency theory and structural formulas in the mid-19$^{\text{th}}$ century,[12–14] which provided a systematic framework for chemical bonding. Around the same time, Mendeleev's periodic table classified elements by atomic weight and recurring properties, providing a more structured theoretical foundation of chemistry.[15] Still, atomic theory remained speculative, and experimental chemistry continued to drive discovery.

A fundamental shift occurred at the turn of the 20$^{\text{th}}$ century with the realization that atoms themselves were composed of subatomic particles. J. J. Thomson's discovery of the electron[16] laid the foundation for modern chemistry by demonstrating that atoms were not indivisible, and further experimental breakthroughs revealed that electrons play a central role in determining chemical properties. Physicists, who later became widely renowned, developed the concepts of energy quantization, discrete electron orbitals, and the wave-particle duality, which culminated in the formulation of quantum mechanics, fundamentally altering the way chemists and physicists described molecular systems.[17–21] At the heart of this new framework stood Schrödinger's equation,[22] which describes how the energy E of a quantum system is related to its wavefunction Ψ and the Hamiltonian operator $\hat{H}$:[†]

$$\hat{H}\Psi = E\Psi. \tag{1.1}$$

Quantum mechanics provided a fundamental mathematical framework for understanding the behavior of electrons in atoms and molecules, enabling theoretical models to predict and complement experimental

[†] Shown here in its non-relativistic, time-independent formulation.

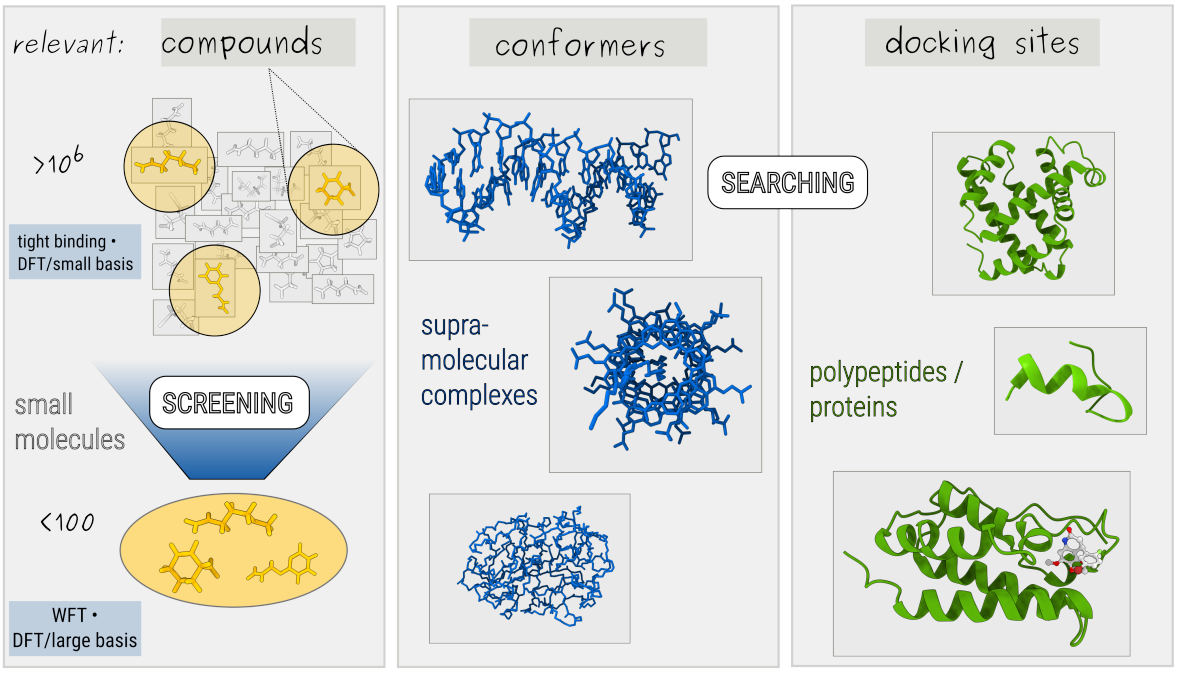

Figure 1.1: Molecular screening applications. *Left:* High-troughput screening (HTS) of small molecule compounds toward a small number of reasonable candidates. *Middle:* Conformer ensembles of supramolecular complexes. *Right:* Docking-site screening in polypeptides or proteins.

findings.[23] The emergence of modern digital computers in the mid-20$^{\text{th}}$ century transformed theoretical chemistry from a conceptual discipline into a practical and predictive field, now known as computational chemistry.[24,25]

The primary quantity derived from Eq. (1.1) is the total electronic energy – that is, the energy of assembling the molecule from separated nuclei and electrons. Relative energy differences provide insights into reaction energies, barrier heights, and conformational stability. Minimizing the energy as a function of the atomic coordinates yields optimized geometries, while incorporating temperature and propagating the atomic coordinates in time according to Newton's equations of motion enables molecular dynamics simulations, revealing time-dependent and thermostatistical properties. Higher-order derivatives of the energy provide access to further properties, such as infrared (IR) and Raman spectra or nuclear magnetic resonance shielding constants.[26] Today, these and other computational chemistry concepts are used to not only complement experiments but also to drive discoveries in materials science,[27–31] drug design,[32–34] and reaction mechanisms.[35–38] In recent years, the scope of computational chemistry has expanded beyond case studies of selected small molecules in specific reactions to include large-scale screenings across diverse molecule datasets. This encompasses screening of thousands or millions of small molecule compounds to find the right candidate, exploring the conformational space of large supramolecular complexes, or searching for docking sites in biomolecules, as illustrated schematically in Fig. 1.1.

If we do not want to revert to Dalton's purely atomistic picture, as implemented in common force fields (FFs) and novel machine learning potentials (MLPs), electronic structure theory must evolve to meet the new challenges – this is the goal of this thesis. To that end, the following sections of this first Chapter 1 outlines the historical development and contemporary state of electronic structure methods, i.e., systematic

algorithms for solving Eq. (1.1). Chapter 2 discusses the motivation for more approximate solutions to Schrödinger's equation and introduces the concept of semiempirical quantum mechanical (SQM) methods. Building on the theoretical foundation of this chapter, we trace their evolution from early formulations to state-of-the-art methods widely used in modern computational chemistry. The final section of the chapter reflects on the progress of semiempirical methods over the past century, identifying key limitations and potential improvements. The subsequent chapters present novel approaches to approximate electronic structure theory, from a less empirical approach in Chapter 3 to an advanced SQM model in Chapter 4. Chapters 5 and 6 lay the groundwork for a refined semiempirical solution that integrates many of these improvements and remains under development at the time of writing. Finally, Chapter 7 summarizes the contributions of this thesis and provides an outlook on future developments in the field.

1.1 Electronic Structure Theory

The hydrogen atom is a unique case where Eq. (1.1) can be solved analytically. Here, the Hamiltonian $\hat{H}$ consists only of the kinetic energy of the electron, $\hat{T}_\mathrm{e}$, and the Coulomb potential energy between the negatively charged electron and the positively charged nucleus, $\hat{V}_\mathrm{ne}$:

$$\hat{H} = \hat{T}_\mathrm{e} + \hat{V}_\mathrm{ne}. \tag{1.2}$$

For a multi-electron system with more than one nucleus, the electronic Hamiltonian generalizes to:

$$\hat{H}_\mathrm{e} = \hat{T}_\mathrm{e} + \hat{V}_\mathrm{ne} + \hat{V}_\mathrm{ee}. \tag{1.3}$$

The individual terms are defined in atomic units as follows:[†]

$$\hat{T}_\mathrm{e} = -\sum_{i}^{N_\mathrm{elec}} \frac{1}{2}\nabla_i^2, \tag{1.4a}$$

$$\hat{V}_\mathrm{ne} = -\sum_{i}^{N_\mathrm{elec}} \sum_{A}^{M_\mathrm{atoms}} \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|}, \tag{1.4b}$$

$$\hat{V}_\mathrm{ee} = \sum_{i}^{N_\mathrm{elec}} \sum_{j>i}^{N_\mathrm{elec}} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{1.4c}$$

To arrive at the complete Hamiltonian, the nucleus-nucleus potential $\hat{V}_\mathrm{nn}$ has to be added to $\hat{H}_\mathrm{e}$ in Eq. (1.3). Within the Born-Oppenheimer approximation, $\hat{V}_\mathrm{nn}$ reduces to a constant energy for a specific arrangement of atoms.[39] The most challenging term in Eq. (1.3) is the electron-electron repulsion $\hat{V}_\mathrm{ee}$, which complicates the extension to multi-electron systems beyond the hydrogen atom. Since each electron's position depends on all others, the wavefunction $\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{N\mathrm{elec}})$ in Eq. (1.1) cannot be factorized into a product of single-electron wavefunctions without introducing approximations, preventing an analytic solution. Various approximations have been developed to make the equation tractable, as outlined in the following sections. Among the practical approaches are mean-field methods, with Hartree-Fock (HF) and density functional theory (DFT) being the two most significant over the past century. Despite their fundamentally different theoretical foundations, their practical implementations share clear similarities.

[†] Here, ∇_i^2 is the Laplacian acting on the i-th electron, Z_A is the nuclear charge of nucleus A, and $\mathbf{r}_i$ and $\mathbf{R}_A$ denote the position vectors of electron i and nucleus A, respectively, in a system of N_elec electrons and M_atoms nuclei.

1.1.1 Hartree-Fock Theory

Hartree approximated the many-electron wavefunction, which depends on all electronic coordinates $\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{N_{\text{elec}}})$, as a product of single-electron wavefunctions $\phi_i(\mathbf{r}_i)$:[40]

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{N_{\text{elec}}}) \approx \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)\cdots\phi_N(\mathbf{r}_{N_{\text{elec}}}). \tag{1.5}$$

This ansatz implies the mean-field approximation, in which each electron moves in the Coulomb potential asserted by the average electron distribution.[41] However, this formulation did not account for the Pauli exclusion principle, introduced only a few years earlier, which requires fermionic wavefunctions to be antisymmetric under particle exchange.[21] Fock resolved this issue by incorporating antisymmetrized wavefunctions using Slater determinants,[42] ensuring the correct permutation behavior for electrons as fermions:[43]

$$\Psi \approx \Phi_{\text{SD}}(1, 2, \cdots, N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) & \cdots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \cdots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(N) & \phi_2(N) & \cdots & \phi_N(N) \end{vmatrix}. \tag{1.6}$$

The energy expectation value of a normalized, approximate wavefunction Ψ is given by the electronic Hamiltonian $\hat{H}_{\text{e}}$:

$$\langle\Psi|\hat{H}_{\text{e}}|\Psi\rangle = E\langle\Psi|\Psi\rangle = E. \tag{1.7}$$

Inserting the Hamiltonian operator from Eq. (1.3) and the antisymmetrized wavefunction Φ_{SD} into the expectation value of the energy, leads to the following HF energy expression:[†]

$$E_{\text{HF}} = \sum_i^N \epsilon_i - \frac{1}{2}\sum_i^N\sum_j^N (J_{ij} - K_{ij}) + \underbrace{E_{\text{nn}}}_{\text{constant}}, \tag{1.8}$$

where ϵ_i represents the contribution of the one-electron operators $\hat{T}_{\text{e}}$ and $\hat{V}_{\text{ne}}$. While both J and K originate from the electron-electron interaction $\hat{V}_{\text{ee}}$, J_{ij} represents the repulsive Coulomb interaction between electrons i and j. K_{ij} enforces the antisymmetrization of the wavefunction by compensation for part of the Coulomb repulsion leading to an attractive exchange interaction between electrons of the same spin. In Eq. (1.8), the sum over J_{ij} includes the unphysical case where an electron interacts with itself. Although HF is not an exact theory, it ensures exact cancellation of this self-interaction through the exchange term K_{ij}. Incomplete cancellation of self-interaction, as in DFT, where exchange is only approximated (*vide infra*), leads to self-interaction errors (SIEs).

As an approximate mean field method, HF obeys the variational principle, which ensures that the energy of an admissable trial wavefunction remains bounded from below by the exact energy.[44,45] The goal is to determine a set of single-electron wavefunctions [also called molecular orbitals (MOs)], ϕ_i, that minimize the HF energy E_{HF} under the constraint of orbital orthonormality:

$$\langle\phi_i|\phi_j\rangle = \delta_{ij}. \tag{1.9}$$

To enforce this constraint, a Lagrangian functional is introduced:

$$\mathcal{L} = E_{\text{HF}} - \sum_{i,j} \lambda_{ij}\left(\langle\phi_i|\phi_j\rangle - \delta_{ij}\right), \tag{1.10}$$

where λ_{ij} are Lagrange multipliers. Minimizing $\mathscr{L}$ with respect to $|\phi_i\rangle$ leads to:

$$\frac{\delta \mathscr{L}}{\delta \langle \phi_i|} = \hat{F}|\phi_i\rangle - \sum_j \lambda_{ij}|\phi_j\rangle = 0. \tag{1.11}$$

Since λ_{ij} forms a Hermitian matrix, it can be diagonalized. By choosing a basis of MOs $\{|\phi_i\rangle\}$ that diagonalizes λ_{ij}, the equation simplifies into the eigenvalue form:

$$\hat{F}|\phi_i\rangle = \epsilon_i|\phi_i\rangle, \tag{1.12a}$$

where ϵ_i are the orbital energies, and the Fock operator $\hat{F}$ is given by:

$$\hat{F} = \hat{h} + \sum_j \left[\hat{J}_j - \hat{K}_j\right]. \tag{1.12b}$$

Here, $\hat{h}$ is the core Hamiltonian, $\hat{J}_j$ is the Coulomb operator, and $\hat{K}_j$ is the exchange operator, as defined in Eq. (1.13). The sum over $\hat{J}_j$ and $\hat{K}_j$ represents the effective mean-field interaction in the system.

$$J_i\,|\phi_j(2)\rangle = \langle \phi_i(1)\,| \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}\, |\,\phi_i(1)\rangle\,|\phi_j(2)\rangle, \tag{1.13a}$$

$$K_i\,|\phi_j(2)\rangle = \langle \phi_i(1)\,| \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}\, |\,\phi_j(1)\rangle\,|\phi_i(2)\rangle. \tag{1.13b}$$

In these expressions, the Coulomb operator J_i multiplies $\phi_j(2)$ by the classical electrostatic potential generated by ϕ_i, whereas the exchange operator K_i effectively replaces $\phi_j(2)$ with $\phi_i(2)$ under the same potential, reflecting the antisymmetry requirement. Starting from an initial guess for the orbitals, the Fock operator $\hat{F}$ is constructed based on the current MOs via $\hat{J}_j$ and $\hat{K}_j$, as highlighted in Eq. (1.13). The Fock operator is then diagonalized to obtain updated MOs and orbital energies. This iterative procedure continues until self-consistency is achieved, meaning that the orbitals used to construct the Fock matrix [left side of Eq. (1.12a)] are also its eigenfunctions. Convergence is typically determined by changes in energy or electron density. This iterative procedure is known as the self-consistent field (SCF) approach.

The mean-field approximation in HF inherently leads to a systematic overestimation of energies, as the interaction between electrons is only captured on an average basis, neglecting the instantaneous, dynamic response of one electron to the movement of another. The difference between the HF energy and the true ground-state energy is known as correlation energy. HF fails to account for dynamic electron–electron response (often referred to as "dynamic correlation"). Additionally, it neglects near-degeneracy (multi-reference) effects whenever multiple electronic configurations (i.e., Slater determinants) contribute significantly to the ground state – commonly termed "static correlation". Typical cases of static correlation are bond dissociation processes. Methods based on HF that recover the full or partial dynamic correlation energy include correlated wavefunction methods such as configuration interaction (CI)[46,47] and coupled-cluster (CC).[48–52] Both represent the correlated wavefunction as a combination of the HF reference and additional excited configurations. Alternatively, perturbative approaches such as Møller–Plesset perturbation theory treat electron correlation as a perturbation from the HF reference wavefunction.[53] Static correlation is addressed by methods that incorporate multiple Slater determinants, allowing for a proper description of mixed-configuration states. Examples include complete active space self-consistent field (CASSCF)[54–56] and multi-reference CI.[57,58] The aforementioned CI method in its full CI limit is capable of capturing both dynamic and static correlation effects missing in HF. For further details on the working principles of these

methods, the reader is referred to the cited literature.

1.1.2 Kohn-Sham Density Functional Theory

An alternative approach to solving the many-electron Schrödinger equation was introduced by Hohenberg and Kohn (HK) in the 1960s, establishing a direct link between the electron density and the total energy of a system.[59,60] The electron density $\rho(\mathbf{r})$ is formally defined in terms of the many-electron wavefunction as:

$$\rho(\mathbf{r}) = N \int |\Psi(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_{N_{\text{elec}}})|^2 \, d\mathbf{r}_2 \dots d\mathbf{r}_{N_{\text{elec}}}. \tag{1.14}$$

This expression gives the probability of finding any of the N electrons at position $\mathbf{r}$ and serves as the fundamental variable in DFT. HK proved that the ground-state energy is a functional of the electron density and that, up to an additive constant in the external potential, the ground-state density uniquely determines this energy functional. This allows electronic structure to be described entirely in terms of $\rho(\mathbf{r})$, without requiring explicit knowledge of the many-electron wavefunction Ψ. In contrast, earlier density-based theories by Thomas, Fermi, and Slater lacked general applicability and rigorous physical foundation for their application in molecular systems.[61–63] Although the HK theorem guarantees the existence of an exact energy functional $E[\rho]$ of the electron density, its explicit form remains unknown. This particularly applies to the kinetic energy functional $T[\rho]$, which lacks an analytically known expression in terms of the density alone.

A breakthrough came in 1965 with the work of Kohn and Sham, who introduced an auxiliary system of non-interacting electrons described by so-called Kohn-Sham (KS) orbitals, for which the kinetic energy could be calculated exactly.[64] Inspired by HF theory, a single Slater determinant represents this *non-interacting* system, whose ground-state electron density $\rho_0(\mathbf{r})$ is (by construction) equal to that of the real, *interacting* electron system:

$$\rho(\mathbf{r}) = \sum_i^{N_{\text{occ}}} |\phi_i(\mathbf{r})|^2 = \rho_0(\mathbf{r}), \tag{1.15}$$

assuming an N-electron system described by a set of single-electron wavefunctions (MOs) $\phi_i(\mathbf{r})$. The electronic energy can then be partitioned into known and unknown contributions:

$$E[\rho, \phi] = \underbrace{T_s[\phi] + J[\rho] + V_{\text{ne}}[\rho]}_{\text{known}} + \underbrace{E_{\text{xc}}[\rho]}_{\text{unknown}}. \tag{1.16}$$

Here, $T_s[\phi]$ (the index s represents the Slater determinant origin) is the kinetic energy of a system of *non-interacting* electrons, which can be computed exactly from the KS orbitals:

$$T_{\text{s}}[\phi] = \sum_i^{N_{\text{occ}}} \langle \phi_i | -\frac{1}{2}\nabla^2 | \phi_i \rangle, \tag{1.17}$$

The terms $J[\rho]$ and $V_{\text{ne}}[\rho]$ represent the classical Coulomb and electron-nucleus interaction, respectively, while the remaining exchange-correlation functional $E_{\text{xc}}[\rho]$ incorporates all unknown effects that cannot be described exactly. This framework of reintroducing orbitals to DFT is known today as KS-DFT.[†] In the following, this thesis always refers to KS-DFT when mentioning "DFT". The KS electronic energy reads

[†] The KS system is a mathematical tool to reproduce the exact ground-state density. KS orbitals and eigenvalues have no direct physical meaning, which is, however, widely ignored in practice.

explicitly:

$$E_{\mathrm{KS}}[\rho] = \sum_{i}^{N_{\mathrm{occ}}} \langle \phi_i | -\frac{1}{2}\nabla^2 + \frac{1}{2}\int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \hat{V}_{\mathrm{ne}} | \phi_i \rangle + E_{\mathrm{xc}}[\rho]. \tag{1.18}$$

All many-body effects of electron interaction are contained in the exchange-correlation functional $E_{\mathrm{xc}}[\rho]$, which absorbs both the difference between the true (*interacting*) and *non-interacting* kinetic energy as well as the non-classical part of the electron-electron interaction (exchange and correlation).[60] In the following, $T[\rho]$ represents the true kinetic energy of the *interacting* system and $E_{\mathrm{ee}}[\rho]$ describes the correlated interaction of the electrons:

$$E_{\mathrm{xc}}[\rho] = (T[\rho] - T_{\mathrm{s}}[\rho]) + (E_{\mathrm{ee}}[\rho] - J[\rho])\,. \tag{1.19}$$

At this point, KS-DFT remains exact, as no approximations have been introduced. If the exact form of $E_{\mathrm{xc}}[\rho]$ was known, both dynamic and static correlation effects would be fully captured. Thus, all deviations from the exact solution originate from the choice of $E_{\mathrm{xc}}[\rho]$. Given an appropriate $E_{\mathrm{xc}}[\rho]$, the KS equations can be solved iteratively under the orthonormality constraint [Eq. (1.9)], analogous to the Fock equations [Eq. (1.12a)]:

$$\left[-\frac{1}{2}\nabla^2 + \hat{V}_{\mathrm{eff}}(\mathbf{r})\right] \phi_i(\mathbf{r}) = \epsilon_i \phi_i(\mathbf{r}), \tag{1.20}$$

where the effective potential is defined as:

$$\hat{V}_{\mathrm{eff}}(\mathbf{r}) = \hat{V}_{\mathrm{ne}}(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \hat{V}_{\mathrm{xc}}(\mathbf{r}). \tag{1.21}$$

Conceptual differences between HF and KS-DFT lie in their points of approximation. HF approximates the many-electron wavefunction with a single Slater determinant, while being formally correct and thus avoiding SIE but neglecting electron correlation. In contrast, KS-DFT is based on the electron density, which, in principle, fully describes the many-electron system, as the Slater determinant serves only as an auxiliary construct. However, it approximates the Hamiltonian through a non-exact exchange-correlation (XC) functional, introducing errors such as SIE. The entire challenge of KS-DFT lies thus in the choice of the XC functional $E_{\mathrm{xc}}[\rho]$, for which the exact general expression is still unknown. Functional development follows two primary philosophies: One approach systematically derives $E_{\mathrm{xc}}[\rho]$ by enforcing physical constraints while minimizing the number of empirically fitted parameters, ensuring formal correctness to a given extent.[65–67] For an in-depth discussion of the various physical constraints on $E_{\mathrm{xc}}[\rho]$, the reader is referred Refs. [60, 65, 68–70]. Alternatively, semiempirical optimization adjusts the functional form and associated parameters to best reproduce reference data of target properties.[71†] Therefore, the mathematical forms of many approximations for $E_{\mathrm{xc}}[\rho]$ are often pragmatically designed for performance rather than physical insight and do not necessarily follow first principles in a strict way.[60,69,70] However, many functionals follow a compromise between both strategies.

In practice, $E_{\mathrm{xc}}[\rho]$ is commonly decomposed into an exchange term $E_{\mathrm{x}}[\rho]$ and a correlation term $E_{\mathrm{c}}[\rho]$ to distinguish their respective contributions. A key challenge arises from the approximation of $E_{\mathrm{x}}[\rho]$, which can lead to incomplete cancellation of the unphysical self-interaction in the Coulomb term [Eq. (1.16)], resulting in SIEs (*vide supra*). Several approximations to the exact functional $E_{\mathrm{xc}}[\rho]$ exist, collectively referred to as density functional approximations (DFAs), or simply *functionals*. These can be categorized conceptually,

† This should not be confused with semiempirical quantum mechanical methods, which replace physically defined integrals with empirical expressions (see Sec. 2.2). Nevertheless, modern KS-DFT functionals are also not purely *ab initio*, as defined by Thiel in Ref. [72].

following Perdew's "Jacob's ladder" framework,[73,74] which will also be used in this work.

Local Density Approximation

The local density approximation (LDA) is constructed based on the hypothetical system of a uniform electron gas, and resembles the lowest rung on "Jacob's ladder". Its key approximation assumes that the XC energy at any given point depends only on the absolute, *local* electron *density* at that point. Thus, the density at each position is treated as if it were part of a homogeneous system. The exchange term in LDA follows from the analytical expression for the exchange energy of a uniform electron gas, as described in Slater's density-based correction to HF theory.[63] Interpolation formulae connecting the analytic low- and high-density limits of the correlation energy have been proposed by Vosko, Wilk, and Nusair, as well as by Perdew and Wang.[75,76] Despite neglecting all density inhomogeneities in real systems, LDA often yields reasonable predictions for molecular structures and charge moments, but tends to overestimate covalent bonding.[†] However, it usually suffers strongly from SIEs. Its overall accuracy is comparable to HF, but their errors tend to go in opposite directions: LDA generally overestimates binding energies, whereas HF underestimates correlation effects. Notably, LDA remains highly accurate for the uniform electron gas and, by construction, serves as a reasonable approximation for systems with slowly varying electron densities like metallic systems.

(Meta)-Generalized Gradient Approximation

Generalized gradient approximation (GGA) DFAs improve upon LDA by incorporating the gradient of the electron density, $\nabla\rho$, to account for inhomogeneities in real systems, i.e., a non-uniform electron gas. This is typically achieved by applying enhancement factors that scale the LDA E_{xc} based on the local density gradient. While GGA functionals are sometimes referred to as *non- or semi-local*, this is not strictly precise, as they, evaluate the electron density and its first derivative only at a given point. Popular GGAs include B88[77] for the exchange term, and LYP[78,79] or PW91[80] for the correlation term. The PBE functional[81] provides both an exchange and a correlation term. Meta (m)-GGA functionals further extend this idea by taking into account the kinetic energy density, $\tau(\mathbf{r})$, or the Laplacian of the electron density, $\nabla^2\rho$.[82,83] Thereby, XC effects can be captured more accurately in many cases. Examples for m-GGA DFAs are TPSS,[84] B97M-V,[85] and r^{2}SCAN.[66,67] Especially the more recent B97M-V (based on optimized parameterization) and the strongly constrained r^{2}SCAN provided significant improvements over GGAs, for instance by noticeably reducing the amount of SIE. GGA and m-GGA functionals correspond to the second and third rung on "Jacob's ladder".

Hybrid Methods

Sec. 1.1.2 outlined that KS-DFT employs a non-interacting reference system, represented by a single Slater determinant, to reproduce the exact ground-state electron density of the real (interacting) system. For any such hypothetical single-determinant wavefunction, HF provides the exact solution and, crucially, an exact expression for the exchange energy. Hybrid DFAs bridge the gap between pure density functionals, which approximate E_{x} solely from the electron density, and methods incorporating *exact* exchange[‡] from HF theory, aiming to combine the strengths of both approaches. Due to a more accurate cancellation of $J[\rho]$ [Eq. (1.16)], the amount of SIE can effectively be reduced in this way. This approach is theoretically justified by the adiabatic connection formula (ACF), which connects the non-interacting and interacting system via the

[†] There might be exceptions to these rule of thumbs, in particular due to SIE.
[‡] Also referred to as *Fock* exchange.

integration over intermediate interaction strenghts, indicated by the coupling parameter λ.[86] In the non-interacting limit ($\lambda = 0$), the exchange energy is exactly given by the K integrals from HF theory.[†] The simplest approximation suggests a linear interpolation for the exchange energy, leading to the so-called *half-and-half* approach.[69,87]

$$E_\mathrm{x} = \frac{1}{2}E_\mathrm{x}^\mathrm{HF} + \frac{1}{2}E_\mathrm{x}^\mathrm{LDA/(m)GGA}. \tag{1.22}$$

Later realizations of this ansatz introduced semiempirically determined coefficients for the individual contributions. A prominent example is the B3LYP functional:[78,88,89]

$$E_\mathrm{xc,B3LYP} = (1-a_\mathrm{x})E_\mathrm{x}^\mathrm{LDA} + a_\mathrm{x}E_\mathrm{x}^\mathrm{HF} + b\Delta E_\mathrm{x}^\mathrm{B88} + cE_\mathrm{c}^\mathrm{LYP} + (1-c)E_\mathrm{c}^\mathrm{LDA}, \tag{1.23}$$

where the empirical parameters a_x, b, and c are set to 0.20, 0.72, and 0.80, respectively. Other established hybrid DFAs are PBE0[90] and TPSSh[91,92]. While empirical studies found that the admixture of exact exchange improves the accuracy of KS-DFT on average, the optimal specific amount to be added (a_x) is rather problem-specific.[69,70,93] Connecting to the remark in Sec. 1.1.2, hybrid DFAs, which correspond to the fourth rung on "Jacob's ladder",[‡] are truly *non-local* to some extent as the expression for the exchange energy in HF, K, is a non-local function. This is in contrast to lower-rung density functionals, which rely solely on the density and its derivatives at a given point.

Range-separated hybrid (RSH) functionals address systematic errors in KS-DFT like SIE and the question of the amount of exact exchange by partitioning the two-electron operator and more specifically E_x into short- and long-range components:[95–98]

$$\frac{1}{r_{12}} = \frac{1-\mathrm{erf}(\omega r_{12})}{r_{12}} + \frac{\mathrm{erf}(\omega r_{12})}{r_{12}}. \tag{1.24}$$

These functionals use a DFA-based exchange expression (E_x^KS) for the short-range and the exact Fock exchange (E_x^HF) for the long-range part. The proper handling of the long-range limit reduces SIEs and resolves issues like artificial charge transfer (CT) in dissociation and zwitterionic structures. The range-separation parameter ω is often tuned empirically or non-empirically.[99] Example DFAs are CAM-B3LYP[100], the ωB97 family,[101–104] or HSE.[105,106] Hybrid functionals that lack such a partitioning are often referred to as *global* hybrids.

Dispersion-Corrected DFT

Like all mean-field methods, approximate KS-DFT in its common form cannot capture long-range correlation effects, particularly dispersion interactions, that is, the attractive part of the van der Waals interaction.[107§] The first investigations of this issue were presented by Pulay, Gordon, Becke, Hobza, and co-workers in the 1990s.[109–112] These interactions arise from instantaneous charge fluctuations and are crucial for accurately modeling non-covalent interactions, molecular crystals, and large biomolecular systems.[107,113–116]

Since standard DFAs inherently lack the explicit treatment of dispersion, various correction schemes have been introduced to remedy this deficiency. Most of them add an additional term to the total energy.

$$E_\mathrm{total} = E_\mathrm{KS\text{-}DFT} + E_\mathrm{disp}. \tag{1.25}$$

[†] See Eq. 1.8 and 1.13 for reference.

[‡] Fifth-rung functionals on "Jacob's ladder" (often termed *double-hybrid* DFAs) incorporate virtual orbitals, partially restoring explicit wavefunction-based correlation, for example through second-order perturbation theory.[94]

[§] Double-hybrid DFAs can capture larger amounts of long-range correlation effects than lower-rung functionals but still benefit from semiclassical dispersion corrections.[108]

A widely used semiempirical approach is DFT-Dn. The energy expression for the most recent versions D3(BJ)$^{\mathrm{ATM}}$ and D4 reads:

$$\begin{aligned} E_{\mathrm{disp}}^{\mathrm{D3^{ATM}/D4}} = &- \sum_{AB} \sum_{n=6,8} s_n \frac{C_n^{AB}}{R_{AB}^n} f_{\mathrm{damp}}^{(n)}(R_{AB}) \\ &- \sum_{ABC} s_9 \frac{C_9^{ABC}\,(3\cos\theta_a \cos\theta_b \cos\theta_c + 1)}{(R_{AB}R_{AC}R_{BC})^3} \\ &\times f_{\mathrm{damp}}^{(9)}\left(\bar{R}_{ABC}\right), \end{aligned} \tag{1.26}$$

with the dispersion energy modeled as a sum over atom pairs and triples.[117–120] Here, R_{AB} denotes interatomic distances, and f_{damp} is the Becke-Johnson (BJ) damping function that prevents divergence at short range:[118]

$$f_{\mathrm{damp,BJ}}^{(n)}(R_{AB}) = \frac{R_{AB}^n}{R_{AB}^n + \left(a_1 R_0^{\mathrm{AB}} + a_2\right)^n}, \tag{1.27}$$

which contains the cut-off radius $R_0^{AB} = \sqrt{\frac{C_8^{AB}}{C_6^{AB}}}$. C_n^{AB} are the dispersion coefficients, which are defined semiempirically from tabulated electric dipole polarizabilities of reference compounds. The effective atomic dynamic polarizabilites used for calculation of the dispersion coefficients C_n^{AB} incorporate information about the chemical environment of each atom by a purely geometry-dependent coordination number. In D4, they additionally depend on the atomic charges derived by a classical electronegativity equilibration (EEQ) model. The second term in Eq. (1.26) is an approximation for the three-body dispersion contribution arising from the interaction of three dipoles, known as Axilrod–Teller–Muto (ATM), with $\bar{R}_{ABC}$ being the geometric mean of all three pairwise distances, and likewise C_9^{ABC} being approximated from the pairwise C_6 coefficients.[107] $f_{\mathrm{damp}}^{(9)}$ denotes an additional damping function for the three-body contribution.[101] s_6, s_8, s_9, a_1, and a_2 are empirical parameters specific to each functional.

Examples for other dispersion corrections in DFT are the Tkatchenko-Scheffler (TS/TS+SCS) methods that derive C_6 values from Hirshfeld-partitioned electron densities, with TS+SCS incorporating self-consistent screening for polarization effects. Many-body dispersion (MBD), such as TS-MBD, goes beyond pairwise and triple interactions by accounting for collective electronic fluctuations.[121,122] The exchange-hole dipole moment (XDM) model computes C_6 coefficients from the electron density's exchange hole.[123] Non-local functionals (e.g., vdW-DF[124], VV10[125] as in ωB97M-V[104]) avoid pairwise schemes by integrating non-local correlation functionals for the treatment of dispersion. For an extensive review on dispersion-corrected mean-field methods, the reader is guided to Ref. [107].

Dispersion corrections have made DFT accurate for inter- and intra-molecular non-covalent interactions (NCIs), and have become indispensable in modern computational chemistry.[126]

1.2 The Basis Set Expansion

In practical applications of quantum chemistry (QC) methods, the wavefunction Ψ or the electron density ρ must be expressed in a mathematical form suitable for computation. While in theory, both HF and DFT can be evaluated purely numerically on a grid, such an expression is usually achieved through a basis

set expansion, where MOs ϕ_i are represented as a linear combination of predefined basis functions χ_μ.[†] This ansatz transforms an otherwise highly non-linear orbital optimization problem into a computationally manageable linear algebra problem. Since in molecular calculations usually atom-centered basis functions, which originally approximate AOs, are employed, the approach is also called linear combination of atomic orbitals (LCAO):

$$\phi_i(\mathbf{r}) = \sum_{\mu}^{N_{\text{bf}}} C_{i\mu}\chi_\mu(\mathbf{r}) = \sum_{\mu}^{N_{\text{bf}}} C_{i\mu}\,|\chi_\mu\rangle. \tag{1.28}$$

Here, $C_{i\mu}$ are the expansion coefficients to be determined during the SCF calculation. Using the definition from Eq. (1.28), the eigenvalue problem from Eq. (1.12a) can be reformulated. While this is demonstrated on the example of HF, the principle stays similar for KS-DFT. The Fock operator acting on MO ϕ_i then reads:

$$\hat{F}\sum_{\mu}^{N_{\text{bf}}} C_{i\mu}|\chi_\mu\rangle = \epsilon_i \sum_{\mu}^{N_{\text{bf}}} C_{i\mu}|\chi_\mu\rangle. \tag{1.29}$$

To obtain a generalized eigenvalue problem in the MO coefficients $C_{i\mu}$, we project both sides onto a basis function χ_ν by integrating over all space:[‡]

$$\langle\chi_\nu|\hat{F}\sum_{\mu}^{N_{\text{bf}}} C_{i\mu}|\chi_\mu\rangle = \epsilon_i\langle\chi_\nu|\sum_{\mu}^{N_{\text{bf}}} C_{i\mu}|\chi_\mu\rangle. \tag{1.30}$$

By introducing the matrix elements of the Fock matrix $\mathbf{F}$ and the overlap matrix $\mathbf{S}$,

$$F_{\nu\mu} = \langle\chi_\nu|\hat{F}|\chi_\mu\rangle \quad \text{and} \quad S_{\nu\mu} = \langle\chi_\nu|\chi_\mu\rangle. \tag{1.31}$$

Eq. (1.29) can be rewritten for all MOs ϕ_i as the generalized eigenvalue problem:

$$\mathbf{FC} = \mathbf{SC}\boldsymbol{\epsilon}, \tag{1.32}$$

which is known as the Roothaan-Hall equation.[127,128] The elements of the Fock matrix $\mathbf{F}$ can now be defined as follows:[§]

$$F_{\nu\mu} = \underbrace{\langle\nu|\hat{h}|\mu\rangle}_{h_{\nu\mu}} + \underbrace{\sum_{\lambda,\sigma} P_{\lambda\sigma}\left[(\nu\mu\mid\lambda\sigma) - \frac{1}{2}(\nu\lambda\mid\mu\sigma)\right]}_{J_{\nu\mu}-K_{\nu\mu}}, \tag{1.33}$$

where $\mathbf{P}$ is the so-called density matrix:

$$P_{\lambda\sigma} = \sum_j n_j C_{\lambda j} C_{j\sigma}. \tag{1.34}$$

[†] Lower-case Latin letter indices denote MOs (ϕ); lower-case Greek letter indices denote atomic orbitals (AOs) (χ), i.e. basis functions.

[‡] Real (not complex) functions are assumed throughout.

[§] Here, we use the chemist's notation as follows:

$$(ij\mid kl) = \iint \frac{\chi_i(\mathbf{r})\chi_j(\mathbf{r})\chi_k(\mathbf{r}')\chi_l(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}'.$$

Here, n_j denotes the occupation number, which takes values of either 0 or 2 in (restricted) closed-shell calculations.[†] Eq. (1.33) highlights that only the contribution of the two-electron integrals depends on the density, but not that of the one-electron integrals $h_{\mu\nu}$. Thus, in the iterative solution of HF and DFT, the density matrix $\mathbf{P}$ is refined until convergence is reached. It provides an efficient and numerically stable way to describe electron distributions in molecular systems. The detailed derivation of Eq. (1.33) and (1.34) is outlined in Sec. A.1. A new semiempirical method to predict $\mathbf{P}$, circumventing costly *ab-initio* calculations, is presented in Chapter 4.

1.2.1 Slater and Gaussian-Type Orbitals

Historically, Slater-type orbital (STO) were introduced as basis functions since they closely resemble hydrogen atomic orbitals, making them an efficient representation of molecular wavefunctions.[129] STOs take the form:

$$\chi_\mu^{\mathrm{STO}}(\mathbf{r}, \zeta) \propto Y_{l,m}(\theta, \varphi) r^{n-1} e^{-\zeta r}, \tag{1.35}$$

where ζ is an orbital exponent governing the radial decay, and $Y_{lm}(\theta, \phi)$ are spherical harmonics determining the angular part. n, l, m are the principal, azimuthal, and magnetic quantum number of the basis function μ. While STOs are physically meaningful, their use in QC is limited due to the complexity of evaluating multi-center electronic repulsion integrals. Consequently, Gaussian-type orbitals (GTOs) are commonly employed instead,[130] defined as:

$$\chi_\mu^{\mathrm{GTO}}(\mathbf{r}, \zeta) \propto Y_{l,m}(\theta, \varphi) r^{2n-2-l} e^{-\zeta r^2}. \tag{1.36}$$

GTOs enable efficient integral evaluation using the Gaussian product theorem,[131] making them the standard choice in modern (molecular) electronic structure calculations.[132]

1.2.2 Contracted Basis Sets

To improve computational efficiency while maintaining accuracy, many basis sets are designed as *contracted basis sets*. Instead of treating each basis function as an independent primitive Gaussian, contracted basis functions are predefined linear combinations of primitives:

$$\chi_\kappa = \sum_{\alpha \in \kappa}^{N_{\mathrm{pr}}} c_\alpha \chi_\alpha(\zeta_\alpha). \tag{1.37}$$

c_α are fixed contraction coefficients that define how the primitive Gaussians χ_α with distinct exponents ζ_α are combined to form the contracted basis function χ_κ with N_{pr} primitive functions. This reduces the *effective* number of basis functions (i.e., AOs) in the SCF procedure, thereby cutting the computational cost for Fock-matrix construction and subsequent matrix operations. Moreover, combining multiple primitives into a single AO improves its shape, better resembling the true AO compared to a single primitive Gaussian. Although modern codes typically compute integrals at the primitive level, only the final contracted integrals are stored and used, thereby simplifying the MO optimization problem. In practice, the contraction of primitive functions into a single contracted AO is usually applied for describing the energetically important but chemical unimportant core electrons.[69]

[†] In unrestricted KS or HF, each spin orbital (α/β) is occupied with only one electron.

1.2.3 Effective Core Potentials

For heavy elements (often referred to as $Z > 36$),[70] fully relativistic QC calculations treating all electrons explicitly become computationally challenging due to: *(i)* the large number of core electrons requiring high angular momentum basis functions, *(ii)* significant scalar relativistic effects, and *(iii)* spin-orbit coupling that cannot be neglected for heavy atoms.[133] To address this, effective core potentials (ECPs) as first used by Hellmann and developed by Phillips and Kleinman replace the explicit treatment of chemically inert core electrons with an effective potential that reproduces their quantum mechanical influence on valence (and semi-core) electrons through a semi-local operator.[134–137]

$$\hat{V}_{\mathrm{ECP}}(\mathbf{r}) = \underbrace{-\frac{Z_{\mathrm{eff}}}{r}}_{\text{local term}} + \underbrace{\sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} |Y_{lm}\rangle \, V_l(r) \, \langle Y_{lm}|}_{\text{non-local term}}, \tag{1.38}$$

where $Z_{\mathrm{eff}} = Z - N_{\mathrm{core}}$ is the effective nuclear charge (N_{core}: number of core electrons), $l_{\max}$ is the maximum angular momentum of the core, and $V_l(r)$ are parameterized radial potentials for each angular momentum channel l.[136] The spherical harmonics Y_{lm} enforce orthogonality between valence and core orbitals via the projection operator.[138] The use of ECPs significantly reduces computational cost by eliminating core orbitals from the basis set expansion and by avoiding an explicit treatment of relativistic effects. In modern DFT calculations, basis sets as the def2 family were often developed for use with matching ECPs, here the Stuttgart-Cologne ECPs.[139–142] Other widely used pseudopotentials include LANL2DZ,[143] and CRENBL.[144]

1.2.4 Balancing Accuracy and Computational Effort

The basis set directly determines the functional flexibility of MOs, dictating how well they can represent the underlying electronic structure. The number of basis functions per angular momentum of the occupied AO shell determines the so-called cardinal number ζ.[†] Thus, a cardinal number of 1 corresponds to a single-ζ or minimal basis (MB) set. An established minimal basis (MB) is STO-nG, where each STO is approximated by a combination of n GTOs.[145] A common standard for DFT calculations are basis sets of at least double-ζ (DZ) or triple-ζ (TZ) quality.[70] Beyond merely increasing the number of basis functions, additional flexibility is introduced through polarization and diffuse functions. Polarization functions (often denoted by a suffix "P") extend the angular momentum beyond what is required for the atomic ground state, improving the description of electron density distortions in bonding and external fields. Diffuse functions (often denoted by a suffix "D") on the other hand, employ low-exponent basis functions that extend the spatial reach of the wavefunction, which is particularly important for accurately modeling anions and weakly bound electronic states. While extended basis sets allow for greater adaptability and accuracy in describing complex electronic environments, the computational cost grows rapidly with the size of the basis set, as the number of one- and two-electron integrals that must be evaluated increases polynomially. One of the first extended basis sets for HF and DFT calculations was Pople's 6-31G basis set.[146] Later, the def2 basis set family by Weigend and Ahlrichs has become a standard for today's DFT calculations.[70,139] Nowadays, different basis sets exist, often specific to calculation types and applications.[147–152]

Basis set errors (BSEs) stem from a finite, that is, incomplete basis set expansion and can be classified into two different categories:[69,153]

[†] Although both variables share the Greek symbol ζ, the cardinal number should not be confused with the exponent of a primitive function, ζ_α.

(i) Basis set incompleteness error (BSIE) is the difference in energy between the system calculated in an infinite basis expansion and a finite basis set.[69,153] BSIEs are inherently unsystematic in relative energies but lead to systematically increased absolute energies.

(ii) Basis set superposition errors (BSSEs) stem from the *unphysical* use of basis functions of another part of the molecular system. A fragment appears artificially stabilized in the presence of a neighboring moiety, not due to genuine interactions but because the additional basis functions compensate for the incompleteness of its own basis set expansion. This leads for instance to a systematic overestimation of interaction energies in non-covalent complexes, caused by an asymmetry in the effective BSE of monomer and dimer.[154,155†] Methods to address this error are discussed in Sec. 2.1 and Chapter 3.

Consequently, both types of BSEs converge to zero in an infinite basis set expansion. It has been shown that the HF and DFT energy converges square-root exponentially to the complete basis set (CBS) limit:[156,157‡]

$$E_{\mathrm{HF/DFT}}(X) = E(\infty) + Ae^{-B\sqrt{X}}, \tag{1.39}$$

with X being a measure for the basis set completeness (e.g., the cardinal number), and A and B arbitrary prefactors. Practically, convergence is mostly achieved with quadruple-ζ (QZ) or larger basis sets.[69,70,157]

Beyond the formal size of a basis set – often defined by its cardinal number ζ or the total number of basis functions – several other factors critically influence its ability to represent the electronic wavefunction while determining the associated computational complexity. One key aspect is the contraction depth, as outlined in Sec. 1.2.2. Another decisive factor in the cost-vs-accuracy trade-off is the procedure used to determine the exponents ζ_α (and coefficients c in contracted basis sets). Traditionally, basis set parameters were optimized variationally with respect to atomic energies, which may not be optimal for describing MOs in molecular systems. Additionally, polarization functions are typically unoccupied in the atomic ground state, necessitating specialized techniques for their consistent and reliable optimization. VandeVondele, Hutter, and Jensen introduced basis sets optimized directly in molecular calculations, deviating from the traditional atomic-based optimization approach.[156,159,160] This strategy is particularly relevant for deeply contracted basis sets, where much of the optimization process shifts from the LCAO approach in the SCF to the pre-defined contracted AOs and their coefficients. Recently, Shaw and Hill developed a framework for semi-automated molecular basis set optimization.[161]

A discussion of the limitations of small or MBs, along with two case studies, is provided in Sec. 2.3.

† Although BSSE is often explained in this way for simplicity, it is not limited to NCIs but is also relevant in intramolecular cases.

‡ Correlated wave function theory (WFT) methods, including double-hybrid DFT methods, exhibit inverse cubic convergence toward the CBS limit.[158]

CHAPTER 2

Semempirical Approaches in Quantum Chemistry

The *ab initio* methods HF and – particularly since the 1990s – DFT have become the standard for simulating the electronic structure of chemical systems.[8,70,126] However, even with modern high-performance computing (HPC) infrastructure, the computational cost of these methods limits their applicability to large systems and HTSs.[162,163] In particular, molecular structures with significantly more than a few hundred atoms remain impractical for routine calculations using standard DFT methods.[116] Arguably even more critical is the ability to perform hundreds or thousands of calculations on small- to mid-sized molecules efficiently.[164] Key applications include conformer searches,[8,165] molecular docking[166] (e.g., for explicit solvation[167]), metadynamics (MTD) or molecular dynamics (MD) simulations over long time scales or for extended molecules,[8,168,169] HTS for candidate molecules,[31,170] and the exploration of reaction networks.[171]

The computational complexity of HF/DFT methods is dominated by two steps:

(i) Construction of the Fock matrix[†] and contraction of **P** with the integrals, including the computation of four-center two-electron integrals, which formally scales as $\mathcal{O}\left(N_{\mathrm{bf}}^4\right)$.[69‡] Apart from the effective scaling, the computational prefactor is relatively high, making this step the dominant bottleneck in modern direct SCF approaches.[182]

(ii) Solving the Fock or KS equation via matrix diagonalization, which scales as $\mathcal{O}\left(N_{\mathrm{bf}}^3\right)$.[§]

Since both computationally demanding steps scale with the number of basis functions N_{bf}, employing standard TZ or larger basis sets significantly increases computational complexity. One approach to mitigate this within first-principles methods is to reduce the basis set size. However, smaller basis sets limit the flexibility of MOs, leading to BSEs. To address this trade-off, small basis set composite methods incorporating (semi)empirical corrections for limitations of the basis set have been developed, as will be discussed in Sec. 2.1. A more drastic strategy for reducing computational cost is to bypass the costly two-electron integrals entirely, replacing them with empirical expressions for electron-electron interactions. This effectively eliminates the computational

[†] In KS-DFT often termed KS matrix.

[‡] By employing integral screening techniques alone,[172,173] the scaling reduces to $\mathcal{O}\left(N_{\mathrm{bf}}^2\right)$ asymptotically. With integral approximations such as the resolution of the identity (RI)[174–177] and the chain-of-spheres algorithm for exchange matrix computations (COSX),[178,179] or related methods,[180,181] the scaling of Hamiltonian construction in HF and DFT can be reduced to $\mathcal{O}\left(N_{\mathrm{bf}}^x\right)$, with x ranging from 1 (asymptotically *linear scaling*) to 3 (even in relatively small systems) in modern implementations.

[§] As with integral evaluation, advanced techniques have been developed to circumvent the cubic scaling of standard diagonalization methods for large systems.[183–186]

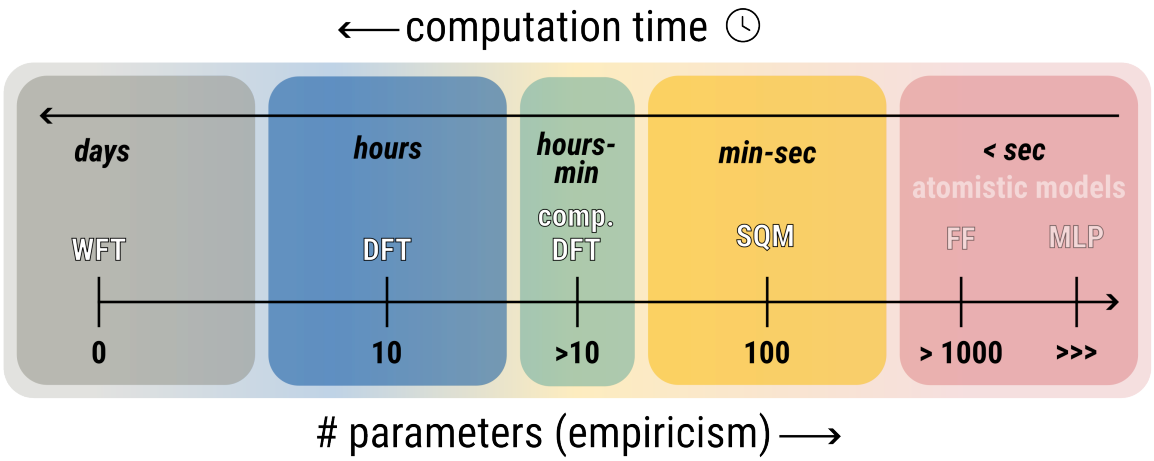

Figure 2.1: Schematic illustration of the relationship between computational efficiency and the degree of empiricism in different classes of computational chemistry methods. *comp. DFT* denotes small basis set composite DFT methods.[†]

burden associated with Hamiltonian construction. Such methods are referred to as semiempirical quantum mechanical (SQM) methods in the following and will be outlined in Sec. 2.2.

Consequently, increasing computational efficiency typically comes at the expense of greater empiricism. The degree of empiricism can be loosely quantified, for example, by the number of empirical parameters incorporated into a method. This trade-off between computational cost and empiricism is schematically illustrated in Fig. 2.1. WFT methods, such as correlated post-HF methods, are purely *ab initio*, relying exclusively on first principles. In contrast, standard KS-DFT often includes few semiempirical parameters that are not strictly dictated by physical laws. This reliance on empirical input is even more pronounced in composite DFT methods. SQM methods extend this trend further by parameterizing entire Fock matrix contributions, such as two-electron interactions, instead of computing them from first principles. Even greater efficiency can be achieved using purely empirical atomistic models that do not incorporate the quantum mechanical (QM) description of electrons. A conventional example is classical FFs, which describe molecular systems using analytical potential energy functions to approximate bonded and nonbonded interactions.[187–189] More recently, MLPs have emerged, learning energies and forces from reference data (e.g., from DFT) and approximating these properties using machine learning (ML) techniques such as neural networks.[190–193] While atomistic models fall outside the scope of this work, they are included here for reference as they also yield molecular geometries and relative energies relevant to computational chemistry.

2.1 Small-Basis-Set Corrections

Employing smaller basis sets accelerates calculations by reducing the number of required integrals, the matrix dimensions and, consequently, the complexity of the optimization problem. However, this comes at the expense of reduced accuracy, as discussed in Sec. 1.2.4.

Various approaches exist to address basis set deficiencies, with most focusing on systematic BSSEs. One

[†] The schematic relationships make no claim to strict correctness and are only intended to reflect general trends. For instance, the number of empirical parameters, as well as the relationship between the computational efficiency of FFs and MLPs methods, strongly depends on the specific approach and is not further classified in this work.

of the earliest methods is the Boys-Bernardi counterpoise (CP) correction, which compensates for BSSE by computing each monomer in a complex using the dimer's basis set.[194] This accounts for the *artificial* use of basis functions, and the resulting energy difference is added to the interaction energy. Kruse and Grimme developed the geometrical counterpoise (gCP) correction, which extends this approach to intramolecular BSSE.[195] Building on gCP, Head-Gordon and co-workers introduced an empirical correction, DFT-C, designed to provide near-complete-basis DFT results using the diffuse DZ basis set def2-SVPD.[139,149,153] DiLabio and co-workers leveraged the functional form of ECPs to enhance the description of NCIs.[196,197] Sure and Grimme proposed an empirical short-ranged basis (SRB) correction to counteract the systematic overestimation of covalent bond lengths in MBs.[198] Further examples specifically for WFT and double-hybrid DFT methods include density-based basis set corrections and the R12/F12 methods that depend explicitly on the interelectronic distance.[199–202]

2.1.1 "3c" Composite Methods

The philosophy behind the "**3 c**orrections" composite methods[†] is to combine HF or DFT functionals with smaller basis sets while incorporating (semi)empirical corrections to address BSEs (as outlined above) along with the D3 or D4 dispersion correction. The first method in this series, HF-3c, combines HF with an MB[‡] and focuses primarily on accurate geometries and NCIs. It employs the gCP correction for BSSE, the SRB correction for bond lengths, and the established D3 dispersion correction.[198] Subsequent methods in this series were based on DFT, including PBEh-3c, which utilizes a specially adapted PBE hybrid DFA.[115] PBEh-3c is designed to improve accuracy by incorporating electron correlation effects and using a DZ basis set. A closely related variant is HSE-3c.[212] B97-3c follows a similar philosophy but, for the first time, employs an adapted TZ basis set and is based on a GGA functional.[213] The latest evolution, r^{2}SCAN-3c, is an m-GGA method incorporating the more recent D4 correction. In addition to the aforementioned properties, it targets conformational energies and aims to improve thermochemical accuracy.[214] Chapter 3 presents ωB97X-3c as a new addition to this family.

2.2 Semiempirical Quantum-Mechanical Methods

A defining feature of all SQM methods is their goal to emulate DFT or HF, while circumventing their computational complexity. This is mostly achieved by replacing the costly two-electron integrals with parametric expressions, reducing the computational burden so that the Hamiltonian matrix diagonalization becomes the primary bottleneck. This simplification allows SQM methods to achieve significant efficiency gains of several orders of magnitude while maintaining reasonable accuracy for many applications.[72,215,216] SQM methods have a long tradition, beginning with the famous Hückel theory already in the 1930s.[217] Further cornerstones for today's SQM methods were laid in the 1950s and 60s by the Pariser–Parr–Pople (PPP)[218–220] method as well as extended Hückel theory (EHT).[221] Modern SQM methods can be separated into HF- and DFT-based branches. The evolution of SQM methods is illustrated in Fig. 2.2 and will be outlined in this section.

[†] The term composite method in QC is also associated with the G*n* wave function theory developed by Pople and co-workers,[203–206] as well as the Weizmann protocols[207–209] and related CBS schemes.[158,210] However, these approaches pursue a different goal.

[‡] The MB MINIS basis set is used for light elements, while for heavier elements with $Z > 18$, split-valence DZ basis sets of the Ahlrichs type are employed.[139,211]

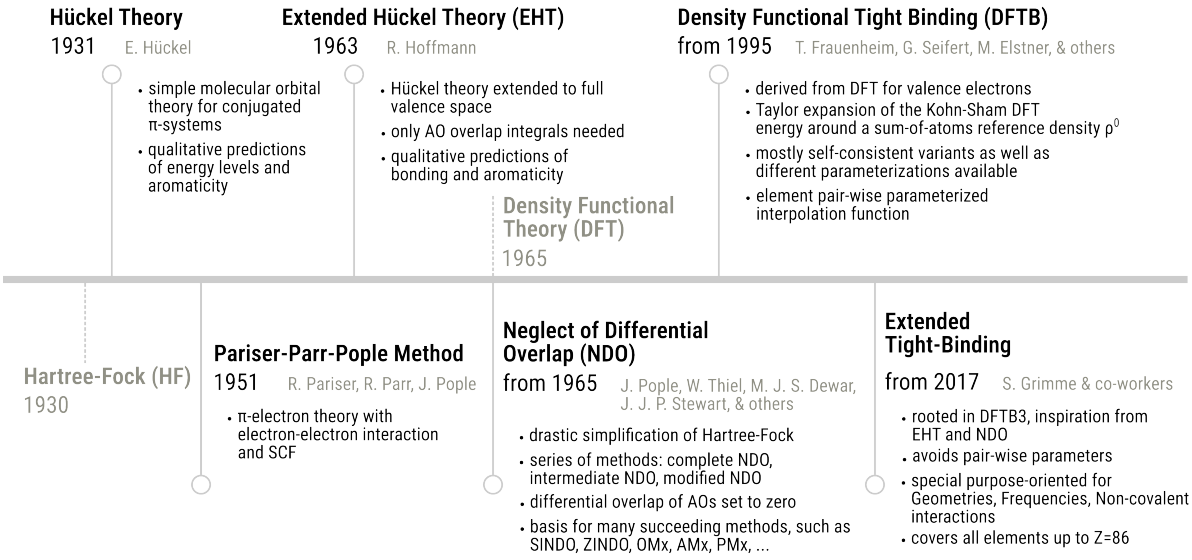

Figure 2.2: Overview of the historical evolution of SQMs methods.

2.2.1 The Beginnings

Hückel Theory

Hückel theory, introduced by Erich Hückel in 1931,[217] is a semiempirical method for modeling π-electron systems in conjugated hydrocarbons. It simplifies the MO problem by considering only π-electrons while treating the potential of all other electrons as a constant, which is neglected.[222] MOs are constructed as linear combinations of atomic p_z orbitals following the LCAO approach [Eq. (1.28)]. Assuming orthonormal AOs ($S_{\mu\nu} = \delta_{\mu\nu}$) simplifies the generalized eigenvalue problem from Eq. (1.32) into a standard eigenvalue problem. The Hamiltonian matrix consists of *on-site* (α) and *interaction* (β) elements: α represents the on-site energy of a π electron localized on a single atom and depends on the specific atom type, while β describes the interaction between adjacent p_z orbitals on neighboring atoms. Consequently, the effective Hamiltonian matrix H_{eff} is topology-dependent and consists solely of empirical elements, independent of the molecular geometry:[†]

$$H_{\text{eff}} = \begin{bmatrix} \alpha & \beta & 0 & \cdots & 0 \\ \beta & \alpha & \beta & \cdots & 0 \\ 0 & \beta & \alpha & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha \end{bmatrix}. \tag{2.1}$$

As H_{eff} does not depend on the electron density, the eigenvalue problem can be solved non-iteratively:

$$\mathbf{H}_{\text{eff}}\mathbf{C} = \boldsymbol{\varepsilon}\mathbf{C} \quad \text{with} \quad \mathbf{C} = \begin{pmatrix} | & | & & | \\ \mathbf{c}^{(1)} & \mathbf{c}^{(2)} & \cdots & \mathbf{c}^{(n)} \\ | & | & & | \end{pmatrix}. \tag{2.2}$$

[†] For Hückel theory, it is convenient to define H_{eff} as done in Eq. (2.1). However, this definition is a special case rather than a general way of constructing a Hamiltonian matrix.

Eq. (2.2) yields the orbital energies $\boldsymbol{\varepsilon}$ and expansion coefficients $\mathbf{c}$ of the MOs. While Hückel theory provides qualitative MO insights and explains aromaticity according to Hückel's $4n + 2$ rule, its primary application today is in education.[69]

A major advancement of Hückel theory was the incorporation of electron-electron Coulomb repulsion, requiring iterative solutions of the effective Hamiltonian.[223] This refinement was independently introduced in the 1950s by Pariser and Parr,[218,219] as well as by Pople,[220] leading to the development of PPP. Starting from the Fock matrix expression in Eq. (1.33) and incorporating all approximations from Hückel theory, PPP can be derived by introducing the famous zero differential overlap (ZDO) approximation: products of AOs depending on the same electron coordinates when located on different atoms are neglected.[69†] Importantly, all three- and four-center integrals vanish completely, implying that non-zero $(\mu\nu|\lambda\sigma)$ occur only if $\nu = \mu$ and $\lambda = \sigma$ or if all four indices are equal (the lower-case Greek letters denote AOs). The remaining matrix elements are not computed explicitly, but are typically replaced by empirical expressions.[223,225,226‡] By construction, PPP reduces to Hückel theory in the limit of non-interacting electrons. Today, PPP remains relevant for modeling π-conjugated systems, particularly for screening singlet-triplet gaps.[225,227,228] A thorough derivation of PPP from HF is provided in Sec. B.1.

Extended Hückel Theory

Introduced by Hoffmann in 1963, EHT is a generalization of Hückel theory that incorporates both π- and σ-electrons.[221] It extends Hückel's approach by including all valence orbitals and explicitly accounting for orbital overlap. The effective EHT Hamiltonian matrix elements are defined as:

$$H_{\mu\nu} = \begin{cases} \dfrac{H_{\mu\nu} + H_{\nu\nu}}{2} K S_{\mu\nu}, & \text{if } \mu \neq \nu, \\ H_{\mu\mu}, & \text{if } \mu = \nu. \end{cases} \tag{2.3}$$

In the original implementation, $H_{\mu\mu}$ corresponds to the atomic orbital ionization energy, while K is an empirical scaling factor. Off-diagonal elements $H_{\mu\nu}$ are computed as the weighted average of the diagonal entries, scaled by the overlap integral $S_{\mu\nu}$. As a result, differences in the chemical environment between atoms of the same type are neglected.[69] Similar to the Hückel theory eigenvalue problem in Eq. (2.2), H_{eff} is independent of the electron density, so the result is obtained through a single diagonalization. Since EHT primarily provides qualitative MOs and orbital energies, it is commonly used to generate an initial density matrix guess in HF or DFT calculations.[229,230] Novel advancements related to EHT are discussed in Chapters 4-6.

2.2.2 Hartree-Fock-based Approaches

Semiempirical approximations to HF were primarily pioneered by Pople in the 1960s.[220,231,232] The following paragraphs provide an overview of HF-based SQM methods, categorized by the extent of integral neglect.[69] For more comprehensive insights, the reader is referred to Refs. [69, 223].

Complete Neglect of Differential Overlap Complete neglect of differential overlap (CNDO)[233,234] follows a similar philosophy to PPP, particularly in its use of the ZDO approximation. However, CNDO goes beyond

† Some definitions of the ZDO approximation do not necessarily imply the condition $S_{ij} = \delta_{ij}$.[224] Here, we consistently adopt the definition from Refs. [69, 223], implying $\mathbf{S}$ to be the unit matrix.

‡ Encompassing elements $(\mu\mu|\sigma\sigma)$, $(\mu\mu|\mu\mu)$, as well as, $H^{\text{PPP}}_{\mu\mu}$ and $H^{\text{PPP}}_{\mu\sigma}$. One-center exchange integrals are generally neglected.

π electrons and considers all valence electrons in an MB. Interatomic interactions are treated atom-pair-specifically rather than distinguishing between different AO shells (s, p, d, ...). While CNDO provided the theoretical foundation for later HF-based SQM methods, it has largely been superseded by more advanced approaches.[72,223] Similar to PPP, a key limitation of CNDO is its neglect of one-center exchange integrals.[232] An in-depth derivation and explanation is provided in Sec. B.2.

Neglect of Diatomic Differential Overlap Neglect of diatomic differential overlap (NDDO) refines CNDO by retaining same-atom differential overlap, preserving more electron-electron integrals.[†] The modified neglect of diatomic overlap (MNDO) parameterization of NDDO laid the foundation for a series of semiempirical methods that remain widely used today.[235] Notable examples include AM1,[236] as well as the PMx methods developed by Dewar, Stewart, and co-workers.[237,238] These methods primarily differ in their parameterization strategies, the treatment of multipole expansions for two-center two-electron integrals, and the handling of core-core repulsion.[69] Thiel and co-workers introduced methods that incorporate orthogonalization effects, leading to the development of the OMy series.[239–242] A new formulation going beyond NDDO in terms of the considered integrals was presented by Neese and co-workers.[243] The corresponding Fock matrix expressions are given in Sec. B.3.

Intermediate Neglect of Differential Overlap Intermediate neglect of differential overlap (INDO), independently proposed by Pople[232] and Dixon,[244] represents a compromise between CNDO and NDDO in terms of both empiricism and computational complexity. Differential overlap is retained only for one-center integrals. The primary distinction from CNDO is the explicit inclusion of one-center exchange integrals of the type $(\mu\nu|\mu\nu)$. The original INDO method was soon replaced by several improved variants, such as MINDO. Further advancements led to the development of ZINDO, which focused on molecular spectra, as well as further variants such as SINDO and MSINDO.[223,245–249] Sec. B.4 includes the general Fock matrix expression of the INDO method.

2.2.3 The Tight-Binding Approximation for Semiempirical Methods

The tight-binding (TB) approximation was historically used independently of DFT, primarily for predicting electronic band structures in solid-state physics and dates back to the 1950s.[250–255] At its most fundamental level, it assumes that electrons remain localized around their parent atoms, described by AOs, and interact only weakly with neighboring atoms.[‡] Thus, the principal principal origins of TB theory are closely linked to other methods for describing the molecular wavefunction from atomic contributions such as the LCAO approach and (extended) Hückel theory.[250,251,255]

Density Functional Tight-Binding

For chemical applications, TB became useful through the combination with DFT as introduced by Porezag et al. in 1995, and termed density functional tight-binding (DFTB).[257§] A key distinction from the HF-based approaches outlined in Sec. 2.2.2 is that DFTB and its variants do *not* employ the ZDO approximation. That

[†] This means that $(\mu\nu|\kappa\lambda)$ is computed explicitly for $\mu, \nu \in A$ and $\kappa, \lambda \in B$. A and B may or may not represent the same atom.

[‡] This contrasts with the nearly-free electron model.[256]

[§] An important preceding work by Seifert already drew a connection to LDA and highlighted the relation to EHT but focused on band structure energies.[258]

is, all elements of **S** are explicitly considered, and the generalized eigenvalue problem is solved. Furthermore, DFT includes electron-electron correlation effects to some extent, making DFT as the underlying and targeted theory generally more accurate than HF.[93,259] Since eXtended Tight-Binding (xTB) (*vide infra*) and the novel density matrix TB method presented in Chapter 4 build upon DFTB, its underlying concept will be outlined in detail in the following.

In their foundational work, Porezag et al. outlined a TB-like scheme derived entirely from LDA DFT instead of experimental data.[†] Employing the two-center approximation, where higher-order multi-center terms are formally neglected, the following expression for the effective Hamiltonian results:[257,263]

$$H_{\mu\nu} = \begin{cases} \varepsilon_\mu^{\text{free atom}}, & \text{if } \mu = \nu, \\ \langle \phi_\mu^A \mid \hat{T} + V_0^A + V_0^B \mid \phi_\nu^B \rangle, & \text{if } A \neq B, \\ 0, & \text{otherwise}, \end{cases} \tag{2.4}$$

where $V_0^A + V_0^B$ approximates the effective potential $V_{\text{eff}}(\mathbf{r})$ from first principles LDA via superposition of atomic potentials:

$$V_{\text{eff}}(\mathbf{r}) = \sum_A V_0^A(|\mathbf{r} - \mathbf{R}_A|). \tag{2.5}$$

$\varepsilon_\mu^{\text{free atom}}$ are the Hamiltonian eigenvalues of the free atom. For generating a minimal set of short-ranged atom-centered basis functions μ, one typically solves an atomic KS problem in which the atom is placed in a confining potential.[257] The short-range nature of the confined AOs ensures a fast decay with increasing distance R_{AB}. By evaluating the integral of the kinetic energy operator and $V_{\text{eff}}(\mathbf{r})$ with the confined atom-centered functions, the effective potential is projected on them. For each pair A and B, LDA calculations of the diatomic molecule A-B are carried out over a grid of bond distances. This procedure yields two-center integrals, $H_{\mu\nu}(R) = \langle \phi_\mu | \hat{H}_{\text{KS}} | \phi_\nu \rangle$ and $S_{\mu\nu}(R) = \langle \phi_\mu | \phi_\nu \rangle$ that depend only on the interatomic distance R. These interactions can be tabulated and interpolated for given interatomic distances, following the approach introduced by Slater and Koster in 1954, which has since been widely adopted.[250,264–266] In doing so, essential DFT physics were captured while keeping the TB framework of localized AOs. To obtain total energies, a simple short-ranged repulsive term

$$E_{\text{rep}} = \sum_{A<B} V_{\text{rep}}^{AB}(R_{AB}), \tag{2.6}$$

consisting of a sum of polynomials is fitted to reproduce DFT reference data, as usual in TB theory. After solving the generalized eigenvalue problem [cf. Eq. (1.32)], the total energy is obtained as

$$E_{\text{tot}} = \underbrace{\sum_i^{\text{occ}} n_i \varepsilon_i}_{\langle \Psi_i | \hat{H} | \Psi_i \rangle} + E_{\text{rep}}, \tag{2.7}$$

where $\hat{H}$ is the Hamiltonian defined in Eq. (2.4). Since the Hamiltonian elements do not depend on the occupation, i.e., no charge redistribution is allowed, the final energy can be obtained in a single diagonalization. The derivation of empirical parameters from DFT rather than empirical adjustment represented a major milestone in the development of SQM methods.[263]

Later, self-consistent variants called self-consistent charge (SCC)-DFTB (also termed DFTB2) and DFTB3 were introduced, which were derived directly from KS-DFT theory.[267–269] For that, we recall Eq. (1.18) and replace the charge density $\rho(\mathbf{r}')$ by a reference density $\rho_0(\mathbf{r}')$, which usually represents that of the neutral

[†] For related approaches that also tried to derive TB expressions in a less empirical way, see Refs. [260–262].

unperturbed atom, and a small fluctuation due to the perturbation by the environment $\delta\rho(\mathbf{r}')$:

$$\rho(\mathbf{r}') = \rho_0(\mathbf{r}') + \delta\rho(\mathbf{r}'). \tag{2.8}$$

Inserting into Eq. (1.18) and Taylor expanding E_{xc} up to the third order leads to Eq. (2.9), which additionally contains all contributions to V_{ee} including electron-electron coulomb repulsion as well as semi-local XC.[269]

$$\begin{aligned} E[\rho_0 + \delta\rho] = {} & \frac{1}{2}\sum_{AB}\frac{Z_A Z_B}{R_{AB}} - \frac{1}{2}\iint\frac{\rho_0(\mathbf{r})\rho_0(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' - \int V_{\mathrm{xc}}[\rho_0]\rho_0(\mathbf{r})\,d\mathbf{r} + E_{\mathrm{xc}}[\rho_0] \\ & + \sum_i n_i\langle\psi_i|\hat{H}[\rho_0]|\psi_i\rangle + \underbrace{\frac{1}{2}\iint\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|} + \left.\frac{\delta^2 E_{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')}\right|_{\rho_0}\right)\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\,d\mathbf{r}d\mathbf{r}'}_{E^2} \\ & + \underbrace{\frac{1}{6}\iiint\left.\frac{\delta^3 E_{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\delta\rho(\mathbf{r}'')}\right|_{\rho_0}\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\delta\rho(\mathbf{r}'')\,d\mathbf{r}d\mathbf{r}'d\mathbf{r}''}_{E^3} \\ = {} & E^{0+1}[\rho_0,\delta\rho] + E^2[\rho_0,(\delta\rho)^2] + E^3[\rho_0,(\delta\rho)^3], \end{aligned} \tag{2.9}$$

wherein $\hat{H}$ represents the one-electron contributions by the kinetic energy and the external potential. The non-self-consistent DFTB1 introduced before can be derived from Eq. (2.9) via neglecting the last two (second and third order) terms.[267] This is a good approximation for all cases, in which the real density of the system is close to the superposition of atomic densities, that is, no significant CT effects or polarized bonds are present. SCC-DFTB considers E^2 in addition to the terms contained in DFTB1. Decomposing E^2 into atom-centered contributions and treating them as monopoles in a first approximation, leads to

$$E^2 \approx \frac{1}{2}\sum_{A,B}^{N}\Delta q_A \Delta q_B \gamma_{AB}(r_{AB}), \tag{2.10}$$

where γ_{AB} is chosen such that $\lim_{R\to 0}\gamma_{AB} = U_\alpha$ and $\lim_{R\to\infty}\gamma_{AB} = \frac{1}{r_{AB}}$ holds, with the so-called Hubbard parameter U_α being related to the chemical hardness.[270†] Thus, in the limit of large interatomic distances, the interaction of two charge fluctuations reduces to be of pure Coulomb type since the LDA XC contributions vanish. The final energy in second order then reads

$$\begin{aligned} E_{\mathrm{DFTB2}}[\rho_0 + \delta\rho] &= \sum_i^{\mathrm{occ}}\underbrace{\langle\Psi_i|\hat{H}_0|\Psi_i\rangle}_{n_i\varepsilon_i} + E_{\mathrm{rep}} + \frac{1}{2}\sum_{A,B}^{N}\gamma_{AB}\Delta q_A\Delta q_B \\ &= E^{0+1} + E^2, \end{aligned} \tag{2.11}$$

where ε_i are the eigenvalues of the atomic Hamiltonian.

DFTB3[268,269,271] incorporates the full expansion from Eq. (2.9) up to third order and adds the following term to the total energy expression from Eq. (2.11), appyling the same approximations as for the second-order

† $\gamma_{AA} \approx I_A - E_A \approx 2\eta_A \approx U_A$, where I_A and E_A are the ionization potential and the electron affinity of A. Thereby, the XC contribution is implicitly considered in the second-order term.

contribution:

$$E^3 \approx \frac{1}{6} \sum_{ABC} \Delta q_A \Delta q_B \Delta q_c \left. \frac{d\gamma_{AB}}{dq_C} \right|_{q_C^0}$$
$$\approx \frac{1}{3} \sum_{A,B}^{N} \Delta q_A^2 \Delta q_B \Gamma_{AB}(r_{AB}), \tag{2.12}$$

which contains a derivative of the γ function with respect to the atomic charge, effectively making the chemical hardness and the electron-electron repulsion charge-dependent. This derivative of the second-order interaction kernel is termed Γ and can be estimated from first-principles calculations.

However, DFTB faced a significant limitation: its Hamiltonian construction relies on element pairwise parameters the number of which scales quadratically as $\mathcal{O}(N^2_{\text{elem.}})$, which makes empirical parameter optimization challenging. Numerous DFTB parameter sets exist, tailored for both general applications and specific element combinations.[272–274] A few sets even cover large portions of the periodic table.[275,276]

Extended Tight-Binding Methods: GFN*n*-xTB

The xTB methods were introduced in 2017 to address the parameterization challenges of DFTB, shifting the focus specifically toward Geometries, Frequencies, and NCIs (GFN).[259,277] The GFN family began with GFN1-xTB in 2017,[277] later expanding to include GFN2-xTB[278] and the non-self-consistent variant GFN0-xTB.[279] Like DFTB3, their theoretical foundation lies in the Taylor expansion of the XC part of the KS-DFT energy around a reference density ρ_0, using small charge fluctuations $\delta\rho$ [see Eqs. (2.8) and (2.9)]. The key differences between xTB and DFTB are:

(i) The semiempirical xTB Hamiltonian elements are explicitly calculated from element-wise parameters and are not interpolated from tabulated pair-wise DFT data. Thus, element pair-wise parameters are almost completely avoided, overcoming the parameterization challenges of DFTB. The GFN*n*-xTB methods are parameterized for all elements up to radon ($Z = 86$), ensuring broad applicability.

(ii) The energy expressions derived for zeroth and first order vary significantly from DFTB, primarily due to point *(i)*. xTB is based on an atom pair-wise function for the repulsion energy and an EHT-type Hamiltonian.

(iii) The (semi)empirical parameters defining the effective Hamiltonian were not obtained from first-principles DFT calculations but were optimized to minimize the root mean square error (RMSE) of xTB-computed target properties with respect to reference data from WFT and DFT,[259] with a focus on the GFN properties.

(iv) Instead of AOs from atomic DFT calculations, xTB employs a minimal set of STOs expanded in contracted GTO basis functions.[145†]

The energy contributions in GFN1- and GFN2-xTB are summarized in Eq. (2.13).[‡] However, not all terms occur in both methods, which is why the contributions are not strictly grouped according to their order in the

[†] GFN1-xTB includes an additional *s* function for hydrogen beyond the MB scheme to improve hydrogen bonding.
[‡] The non-self-consistent GFN0-xTB[279] differs significantly from this scheme and shall not be discussed in this thesis.

Taylor expansion from Eq. (2.9), as indicated by the superscript index.

$$E_{\text{GFN}n\text{-xTB}} = \underbrace{E_{\text{rep}}^{(0)} + E_{\text{disp}}^{(0)[,(1),(2)]} + E_{\text{EHT}}^{(1)} + E_{\text{IES+IXC}}^{(2)} + E_{\text{IXC}}^{(3)}}_{\text{GFN1+GFN2}} + \underbrace{E_{\text{XB}}^{(0)}}_{\text{GFN1}} + \underbrace{E_{\text{AES+AXC}}^{(2)}}_{\text{GFN2}}. \tag{2.13}$$

The repulsion energy [cf. Eq. (2.6) in DFTB] is defined in an atom pair-wise fashion

$$E_{\text{rep}} = \frac{1}{2}\sum_{A,B}\frac{Z_A^{\text{eff}} Z_B^{\text{eff}}}{R_{AB}} e^{-\sqrt{\alpha_A \alpha_B}(R_{AB})^{k_f}}, \tag{2.14}$$

where α_A and Z_A^{eff} are empirical, element-specific parameters, with the latter roughly corresponding to effective nuclear charges. k_f is a global parameter. The EHT-type first-order term is in principle defined as follows

$$E_{\text{EHT}} = \sum_{\mu\nu} P_{\mu\nu} H_{\mu\nu}^{\text{EHT}} \quad \text{with} \quad H_{\mu\nu}^{\text{EHT}} = \frac{1}{2} K_{AB}^{ll'} S_{\mu\nu} (H_{\mu\mu} + H_{\nu\nu})[\cdots], \tag{2.15}$$

where $S_{\mu\nu}$ is the corresponding overlap integral element, and $K_{AB}^{ll'}$ is an empirical, shell-specific parameter for shells l and l' (s, p, ...).[†] The parameters $H_{\mu\mu}$ correspond to the diagonal elements of atomic shell energy levels, which depend on the chemical environment, e.g., via the coordination number (CN). The dots indicate further parameterization of $H_{\mu\nu}^{\text{EHT}}$ specific to GFN1 and GFN2. Detailed definitions are provided in Sec. B.5.1. $H_{\mu\nu}^{\text{EHT}}$ serves as the analogue to the effective Hamiltonian in DFTB1 (Eq. (2.4)), where a significant portion of the empirical parameters in GFNn-xTB is embedded, even though their origin is fundamentally different. The isotropic electrostatics (ES) and XC energies in second and third order remain qualitatively similar to their DFTB counterparts [Eqs. (2.10) and (2.12)]. However, the γ function used for damping Coulomb interactions at short distances follows the Ohno-Klopman form:[226,280]

$$\gamma_{AB,ll'} = \frac{1}{\sqrt{R_{AB}^2 + \eta_{AB,ll'}^{-2}}}. \tag{2.16}$$

In GFN1-xTB, $\eta_{AB,ll'}$ is defined as the harmonic mean (GFN2-xTB: arithmetic average) of effective shell hardness values of the form $\eta_A\left(1 + k_A^l\right)$, where η_A is an empirical atomic hardness and k_A^l is a shell-specific scaling parameter. The third-order term [see Eq. (2.12)] appears only as an on-site contribution, making Γ_A an atomic parameter, which in GFN2-xTB is additionally scaled shell-wise. Beyond the isotropic second-order interactions, GFN2-xTB also incorporates anisotropic ES and XC interactions by expanding the atom-centered contributions up to second order in a multipole expansion.[‡] This partially compensates for BSEs caused by the absence of polarization functions, eliminating the need for explicit hydrogen or halogen bond corrections. Similar to E_{IES}, E_{AES} employs a damping function to prevent divergence at short distances. To improve the description of weak halogen bonds in view of the missing anisotropic contributions in GFN1-xTB, an explicit zeroth-order (purely geometry-dependent) correction term E_{XB} is introduced (cf. Ref. [281] for a related approach). The explicit formulation of E_{AES} and E_{XB} is provided in Sec. B.10. E_{disp} represents the semiclassical dispersion correction in both methods, as introduced in Sec. 1.1.2. GFN1-xTB employs the D3(BJ) dispersion correction, which is a zeroth-order contribution since it is applied as a post-SCF correction. In contrast, GFN2-xTB uses the charge-dependent D4 correction, which is incorporated self-consistently via

[†] For certain elements, $K_{AB}^{ll'}$ in GFN1-xTB depends on the specific element pair AB, representing an exception to the element-wise parameterization philosophy outlined above.

[‡] This introduces charge-dipole, charge-quadrupole, and dipole-dipole interactions, in addition to the charge-charge interaction in the E_{IES} term.

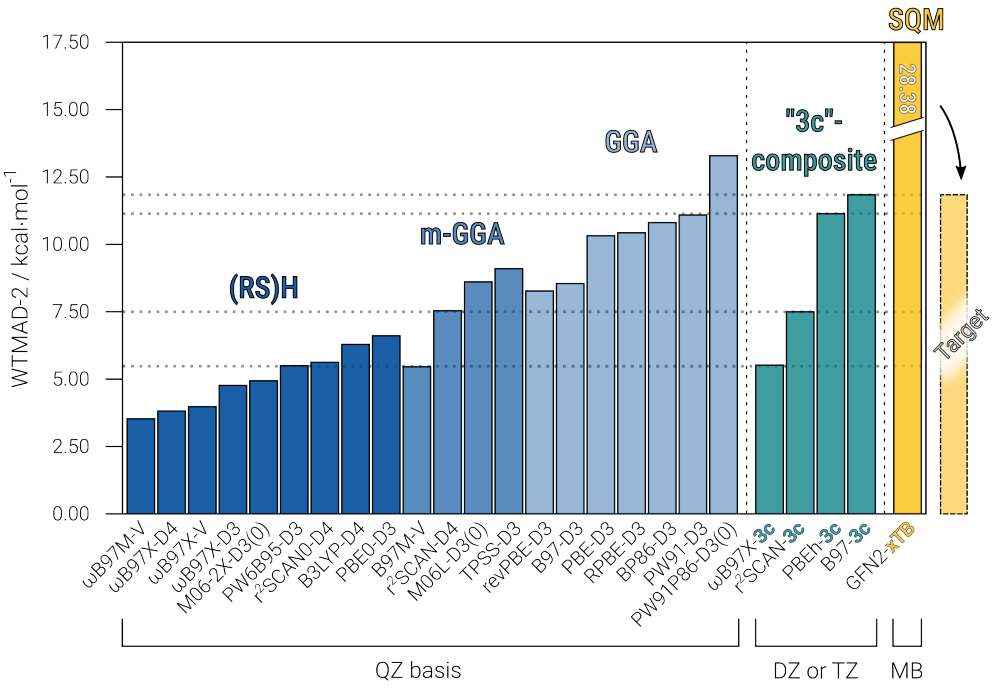

Figure 2.3: Weighted total mean absolute deviation (WTMAD-2) with respect to reference data for various DFAs including "3c" composite methods, and GFN2-xTB on the GMTKN55 benchmark set for general main-group thermochemistry.[93] The approximate target accuracy ("*DFT accuracy at TB speed*") is shown in transparent color. See Ref. [93] for details on the WTMAD-2 measure. The dashed horizontal lines mark the performance of the "3c" composite DFT methods.

Mulliken charges[282] derived from the density matrix.

2.3 Interim Conclusion on the Shortcomings of SQM Methods

A substantial number of benchmark studies in recent years has demonstrated that SQM methods generally do not achieve the accuracy of standard DFAs, which represent the workhorses of modern computational chemistry.[8,283–285] While SQM methods can describe geometries and NCIs of neutral systems with reasonable accuracy,[116,286–289] they struggle significantly with relative conformer rankings, thermochemistry involving bond dissociation, the description of spin states, and the treatment of charged or strongly polarized systems.[8,283–285] As an example for this challenge, Fig. 2.3 compares the errors of various DFAs with those of the semiempirical GFN2-xTB method on the well-established general main-group thermochemistry, kinetics, and noncovalent interactions (GMTKN)55 benchmark set. GFN2-xTB, one of the most recent SQM developments, exhibits errors that are, on average, two to ten times larger than those of standard DFAs. Even lower-rung GGA functionals outperform GFN2-xTB in accuracy. Extensive benchmarking over the past years has shown that this is not unique to GFN2-xTB but extends to various SQM methods, regardless of whether they originate from HF or DFT.[8,283,290–292] Rather than delving into the individual strengths and weaknesses of specific semiempirical methods, we emphasize that they have yet to achieve the overarching goal of "*DFT accuracy at semiempirical speed*" for the applications described above.

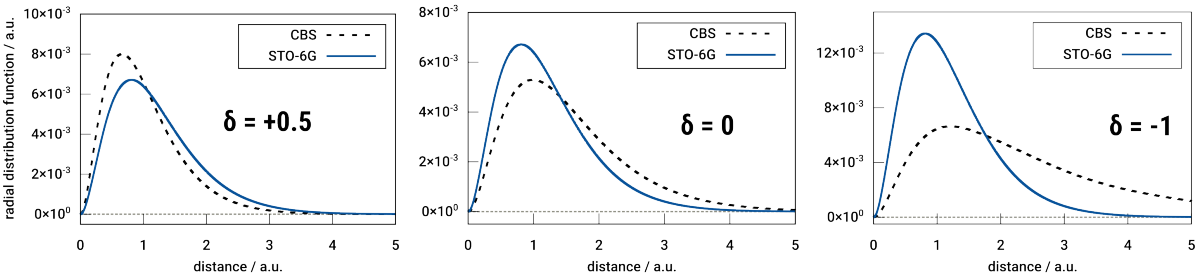

Figure 2.4: Radial distribution function of the electron density for hydrogen with charges of +0.5 (*left*), 0 (*middle*), and −1 (*right*), computed using the RSH DFA ωB97X-V with different basis sets.[103] *CBS* refers to an uncontracted 14*s* basis set from the TURBOMOLE basis set library.[294,295]

Having reviewed the historical evolution of SQM methods up to the present, we can now identify their remaining shortcomings. Some of these limitations apply to all SQM methods, while others are specific to the more recent DFTB-based variants.

Choice of Basis Set All SQM methods discussed in Sec. 2.2 are formulated using some form of MB set. MBs impose significant limitations on the MO expansion by restricting it to one fixed Gaussian function per AO, regardless of the underlying electronic structure theory. In other words, the accuracy of the results is fundamentally constrained by the choice of basis set.

Fig. 2.4 demonstrates the limitation of using a fixed single Gaussian function per AO through a simple example: the radial distribution function of the electron density for a hydrogen atom in different charge states. The employed MB, STO-3G,[293] fails to capture the contraction and expansion of the electron density in response to positive and negative charges, respectively. Beyond charge effects, additional environmental influences that modify AO shapes – such as compression due to spatial proximity of neighboring atoms (e.g., in covalent bonds) – cannot be accounted for in MBs. The impact of using a MB compared to various multiple-ζ basis sets on a calculated barrier height is illustrated in Fig. 2.5. The MB sets MINIS[296] and STO-3G[293] fail to describe the transition state correctly, erroneously predicting it as a local minimum. While basis sets of TZ quality and larger converge toward the reference barrier height, even the simplest DZ basis set already yields a positive barrier. Adding polarization functions, as in SVP, further reduces the error, highlighting their importance. This trend is supported by the mean absolute errors (MAEs) over a representative benchmark set, where the most significant improvement is observed when transitioning from MBs to a (polarized) DZ basis set. These limitations of MBs partially explain the insufficient accuracy of GFN2-xTB in the GMTKN55 benchmark set in Fig. 2.3.

A central objective of this thesis is the development of solutions to address basis set deficiencies in SQM methods. Two strategies are conceivable.

First, the development and implementation of a multiple-ζ basis set within an SQM framework. This approach is pursued in Chapters 3 and 4, where a new small-basis-set composite DFT method is introduced as a spin-off toward a multiple-ζ SQM method, which is realized in form of the PTB method. However, MBs have been employed in SQM methods for almost a century for good reasons: Introducing more than one AO shell complicates the parameterization of AO interactions, as multiple interaction terms per element and angular momentum combination must be considered. Furthermore, element pair-wise parameterized interactions that do not differentiate between AO shells do not benefit from a multiple-ζ basis set at all.

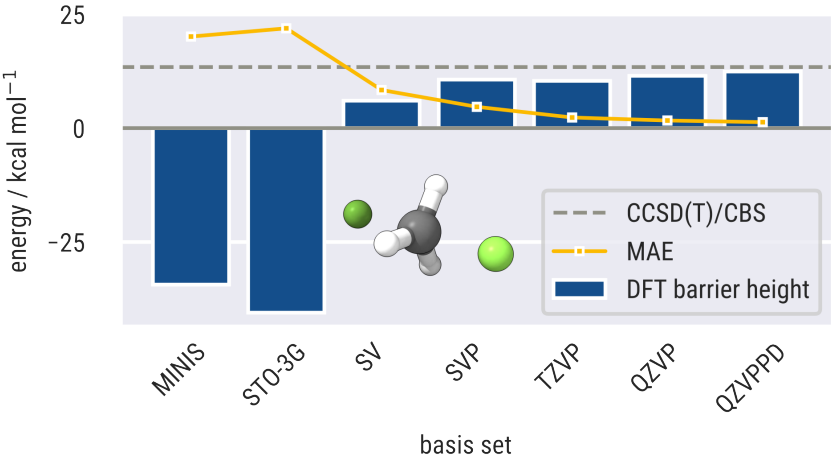

Figure 2.5: Barrier height of the S_N2 fluorine exchange reaction in fluoromethane (CH_3F) computed with the RSH DFA ωB97M-V[104] using different basis sets (blue bars).[146,211,295,296] The SVP, TZVP, QZVP, and QZVPPD basis sets correspond to the def2 family.[139,149] The theoretical reference at the W2-F12 level is indicated by the gray dashed line.[297] The orange line represents the MAE across the entire BH76 benchmark set for reaction barrier heights in small organic molecules, which also includes the shown example reaction.[93,298]

Chapters 5 and 6 describe an alternative approach, where a consistent and easily applicable framework for an improved MB is developed that retains the accuracy of multiple-ζ basis sets.

Reference Data and Target Properties In general, the predictive accuracy of semiempirical methods is ultimately constrained by the quality and consistency of the reference methods used during parameterization. Early SQM approaches relied on experimental data, either as targets or directly as parameters. While experimental data may seem ideal, they present significant drawbacks. First, their availability is limited, and generating new data is costly in both time and resources. Second, isolating the electronic energy from other contributions, such as solvation effects, free energy corrections, and finite-temperature effects, is highly challenging, making a *pure* representation of the electronic energy difficult.[†] Later approaches sought to emulate either HF or DFT interactions and properties, circumventing these complications. DFTB even replaces entire integrals with tabulated DFT data. Meanwhile, xTB employs theoretical reference data of varying accuracy during parameterization, ranging from WFT to GGA DFT, depending on the property.

In Chapter 4, we introduce a novel approach in which an SQM method is specifically designed to reproduce the results of a consistent and highly accurate RSH DFT calculation. Instead of directly fitting parameters to minimize deviations from target properties, this method aims to replicate the converged electron density matrix of the reference method. Thus, the work also represents an attempt to incorporate non-local information into a DFTB-based framework. While some HF-based methods discussed in Sec. 2.2.2 account for one-center exchange integrals, DFTB and xTB effectively mimic only GGA DFT.

Another key aspect is incorporating data beyond conventional chemical space. This can be achieved by expanding parameterization and validation datasets to include not only chemically relevant molecules but also "mindless" molecules – automatically generated structures that do not necessarily adhere to typical bonding motifs.[301] This strategy enhances transferability and is explored to varying degrees in Chapters 4, 5, and 6.

[†] This distinction is crucial to ensure that temperature-dependent corrections[299] and environmental effects (e.g., solvent models)[300] can be applied independently without requiring reparameterization of the electronic structure method.[164]

CHAPTER 3

ωB97X-3c: A Composite Range-Separated Hybrid DFT Method with a Molecule-Optimized Polarized Valence Double-ζ Basis Set

Marcel Müller,[†] Andreas Hansen,[†] Stefan Grimme[†]

Received: November 01, 2022
Published online: January 03, 2023

Reprinted in Appendix C with permission[§] from
M. Müller, A. Hansen, and S. Grimme, *ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set*, The Journal of Chemical Physics **158** *1* (2023) 014103, DOI: 10.1063/5.0133026

Own contributions

- Conceptualization of the composite DFT method
- Implementation into the ORCA QM code, support of the TURBOMOLE implementation, and proliferation of the vDZP basis set
- Performing all calculations
- Interpretation of the results
- Writing of the manuscript

[†] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany
[§] Permission requests to reuse material from this chapter should be directed to AIP Publishing.

Originally, the "3c" composite methods were designed for specific properties such as geometries and interaction energies. With r^{2}SCAN-3c,[214] employing a modern meta-GGA DFA,[66,67] broader applications, including main group thermochemistry and accurate conformer energies, have become feasible. However, as an m-GGA DFA, r^{2}SCAN-3c exhibits significant amounts of SIE, limiting its accuracy for certain applications, such as polar molecules, barrier heights, and large biomolecules.

ωB97X-3c addresses this issue by providing an efficient small-basis-set method capable of accurately handling such cases.[1] While following the "3c" paradigm, it incorporates several key innovations. The first novelty is the use of the RSH DFA ωB97X-V, which ranks among the most accurate DFAs, particularly for main-group thermochemistry.[93,103] The three components of this small-basis-set variant of ωB97X-V are: *(i)* the novel polarized valence double-ζ (vDZP) basis set, *(ii)* matching large-core ECPs, and *(iii)* the established D4 dispersion correction, replacing the default VV10 dispersion correction. For the first time in a "3c" approach, vDZP has been developed entirely from scratch, rather than modifying existing basis sets from the MINIS[296] or def2[139] families. Unlike most conventional Gaussian basis sets, its exponents and coefficients were variationally optimized in molecular DFT calculations for both neutral and ionic systems, rather than in atomic calculations. This molecular optimization, combined with a deep contraction of the AOs, significantly improves the description of the electron density in bound and charged species, as demonstrated for H^-. Additionally, this strategy drastically reduces BSSE to levels comparable to conventional triple-ζ basis sets. Small residual BSSE effects were absorbed into the D4 parameterization, which was optimized on a broad set of NCI databases.[93,285,292,302–306] As a by-product, we also present new D4 parameters for ωB97X-V (replacing the VV10 contribution) in a converged basis set, significantly improving previous parameterizations.[307] The large-core ECPs reduce computational cost substantially, particularly for heavy elements. For an appropriate description of metallic systems, vDZP includes additional semi-core electrons (*s* and *p* shell) for groups 1-12, while main groups 13-18 retain a strict valence-only basis. Originally, the combination of the deeply contracted vDZP basis set and matching large-core ECPs was developed for the SQM method PTB (see Chapter 4), where Hamiltonian diagonalization – scaling with the number of contracted basis functions – constitutes the computational bottleneck. By combining ωB97X-V with the accurate vDZP basis set, ωB97X-3c eliminates the need for the more empirical gCP and SRB corrections employed in earlier "3c" methods.

ωB97X-3c rivals standard hybrid DFT/QZ methods on thermochemistry benchmarks such as GMTKN55, achieving a WTMAD-2 of 5.5 kcal·mol^{-1} at a fraction of the computational cost, resulting in a speed-up by a factor of three to five. For NCI benchmarks such as S30L[306] and ACONFL,[292] it matches or surpasses the accuracy of the parent functional in the complete basis set.[93,307,308] By mitigating SIE, ωB97X-3c enhances reliability for supramolecular complexes and large biomolecules, where (semi)local DFT methods often fail due to self-interaction errors and associated convergence issues. For the same reason, transition states as typical SIE cases represent another key application of ωB97X-3c. Furthermore, since memory requirements in DFT codes scale with the number of AOs, ωB97X-3c's vDZP basis saves noticeable amounts of memory for large-scale computations. A minor limitation arises in transition metal thermochemistry, where the use of large-core ECPs results in slightly lower accuracy for these systems.

Thanks to its ability to accurately describe NCIs in large systems, ωB97X-3c has been employed to benchmark SQM and FF NCI energies for structures containing up to 2000 atoms.[116] In a recent study, it has also been applied to investigate reaction paths and transition states in the context of mass spectrometry simulations.[284] ωB97X-3c completes the "3c" family by addressing thermochemistry for large systems. It is particularly advantageous in cases where an efficient yet accurate non-local DFT method virtually free of SIE is essential. ωB97X-3c is implemented in several QM packages, such as ORCA,[309] TURBOMOLE,[294] and Psi4,[310] while the vDZP basis set is additionally distributed on an open-source basis set platform.[152]

CHAPTER 4

A Non-Self-Consistent Tight-Binding Electronic Structure Potential in a Polarized Double-ζ Basis Set for all *spd*-Block Elements up to Z=86

Stefan Grimme,[†] Marcel Müller,[†] Andreas Hansen[†]

Received: December 06, 2022
Published online: March 22, 2023

Reprinted in Appendix D with permission[§] from
S. Grimme, M. Müller, and A. Hansen, *A non-self-consistent tight-binding electronic structure potential in a polarized double-ζ basis set for all spd-block elements up to $Z = 86$*, The Journal of Chemical Physics **158** *12* (2023) 124111, DOI: 10.1063/5.0137838

Own contributions

- Translating the development code into equations
- Implementation into the `xtb` code
- Performing all calculations and generation of the results
- Interpretation of the results
- Writing of the manuscript

[†] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany
[§] Permission requests to reuse material from this chapter should be directed to AIP Publishing.

The xTB methods, introduced in the late 2010s, have become widely used, robust, and established tools for a variety of applications, including geometry optimizations,[171,311,312] the calculation of thermostatistical corrections,[164,313] and conformer searches.[8,30,165] Despite their versatility, their accuracy for typical thermochemistry problems involving bond breaking or transition states remains distinctly lower and is often insufficient.[259,283] Additionally, significant inaccuracies in IR intensities were reported for many test cases.[314] A key limitation underlying these issues is the use of an MB set in all xTB methods. Developing a TB method applicable to thermochemical problems and capable of completely replacing low-cost DFT has thus become a primary target.

To address these limitations, PTB (density matrix, **P**, Tight-Binding) has been developed with a strategy that diverges significantly from previous approaches: Rather than fitting empirical Hamiltonian parameters to reproduce target properties such as geometries, frequencies, and NCI energies, the PTB parameters were optimized to emulate the converged one-particle density matrix of the reference DFT method in the same basis set. This approach builds on the concept of separating energy and potential, as pioneered by Burke *et al.* in density-corrected DFT.[315] The empirical Hamiltonian parameters were fitted to electronic properties directly related to **P**, such as shell populations, atomic charges, bond orders, dipole and second moments, and their derivatives. The fitting loss function also included the energy difference between a DFT energy expression evaluated with the PTB density and the converged ωB97X-3c density. A further innovation in PTB is its reference method. Instead of local DFT, PTB aims to approximate the ωB97X-3c RSH composite method, as presented in Chapter 3. Like ωB97X-3c, the PTB Hamiltonian is expanded in the vDZP basis, marking the first time an SQM method employs a basis set with more than one basis function per AO shell (i.e., not an MB set). PTB operates as a non-self-consistent method, requiring a fixed number of two Hamiltonian matrix diagonalizations (three for response properties), eliminating any SCF convergence issues. We found that self-consistent solutions within TB-SQM theory remain robust only when using MBs, as extended basis sets like vDZP introduce flexibility that cannot be reliably exploited by approximate SQM Hamiltonians. The PTB Hamiltonian follows the construction principles of xTB and DFTB, but incorporates additional non-local terms dependent on **P** to account for the RSH reference DFA. Notably, it requires only AO overlap integrals, following the tradition of the xTB methods. Via an additional diagonalization, the wave function's response to static external electric fields is approximated, enabling the calculation of polarizabilities and derived properties such as Raman activities. PTB has been parameterized for all elements up to $Z = 86$, excluding the lanthanides.

In thorough tests, the method achieves absolute errors of only 0.02 e^- in atomic charges and relative errors of approximately 5-10 % in dipole moments and polarizabilities. This performance is consistent even for "mindless" molecules (MLMs) consisting of electronically more complex transition metals and for highly charged systems. Compared to the xTB methods, PTB demonstrates significantly higher accuracy for derived electronic properties, such as IR and Raman intensities. Coupling PTB intensities with GFN2-xTB frequencies yields vibrational spectra approaching DFT quality. Its speed-up of 10^3 to 10^4 compared to conventional DFT enables efficient calculation of Raman spectra for large systems, such as Crambin, a protein with 327 atoms. Despite requiring only a few hours instead of several days, PTB closely matches the ωB97X-3c reference spectrum. While PTB can successfully emulate the ωB97X-3c one-particle density matrix, a dedicated TB energy expression could not be established. However, early tests indicated that combining the PTB density matrix with simple DFA energy expressions, such as RPBE,[316] can produce results comparable to those of the converged density.

PTB represents a significant advancement in exploring the potential of TB methods with extended basis sets and in emulating DFT electron density. It is natively implemented in the open-source `xtb` package.[259]

CHAPTER 5

An Atom-in-Molecule Adaptive Polarized Valence Single-ζ Atomic Orbital Basis for Electronic Structure Calculations

Marcel Müller,[†] Andreas Hansen,[†] Stefan Grimme[†]

Received: August 15, 2023
Published online: October 25, 2023

Reprinted in Appendix E with permission[§] from
M. Müller, A. Hansen, and S. Grimme, *An atom-in-molecule adaptive polarized valence single-ζ atomic orbital basis for electronic structure calculations*, The Journal of Chemical Physics **159** *16* (2023) 164108,
DOI: 10.1063/5.0172373

Own contributions

- Conceptualization of the minimal basis set approach
- Development of an `ORCA` input generator for q-vSZP
- Initial implementation of the Charge Extended Hückel (CEH) charge model into `tblite`
- Performing all calculations and generation of the results
- Interpretation of the results
- Writing of the manuscript

[†] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany
[§] Permission requests to reuse material from this chapter should be directed to AIP Publishing.

Chapter 4 discussed that extended basis sets containing more than one basis function per AO shell are incompatible with self-consistent SQM methods.[2] However, a fundamental limitation of conventional MBs compared to extended basis sets is their inability to "breathe" in response to the chemical environment. For example, DZ basis sets allow the LCAO approach to flexibly combine more diffuse or compact AOs of the same angular momentum on a given atom, enabling the atomic electron density to adapt dynamically. This flexibility is crucial for achieving higher accuracy, particularly in thermochemical applications.

The seeming contradiction between the benefits of extended basis sets and the necessity of MBs in SQM methods inspired the development of the charge-dependent (polarized) valence single-ζ (q-vSZP) basis. This basis set aims at achieving the accuracy of DZ or larger basis sets while maintaining a formally minimal structure, limiting the degrees of freedom within the LCAO during SCF procedures. Despite its single-ζ character, each AO in q-vSZP is deeply contracted, consisting of 4 to 8 primitive functions for valence orbitals. q-vSZP emulates the flexibility of extended basis sets by incorporating the CN and the atomic partial charge (q) as descriptors, which account for the chemical environment of each atom. These descriptors are used to modify the AOs of each symmetry-unique atom individually, making q-vSZP atom-in-molecule-specific. The atomic partial charge is computed using a novel semiempirical charge model called Charge Extended Hückel (CEH), while the CN is a purely geometric property. CEH is based on an EHT-like Hamiltonian, fitted to reproduce RSH DFT[104] Hirshfeld[317] charges, and will be introduced in detail in Chapter 6. A simple formula employing three empirical prefactors adjusts the effective contraction coefficients of the AOs, which incorporates linear dependencies on q and the CN, a quadratic q term, and a linear CN-q cross-term. Crucially, only the contraction coefficients are parameterized, not the Gaussian exponents, ensuring that the range of exponents within the contracted AOs remains fixed. This approach avoids excessively compact or diffuse functions due to extreme charges or CNs. Several features of the vDZP basis set presented in Chapter 3 were incorporated into q-vSZP. All basis set parameters, including Gaussian exponents, coefficients, the coefficients' dependency on the effective charge, and empirical parameters defining the effective charge from CN and q, were variationally optimized in molecular DFT calculations, including ions. In line with the valence-only approach of TB methods, q-vSZP employs large-core ECPs. Unlike vDZP, groups 1–12 are also treated as valence-only due to this approach. The q-vSZP basis set has been published for all *spd* block elements up to radon, with a later update (see Chapter 6) extending support to Fr, Ra, the lanthanide series, and the actinides.

The comparison of radial electron densities for H and H^- illustrates the advantage of q-vSZP over conventional MBs. While MBs yield identical radial distribution functions for both charge states, q-vSZP adapts to the negative charge, producing a significantly expanded distribution for H^-. Another benefit of the charge-dependent parameterization of basis functions is their response to external electric fields. The CEH model's charges are electric field-dependent, influencing the effective AOs and enabling q-vSZP to achieve isotropic dipole polarizabilities approaching QZ level. The variational optimization for molecular systems ensures well-defined AOs, resulting in low BSSE comparable to the vDZP basis set and close to that of standard triple-ζ basis sets. For the established GMTKN55 database,[93] q-vSZP demonstrates a reasonably accurate WTMAD-2 of 12.6 kcal·mol^{-1}, comparable to the standard split-valence def2-SVP[139] basis set. In contrast, conventional MBs achieve significantly higher WTMAD-2 values, ranging from 30 to 50 kcal·mol^{-1}. Similar improvements are observed in molecular structure determination.

q-vSZP, as the first generally applicable atom-in-molecule adaptive MB available for up to $Z = 103$, paves the way for more accurate TB methods in the future. Given efficient analytical gradient implementations for environment dependent Gaussian basis sets, q-vSZP will be a relevant tool for QM/QM workflows with mixed basis sets even in DFT applications.

CHAPTER 6

Advanced Charge Extended Hückel (CEH) Model and a Consistent Adaptive Minimal Basis Set for the Elements Z=1–103

Marcel Müller,[†] Thomas Froitzheim,[†] Andreas Hansen,[†] Stefan Grimme[†]

Received: October 15, 2024
Published online: November 25, 2024

Reprinted in Appendix F with permission[§] from
M. Müller, T. Froitzheim, A. Hansen, and S. Grimme, *Advanced Charge Extended Hückel (CEH) Model and a Consistent Adaptive Minimal Basis Set for the Elements Z = 1–103*, The Journal of Physical Chemistry A **128** *49* (2024), Publisher: American Chemical Society 10723, DOI: 10.1021/acs.jpca.4c06989

Own contributions

- Conceptualization of improvements
- Performing all calculations and generation of the results
- Interpretation of the results
- Writing of the manuscript

[†] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany

Models for atomic partial charge assignment have a long history, with Mulliken population analysis,[282] Hirshfeld density partitioning,[317] and natural population analysis (NPA)[318] being prominent examples.[319] These methods traditionally relied on computationally expensive analyses of converged wavefunctions from DFT or WFT. However, efficient atomic partial charge determination by classical or SQM methods is essential for applications such as semi-classical dispersion corrections,[119,120,320] polarizable FFs,[189,321,322] and as descriptors for ML applications.[323] Chapter 5 presented another computational chemistry application: the charge-dependent, atom-specific parameterization of basis functions. For this application, highly efficient charge assignment is crucial to ensure that the setup of the basis set does not become a computational bottleneck. While SQM approaches like xTB or PTB offer robustness and relatively accurate atomic charges,[2] their computational cost made them unsuitable for this purpose. Conversely, classical charge equilibration (QEq) models, such as EEQ,[120,320] often suffer from artificial CT in NCI complexes with significant electronegativity (χ) differences.

To bridge this gap, we developed the Charge Extended Hückel (CEH) model, as noted in Chapter 5. With a single diagonalization step in an MB, CEH is approximately 10 to 20 times faster than GFN2-xTB while adhering to basic QM principles. Thereby, it follows the philosophy of an earlier method called sTDA-xTB.[324] The CEH Hamiltonian is based on an EHT-type framework and is expanded in Stewart's Gaussian expansion of Slater-type orbitals.[145] A feature adapted from past SQM methods is the transformation of AO overlap integrals into the diatomic frame, with integral elements scaled based on interaction type.[247,325,326] The Hamiltonian's atomic shell energy levels are parameterized using a χ-weighted CN to partially account for electrostatic screening effects. As a significant improvement over the initial CEH model presented in Chapter 5, we introduce a purely geometry-dependent approximation for a pseudo self-consistent solution in this work: pseudo atomic charges, which do not necessarily sum up to the total molecular charge, are assigned based on χ-weighted CN values. Additionally, the total charge Q_{tot} is distributed equally. This enables the inclusion of second- and third-order TB ES in the Hamiltonian, capturing long-range effects absent in the zeroth-order formulation. After diagonalization, final atomic charges are computed from the density matrix via Mulliken population analysis.[282] To emulate mild static correlation effects, valence MOs are populated according to Fermi-Dirac statistics at an elevated temperature of 4000 K, effectively mitigating artificial CT between fragments with close-lying energetic states and thus a small fundamental gap. The CEH model was parameterized against reference DFT Hirshfeld charges[317] at the ωB97M-V[104]/def2-TZVPPD[149] level for elements up to lawrencium ($Z = 103$). Remarkably, the *f* electrons are treated explicitly as part of the valence space for the actinides, a rare feature for SQM methods.

Validation against randomly selected molecules from the PubChem database demonstrates an MAE of 0.019 e^- for CEH, outperforming all compared methods, including the computationally more demanding xTB methods. For MLMs, which provide a stringent transferability test as they do not follow any chemical bias, the MAE increases to 0.077 e^- but remains the lowest among all tested methods. For actinide complexes from the "lnqm" database,[320] CEH exhibits a higher correlation with reference charges than the established EEQ model. Comparisons of the fit RMSE over comparable training data indicates a consistent accuracy across the periodic table.

CEH is a valuable general charge model, particular in cases, for which classical QEq models fail due to their inherent limitations. With its implementation in the openly accessible `tblite` library, CEH provides convenient access to robust atomic charges. However, the computationally complex analytic derivative of the charges with respect to the nuclear positions due to the non-self-consistent QM approach represents a shortcoming of CEH in comparison to QEq models.

CHAPTER 7

Summary and Outlook

Over the past decades, discovery and innovation in chemistry have expanded beyond case studies of individual molecules and reactions to large-scale and high-troughput screenings (HTSs). This transformation, particularly in theoretical research, underscores the demand for computational methods that balance accuracy with efficiency. The work presented in this dissertation has contributed to this goal by advancing electronic structure methods that optimize this trade-off. Fig. 7.1 introduces a new perspective on these developments, illustrating the relationship between computational cost and accuracy in electronic structure theory and complementing Fig. 2.1. For relatively small systems, *the right answer for the right reason* can be obtained using coupled-cluster or related wave function theory (WFT) techniques. Kohn-Sham-density functional theory (DFT) offers a computationally more efficient alternative and is routinely applicable to systems of up to a few hundred atoms. While it has become the workhorse of computational chemistry[126] over the past 30 years, for larger systems or HTS approaches, access to large-scale high-performance computing resources becomes necessary or DFT is not applicable anymore at all. This is where semiempirical quantum mechanical (SQM) methods play a critical role, which have been a cornerstone of computational chemistry for nearly a century. While their accuracy is typically lower than DFT, they achieve speed-ups of three to four orders of magnitude compared to first-principles methods. The increased efficiency enables calculations on systems containing tens of thousands of atoms or databases comprising millions of small to mid-sized molecules in a matter of days. Novel machine learning potentials (MLPs) are a relatively recent addition to this landscape and are included for reference, as well as conventional force fields (FFs). Consequently, the different method categories in Fig. 7.1 form a *Pareto front*,[†] balancing accuracy against computational cost. Thus, the overarching goal of this thesis has been to advance methodologies that push the *Pareto front* closer to the *heaven of computational methods*.

After an introduction into the relevant basics of electronic structure theory in Chapter 1, Chapter 2 discussed how the Gaussian basis functions in the standard linear combination of atomic orbitals (LCAO) expansion represent a key limiting factor in both small basis set composite DFT methods, such as the "3c" series, and even more so in SQM methods. While composite DFT methods can be refined through a more sophisticated balance between computational efficiency and mitigating excessive basis set errors, SQM methods have traditionally relied on conventional minimal bases (MBs), restricting the LCAO expansion to a fixed atomic

[†] The Pareto front represents the set of Pareto efficient solutions that achieve the best possible balance between accuracy and computational cost. A solution is Pareto efficient if no improvement in one dimension (e.g., accuracy) can be achieved without worsening another, such as computational affordability.[327]

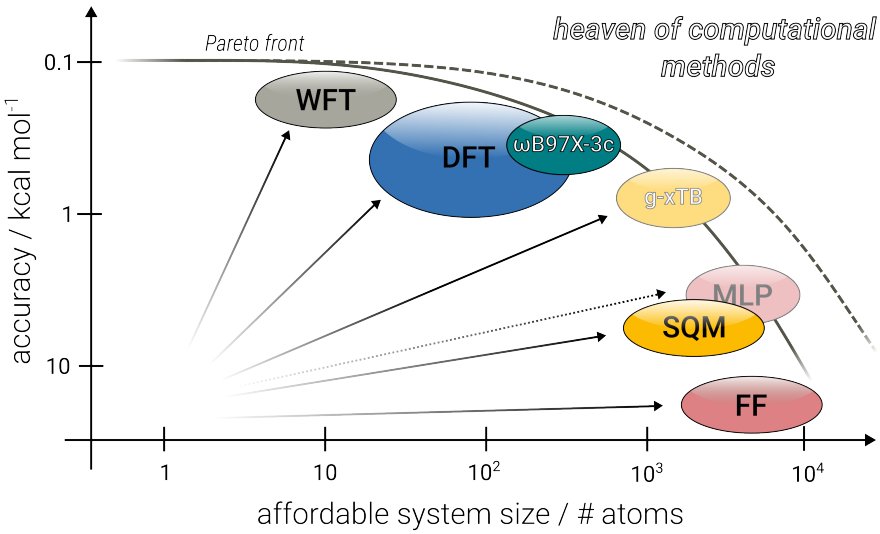

Figure 7.1: Accuracy vs. affordability of different categories of electronic structure and simulation methods. MLPs are, like FFs, not able to describe the electronic structure of a molecule but were included here due to their comparable field of application.[†]

orbital (AO) per shell. Fundamentally, most developments that were presented in this work originate from the central idea of enhancing the basis functions used in wavefunction expansions.

Most conventional basis sets were optimized for atomic situations, overlooking the fact that optimal AOs for molecular calculations – the primary application – can differ significantly. This limitation is particularly pronounced in small or reduced basis sets, such as those used in previous composite DFT and SQM methods, where the molecular orbital (MO) optimization space is inherently constrained. Crucially, improving basis set optimization does not introduce additional computational complexity, as it merely refines predetermined parameters. In fact, it can reduce resource demands by eliminating the need for formally larger basis sets, as better-fitted AOs provide enhanced accuracy with fewer functions. To leverage this potential, we developed the polarized valence double-ζ (vDZP) basis set, designed for both DFT and SQM applications. vDZP features deep contraction and variational optimization in molecular DFT calculations, ensuring well-matched AOs that accurately reflect typical molecular environments. Its first application was in the novel ωB97X-3c composite DFT method (Chapter 3), where vDZP enabled accuracy in thermochemistry and non-covalent interactions (NCIs) previously achievable only with triple-ζ or larger basis sets. A key advantage of vDZP is its extremely low basis set superposition error, which is particularly beneficial for accurately describing NCIs. Beyond reducing computational complexity and wall time, ωB97X-3c also lowers memory requirements. Its integration of large-core effective core potentials and a specifically parameterized D4 dispersion correction aligns it with the "3c" composite DFT family, while simultaneously reducing empiricism by eliminating the need for more empirical corrections like the short-ranged basis and geometrical counterpoise corrections. ωB97X-3c has been successfully applied to large systems, including supramolecular complexes[116], and barrier height screenings in mass spectrometry predictions.[284] Overall, ωB97X-3c extends the practical reach

[†] The axes illustrate relative trends rather than absolute values. Specifically trained MLPs can achieve DFT-level accuracy for certain chemical systems but may fail outside their targeted chemical space, leading to their intermediate placement between traditional SQM methods and DFT.

of DFT to very large molecular systems while maintaining high accuracy, effectively shifting the *Pareto front* of DFT toward larger system sizes.

Building on vDZP, we extended our focus to improving SQM methods through enhanced basis sets. Chapter 4 introduced PTB, a novel tight-binding (TB) method rooted in density functional tight-binding and eXtended Tight-Binding (xTB). PTB represents a milestone in SQM development as the first method expanded in a double-ζ (DZ) basis set. A key advantage of vDZP is its deep contraction, which is particularly beneficial for SQM methods. Since their computational bottleneck is Hamiltonian diagonalization, which scales with the number of AOs, the increased number of primitive functions per AO comes at almost no cost. Additionally, we devised an innovative parameterization strategy for PTB: rather than minimizing the deviation of target properties, PTB's parameters were optimized to reproduce the density matrix of the reference range-separated hybrid DFT calculation in the same basis set – ωB97X-3c. This allows PTB to effectively emulate the converged MOs of the reference method. By combining the vDZP basis with this density matrix-driven approach, PTB achieves near-DFT accuracy for key electronic properties such as infrared (IR) and Raman intensities, which correspond to derivatives of the dipole moment and dipole polarizability, respectively. Additionally, properties directly derived from the density matrix, such as atomic charges and bond orders, closely match those from the reference density functional approximation. However, we found that developing a consistent and reliable energy expression within the xTB framework (see Sec. 2.2.3) remains challenging and, to date, infeasible. This realization – that a robust TB energy expression is not viable with extended basis sets – is a key finding of this thesis. Nevertheless, the highly accurate electron density obtained with PTB presents new opportunities. Early tests suggest its potential for improving density-corrected DFT, either by enhancing self-consistent field (SCF) convergence or enabling non-self-consistent DFT evaluations. Future research should explore its broader applications, including its use in machine-learning-driven predictive techniques.

As discussed in Sec. 2.3, conventional MBs pose a major limitation in electronic structure methods, particularly in SQM approaches where they remain widely used. However, the previous attempt to improve SQM methods with multiple-ζ basis sets has so far failed to yield a reliable energy expression. To address this dilemma, we developed the charge-dependent (polarized) valence single-ζ (q-vSZP) basis set as an advanced, molecular environment-adaptive MB (Chapter 5). The key shortcoming of conventional MBs, mostly developed 40 to 50 years ago,[293,296] is their inability to "breathe" in response to the chemical environment due to the restriction of a single basis function per AO shell. Unlike multiple-ζ basis sets, which inherently provide such flexibility, q-vSZP overcomes this limitation by incorporating atom-specific expansion and contraction based on the respective coordination number (CN) and atomic charge q. This adaptation allows the basis functions to dynamically respond to the molecular environment. For instance, strongly negatively charged moieties requiring more diffuse Gaussian functions benefit considerably from this adaptability. Additionally, q-vSZP inherits the key improvements from the vDZP basis set, particularly the variational optimization of deeply contracted AOs in molecular calculations. Although designed for future SQM methods, q-vSZP was rigorously validated through DFT calculations, providing an unbiased assessment of its quality. In these tests, q-vSZP achieved high accuracy on thermochemistry benchmarks, such as the GMTKN55 database, reducing errors by a factor of two to four compared to existing MBs.[93] Thus, q-vSZP provides an ideal starting point for developing a next-generation SQM method, offering a fundamentally improved MB, and consequently enabling access to a robust energy expression.

However, q-vSZP introduces additional preparational steps beyond those required for conventional basis sets. In standard electronic structure calculations, atomic charges are *output* quantities rather than *input* descriptors. In contrast, q-vSZP requires precomputed atomic charges with sufficient accuracy to ensure a reasonable AO setup. At the same time, this charge determination must be significantly faster than the

subsequent (S)QM calculation, excluding computationally more demanding semiempirical methods such as PTB and GFN2-xTB. Thus, the final calculation's accuracy directly depends on the quality of the input descriptors q and CN. While the CN is purely geometric, determining atomic charges for molecular systems is non-trivial. To address this challenge, we developed a new charge model rooted in extended Hückel theory that requires only a single diagonalization, as described in Chapter 6.[†] Focusing solely on atomic charge prediction, this Charge Extended Hückel (CEH) model achieves highly accurate atomic charges, outperforming more complex (but general-purpose) models such as GFN2-xTB. Yet, it is 10–20 times faster than GFN2-xTB due to the absence of iterative SCF steps.

Over the past two decades, density functionals and SQM methods have significantly improved in accuracy and applicability. However, the basis sets used in molecular ground-state DFT calculations and SQM methods have remained largely unchanged. Given their fundamental role in determining how well MOs approximate the true wavefunction, the advancements introduced with q-vSZP and vDZP represent important steps forward and will likely shape future developments in the field (see for example Ref. [161]). A logical extension is the adaptation of the environment-dependent q-vSZP scheme to a DZ basis set, yielding "q-vDZP". This approach could allow for a less deep contraction, which is particularly beneficial for DFT applications, as the number of integral evaluations is reduced. Furthermore, this approach can overcome inherent limitations of the MB design that environment-dependent contraction coefficients alone cannot resolve. In q-vSZP, AOs expand or contract only isotropically, lacking the anisotropic flexibility of DZ bases, which is relevant for accurately describing directional covalent bonds, for example.

The "3c" series also presents a clear path for further development: r^{2}SCAN-3c excels in tasks such as geometry optimizations and conformer ranking, provided that self-interaction error (SIE) is not an issue, while ωB97X-3c offers distinct advantages for highly accurate NCI energies and systems prone to SIE. However, ωB97X-3c comes with a noticeable increase in computational cost compared to r^{2}SCAN-3c. Earlier "3c" methods that do not suffer from SIE and are relatively efficient, such as HF-3c or PBEh-3c, fall short in thermochemical applications. This creates a strong demand for a true successor to HF-3c or PBEh-3c – one that is (almost) free of SIE, significantly more accurate than its predecessors, yet as efficient as r^{2}SCAN-3c. A new basis set from the *(q-)vXZP* series, developed during my PhD, provides a natural starting point for such a method.

With the insights gained from previous method developments and the highly accurate q-vSZP basis set, the foundation for a next-generation SQM method, a true successor to GFN*n*-xTB, has been established. A key lesson from developing and testing PTB and CEH is the inclusion of reference data beyond the usual chemical space. An essential component of such datasets is "mindless" molecules (MLMs), molecules that do not adhere to typical bonding motifs and can be generated automatically by random atomic placement and subsequent optimization. By enhancing transferability and minimizing bias from curated reference data, MLMs are poised to play a critical role in future method development, even beyond this project. At the time of writing, this GFN*n*-xTB successor – tentatively named *general* (g)-xTB – delivers highly promising results across all relevant application areas. While retaining and in some cases improving upon the performance of specialized xTB methods for geometries, frequencies, and NCIs, g-xTB extends its accuracy to thermochemical properties, including reaction energies, barrier heights, and isomerizations, reaching the precision of lower-rung DFT methods. Thus, it fulfills the target set in Fig. 2.3. Moreover, unlike preceding xTB variants, g-xTB also provides reliable predictions for IR and Raman spectra (following PTB in this

[†] Classical charge equilibration (QEq) models, such as the EEQ model used in the D4 dispersion correction, suffer from artificial charge transfer because charges are distributed globally, largely dictated by parameterized electronegativities. A study not yet published at the time of writing assessed a bond capacitor-based QEq model that mitigates this issue while remaining significantly faster than QM-based approaches.

regard), as well as for properties of electronically excited states such as the singlet-triplet energy gap (S_1-T_1).

In the context of Fig. 7.1, this advancement shifts the *Pareto front* for SQM methods significantly upwards, bringing computational chemistry closer to the long-standing goal of "*DFT accuracy at SQM speed*". The dashed *Pareto front* illustrates the desired scenario where more advanced composite DFT methods such as ωB97X-3c and the next-generation TB method g-xTB define the frontier of accuracy and efficiency. While modern MLPs can achieve DFT-level accuracy for specific chemical spaces by learning energies and forces from data, their lack of embedded QM principles currently prohibits capturing discrete energy levels, electronic excitation energies, and spin-state preferences, limiting their applicability. By contrast, SQM methods rooted in well-defined physical approximations provide an increased level of *interpretability* and *transferability*, ensuring robustness even for systems outside the training domain. Given these considerations, SQM methods are expected to remain indispensable, with an even brighter future ahead.

As electronic structure methods continue to refine the balance between computational cost and accuracy, the primary bottleneck in practical molecular simulations is shifting away from solving the single-determinant Schrödinger equation toward more complex challenges that lack broadly accessible solutions yet. These challenges include accurately modeling environmental effects such as solvent interactions in condensed-phase systems,[300,328,329] finite-temperature effects,[330,331] as well as tackling non-adiabatic transitions involving electron-nuclear coupling[332–334] and routine methods for multiconfigurational electronic states.[55,335,336]

Highly efficient *and* accurate SQM and DFT methods are key to addressing some of these and other emerging scientific challenges. This thesis has contributed to advancing the field by developing novel approaches that push the boundaries of accuracy and efficiency. Thus, we can anticipate groundbreaking advancements in computational chemistry that, even today, seem extraordinary.

Acknowledgements

During the past four years, many people have accompanied me on my PhD journey. Before naming some of them, I want to express how deeply honored I feel to have conducted my doctoral studies in such an inspiring environment – both scientifically and personally.

First and foremost, I want to thank my supervisor, Prof. Stefan Grimme, to whom I am grateful in many ways. Not only did he give me the opportunity to work in his research group, but he also placed his trust in me to collaborate with him on advanced semiempirical methods and related projects. I deeply appreciate the countless discussions on electronic structure theory, whether they involved fundamental concepts or seemingly subtle decisions during development of a method. In all cases, I have learned a lot from him – not only, but also about quantum chemistry.

I am also very grateful for the opportunity to have attended numerous conferences, some of which will remain unforgettable. One particularly fond memory is the trip to Vancouver for WATOC 2022, together with Stefan, Dr. Andreas Hansen, and Dr. Julius Kleine Büning, whose company I enjoyed very much. Other highlights include the European Summer School in Quantum Chemistry 2022 and the various meetings of the Priority Programme "Molecular Machine Learning", for which I am grateful to have been part of. Beyond the great company on several scientific meetings and proof-reading many of my manuscripts, I want to thank Andreas for always having an open ear and giving me important and insightful advices.

For the collaborations on scientific projects – both completed and ongoing – I want to express my deep gratitude to several people beyond those already mentioned: Thomas Froitzheim for his great expertise in electronic structure theory and for working with me on several projects, of which g-xTB is probably the biggest one. I have no doubt that Thomas will bring this project to great success, and I feel privileged to be part of it alongside Stefan and Andreas. Christian Hölzer, for our joint work on new ligand descriptors for Bayesian optimization, in collaboration with colleagues from the industry side (*vide infra*). Dr. Philipp Pracht, for giving me the opportunity to contribute to the latest CREST publication. Dr. Nico Fleck, for entrusting me with the theoretical studies for our joint publication. Abylay aka Albert Katbashev, for pioneering the collaboration with Prof. Thomas Kühne on mitigating the diagonalization bottleneck in SQM methods, using PTB as an example. Thomas Gasevic, for combining the efforts regarding a new publication on "mindless" molecules. In this context, I'd also like to thank Jonathan Schöps, who contributed a lot to this project during his focusing laboratory course. Christoph Plett, for the work on the periodic table-wide benchmark set, together with Robin Dahl and Vanessa Kniebes. Benedikt Bädorf, for continuing the development of environment-adaptive and molecular-optimized basis sets. Leopold Seidler, for his valuable assistance with parallelization techniques in Python. And finally, Prof. Gerhard Erker, for welcoming me into two of his studies on the stereochemical behavior of frustrated Lewis pairs and for supporting me in my application for the Kekulé fellowship. At this point, I also want to sincerely thank the "Fonds der Chemischen Industrie (FCI)" for awarding me the fellowship, which co-funded my doctoral studies.

Of course, I will also have special memories of my office mates Dr. Philipp Pracht, Dr. Marcel Stahn, and Tim Schramm, whose company I truly enjoyed. In addition, I want to thank Philipp and Dr. Sebastian Spicher for supervising my very first projects. At this point, I extend my gratitude to all colleagues with whom I shared my time at the Mulliken Center for Theoretical Chemistry, making this journey unforgettable, especially those not yet mentioned: Fabian Bohle, Marvin Friede, Johannes Gorges, Lukas Kunze, Sarah Löffelsender, Thomas Rose, Christian Selzer, Lukas Wittmann, Dr. Markus Bursch, Dr. Sebastian Ehlert, Dr. Julia Kohn, Dr. Jeroen Koopman, Dr. Hagen Neugebauer, Dr. Zheng-Wang Qu, Dr. Hui Zhu, and Prof. Marc de Wergifosse. I am particularly glad that the pub quiz tradition will likely outlast my time here, and I hope the same holds for our table tennis tournaments, which kept us busy during countless coffee breaks – especially Marvin, Thomas G., Thomas R., Benedikt, Christian S., and Christoph (please forgive me if I underestimated some people's passion for table tennis at this point). For administrative and technical support, I want to thank Claudia Kronz and Jens Mekelburger. A special thanks goes to Dr. Jan-Michael Mewes, who loves passionate discussions – scientific or otherwise – as much as I do. I also want to thank Dr. Joachim Laun for his great company at many of our events. Finally, heartfelt thanks to Christoph, Jan, Julius, and Thomas F. for dedicating several hours to improving this thesis through their meticulous proofreading.

Beyond the purely academic side, I am deeply grateful for the collaboration with the Digital Chemistry department at Merck KGaA, which introduced me to outstanding scientists and people. I especially want to thank Dr. Jan Gerit Brandenburg, who also mentored me in later stages, as well as Dr. Martin Fitzner and Dr. Marko Hermsen, whose company I greatly enjoyed and from whom I learned a lot. I am also grateful for the opportunity to have worked in Darmstadt for a few months, which gave me enriching insights into digital chemistry from an industrial perspective.

I would like to thank my doctoral committee in addition to my supervisor: Prof. Thomas Bredow for serving as the second reviewer of my dissertation, Jun.-Prof. Patrycja Kielb for assuming the role of chair, and PD Dr. Gregor Hagelüken as the fourth member. I sincerely appreciate the time and effort all four of them dedicated to my evaluation.

Beyond the scientific community, I want to express my deepest gratitude to my beloved family and friends for their lasting support – without which I would not be in this position today. A special thanks goes to Marvin Schumacher, without whom I would probably have been far less successful in my chemistry studies, especially aside from the theoretical parts. Finally, I am profoundly grateful to Laura for her continuous support and for always standing by my side throughout this journey. Her patience, encouragement, and the safe space she has provided have made even the most demanding moments easier. Without her by my side, this journey would have been a significantly poorer one.

APPENDIX A

Supporting Information to Chapter 1: From Observation to Simulation

A.1 Supporting Information to Sec. 1.2: The Basis Set Expansion of Hartree-Fock

Using the definition of the Fock operator from Eq. 1.12b, the elements of the Fock matrix can be defined as follows:

$$F_{\nu\mu} = H_{\nu\mu} + J_{\nu\mu} - K_{\nu\mu}. \tag{A.1}$$

To arrive at a short-hand notation for the Fock matrix elements based on the density matrix $\mathbf{P}$ and the two-electron integrals, we will derive exemplarily the definition of $J_{\nu\mu}$:

$$J_{\nu\mu} = \int \chi_\nu(\mathbf{r}) \left[\hat{J}\left(\chi_\mu(\mathbf{r})\right)\right] d\mathbf{r}. \tag{A.2}$$

Here, the Coulomb operator $\hat{J}$ acts on the basis function χ_μ, and the result is projected onto another basis function χ_ν. The Coulomb operator acting on a single basis function is defined as follows (utilizing the basis set expansion of Eq. 1.28):

$$\hat{J}(\chi_\mu(\mathbf{r})) = \sum_{\lambda,\sigma}\sum_{j} C_{j\lambda}C_{j\sigma} \int \frac{\chi_\lambda(\mathbf{r}')\chi_\sigma(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \chi_\mu(\mathbf{r}). \tag{A.3}$$

We can use the definition of the density matrix $\mathbf{P}$ to simplify the expression:

$$P_{\lambda\sigma} = \sum_j n_j C_{j\lambda} C_{j\sigma}. \tag{A.4}$$

Inserting $\mathbf{P}$ from Eq. 1.34 yields:

$$\hat{J}(\chi_\mu(\mathbf{r})) = \sum_{\lambda,\sigma} P_{\lambda\sigma} \int \frac{\chi_\lambda(\mathbf{r}')\,\chi_\sigma(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \chi_\mu(\mathbf{r}). \tag{A.5}$$

Inserting Eq. A.5 into Eq. A.2 leads to:

$$J_{\nu\mu} = \int \chi_\nu(\mathbf{r}) \left[\sum_{\lambda,\sigma} P_{\lambda\sigma} \int \frac{\chi_\lambda(\mathbf{r}')\chi_\sigma(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \chi_\mu(\mathbf{r})\right] d\mathbf{r}, \tag{A.6}$$

which can be rewritten as:

$$J_{\nu\mu} = \sum_{\lambda,\sigma} P_{\lambda\sigma} \iint \frac{\chi_\nu(\mathbf{r})\,\chi_\mu(\mathbf{r})\,\chi_\lambda(\mathbf{r}')\,\chi_\sigma(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}'. \tag{A.7}$$

For the following simplification, we use the so-called "chemists notation":[69]

$$(ij \mid kl) = \iint \frac{\chi_i(\mathbf{r})\;\chi_j(\mathbf{r})\;\chi_k(\mathbf{r}')\;\chi_l(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}'. \tag{A.8}$$

The expression for $J_{\nu\mu}$ then reads:

$$J_{\nu\mu} = \sum_{\lambda,\sigma} P_{\lambda\sigma}\,(\nu\mu \mid \lambda\sigma)\,. \tag{A.9}$$

Applying analogous steps for $K_{\nu\mu}$ and $H_{\nu\mu}$ leads to the definition of the Fock matrix:

$$F_{\nu\mu} = \underbrace{\langle\nu|\hat{h}|\mu\rangle}_{H_{\mu\nu}} + \underbrace{\sum_{\lambda,\sigma} P_{\lambda\sigma}\left[(\nu\mu \mid \lambda\sigma) - \frac{1}{2}(\nu\lambda \mid \mu\sigma)\right]}_{J_{\nu\mu}-K_{\nu\mu}}. \tag{A.10}$$

APPENDIX **B**

Supporting Information to Chapter 2: Semempirical Approaches in Quantum Chemistry

B.1 Supporting Information to Sec. 2.2.1: Hückel Theory

The approximations introduced in the PPP method compared to the non-empirical HF from Sec. 1.1.1 can be summarized as follows:

(i) Only π electrons are taken into account. The σ electrons are assumed to be an unpolarizable core and their effect can be absorbed into the one-electron contributions (*vide infra*).

(ii) A set of only one p_z orbital per atom is considered, which are assumed to be orthonormal.

(iii) ZDO: That is, products of AOs depending on the same electron coordinates when located on different atoms are neglected.[69] Given *(ii)*, $\mathbf{S}$ becomes the unit matrix, i.e. $S_{ij} = \delta_{ij}$. This transforms the generalized eigenvalue problem from Eq. (1.32) into a special eigenvalue problem.[†]

(iv) As a consequence from *(iii)*, all three- and four-center integrals vanish completely. Non-zero $(\mu\nu|\lambda\sigma)$ occurs only if $\nu = \mu$ and $\lambda = \sigma$ or if all four indices are equal.

The resulting Fock matrix looks as follows [cf. Eq. (1.33)]:[‡]

$$F^{\mathrm{PPP}}_{\mu\mu} = H^{\mathrm{PPP}}_{\mu\mu} + \sum_{\nu} P_{\nu\nu}(\mu\mu|\nu\nu) - \frac{1}{2}\sum_{\mu} P_{\mu\mu}(\mu\mu|\mu\mu), \tag{B.1a}$$

$$F^{\mathrm{PPP}}_{\mu\sigma} = H^{\mathrm{PPP}}_{\mu\sigma} - \tfrac{1}{2}P_{\mu\sigma}(\mu\mu|\sigma\sigma). \tag{B.1b}$$

Typically, $(\mu\mu|\sigma\sigma)$ and $(\mu\mu|\mu\mu)$, as well as, in part, $H^{\mathrm{PPP}}_{\mu\mu}$ and $H^{\mathrm{PPP}}_{\mu\sigma}$, are replaced by empirical expressions, such as those proposed by Ohno.[223,225,226] This empirical treatment allows the one-electron terms to effectively incorporate contributions from the neglected σ-electron system. Consistent with Hückel theory, $H^{\mathrm{PPP}}_{\mu\sigma}$ is often neglected when the atomic centers A and B are not adjacent. However, unlike Hückel and the EHT approach introduced earlier, the effective Fock matrix in the PPP model depends explicitly on the density

[†] Some definitions of the ZDO approximation do not necessarily imply the condition $S_{ij} = \delta_{ij}$.[224] Here, we consistently adopt the definition from Refs. [69] and [223], implying $\mathbf{S}$ to be the unit matrix.

[‡] Due to (*(ii)*), μ is at atom A, σ at atom B, ν at atom C.

matrix **P**. Consequently, similar to HF and DFT, the resulting Roothaan equation [cf. Eq. (1.32)] must be solved iteratively within an SCF framework. A key implication of the ZDO approximation is the neglect of interatomic exchange interactions, as integrals of this type are generally omitted.

B.2 Supporting Information to Sec. 2.2.2: Complete Neglect of Differential Overlap

The CNDO Fock matrix takes the form:[223,233]

$$F_{\mu\mu} = H_{\mu\mu} + \underbrace{\left(P_{AA} - \frac{1}{2}P_{\mu\mu}\right)\gamma_{AA} + \sum_{B\neq A} P_{BB}\gamma_{AB}}_{J_{\mu\mu}-K_{\mu\mu}}, \tag{B.2a}$$

$$F_{\mu\nu} = H_{\mu\nu} - \underbrace{\frac{1}{2}P_{\mu\nu}\gamma_{AB}}_{J_{\mu\nu}-K_{\mu\nu}}, \tag{B.2b}$$

$$\text{with} \quad \gamma_{AB} \simeq \gamma_{\lambda\mu} \simeq (\lambda\lambda|\mu\mu) \quad \text{and} \quad P_{BB} = \sum_{\nu\in B} P_{\nu\nu}. \tag{B.2c}$$

The underbraces in Eq. (B.2) indicate the principal origin of each term when compared to the HF expression in Eq. (1.33). The term γ_{AB} represents the approximated isotropic Coulomb interaction between atoms A and B. The one-electron contributions H are further simplified and replaced with empirical expressions. Additionally, all one-center off-diagonal elements $H_{\mu\nu}$ ($\mu \neq \nu$ and $A = B$) are completely neglected. A key consequence of this approximation is that, for all interatomic terms, different atomic shells (s, p, d, ...) are treated equivalently (γ_{AB}), ensuring rotational invariance. Eq. B.3 is obtained by replacing the one-electron contributions H with largely empirical expressions and making the interaction terms γ fully independent of the shell type. The final CNDO Fock matrix reads:

$$F_{\mu\mu} = \underbrace{U_{\mu\mu} - \sum_{B\neq A} V_{AB}}_{H_{\mu\mu}} + \underbrace{\left(P_{AA} - \frac{1}{2}P_{\mu\mu}\right)\gamma_{AA} + \sum_{B\neq A} P_{BB}\gamma_{AB}}_{J_{\mu\mu}-K_{\mu\mu}}, \tag{B.3a}$$

$$F_{\mu\nu} = \underbrace{\beta^0_{AB}S_{\mu\nu}}_{H_{\mu\nu}} - \underbrace{\frac{1}{2}P_{\mu\nu}\gamma_{AB}}_{J_{\mu\nu}-K_{\mu\nu}}, \tag{B.3b}$$

$$\text{with} \quad \gamma_{AB} \simeq \gamma_{\lambda\mu} \simeq (\lambda\lambda|\mu\mu) \quad \text{and} \quad P_{BB} = \sum_{\nu\in B} P_{\nu\nu}. \tag{B.3c}$$

$H_{\mu\mu}$ is further simplified by separating same- ($U_{\mu\mu}$) and different-atom contributions (V_{AB}). $H_{\mu\nu}$ elements with μ and ν being centered on different atoms are replaced by the empirical expression β^0_{AB}, which is scaled with the corresponding overlap element $S_{\mu\nu}$, representing an exception from the ZDO approximation.

B.3 Supporting Information to Sec. 2.2.2: Neglect of Diatomic Differential Overlap

NDDO differs from CNDO by retaining same-atom differential overlap. Additionally, all terms in the expression $H_{\mu\nu} = U_{\mu\nu} - \sum_{B\neq A}(\mu|V_B|\nu)$ are kept. The Fock matrix entries in NDDO generalize to the

following definition, compared to CNDO:[223,233]

$$F_{\mu\nu} = H_{\mu\nu} + \sum_{\sigma}^{A}\sum_{\lambda}^{A} P_{\lambda\sigma}\left((\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\sigma|\nu\lambda)\right) + \sum_{\rho}^{B}\sum_{\tau}^{B} P_{\rho\tau}(\mu\nu|\rho\tau), \tag{B.4a}$$

$$F_{\mu\rho} = H_{\mu\rho} - \frac{1}{2}\sum_{\sigma}^{A}\sum_{\lambda}^{B} P_{\lambda\sigma}(\mu\sigma|\rho\tau), \tag{B.4b}$$

with $(\mu, \nu, \sigma, \lambda \in A)$ and $(\rho, \tau \in B)$. μ and ν can but do not have to be the same basis function.

B.4 Supporting Information to Sec. 2.2.2: Intermediate Neglect of Differential Overlap

The expression for the Fock matrices for diagonal, one-center off-diagonal, and two-center off-diagonal expression reads:[223]

$$F_{\mu\mu} = H_{\mu\mu}^{\mathrm{CNDO}} + \sum_{\nu}^{A} P_{\nu\nu}\left[(\mu\mu|\nu\nu) - \frac{1}{2}(\mu\nu|\mu\nu)\right] + \sum_{B\neq A}\sum_{\sigma}^{B} P_{\sigma\sigma}\gamma_{AB}, \tag{B.5a}$$

$$F_{\mu\nu} = P_{\mu\nu}\left[\frac{3}{2}(\mu\nu|\mu\nu) - \frac{1}{2}(\mu\mu|\nu\nu)\right], \tag{B.5b}$$

$$F_{\mu\sigma} = H_{\mu\sigma}^{\mathrm{CNDO}} - \frac{1}{2}P_{\mu\sigma}\gamma_{AB}, \tag{B.5c}$$

with $(\mu, \nu \in A)$ and $(\sigma \in B)$.

B.5 Supporting Information to Sec. 2.2.3: Extended Tight-Binding Methods: GFN*n*-xTB

B.5.1 The Extended Hückel Theory Contribution in xTB

The full EHT Hamiltonian matrix in GFN1-xTB (A are atoms, l shells, and μ AOs) reads:

$$H_{\mu\nu}^{\mathrm{EHT}} = \frac{1}{2}K_{AB}^{ll'}S_{\mu\nu}(H_{\mu\mu} + H_{\nu\nu})\left(1 + k_{\mathrm{EN}}\Delta\mathrm{EN}_{AB}^{2}\right)\left(1 + k_{A,l}^{\mathrm{poly}}\left(\frac{R_{AB}}{R_{\mathrm{cov},AB}}\right)^{\frac{1}{2}}\right)$$
$$\times\left(1 + k_{B,l'}^{\mathrm{poly}}\left(\frac{R_{AB}}{R_{\mathrm{cov},AB}}\right)^{\frac{1}{2}}\right) \quad \text{with} \quad \mu \in l(A), \quad \nu \in l'(B). \tag{B.6}$$

k_{EN} is global parameter and $k_{A,l}^{\mathrm{poly}}$ an atom- and shell-specific parameter. $R_{\mathrm{cov},AB}$ corresponds to the sum of covalent radii taken from Ref. [337]. The distance-dependent polynomial scaling function is a common feature of GFN1 and GFN2, in addition to the intrinsic distance dependence of $\mathbf{S}$. $\Delta\mathrm{EN}_{AB}$ is the difference of the standard Pauling electronegativities of atoms A and B. The GFN2-xTB EHT matrix is multiplied with an additional term depending on the STO exponents ζ_l^A:

$$\mathbf{Y}(\zeta_l^A, \zeta_{l'}^B) = \left(\frac{2\sqrt{\zeta_l^A\zeta_{l'}^B}}{\zeta_l^A + \zeta_{l'}^B}\right)^{\frac{1}{2}}. \tag{B.7}$$

The product of two basis function exponents induces effects similar to the kinetic energy integral in DFT or HF, which also increases with larger exponents. The diagonal elements $H_{\mu\mu}$ are defined as follows:

$$H_{\mu\mu} = \begin{cases} h_A^l \left(1 + k_{\mathrm{CN},l_a} \mathrm{CN}_A\right), & \text{for GFN1-xTB,} \\ h_A^l - \delta h^l_{\mathrm{CN}'_A} \mathrm{CN}'_A, & \text{for GFN2-xTB.} \end{cases} \tag{B.8}$$

h_A^l and $h^l_{\mathrm{CN}'_A}$ are atom- and shell-specific parameters, while k_{CN,l_a} depends only on the angular momentum of the respective shell (s, p, d). CN_A is taken from the D3 model[117]. CN'_A indicates a modified form in case of GFN2-xTB.

B.5.2 Anisotropic Electrostatics and Exchange-Correlation Contributions and Halogen Bond Corrections

The anisotropic ES and XC contributions following from the multipole expansion are provided in Eq. B.9 and are clustered according to their interaction type.

$$\begin{aligned} E_{\mathrm{AES}} &= E_{q\mu} + E_{q\Theta} + E_{\mu\mu} \\ &= \frac{1}{2} \sum_{A,B} \Big\{ f_3(R_{AB}) \left[q_A(\boldsymbol{\mu}_B^T \mathbf{R}_{BA}) + q_B(\boldsymbol{\mu}_A^T \mathbf{R}_{AB}) \right] \\ &\quad + f_5(R_{AB}) \left[q_A \mathbf{R}_{AB}^T \boldsymbol{\Theta}_B \mathbf{R}_{AB} + q_B \mathbf{R}_{AB}^T \boldsymbol{\Theta}_A \mathbf{R}_{AB} \right] \\ &\quad - 3(\boldsymbol{\mu}_A^T \mathbf{R}_{AB})(\boldsymbol{\mu}_B^T \mathbf{R}_{AB}) + (\boldsymbol{\mu}_A^T \boldsymbol{\mu}_B) R_{AB}^2 \Big\}. \end{aligned} \tag{B.9}$$

$\boldsymbol{\mu}_A$ stands for the cumulative dipole moment of atom A and $\boldsymbol{\Theta}_A$ for the corresponding traceless quadrupole moment. The damping function of the E_{AES} term in second-order is related to the original damping function in the D3 dispersion correction:[117]

$$f_n(R_{AB}) = \frac{f_{\mathrm{damp}}(a_n, R_{AB})}{R_{AB}^n} = \frac{1}{R_{AB}^n} \cdot \frac{1}{1 + 6\left(\frac{R_0^{AB}}{R_{AB}}\right)^{a_n}}. \tag{B.10}$$

a_n are global empirically adjusted parameters and the covalent radii in $R_0^{AB} = \frac{1}{2}\left(R_0^{A'} + R_0^{B'}\right)$ depend on the CN (see Ref. [278] for details). The anisotropic XC contribution in GFN2-xTB corresponds to

$$E_{\mathrm{AXC}} = \sum_A \left(f_{\mathrm{XC}}^{\mu_A} |\boldsymbol{\mu}_A|^2 + f_{\mathrm{XC}}^{\Theta_A} \|\boldsymbol{\Theta}_A\|^2 \right), \tag{B.11}$$

in which $f_{\mathrm{XC}}^{\mu_A}$ and $f_{\mathrm{XC}}^{\Theta_A}$ are empirically determined element-specific parameters. The halogen bond correction in form of a pair-wise repulsive Lennard-Jones potential in GFN1-xTB is defined as follows:

$$E_{\mathrm{XB}}^{\mathrm{GFN1}} = \sum_{AXB}^{N_{\mathrm{XB}}} f_{\mathrm{damp},AXB} k_{\mathrm{X}} \left[\left(\frac{k_{\mathrm{XR}} R_{\mathrm{cov},AX}}{R_{AX}} \right)^{12} - k_{\mathrm{X2}} \left(\frac{k_{\mathrm{XR}} R_{\mathrm{cov},AX}}{R_{AX}} \right)^{6} \right] \left[\left(\frac{k R_{\mathrm{cov},AX}}{R_{AX}} \right)^{12} + 1 \right]^{-1}$$

$$\text{with} \quad f_{\mathrm{damp},AXB} = \frac{1}{2} \left(1 - \frac{1}{2} \frac{R_{XA}^2 + R_{XB}^2 - R_{AB}^2}{|R_{XA}||R_{XB}|} \right)^6. \tag{B.12}$$

k_{X2} and k_{XR} are global parameters, while XR depends on the halogen atom X.

APPENDIX C

Appendix: ωB97X-3c: A Composite Range-Separated Hybrid DFT Method with a Molecule-Optimized Polarized Valence Double-ζ Basis Set

Marcel Müller,[†] Andreas Hansen,[†] Stefan Grimme[†]

Received: November 01, 2022
Published online: January 03, 2023

Reprinted in Appendix C with permission[§] from
M. Müller, A. Hansen, and S. Grimme, *ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set*, The Journal of Chemical Physics **158** *1* (2023) 014103, DOI: 10.1063/5.0133026

Own contributions

- Conceptualization of the composite DFT method
- Implementation into the ORCA QM code, support of the TURBOMOLE implementation, and proliferation of the vDZP basis set
- Performing all calculations
- Interpretation of the results
- Writing of the manuscript

[†] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany
[§] Permission requests to reuse material from this chapter should be directed to AIP Publishing.

C.1 Abstract

A new composite density functional theory (DFT) method is presented. It is based on ωB97X-V as one of the best-performing density functionals for the GMTKN55 thermochemistry database and completes the family of "3c" methods toward range-separated hybrid DFT. The method is consistently available for all elements up to Rn (Z=1–86). Its further key ingredients are a polarized valence double-ζ (vDZP) Gaussian basis set, which was fully optimized in molecular DFT calculations, in combination with large-core effective core potentials and a specially adapted D4 dispersion correction. Unlike most existing double-ζ atomic orbital sets, vDZP shows only small basis set superposition errors (BSSEs) and can compete with standard sets of triple-ζ quality. Small residual BSSE effects are efficiently absorbed by the D4 damping scheme which overall eliminates the need for an explicit treatment or empirical corrections for BSSE. Thorough tests on a variety of thermochemistry benchmark sets show that the new composite method, dubbed ωB97X-3c, is on par with or even outperforms standard hybrid DFT methods in a quadruple-ζ basis set, at a small fraction of the computational cost. Particular strengths of the method are the description of non-covalent interactions and barrier heights, for which it is among the best-performing density functionals overall.

C.2 Introduction

Composite electronic structure methods have a long tradition in QC dating back to the so-called Gn WFT developed by Pople and co-workers in the 1980s-90s.[203–206] Popular extensions of this approach combining various theory levels and AO basis sets of different degrees of sophistication in order to save computational time are the so-called Weizmann protocols[207–209] and CBS schemes.[158,210]

Based on the framework of KS DFT, which is nowadays the most widely used electronic structure method,[338] similar but more empirical approaches have been proposed.[153,196,259,268,339] Despite its high efficiency, the limits of conventional DFT are quickly reached for calculations of large systems that contain more than 200-300 atoms. This holds in particular for large-scale screenings, very accurate thermochemical calculations, or geometry optimizations. The emerging need for fast yet accurate low-cost methods paves the way for composite schemes also in a DFT framework. Probably most widely used in computational chemistry nowadays are members of the so-called "3c" family of methods.[339] The first 3c method was the HF theory-based HF-3c[198] that contains three name-giving corrections to improve its accuracy. The same concept was later applied to DFT resulting in the PBEh-3c/HSE-3c/B3LYP-3c[115,212,314,339] hybrid, B97-3c[213] GGA, and recently the m-GGA based r^{2}SCAN-3c[214] method.

While originally focused on an especially good description of molecular geometries and NCIs, the latest "3c" methods aim more generally on accurate thermochemistry and conformational energies.[70] In this context, the related DFT-C approach of Witte and Head-Gordon is mentioned, which is specifically recommended for accurate NCI energy calculations.[153] In the same spirit but using ECPs to account for BSSE, a specialized B3LYP approach in a medium-size AO basis set has been proposed.[196,197]

The mentioned "3c" methods are successfully applied in various fields. To give just a few examples, they have recently contributed to a better understanding of the stereochemical behavior of phosphanyl groups[5,7] and of the ground-state properties of porphyrinic sandwich complexes[340], they served as a reference method for calculation of IR spectra[314] and paved the way for online computer-aided design of catalytic pockets.[341] Furthermore, they play a key role in our multi-level workflow "CRENSO",[164] with which thermodynamic and spectroscopic properties of structure ensembles can be calculated efficiently. However, the existing "3c" methods are not optimal for certain applications such as the calculation of barrier heights or the study of NCIs

Table C.1: Comparison of the hierarchy of efficient composite "3c" electronic structure methods.

	PBEh-3c	B97-3c	r^{2}SCAN-3c	ωB97X-3c
AO basis set	mSV(P)	mTZVP[†]	mTZVPP[†]	vDZP
# param. F_{xc}[‡]	3	10	unaltered	unaltered
Fock ex. [%]	42	0	0	16.7-100
dispersion	D3[117,118]	D3[117,118]	D4[119,120]	D4[119,120]
SRB corr.[198]	no	yes	no	no
gCP corr.[195]	yes	no	yes	no

of very large systems. For the former, no or a small amount of global Fock exchange is often insufficient for the desired target accuracy. Extended molecular systems suffer strongly from BSSE with commonly applied small basis sets and are difficult to SCF converge with (m)-GGA-based methods due to an over-delocalization of electrons.

In this work, we complete the "3c"-family of methods by ωB97X-3c, a variant with an underlying RSH functional which strongly reduces the effect of SIE in DFT calculations (see Tab. C.1). We choose one of the most accurate RSH functionals on the market developed in the group of Head-Gordon termed ωB97X-V.[103] When combined with a large AO basis set of QZ quality, this functional performs extraordinarily well on the huge GMTKN55 thermochemistry benchmark set[93] with 1500+ chemically significant relative energies. For the new "3c" method presented here, ωB97X-V is combined with a newly developed double-ζ basis set, a specially assembled set of effective core potentials, and the established D4 dispersion correction.[120] Detailed explanations of each of these components of ωB97X-3c are given in Sec. C.3.

In addition, we present an improved D4 parameterization for ωB97X-V replacing the density-dependent VV10[125] dispersion treatment employed in the original functional. The first DFT-D4 counterpart for ωB97X-V was presented earlier by Goerigk *et al.*[307] We show that the ωB97X-D4/QZ method with the here refitted D4 parameters performs globally better than ωB97X-V/QZ for the GMTKN55 database and other common thermochemistry benchmarks as well as for NCI benchmarks going beyond those contained in GMTKN55.

C.3 Theory

C.3.1 Basis Set Construction

The direct predecessors of ωB97X-3c, B97-3c[213] and more recently r^{2}SCAN-3c,[214] were constructed with slightly modified contracted Gaussian orbital basis sets of triple-ζ quality originating from the Ahlrichs basis set def-TZVP[342] and its extension def2-TZVPP,[139] respectively. Recently, r^{2}SCAN-3c was adapted for a Slater-type TZ AO expansion in the AMS modeling suite of programs.[343] However, to preserve the computational efficiency of the existing "3c" family of methods, a smaller basis set that still yields comparable or even higher accuracy is desired for ωB97X-3c. This is especially important since RSH-DFT is computationally more demanding than (m)-GGAs like B97-3c and r^{2}SCAN-3c. For this purpose, a new polarized valence double-ζ (vDZP) basis set was constructed, which forms the key part of the new composite method. This basis was developed also with its application in novel semi-empirical TB methods in mind which will be published separately. It is characterized by the following key points:

[†] Modified version of the def2-TZVP basis.
[‡] Exchange-correlation enhancement factor.

- The use of large-core ECPs to reduce the number of explicitly treated electrons and, hence, the required number of basis functions.
- Optimization of the Gaussian-type primitive functions in atomic and molecular calculations for neutral as well as ionic systems at the RSH DFT level.
- Relatively deep contraction, i.e., a large number of primitives for each AO to reduce incompleteness errors but keeping the number of contracted functions small. This reduces the memory demands in the linear algebra parts of the calculation for very large systems as well as the diagonalization bottleneck in TB treatments.

Molecule-optimized basis sets are rarely used in quantum chemistry. We are aware of the MOLOPT sets[160] in the CP2K code[344] and the polarization consistent (pc-*n*) basis sets by Jensen.[156,159] In the latter, only the polarization functions are optimized with respect to molecular energies. A significant advantage of molecular basis set optimizations is that, contrary to atomic optimizations, all angular momentum functions (i.e., polarization functions not occupied in the atomic ground state) can be determined consistently, as already noted by Hutter and VandeVondele.[160] We observed that even contracted polarization functions can be obtained reliably in this way, thus further improving the quality of the employed AOs. Zijlstra *et al.* observed in agreement with results by Hutter and VandeVondele that molecule-optimized Gaussian basis sets outperform atom-optimized ones in general applications.[345] However, Petersson *et al.* noticed that the coupling of different angular momenta in molecular systems makes the basis set optimization converge much more slowly than in atomic optimizations, which is probably the technical reason why molecular optimizations are rarely employed for basis set construction.[346]

Although ωB97X-3c is designed mostly in the spirit of the other "3c" methods, the meaning and definition of the applied "three corrections" have changed over the years. As before, the acronym stands for the dispersion correction, as well as for the specially developed AO basis set, but here for the compilation of ECPs, which are essential for efficiency, as a third modification. As noted already, ωB97X-3c forms the basis of our next-generation TB method, where the application of ECPs to yield an (almost) valence-only treatment is of fundamental importance.

Choice of Effective Core Potentials

Finding the right compromise between accuracy and computational efficiency in a "3c"-treatment was a particular challenge regarding the choice of the ECPs. Consistent sets for all elements up to Rn from one resource were not available and hence, the potentials from various authors namely the Stuttgart-Cologne, CRENBL, and SBKJC type ECPs were used, as summarized in Tab. C.2.

For main group elements from group 13 and higher, the ECPs are of "largest core" type, i.e., only the sp-valence shell is explicitly considered. For the transition metals and group 1 / 2 elements (except Li, Be for which no ECP is applied), the sub-valence sp-shell (semi-core) is included explicitly in the calculation. In this way, the chemically important semi-core polarization effects are mostly accounted for. For the lanthanoids, the 5s5p4f5d electrons (or depending on the element 5s5p4f, see the Supporting Information (SI) for details) are included.

For the ECP-2-MWB of the elements boron and fluorine, we discovered in the course of this project that the exponents of the d-f projector Gaussian functions are very small (< 0.04) leading to artificial potentials in larger molecules. This resulted in inconsistent NCI energies with ECP-caused errors of several

[†] Modified d-f projector for B and F, see text.

Table C.2: Type of effective core potential used and the number of core electrons included in the ECP.

Z (element)	Type of ECP	# core electrons
B–Ne	ECP-2-MWB[141†]	2
Na–Mg	CRENBL[347]	2
Al–Ar	ECP-10-MWB[141]	10
K–Ca	CRENBL[348]	10
Sc–Zn	ECP-10-MDF[137]	10
Ga–Kr	ECP-28-MWB[141,349]	28
Rb–Sr	CRENBL[144]	28
Y–Cd	ECP-28-MWB[350]	28
In–Xe	ECP-46-MWB[141,349]	46
Cs–Ba	CRENBL[351]	46
La	ECP-46-MWB[140,352]	46
Ce–Lu	SBKJC-ECP[353]	46
Hf–Hg	ECP-60-MWB[350]	60
Tl–Rn	ECP-78-MWB[354]	78

kcal·mol^{-1}. For boron, we remedied this problem by increasing the exponent and adapting the corresponding Gaussian-ECP prefactor such that both typical thermochemistry results in GMTKN55 and NCI energies of boron model complexes are optimally described compared to accurate all-electron reference data. Afterward, the corresponding p-f projector was fine-tuned to account for the modified d-f projector. For fluorine, no beneficial effects of modifying the d-f projector were observed compared to its removal, so we deleted it from the fluorine ECP. The slightly modified ECPs for B and F are given together with the complete set of potentials in the SI.

Optimization of the vDZP Basis Set

The design and primitive contraction scheme for all elements up to Rn (Z=1-86) are given in Tab. C.3. Generally, semi-core (sub-valence) shells are described by 4-5 primitives while valence shells contain 2-4 primitive functions. A notable exception is hydrogen where the deep 53 contraction of the two *s*-shells substantially reduces the BSSE compared to, for example, a typical 31 contraction. Similar considerations apply for polarization functions which generally consist of two primitives instead of normally just one.

The contraction coefficients and primitive exponents of the vDZP basis were fully optimized with the help of the Powell optimization algorithm[355] which does not require any objective function value derivatives. For each individually treated element, the objective function to be minimized was the sum of atomic and molecule ωB97X-D3[356] total energies. The ECPs as described above, but without the mentioned modifications for B/F, were used in these calculations. The set of systems usually contains the neutral atom, atomic anion and cation, hydrides XH_n and X_2H_m as well as corresponding anionic/cationic hydrides. The hydrogen basis (which must be generated first in this procedure) was obtained from hydrides of the elements Li-F (including ions) described initially by a standard def2-TZVP basis which was replaced later by the emerging vDZP sets. For some metals (and all lanthanoids) fluorides XF_n and X_2F_m were additionally included. The inclusion of ions (anions in particular) ensures that several primitives in the outer (more diffuse) AO are properly defined and describe "physical" states (see Fig. C.1 in Sec. C.5 for an example). Note that for anions this is only possible in HF or RSH-based optimizations because the correct asymptotics of the exchange potential is

Table C.3: Contraction scheme of the vDZP basis set used in ωB97X-3c. Multiply occurring basis set schemes are given only for their first appearance.

Z (element)	contraction scheme		
H–He	(8s2p)	[2s1p]	{53/2}
Li–Be	(10s3p2d)	[3s1p1d]	{5 32/3/2}
B–Ne	(6s6p2d)	[2s2p1d]	{42/42/2}[†]
Na–Mg	(9s8p2d)	[3s2p1d]	{4/5 32/3/2}
K–Ca	(9s7p2d)	[3s2p1d]	{4/4 32/3/2}[‡]
Sc–Zn	(9s6p6d)	[3s2p2d]	{5/4 22/2/42}[§]
La, Hf–Hg	(8s6p6d)	[3s2p2d]	{4/4 22/2/42}
Ce–Lu	(8s5p5d7f)	[3s2p2d2f]	{4/4/3 22/1/2/52}

important to yield bound electronic states. The number of systems for each element was between 10 and 20 atoms and molecules. Start parameters for the contraction coefficients and primitive exponents were taken from existing AOs in the TURBOMOLE basis set library and at later stages of the optimization procedure, by exponent scaling from neighboring elements. The typical number of objective function evaluations per element for the complete AO optimizations is on the order of 10^4.

C.3.2 Dispersion Correction

Long-range electron correlation (London dispersion) effects are included by the atomic-charge dependent D4 correction calculated according to

$$\begin{aligned} E_{\text{disp}}^{\text{D4}} = &-\frac{1}{2}\sum_{AB}\sum_{n=6,8} s_n \frac{C_{(n)}^{AB}}{R_{AB}^{(n)}} f_{\text{damp}}^{(n)}(R_{AB}) \\ &-\frac{1}{6}\sum_{ABC} s_9 \frac{C_{(9)}^{ABC}}{R_{ABC}^{(9)}} f_{\text{damp}}^{(9)}(R_{ABC}, \theta_{ABC}), \end{aligned} \tag{C.1}$$

where A, B and C denote atoms, s_n are scaling parameters, $C_{(n)}$ are the dispersion coefficients, R_{AB} the interatomic distances, R_{ABC} atomic triangle averaged distances, θ_{ABC} the corresponding angles, and $f_{\text{damp}}^{(n)}$ the BJ damping function $f_{\text{BJ}}^{(n)}$

$$f_{\text{BJ}}^{(n)}(R_{AB}) = \frac{R_{AB}^{(n)}}{R_{AB}^{(n)} + (a_1 R_0^{AB} + a_2)^{(n)}} \tag{C.2}$$

with the functional specific damping parameters a_1 and a_2. The $E_{\text{disp}}^{\text{D4}}$ term is simply added to the dispersion-uncorrected (without VV10 part) "ωB97X-V ($-$VV10)"/vDZP electronic energy. The medium-range scaling parameters s_8 and s_9 as well as the damping parameters a_1 and a_2 were determined as usual by a fit to reference NCI as well as a few conformational energies. The benchmark sets used for training the D4 parameters were the following (sets for which no reference is given were taken from the GMTKN55 database[93]): S22, S66, ADIM6, HEAVY28, HAL59, CARBHB12, NCIBLIND10,[302] L7,[303] ISO34, ISOL24, TAUT15, BSR36, DARC, IDISP, ACONF, ACONFL,[292] SCONF, MCONF, PCONF21, UPU23, 37CONF8,[304] MPCONF196,[305]

[†] Equivalent for Al-Ar, Ga-Kr, In-Xe, and Tl-Rn.
[‡] Equivalent for Rb, Sr, Cs, and Ba.
[§] Equivalent for Y-Cd.

Table C.4: Optimized parameters of the D4 dispersion correction. Additionally, parameters for the ωB97X-V functional earlier published by Najibi and Goerigk (see Ref. [307]) are given.

D4 parameter	$a1$	$a2$	$s6$	$s8$	$s9$
ωB97X-3c	0.2464	4.737	1.0	0.0	1.0
ωB97X-D4 (present work)	0.3306	4.279	1.0	0.4485	1.0
ωB97X-D4 (Goerigk et al.)	0.0662	5.449	1.0	0.5093	1.0

S30L,[306] and IONPI19.[285]

The parameters in the charge-scaling functions β and γ [see Eq. (2) of Ref. [120]] are kept unchanged and $s_6 = 1$ as usual. In general, the small BSSE of TZ quality basis sets like vDZP can be absorbed in the D4 parameterization. A reasonable approach is to use a zero value for s_8 to ensure a slightly more repulsive behavior at intermediate distances (see also Ref. [214]), which is also pursued here as a strategy. Thus, no further corrections like, e.g., gCP[195] or related potentials need to be applied for ωB97X-3c.

During the development and testing of ωB97X-3c, we noticed in some cases a questionable behavior related to the D4 parameters for the existing ωB97X(-V)/QZ functional. Therefore, the presented fitting procedure was applied to determine improved damping parameters also for ωB97X-D4. As usual, this refers to the large def2-QZVPP AO basis set without counter-poise correction. The optimized parameter values are given in Tab. C.4, for comparison together with the parameters suggested by Goerigk *et al.* In the following,"ωB97X-D4" refers only to the D4 parameters presented in this work.

C.4 Technical Details

The ωB97X-3c composite electronic structure method has been implemented in development versions of the TURBOMOLE[294] and ORCA[309,357] programs and is expected to be included in the next official releases of both program packages. The RI approximation for the electronic Coulomb energy was used with the large universal auxiliary basis sets by Weigend.[358] For the semi-local XC part, the medium-sized numerical quadrature grids `m4` (TURBOMOLE) and `DEFGRID2` (ORCA) are used. Semi-numerical treatment of exact exchange (`$senex`[359,360] with `m2` grid and `RIJCOSX`[178] with `DEFGRID2`) is applied per default. An in-depth investigation of the accuracy of semi-numerical treatment of exact exchange is given in the spreadsheet attached to the SI. All other DFT calculations were conducted with a current development version of TURBOMOLE 7.6 or the ORCA 5.0.3[309,357] program package release mostly applying large quadruple-ζ AO basis sets def2-QZVP and def2-QZVPP[139] (together with corresponding auxiliary basis sets). Other technical settings refer to the current defaults for standard DFT calculations with both programs.

C.5 Results and Discussion

With r^{2}SCAN-3c as the most recent member, the focus of the "3c" methods expanded from exclusively structural properties and inter- and intramolecular NCIs to also thermochemical problems, including the formation of strong covalent bonds. With ωB97X-3c, the scope is further extended towards higher accuracy for many systems and towards systems prone to SIE. The reason for the latter is mainly the RSH nature of the method, also circumventing occasional convergence problems of earlier (m)-GGA-based "3c" methods.

As will be shown in the following, ωB97X-3c surpasses the performance of some of the best hybrid DFAs/large-QZ schemes in many applications, at a fraction of the computational costs. Comprehensive

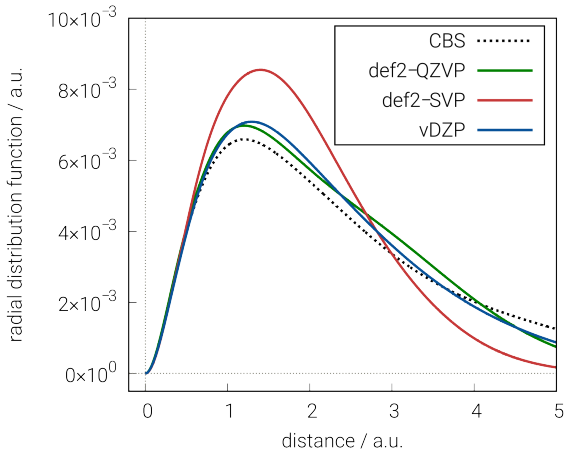

Figure C.1: Radial distribution function of the electron density of H^- by employing ωB97X-V in combination with different basis sets. The annotation "CBS" is used for an uncontracted 14s basis set.

benchmarks for thermochemistry, NCIs, barrier heights, metal-organic reactions, and molecular structures are discussed in the following subsections, proving the general applicability and robustness of the presented method to a wide variety of systems. Definitions of the applied standard statistical measures and the raw data required for the reproduction of all plots are given in the SI.

C.5.1 Quality of the vDZP Basis Set

As pointed out earlier, the key ingredient of the new "3c" method presented in this work is the vDZP basis explained in detail in Sec. C.3.1. Thus, an investigation of the qualitative behavior of the introduced basis set is appropriate to show that the method is based on solid grounds.

For this purpose, the radial distribution function of the electron density of H^- was calculated using ωB97X-V with a virtually complete basis set as the reference ({14s} as available in the TURBOMOLE basis set library), the vDZP basis, as well as the Ahlrichs basis sets def2-QZVP and def2-SVP for comparison (see Fig. C.1). The large def2-QZVP basis set approaches the "CBS" reference closely whereas the double-ζ basis set def2-SVP deviates significantly in the medium- and long-range parts. The highly contracted vDZP basis set behaves very similarly to both def2-QZVP and the "CBS" limit. The good behavior of vDZP underlines its advantages with respect to usual double-ζ basis sets and shows the importance of optimizing the basis for molecular and (an)ionic systems. Fig. S4 in Sec. III-B of the SI shows another example, in which electron densities evaluated with vDZP and def2-SVP / def2-QZVP basis sets are compared.

In the development of ωB97X-3c and its vDZP basis set, another important goal was to avoid significant BSSE, and any related empirical corrections and to closely approach the physically correct solution. In the present approach, this is realized by optimizing the basis set in molecular calculations and by deep contraction of the functions. A quantitative investigation of the amount of remaining intermolecular BSSE is possible by calculating the BB-CP correction which should vanish for complete basis sets. To this end, small organic dimers from S22 and heavy main group dimers from HAL59 (both parts of the common GMTKN55 database) are taken into consideration. In addition to the Ahlrichs basis sets def2-QZVP, def2-TZVP, def-TZVP, and def2-SVP, also the correlation-consistent cc-pVDZ basis set[150] and the popular Pople-style 6-31G* basis set[361] are evaluated. Relative and absolute BB-CP corrections are shown for each case in Fig. C.2. The

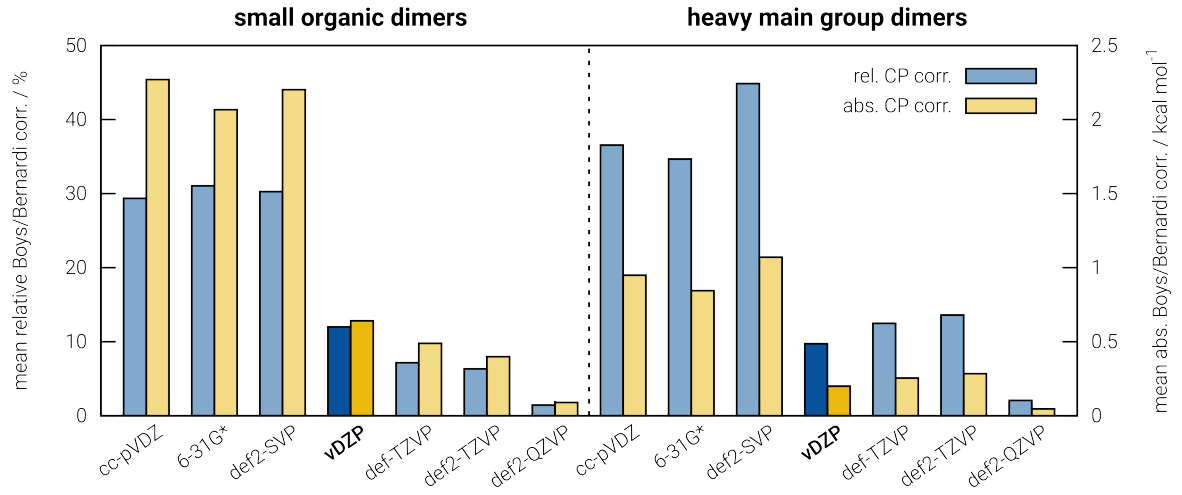

Figure C.2: Boys-Bernardi counter-poise correction evaluated at the ωB97X-V level for different basis sets. Relative values (blue) for the Boys-Bernardi (BB)-CP correction in % are determined from the absolute values in kcal·mol^{-1} (yellow) with respect to the reference dimerization energies. Small organic dimers are taken from the S22 benchmark set, heavy main group dimers from the HAL59 set, both of which are part of GMTKN55[93].

three commonly applied DZ basis sets perform very similarly with relative BB-CP corrections around 30 % (small organic dimers) and 40 % (heavy main group dimers), respectively. As expected, the BSSE decreases significantly when moving to triple- (7 and 13 %, respectively) or even quadruple-ζ basis sets (2 % or lower for both compound classes). Remarkably, vDZP performs close to the Ahlrichs-type TZ basis sets (12 % rel. BB-CP corr.) for small organic dimers. For heavy main group dimers, the picture is even better with vDZP (10 % rel. BB-CP corr.) ranking between def2-TZVP and def2-QZVP. This might also be attributed to the use of large-core ECPs and, consequently, a smaller number of basis functions that potentially could be used unphysically (see Sec III-A of the SI for a supporting study exemplified for the cc-pVDZ basis set). The low amount of remaining BSSE can be absorbed by an appropriately fitted D4 dispersion correction, as described in Sec. C.3.2.

C.5.2 Timings for Large Supramolecular Complexes

The use of a (mostly) valence DZ basis set in combination with large-core ECPs reduces the computational effort substantially in comparison to usually applied all-electron QZ basis sets. The wall times for a single-point calculation on a large supramolecular complex, a boron-nitrogen nanotube containing an organic guest molecule with 381 atoms in total, are discussed as an illustrative example. Wall times relative to r^{2}SCAN-3c (colored bars) and memory consumption (grey bars) are shown in Fig. C.3. The computational effort required by typical hybrid density functionals in quadruple-ζ basis sets is decreased by a factor 4-5 with ωB97X-3c and approaches that of class-leading m-GGA density functionals like B97M-V[85,308] with large QZ basis set. For large systems containing many heavy atoms, this favorable behavior increases further due to the fewer number of basis functions in comparison to methods employing no or only small-core ECPs.

Another aspect that is particularly important for very large systems consisting of hundreds or even thousands of atoms is memory consumption. Often, such kinds of systems cannot be treated by DFT / WFT methods in standard TZ / QZ basis sets anymore since the required memory would exceed the available resources. The

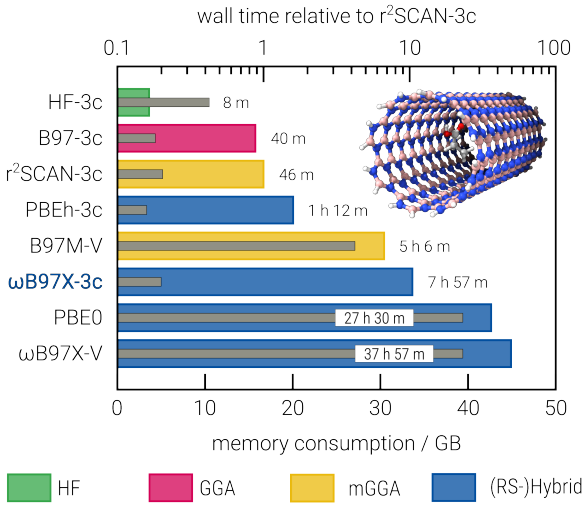

Figure C.3: Wall time required for a single-point calculation of a large boron-nitrogen nanotube containing an organic guest molecule (381 atoms; see Ref. [362] and [116] for details) using `ORCA 5.0.3` on 16 *AMD EPYC 7742* cores. The memory consumption refers to the `ORCA` printout (per CPU core). The methods are ordered by the associated computational effort and the colors encode the different rungs on "Jacob's ladder"[74]. Note that HF cannot be clearly categorized on "Jacob's ladder", which is why we have assigned it its own color. For all calculations except for"3c" methods, the def2-QZVPP basis set was used. Default numerical settings were applied throughout, including the seminumerical COSX (keyword: `RIJCOSX`).

gray bars in Fig. C.3 clearly show that the memory consumption of ωB97X-3c is an order of magnitude lower compared to methods with QZ basis sets. Compared to TZ basis sets, it is still reduced by a factor of ≈ 3.

C.5.3 Main-Group Thermochemistry

GMTKN55[93] is a comprehensive and well-established benchmark database for main group chemistry. With over 1500 individual and chemically important reactions, it provides a broad overview of intra- and inter-molecular NCIs, basic properties, reactions of small and large molecules, and barrier heights. For each reaction, reliable CCSD(T)/CBS reference data are provided. The GMTKN55 benchmark set introduces a weighted measure for mean absolute deviations called WTMAD-2, which levels large differences in reaction energies among the 55 subsets and will be used in the following (see the SI for an explanation of statistical measures).

For comparison within this and all following benchmark sets, we decided to choose a) state-of-the-art DFAs with similar construction, b) the most recent member of the "3c" family, r^{2}SCAN-3c,[214] and c) the commonly used B3LYP hybrid density functional,[78,88,89] each of them employing a dispersion correction. Apart from "3c" methods, def2-QZVP(P) basis sets are used throughout (except for a few subsets of GMTKN55 using specifically augmented def2-QZVP basis sets, see Ref. [93] and [308] for details). WTMAD-2 values for the whole GMTKN55 database and for each of the five subclasses are given in Fig. C.4 for ωB97X-3c (blue shaded), the parent functional ωB97X-V including both VV10 and D4 dispersion corrections, the m-GGA B97M-V,[85] the composite m-GGA DFT method r^{2}SCAN-3c and B3LYP-D4. Results for B97M-V,[308] B3LYP,[93] r^{2}SCAN-3c,[214] and ωB97X-V[93] were taken from literature and re-evaluated for calculation of the WTMAD-2. For B3LYP-D4 and ωB97X-D4, only the dispersion correction contribution was added or

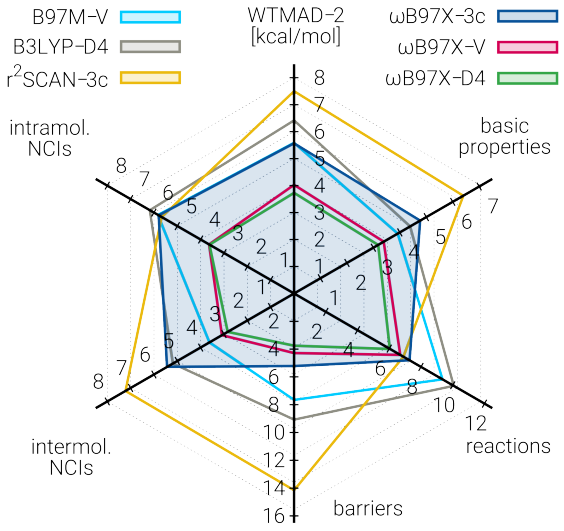

Figure C.4: WTMAD-2 values for the GMTKN55 benchmark set in kcal·mol^{-1}. Except for "3c" methods, QZ basis sets are used throughout (see main text for details). Results for B97M-V[308], B3LYP[93], r^{2}SCAN-3c[214] and ωB97X-V[93] were taken from literature and re-evaluated for calculation of the WTMAD-2. For B3LYP-D4 and ωB97X-D4, only the dispersion contribution was added or replaced, respectively.

replaced, respectively.

ωB97X-3c performs equally well or better than r^{2}SCAN-3c in all subclasses, which is evident from an overall WTMAD-2 of 5.6 compared to 7.5 kcal·mol^{-1}. The very commonly applied hybrid DFT method B3LYP with an extended QZ basis set is significantly behind ωB97X-3c with a WTMAD-2 of 6.4 kcal·mol^{-1}. The best-in-class m-GGA B97M-V[308] is on par with ωB97X-3c. Noticeably, the re-fitted parameters of the D4 dispersion correction for the ωB97X-V functional lead to an improved WTMAD-2 of now 3.7 kcal·mol^{-1} (cf. to ωB97X-D4 by Goerigk et al.[307] with a WTMAD-2 of 4.3 kcal·mol^{-1}), now outperforming ωB97X-V in its original form employing the VV10 nonlocal correlation functional (4.0 kcal·mol^{-1}).

Especially for barrier heights, ωB97X-3c shows a drastic improvement over its predecessor r^{2}SCAN-3c, which is a consequence of employing range-separated Fock exchange thereby eliminating most of the SIE. Noticeably, ωB97X-3c outperforms not only B97M-V but also B3LYP-D4 in this subclass and nearly matches the performance of the parent functional (ωB97X-V/D4), all of which use QZ basis sets. The performance for barrier heights is furthermore investigated in Sec. C.5.6 on the recently published BH9 benchmark set[363].

For basic properties and reactions of small molecules, ωB97X-3c performs also close to the best-performing functionals in this comparison. For sets such as SIE4x4 (self-interaction-error related problems) and YBDE18 (bond dissociation energies in ylides), ωB97X-3c shrinks the error of m-GGAs such as r^{2}SCAN-3c or B97M-V by 30-60 %. For SIE4x4, ωB97X-3c shows even the lowest MAE of all tested functionals, proving the very good applicability for systems prone to SIE.

For reactions of large molecules, B3LYP-D4 and B97M-V yield significantly larger errors than r^{2}SCAN-3c and ωB97X-based functionals, including the "3c" composite scheme. Specifically, isomerization energies, such as those composed in CDIE20 (double-bond isomerizations in cyclic systems) and ISO34 (isomerizations of small and medium-sized organic molecules), are much better described by ωB97X-3c compared to the m-GGAs r^{2}SCAN-3c and B97M-V and to B3LYP-D4.

For inter- and intramolecular NCIs of small- to medium-sized systems, a slightly larger difference of

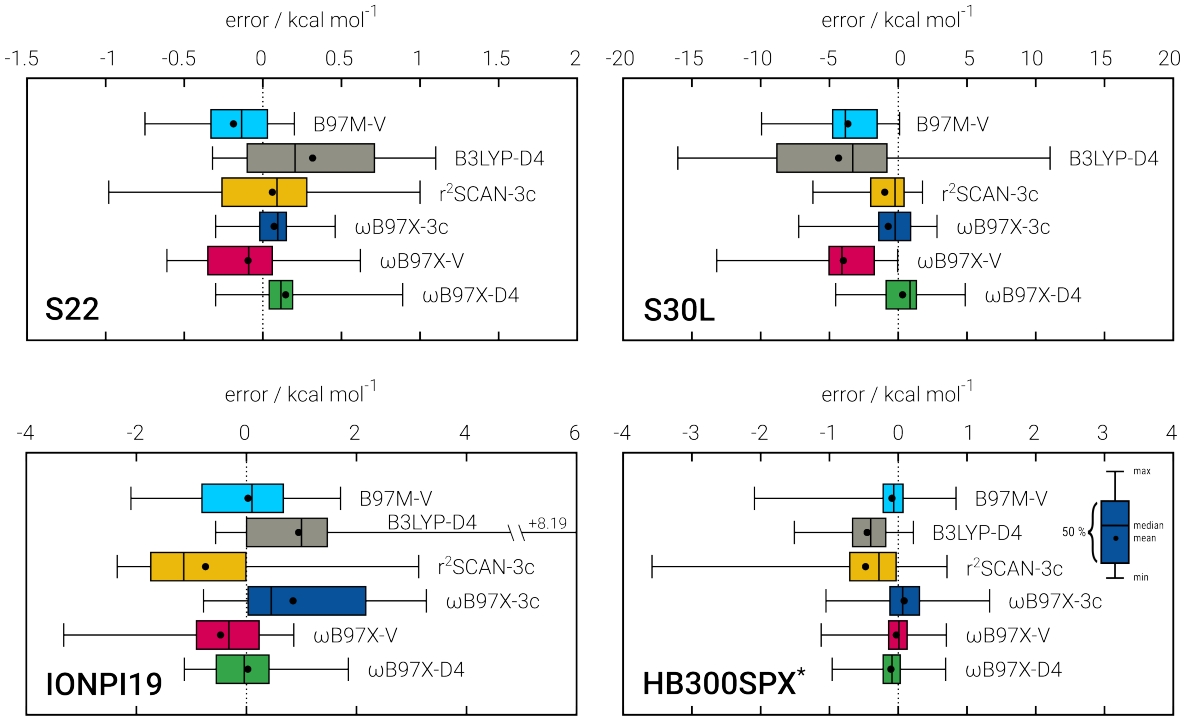

Figure C.5: Error boxplots for the NCIs benchmark sets S22, S30L, IONPI19, and HB300SPX in kcal·mol^{-1}. The whiskers of the boxplots span the entire error range (ErrR) in all cases. *Only equilibrium geometries of HB300SPX were taken into account.

ωB97X-3c from ωB97X-D4 evaluated in the QZ basis set is evident. However, ωB97X-3c is still on par with B3LYP-D4 and better than r^{2}SCAN-3c. From the comparison with functionals evaluated in the QZ basis, it appears that the interaction energies of charged species, such as compiled in the CHB6, AHB21, and WATER27 sets, benefit from larger basis sets. This observation also holds true for interaction energies of n-alkane dimers in ADIM6. For many of the remaining benchmark sets, there is a significant improvement over r^{2}SCAN-3c and comparable performance to the other DFAs tested. For the halogen and hydrogen-bonded complexes in CARBHB12 and HAL59, the performance of ωB97X-3c is similar to that of the parent functional ωB97X-V/D4 evaluated in the QZ basis set and better than that of the other competitors, which is in particular remarkable given the clearly smaller number of basis functions in ωB97X-3c. A more in-depth discussion of the performance for NCIs is provided in the following two sections C.5.4 and C.5.5.

C.5.4 Intermolecular Non-Covalent Interactions of Large Systems

Non-covalent interactions of large systems are particularly challenging for methods using small basis sets, since BSSE and BSIE effects have a greater relative effect here than for thermochemical properties of small systems. On the other hand, the application of large basis sets is often not computationally feasible, which is also due to a massive increase in memory consumption (see also Sec. C.5.2 and Fig. C.3). Moreover, larger basis sets occasionally require distinctly more SCF steps until convergence than small basis sets, and it is sometimes difficult to get the SCF calculation converged at all.

It is therefore all the more desirable that ωB97X-3c is able to describe also large systems accurately. In the following subsection, a variety of benchmark sets covering intermolecular NCIs is discussed: Fig. C.5 shows error distributions for the S22 (small organic dimers, part of GMTKN55), S30L (realistic host-guest complexes), IONPI19 (non-covalent ion-π interactions), and HB300SPX[364] (hydrogen bonds) benchmark sets. Results for further benchmarks such as L7 (large molecule NCIs), R160x6[365] (repulsive NCI contacts), and CHAL336[366] (chalcogen bonding) are shown in the SI.

ωB97X-3c exhibits very small MAEs (hereafter in kcal·mol^{-1} in parentheses) as well as moderate error spreads for all four benchmark sets shown in Fig. C.5. The well-known S22 benchmark set from GMTKN55 impressively demonstrates the capabilities of the vDZP basis set in ωB97X-3c: With an MAE of 0.21 kcal·mol^{-1}, ωB97X-3c is among the best DFAs ever tested, and in this comparison only beaten by ωB97X-D4 (0.15). The ErrR is the smallest of all tested methods in this comparison. For S30L, which assesses host-guest interactions of larger supramolecular complexes, ωB97X-3c (1.7) is very close to the best-performer r^{2}SCAN-3c (1.5) and is even more accurate than ωB97X-D4 (1.9) and by far better than ωB97X-V (4.0), each of which evaluated in the QZ basis. The noncovalent ion-π interactions in IONPI19 are slightly too repulsive with ωB97X-3c(1.0), though still better described than with r^{2}SCAN-3c (1.3) and B3LYP-D4 (1.4). Again, ωB97X-D4 is the best-performing method in this comparison. The MAE and the ErrR for the description of hydrogen bonding interactions in HB300SPX could be reduced by 50 % with ωB97X-3c compared to r^{2}SCAN-3c. Furthermore, there is no systematic error with an mean error (ME) of 0.1 kcal·mol^{-1}. ωB97X-3c (0.29) is almost on par with the best-performing methods ωB97X-V/D4 (0.21/0.20). Also, B97M-V exhibits a very low MAE of 0.23 kcal·mol^{-1}.

Besides the benchmarks shown in Fig. C.5, error box plots for further benchmark sets on NCIs are given in Sec. II of the SI and are discussed in the following. Extended molecular complexes as present in L7, stabilized mainly by dispersion interactions, are insightful test cases as their description is challenging for contemporary computational methods. ωB97X-3c (1.0) outperforms all tested methods except ωB97X-D4 (0.9) and is closely followed by r^{2}SCAN-3c (1.1) (cf. to Fig. S4 in the SI). However, small differences in the MAEs should be treated with caution as the discussion about highly accurate ($\Delta < 0.5$ kcal·mol^{-1}) reference values is still ongoing[367] (see Sec. II of the SI for details on the reference data). Furthermore, the statistical size of seven sample points is at least questionable. Interaction energies of repulsively interacting fragments studied in R160x6 are slightly too repulsive with ωB97X-3c but still improved compared to r^{2}SCAN-3c. The corresponding reference data have been revised during the evaluation of r^{2}SCAN-3c (cf. Ref. [214]). In CHAL336, chalcogen bonds of mainly heavier elements are tested. Again, ωB97X-3c outperforms r^{2}SCAN-3c and has only slightly larger errors than the other methods assessed in the QZ basis. This is particularly noteworthy since polarization effects can play a larger role for heavy atoms, while ωB97X-3c relies on large-core ECPs avoiding the explicit treatment of core electrons.

The excellent results discussed in this section suggest ωB97X-3c as the ideal method for the investigation of large, non-covalently bound complexes. This is particularly noteworthy from the aspects of computational effort, memory requirements and convergence issues mentioned earlier, which are mostly overcome with ωB97X-3c. Therefore, it was already utilized as the reference method for benchmarking (semi)empirical methods on NCIs of very large systems consisting of up to 2000 atoms[116].

Remarkably, ωB97X-D4/QZ employing the revised D4 parameterization presented here yields the lowest MAE of all tested methods in five of the seven discussed benchmark sets for intermolecular NCIs.

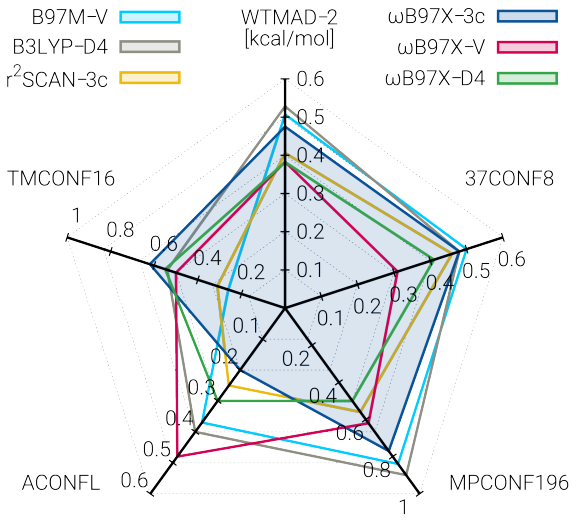

Figure C.6: MAEs for the conformer and intramolecular interaction energy benchmark sets 37CONF8, MPCONF196, ACONFL, and TMCONF16 in kcal·mol^{-1}. In addition, the WTMAD-2 value for the four benchmark sets is shown. The blue shaded error displays the error by ωB97X-3c.

C.5.5 Conformational Energies and Intramolecular Non-Covalent Interactions

Even though the application for conformer ranking is not the main target of ωB97X-3c, further benchmark sets on conformer energies going beyond those in GMTKN55 were evaluated. Specifically, conformational energies of 1) typical organic molecules (37CONF8), 2) peptides and small macrocycles (MPCONF196), 3) long *n*-alkane chains (ACONFL), and 4) transition metal complexes (TMCONF16, subset of TMCONF40[368]) were considered. In Fig. C.6, MAEs of each conformer set are shown together with the WTMAD-2 over all sets. The radar chart impressively demonstrates the performance of r^{2}SCAN-3c for conformational energies. In the overall performance, it is only marginally surpassed by ωB97X-V/D4. ωB97X-3c follows r^{2}SCAN-3c and outperforms B3LYP-D4 and B97M-V. For the individual benchmark sets, the aforementioned ranking consistently holds for ωB97X-3c, except for TMCONF16 and ACONFL. For the latter, ωB97X-3c shows an impressive MAE of 0.2 kcal·mol^{-1}, which is already close to the accuracy of the reference method and proves to be better than the same method evaluated in the QZ basis. Very likely, the molecular optimized vDZP basis set offers its greatest strengths for these or similar kinds of systems. The slightly larger errors for TMCONF16 can be attributed to the use of large-core ECPs as explicit interactions with electrons closer to the nucleus are more relevant for transition metal complexes such as evaluated in this test set.

C.5.6 Barrier Heights

The accurate description of barrier heights is challenging for efficient (m)-GGA methods due to the presence of SIE which is diminished here by the inclusion of range-separated Fock exchange. The consequence is well illustrated in Fig. C.7 showing error distributions for the BH9 benchmark set. BH9 consists of 449 chemical reactions "belonging to nine types common in organic chemistry and biochemistry"[363]. For each reaction, the reaction energy and two barrier heights (forward and backward) are given.

Comparing the performance of ωB97X-3c to that of r^{2}SCAN-3c, the error in reaction energies is very similar whereas the barrier heights are much more accurate with ωB97X-3c in both directions. In fact, the

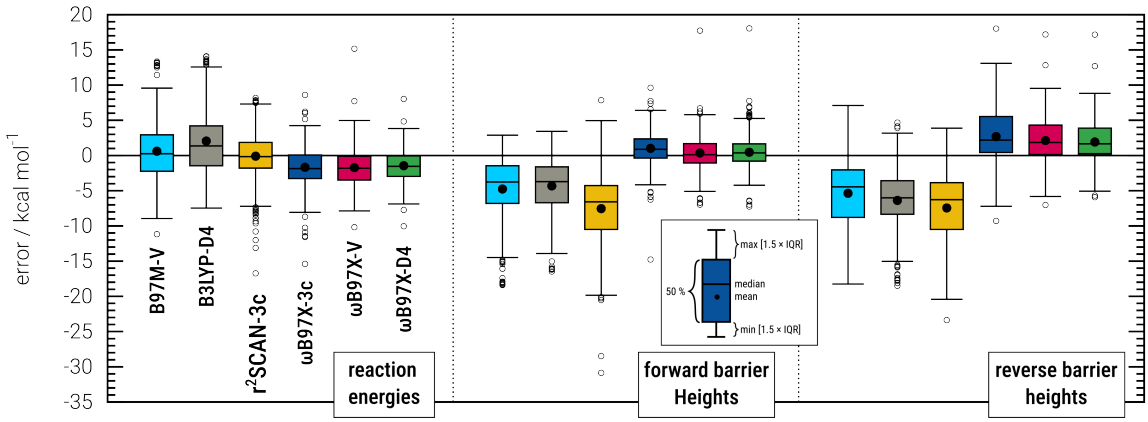

Figure C.7: Boxplots for each tested DFA for reaction energies, forward and backward barrier heights of the organic molecules composed in the BH9 benchmark set (errors given in kcal·mol^{-1}). The inlet displays the utilized statistical boxplot definition, in which 'IQR' corresponds to the interquartile range, the distance between the upper and lower quartiles. Data points outside of the whiskers are labeled as outliers in form of circles.

Table C.5: Statistical evaluation of transition metal thermochemistry in the MOR41, ROST61, MLA24 and revMOBH35 benchmark sets in kcal·mol^{-1}. Except for the "3c" composite methods, the tested methods employ the large def2-QZVPP basis set.

MAE / kcal·mol^{-1}	MOR41	ROST61	MLA24	revMOBH35
B97M-V/QZ	3.9	3.4	2.3	1.8
B3LYP-D4/QZ	4.2	3.3	4.0	2.5
r^{2}SCAN-3c	3.7	2.9	4.2	2.7
ωB97X-3c	3.4	4.6	1.7	3.2
ωB97X-V/QZ	2.2	2.8	1.7	2.3
ωB97X-D4/QZ	2.6	3.2	1.7	2.4

MAE of ωB97X-3c (1.9) for the forward barrier heights is on par with that of ωB97X-V/D4 (each 1.8) and for the backward barrier heights only slightly worse (3.4) than with ωB97X-D4 (2.7) and ωB97X-V (3.1). B3LYP-D4, B97M-V, and r^{2}SCAN-3c show two to three times larger MAEs (4.6-7.6) than ωB97X-3c for the barrier heights in both directions. A similar conclusion holds for the standard deviations. The larger errors of B3LYP-D4, B97M-V, and r^{2}SCAN-3c can be clearly attributed to SIE, as the MAE is almost equal to the negative ME for all three methods. Consequently, barrier heights are systematically underestimated (see also Fig. C.7). Evidently, the constant admixture of 20 % exact exchange as in B3LYP-D4 cannot really remedy this SIE-related shortcoming. Instead, range-separated Fock exchange is required as in the ωB97X-V/D4/3c methods. In line with the aforementioned results, ωB97X-D4 applied with the large QZ basis set is the best-performing method in this comparison.

C.5.7 Metal-Organic Reactions

The benchmark sets shown so far covered mostly only main-group chemistry. In the following, the performance for transition metal thermochemistry is discussed separately. The electronic structure of transition metal

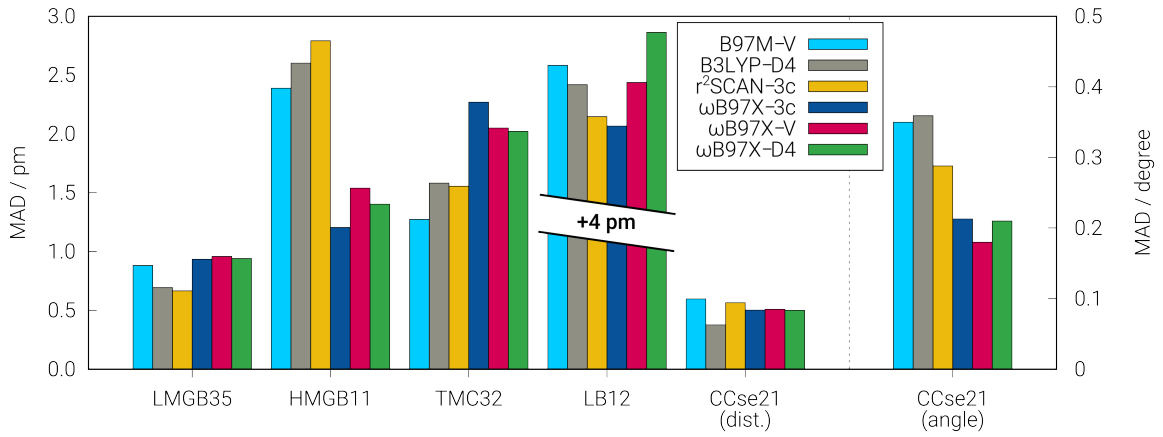

Figure C.8: MAEs of bond lengths (pm) in the molecular geometry benchmark sets LMGB35, HMGB11, TMC32, LB12, and CCse21. In the latter, also bond angles (degree) are evaluated. For geometry optimizations in `ORCA 5.0.3`, the settings `DEFGRID3` and `TightSCF` were applied throughout.

complexes is usually more complex than that of main group compounds. Therefore, slightly larger deviations of ωB97X-3c compared to methods employing converged QZ basis sets are to be expected. An additional challenge for ωB97X-3c arises from the lack of explicit treatment of core electrons due to the use of large-core ECPs as core-valence contributions can have a noticeable impact in transition metal chemistry.

The performance in this field is evaluated with four benchmark sets, all of them employing high-level CCSD(T) reference data. They cover closed-shell metal-organic reactions (MOR41[369]), chemically relevant open-shell transition metal reactions (ROST61[290]), polyethylene chains cross-linked by metal atoms (MLA24[370]), and metal-organic barrier heights (revised MOBH35 benchmark[371,372], termed revMOBH35). Results in the form of MAEs are presented in Tab. C.5 for all four benchmark sets and for each tested method.

Comparing the performance of ωB97X-3c, a difference between closed-shell and open-shell transition metal reactions is visible: For MOR41, a good result, even better than that of B97M-V, r^{2}SCAN-3c, and B3LYP-D4, is observed with ωB97X-3c. This is in agreement with MLA24, for which ωB97X-3c is even the best-performing method, together with ωB97X-V/D4. Remarkably, r^{2}SCAN-3c and B3LYP-D4 show two- to threefold larger errors than the ωB97X-V/D4/3c methods for MLA24. In contrast, ROST61 and revMOBH35 reveal larger deviations for ωB97X-3c. Still, the standard deviation (StdDev), the ErrR, and the ME are similar to the competing methods and neither outliers nor systematic errors were observed. The partly larger errors for transition metal chemistry can be attributed to the missing explicit treatment of interactions between valence electrons and electrons in lower shells, as large-core ECPs are used instead. Apparently, this is more pronounced for open-shell reactions in ROST61 compared to the closed-shell benchmarks. Unlike to the findings in previous sections, ωB97X-V shows slightly smaller errors than ωB97X-D4 for all four benchmark sets. This is in agreement with observations in the MOR41 and ROST61 benchmark sets, for which a similar trend was found in the comparison of B3LYP-D4 and B3LYP-NL.

C.5.8 Molecular Structures

Even though molecular geometries are not the main focus of the method, it can be convenient in some cases to utilize an efficient RSH method without the need for TZ / QZ basis sets to produce accurate structures. This is of special importance if, e.g., CT or SIE effects play a role, for which ωB97X-3c performs very well. To check

the performance for this application, we compare optimized equilibrium distances for five different geometry benchmark sets and also equilibrium bond angles for one of them, each employing experimental data as the reference. LMGB35 contains light main group bonds of very small first and second-row molecules[373], while HMGB11 consists of heavy main group molecules. LB12 covers larger organic molecules with chemically interesting bonding situations. Each of them was compiled during the development of PBEh-3c (see Ref. [115]). Furthermore, TMC32[374] containing 3d transition metal complexes and CCse21[375] (small molecules) were taken into account. Error bars displaying MAEs for each of the sets are given in Fig. C.8.

Remarkably, ωB97X-3c outperforms every tested method, including ωB97X-D4/QZ, for the HMGB11 and LB12 sets, and is on par with ωB97X-V/D4 for LMGB35. This holds not only for the MAE but also for the ME, showing that no systematic over- or underestimation of bond lengths occurs. Among these, the result for HMGB11 is probably the most impressive since heavy atoms (up to Pb_2Me_6) are involved, probing the influence of the large-core ECPs. Evidently, the molecule-optimized and fine-balanced basis set is superior in these cases even to the combination of large QZ basis sets with small-core (def2) ECPs, and without the need for further empirical corrections. CCse21 shows for ωB97X-3c a similar performance to that of ωB97X-V and ωB97X-D4 with only B3LYP-D4 showing a lower MAE. However, the relative deviation to the MAE of the best-performing method (B3LYP-D4) is very small (< 0.1 pm) and potentially below the uncertainty of the reference data. Looking at the bond angles in CCse21, ωB97X-3c together with ωB97X-V/D4 yield the smallest deviations. TMC32 with electronically more difficult systems reveals a slightly different picture, as ωB97X-3c displays larger errors than the best-performing methods B97M-V, B3LYP-D4, and r^{2}SCAN-3c. On the one hand, this is related to findings for transition metal chemistry discussed in Sec. C.5.7, where the large-core ECPs seem to have a more severe influence. On the other hand, the performance of all tested methods derived from ωB97X-V is very similar and thus, this might also be a limitation inherent to the chosen DFA. In contrast to results for the majority of the thermochemistry benchmark sets, ωB97X-D4 does not perform better than ωB97X-V for geometries.

The presented geometry benchmark sets convincingly demonstrate that ωB97X-3c is not only capable of describing thermochemistry at the hybrid DFT/QZ level but also equilibrium geometries of mostly covalently bound systems. This is particularly encouraging given that ωB97X-3c comes without any further empirical corrections such as SRB correction in B97-3c or the gCP correction in r^{2}SCAN-3c (cf. to Tab. C.1).

Besides these systems, also dissociation curves of non-covalently bound fragments were investigated within the S66x8[376,377] and HB300SPX[364] benchmark sets (see Tab. I in the SI). For these sets, slightly too-long equilibrium distances were observed. Nevertheless, the mean absolute difference between association energies (calculated at the same ωB97X-V/QZ level of dimer complex geometries optimized with ωB97X-3c compared to ωB97X-V/QZ is about 0.2 kcal·mol^{-1} or 5 % (see Sec. I-B of the SI for details), which corresponds to the error of very accurate DFAs on fixed geometries. Thus, in principle, ωB97X-3c yields also reliable molecular geometries of non-covalently bound complexes. However, due to the still considerable computational cost associated with ωB97X-3c in the vast majority of cases, m-GGA methods such as r^{2}SCAN-3c are to be preferred due to a better cost-benefit ratio.

C.6 Summary and Conclusions

We presented a new composite DFT method based on one of the leading range-separated hybrid DFAs available on the market. It offers a unique combination of a molecule-optimized polarized valence double-ζ basis set, matching large-core ECPs reducing the computational effort and the conventionally adapted D4 dispersion correction. We show that the vDZP basis exhibits qualitative advantages over usual double-ζ basis

sets as it is virtually free of BSSE and describes typical electron densities very similar to TZ or even large QZ basis sets. Small remaining BSSE contributions are absorbed by the D4 parameterization.

ωB97X-3c clearly outperforms the commonly used dispersion-corrected B3LYP functional in a large basis set at a small fraction of the computational cost. Furthermore, r^{2}SCAN-3c as its direct predecessor and B97M-V as one of the leading m-GGAs are beaten by ωB97X-3c in most cases. ωB97X-3c can often even keep up with the parent functional ωB97X-V/D4 evaluated in an almost converged QZ basis set. This conclusion is based on a huge variety of thermochemistry and geometry benchmark sets, with which we thoroughly investigated the behavior of ωB97X-3c. At this point, we highlight that in all comparisons, only well-performing state-of-the-art DFT methods were utilized in addition to the commonly used B3LYP-D4, which further underlines the high performance of the presented method. We especially suggest ωB97X-3c as an ideal method for investigating the thermochemistry of supramolecular complexes, especially if 1) a higher accuracy than with (m)-GGA methods is desired or electronic over-delocalization prevents (semi-)local DFT methods from SCF convergence, or 2) wall time and/or memory restrictions make the use of (hybrid) DFT methods together with larger basis sets impossible. In addition, the efficient study of barrier heights and reaction energies (especially for large systems) is a particular strength of ωB97X-3c. The quality of, e.g., molecular properties and excited state chemistry with ωB97X-3c for possible further applications is to be investigated in future studies.

Moreover, a new parameterization of the D4 dispersion correction for the replacement of VV10 in ωB97X-V was presented. We show that it is superior to the originally used VV10 variant in most cases and reduces the WTMAD-2 of GMTKN55 by 0.6 kcal·mol^{-1} compared to a previous D4 parameterization. To our knowledge, ωB97X-D4 is the second-best-performing hybrid DFA of all DFT methods ever tested on the GMTKN55 data set.

C.7 Data Availability

The data that supports the findings of this study are available within the article and its supplementary material. Any further information is available upon request from the authors.

C.8 Supplementary Material

In the supplementary material, which is available free of charge, we provide another example of an electron density with the vDZP basis set and further investigations on the influence of ECPs on the BSSE, additional information on the statistical descriptors used throughout, error plots for NCI benchmark sets L7, R160x6, and CHAL336, tabulated statistical data and additional explanations to the non-covalent molecular geometry benchmark sets (S66x8 and HB300SPX), the vDZP basis set and matching ECPs in tabulated form, detailed information on the usage of the dftd4 stand-alone program with the appropriate parameters and a reference input and total energy output (`ORCA 5.0.3`) for a molecular example. Moreover, we include a spreadsheet (`benchmarkdata_wB97X-3c.ods`) with raw data for all given benchmark sets. The vDZP basis set (including the ECPs) is also provided in text file format for ORCA and TURBOMOLE within the `vDZP_basis.zip` archive.

C.9 Acknowledgements

The German Science Foundation (DFG) is gratefully acknowledged for financial support through the SPP 2363 “Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning”. M.M. thanks the Fonds der Chemischen Industrie (FCI) for funding under a Kekulé scholarship and T. Gasevic and S. Ehlert for technical support. Moreover, the authors thank U. Huniar (*COSMOlogig GmbH*) for providing an adapted TURBOMOLE development version.

C.10 Conflict of Interest

The authors have no conflicts to disclose.

APPENDIX D

Appendix: A Non-Self-Consistent Tight-Binding Electronic Structure Potential in a Polarized Double-ζ Basis Set for all *spd*-Block Elements up to Z=86

Stefan Grimme,[†] Marcel Müller,[†] Andreas Hansen[†]

Received: December 06, 2022
Published online: March 22, 2023

Reprinted in Appendix D with permission[§] from
S. Grimme, M. Müller, and A. Hansen, *A non-self-consistent tight-binding electronic structure potential in a polarized double-ζ basis set for all spd-block elements up to $Z = 86$*, The Journal of Chemical Physics **158** *12* (2023) 124111, DOI: 10.1063/5.0137838

Own contributions

- Translating the development code into equations
- Implementation into the `xtb` code
- Performing all calculations and generation of the results
- Interpretation of the results
- Writing of the manuscript

[†] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany
[§] Permission requests to reuse material from this chapter should be directed to AIP Publishing.

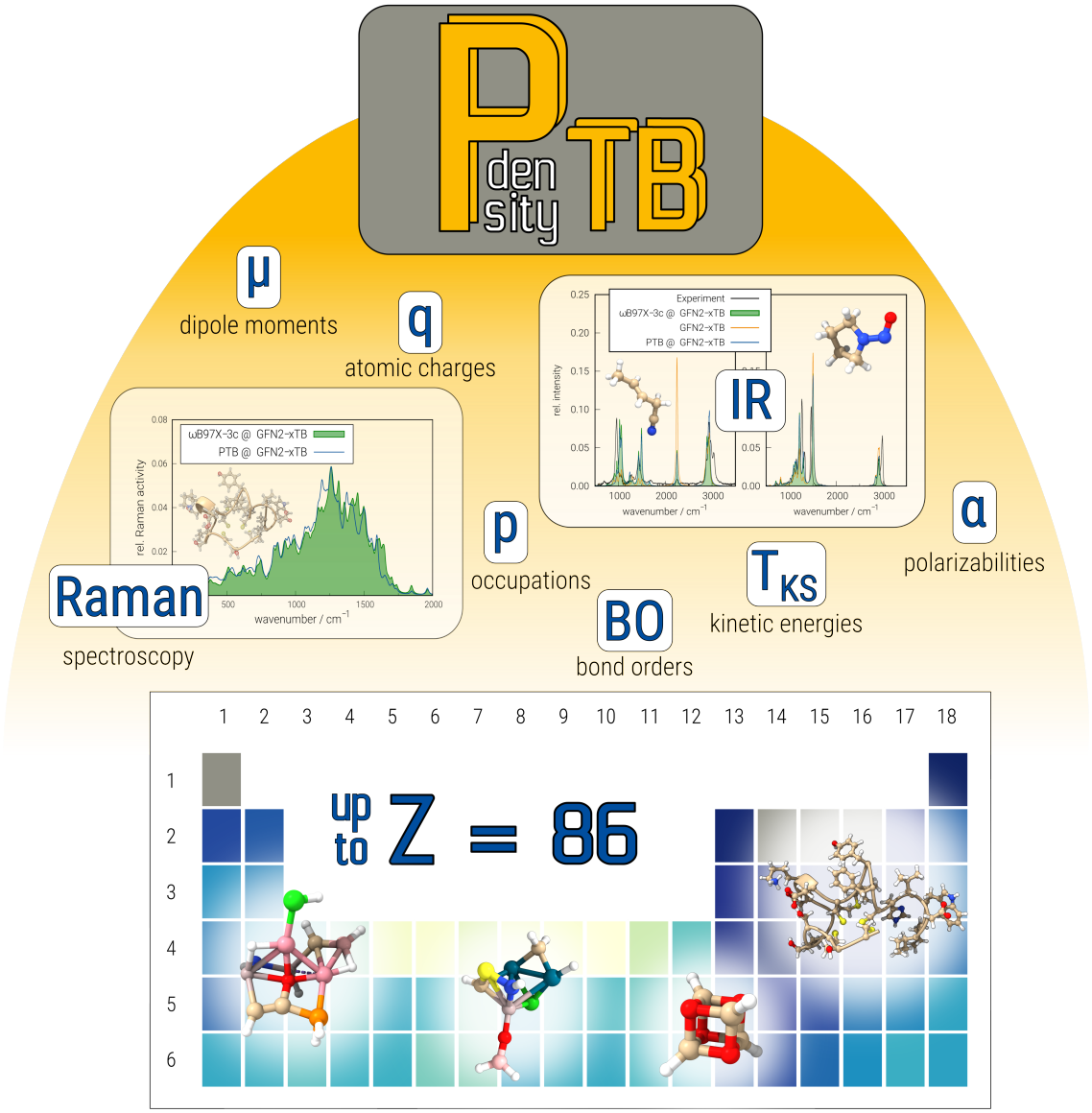

Figure D.1: Cover image for volume 158, issue 12, of the Journal of Chemical Physics.

D.1 Abstract

Existing semi-empirical molecular orbital methods suffer from the usually minimal atomic orbital (AO) basis set used to simplify the calculations. Here, a completely new and consistently parameterized tight-binding electronic structure Hamiltonian evaluated in a deeply contracted, properly polarized valence double-ζ (vDZP) is described. The inner-shell electrons are accounted for by standard, large-core effective potentials and approximations to them. The primary target of this so-called PTB method is to reproduce the one-particle density matrix **P** of a molecular ωB97X-V range-separated hybrid density functional theory (DFT) calculation in exactly the same basis set. Additional properties considered are orbital energies, dipole polarizabilities

as well as dipole moment and dipole polarizability derivatives. Key features of the method are a) non-self-consistency with an overall fixed number of only three required matrix diagonalizations, b) only AO overlap integrals are needed to construct the effective Hamiltonian matrix, c) new **P**-dependent terms emulating non-local exchange are included, and d) only element-specific empirical parameters (about 50 per element) need to be determined. The method globally achieves high accuracy for the target properties at a speed-up compared to the ωB97X-V/vDZP reference of about 3-4 orders of magnitude. It performs robustly for difficult transition metal complexes, highly-charged or zwitterionic systems as well as for chemically unusual bonding situations indicating a generally robust approximation of the (self-consistent) Kohn-Sham potential. As an example application, the vibrational Raman spectrum of an entire protein with 327 atoms with respect to the DFT reference calculation is shown. The method may be used out-of-the-box to generate molecular/atomic features for machine learning applications or as the basis for accurate high-speed DFT methods.

D.2 Introduction

Quantum chemistry-based electronic structure calculations are currently dominated by KS-DFT[126] for system sizes up to a few hundred atoms. For larger systems with 500-1000 or more atoms,[289,378] or for screening of many thousand candidate structures in automatic chemical space exploration workflows,[164,171,323] however, even simplified DFT schemes such as "3c" composite or related methods[153,213,214] become computationally unfeasible. This is one of the reasons why SQM methods[72,379,380] are currently experiencing a renaissance in the community, both regarding the development of improved methods or in large-scale application studies.[381–384] SQM methods provide a well-known alternative route because they are at least two orders of magnitude faster than conventional DFT treatments. They approximate single-reference HF or first-principles DFT and have been investigated extensively already about 3-4 decades ago.[72]

Along the lines of the development of the so-called density functional (DF) TB methods pioneered by Seifert, Elstner, and Frauenheim,[267,268,385–388] we have contributed to the field with the GFN*n*-xTB family of methods.[189,259,277–279] They are rooted in DFTB theory but were developed consistently from the very beginning for all elements up to $Z = 86$ (radon) and secondly, with a focus on the primary target properties Geometries, Frequencies, and Non-covalent interactions (GFN). In particular, the latest GFN2-xTB method including multipole electrostatics as well as density-dependent dispersion terms has found widespread use and is routinely used worldwide in various applications.[323,389–393]

Despite the success of SQM methods also in combination with modern ML approaches,[394–397] there are important challenges ahead regarding the further development of this energy-efficient technology. The most serious issue is the inherently limited accuracy of SQM methods for chemically important properties, namely conformational energies,[278,368] standard thermochemistry, or for electronically difficult systems like transition metals for which DFT accuracy is far away.[398] Furthermore, molecular properties that are sensitive to details of the electronic wavefunction like IR intensities are notoriously inaccurate.[314] The same is true for Raman activities, for which SQM approaches[399,400] are rarely used as they have not yet been accurate and broadly applicable enough for routine application.

One of the fundamental reasons for the mentioned SQM problems (besides the typical integral approximations applied very differently in the various methods) is in our opinion the use of mostly minimal, atom-centered AO basis functions. Although third- and higher-row elements in DFTB and GFN-xTB include d-functions, the most important "core" elements HCNOF lack such polarization AOs and in fact, the typically considered valence shells are minimally described. For good reasons, this somewhat non-physical description of electronic structure was abandoned already decades ago for HFs and DFT treatments. The current view is

that reasonable mean-field calculations should employ at least properly polarized valence double-ζ (vDZP) type basis sets consisting of two valence shells for each occupied one in the atomic ground state. This is not fulfilled by any current SQM method. However, the reasoning for a minimal AO representation in SQM is clearly two-fold: first, it speeds up integral and in particular matrix diagonalization steps, and secondly, often overlooked, it avoids numerically large and complicated one-center matrix elements of the Hamiltonian between AOs of the same angular momentum in a vDZP basis.

Very recently we designed and optimized a completely new vDZP basis set for all elements Z=1-86 including the lanthanides with the goal to use it also for an extended basis set TB method.[1] The basic idea is to avoid the matrix diagonalization bottleneck that appears in many low-cost electronic structure methods (including DFT for system sizes of 5000-10000 atoms) by constructing a compact (medium-sized), but inherently accurate AO basis set. This was achieved by a) relative deep contraction (a large number of Gaussian primitive functions per AOs) and b) full optimization of primitive exponents and contraction coefficients with respect to DFT (ωB97X-D3[356]) energies of representative neutral and charged atoms and molecules of the respective element. This new vDZP basis set performs universally well for thermochemistry and non-covalent interactions and provides small basis set superposition and incompleteness errors with triple-ζ or even quadruple-ζ quality. It represents a fundamental ingredient for the here described method and is used in unmodified form, see Ref. [1] for more discussion and details.

The vDZP basis combined with large-core effective potentials (ECPs) from the literature (ECP-xx-MWB,[140,141,349,350,352,354] ECP-10-MDF,[137] CRENBL,[144,347,348,351] SBKJC-ECP,[353]) the adjusted D4 dispersion correction,[120] and the ωB97X-V RSH density functional[103] [without the -V (VV10[125]) dispersion part] was used in Ref. [401] to create a composite DFT method dubbed ωB97X-3c as a new RSH member of the "3c" family. Comparisons of results with self-consistent vs. non-self-consistent implementations of the VV10 dispersion correction have shown that its influence on the electron density is practically negligible[113] and hence, the dispersion uncorrected part of ωB97X-V is used as the base functional here.

For a wide range of benchmark sets, ωB97X-3c provides thermochemical results closely approaching the accuracy of the best density functionals available evaluated in much larger AO basis sets. As an RSH functional, ωB97X-3c is practically free of SIE-related problems like vanishing orbital energy gap, artificial CT, or charge/spin over-delocalization. It seems to be the ideal choice as a reference method for developing a new TB method evaluated in the large vDZP basis, even though other DFT functionals could in principle also be used for this purpose.

Opposed to many older SQM approaches which try to approximate HFs or DFT results rather "globally" for almost all chemical properties and systems, but in the spirit of the special purpose GFN methods, the approach proposed here aims primarily at reproducing a fully SCF converged ωB97X-3c one-particle density matrix **P** as efficiently as possible. Instead of the usual QC procedure of first developing an approximate energy expression and then deriving the corresponding electronic potential from it, we turn this procedure around and describe here first an approximate potential and consider suitable energy expressions in further work. Additionally, an approximate response treatment is introduced to compute the dipole polarizability and by numerical differentiation Raman intensities. Calculation of first-order properties like dipole or quadrupole moments including numerical derivatives, e.g., for computation of IR intensities is straightforward.

After the presentation of the PTB theory in the following chapter, Sec. D.3.6 and D.3.7 describe the determination of the empirical parameters and the compilation of the fit sets. In Sec. D.4, the results for chemically relevant properties derived from **P** like atomic charges, shell populations, bond orders, dipole moments, and dipole polarizabilities are evaluated with reference to the corresponding ωB97X-3c data. Because this can not be shown comprehensively for the 72 elements which have been parameterized so far (86 minus the 14 excluded lanthanoids, see below), results for selected elements including some difficult and

unusual molecules are shown.

D.3 Theory

The method, termed PTB from now on and described below in full detail, has the following characteristics:

- A full self-consistent field (SCF) treatment is avoided and the target DFT reference density matrix is obtained in just two matrix diagonalization steps starting from semi-classically obtained atomic charges. A third diagonalization is required for a shell-wise population analysis which forms the basis of the electrostatic treatment.
- The number of empirical parameters is kept to a minimum by entirely avoiding pair-specific potentials. Depending on the number of shells for an element in the vDZP basis, the number of empirical parameters varies between 35 (H, He) over 49 (main group) to 63 (d-block). The main reason for the greater number of element-specific parameters compared to existing SQM schemes is the larger AO basis set.
- The method is applicable to diverse chemical systems composed of elements from different corners of the periodic table and ranging from organic molecules to inorganic main group compounds and (transition) metal complexes. Electronically degenerate (multi-configurational) situations are handled by a finite electronic temperature approach as in GFN*n*-xTB, which allows a qualitatively correct description of poly-radicals or dissociation processes in a spin-restricted manner.

That one can mimic the result of an ordinary SCF procedure by a simple two-step diagonalization scheme even for electronically rather difficult cases like transition metal complexes seems hard to believe but is actually possible as will be shown below. Besides the usefulness of PTB in practice, this finding is theoretically the most important result of this work. It demonstrates that the separation of energy and potential in electronic structure theory may be beneficial in QC method development. In this sense, our basic ansatz is related to recent work by a Sim/Burke collaboration which separates the error from approximate density functionals like PBE into a (small) functional and larger density (potential) driven error leading to new methods termed density-corrected (DC) DFT.[315] Along these lines, energy expressions for PTB will be developed that are published separately in the near future.

Turning the argumentation around, it could be the case that traditional SQM with SCF is *only* possible in a minimal basis set with little flexibility so that errors in the potential can not destructively be amplified during the SCF leading overall to rather stable results. We think that extended basis set SQM methods *with* SCF are extremely difficult to establish because the variational freedom is already in a vDZP basis quite large and hence our non-SCF approach is without alternative at this point. For a recent, not fully satisfactory non-SCF attempt in the GFN framework termed GFN0-xTB, see Ref. [279].

The general workflow of a PTB calculation is schematically shown in Fig. D.2 and discussed in detail in the next sections.

D.3.1 The PTB Model Hamiltonian

Here we present details of the new QM model Hamiltonian termed PTB in the following. Its purpose is to provide a reasonably accurate density matrix $\mathbf{P}$ from which atomic charges q_A, orbital shell populations p_A^l, and Wiberg bond orders BO_{AB} are obtained (see Sec. D.3.2) (A refers to atoms or chemical elements, l to an

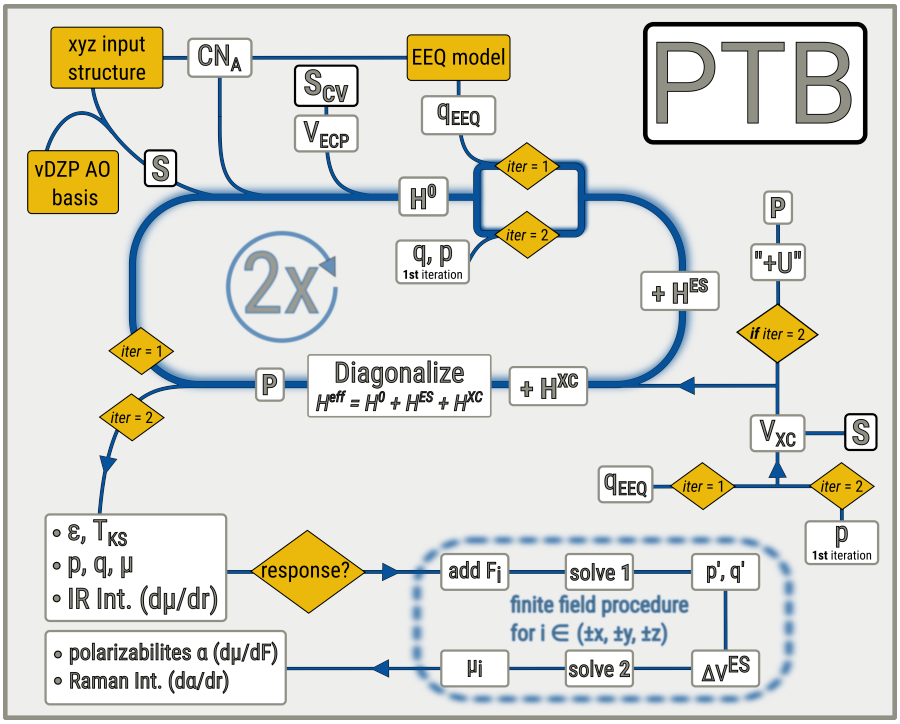

Figure D.2: Schematic representation of the non-self-consistent PTB workflow.

AO shell ($1s, 2s, 2p, ...$), l_a to its angular momentum ($s, p, d, ...$), and AB denotes an atom pair). Moreover, as already noted, derived properties like dipole moments, static dipole polarizabilities, and their geometric derivatives are of interest.

Opposed to most existing semi-empirical methods, it is defined in an extended, standard polarized valence (ECP) double-ζ Gaussian AO basis of [2s1p]/[2s2p1d] quality (vDZP) in the common notation for (non-metallic) first- and second-row elements. This basis set, dubbed vDZP, was presented with our new "3c" method, ωB97X-3c (see Ref. [1]), and detailed explanations, as well as quality checks, are given there. Since ωB97X-3c is evaluated in exactly the same basis set, it is the natural choice for the reference method.

Importantly, due to the use of QC as a basis for the computation of key quantities $\{q, p, BO\}$, artificial CT effects present in alternative semi-classical electronegativity equilibration (EEQ) models[120,402–404] to compute such quantities are completely avoided because discrete states are involved (integer-derivative-discontinuity conditions are obeyed).

More specifically, we diagonalize an effective Hamiltonian $\mathbf{H}^{\mathbf{eff}}$ in the above basis set twice

$$\mathbf{H}^{\mathbf{eff}}\mathbf{C} = \mathbf{S}\mathbf{C}\boldsymbol{\epsilon} \tag{D.1}$$

to obtain the orbital coefficients $\mathbf{C}$ and density matrix $\mathbf{P}$ from the occupied states (including Fermi smearing at finite electronic temperature $T = 300$ K). Before the first diagonalization, the q_A are initialized with values from the EEQ model utilized in the D4 dispersion correction, while the values obtained from the first step are employed in the subsequent diagonalization (two-step SCF as in Ref. [324]). This procedure is robust, and efficient and accounts for the most crucial polarization effects but avoids SCF convergence problems.

The matrix elements of $\mathbf{H}^{\mathbf{eff}}$ can be separated into contributions of extended Hückel theory (EHT) type

$H^0_{\kappa\lambda}$ and in isotropic electrostatic (IES) contributions which include terms of second and third-order (TO) in electron density fluctuations $\delta\rho$. We also include a term akin to one in DFT+U theory.[405] The general expression for the effective Hamiltonian reads

$$\langle\psi_\kappa|H^{\text{eff}}_{\kappa\lambda}|\psi_\lambda\rangle = \langle\psi_\kappa|H^0_{\kappa\lambda}|\psi_\lambda\rangle + \langle\psi_\kappa|H^{\text{IES+U+TO}}_{\kappa\lambda}|\psi_\lambda\rangle \quad \text{with} \quad (\kappa \in l(A),\ \lambda \in l'(B))\,. \tag{D.2}$$

where κ, λ are basis functions and l, l' denote atomic shells ($l = 1s, 2s, 2p, ...$) of atoms A and B, respectively. The H^0 matrix elements contain parameterized valence orbital energy levels (related to eigenstates of the free atom in a zeroth-order approximation in DFTB) which depend in a complicated way on the coordination number CN_A, the interatomic distances R_{AB}, and the type of orbital interaction (inter- or intra-shell). An EHT-type expression is also invoked for approximating the off-diagonal matrix elements. Furthermore, an approximation for the ECPs is included in H^0. $H^{\text{IES+U+TO}}$ adds dependencies on q_A (shell- and element-wise) in second and third-order and on the density matrix $\mathbf{P}$ for the "+U" contribution. "+U" is a technique traditionally applied to treat electron over-delocalization in LDA or GGA DFT by a distinction of "localized" and "delocalized" electrons, realized by a dependence on the orbital occupation.[406,407] The setup of H^{eff} and explicit expressions are given in subsections D.3.3 and D.3.4. In the following, quantities of type k^l_A denote shell-specific and k_A element-specific empirical (fit) parameters. Global fit parameters do not have any of these indices. A list of all empirical parameters, including a brief description of each, can be found in Tab. 1 in the Supporting Information (SI).

D.3.2 Population Analysis

For the setup of H in the second iteration as well as for evaluation of the final density matrix, an appropriate population analysis is required, with which atomic charges q_A and orbital shell populations p^l_A are obtained. For this purpose, we have – depending on the AO basis set and field of application – good experience with the well-known Mulliken or Löwdin schemes.

Given this general application, we propose a mixed approach for the transformation of the density matrix $\mathbf{P}$ expressed in the non-orthogonal AO basis to the orthogonal form $\mathbf{P}^{mix}$ for population analysis according to

$$\mathbf{P^{mix}} = \mathbf{S}^x\mathbf{P}\mathbf{S}^{1-x}, \tag{D.3}$$

where $x = \frac{1}{3}$. Here, $x = 0$ corresponds to the pure Mulliken partitioning, while $x = \frac{1}{2}$ yields the Löwdin approach. The chosen average combines the best properties of each method. Still, it avoids their pitfalls, i.e., unphysical (negative) shell populations with Mulliken or too strong covalency with the Löwdin scheme. The atomic shell charges used to compute the important second-order electrostatic interactions (see Sec. D.3.4) are obtained from $\mathbf{P^{mix}}$ as follows, i.e.,

$$p_{A,l} = \sum_{i\in l}^{\text{AO}} P^{\text{mix}}_{ii}, \tag{D.4}$$

and

$$q_{A,l} = p^{\text{ref}}_{A,l} - p_{A,l}, \tag{D.5}$$

where $p^{\text{ref}}_{A,l}$ is a reference shell population of the neutral atom in its ground state, and $p_{A,l}$ is the actual PTB shell population. The $p^{\text{ref}}_{A,l}$ values are obtained as an average from ωB97X-3c data for the molecules in the fit set where the shell populations are normalized by the total atomic charge to those of a neutral atom and are obtained with the same mixed Mulliken-Löwdin scheme as in Eq. D.3. Atomic charges are obtained as usual

from

$$q_A = Z_A - \sum_{i \in A}^{\mathrm{AO}} P_{ii}^{\mathrm{mix}}. \tag{D.6}$$

D.3.3 $\mathbf{H^0}$ Matrix

The two-center matrix elements of H_0 are given by:

$$H^0_{\kappa_A \lambda_B} = \frac{1}{2}(k_A^{l_a} + k_B^{l_a}) S^{S_{H^0}}_{\kappa\lambda} \left(H_\kappa + H_\lambda\right) \left(1 - \left(\frac{H_\kappa - H_\lambda}{H_\kappa + H_\lambda}\right)^2 k_{\mathrm{pol}}\right) \left(1 + \frac{\left(k_{r,A} + k_{r,B}\right) X}{R_{AB}}\right) + V^{\mathrm{ECP}}_{\kappa\lambda} + V^{\mathrm{XC}}_{\kappa\lambda}$$

$$\text{with } X = \begin{cases} k_{\mathrm{it},r}, & \text{if } n_{\mathrm{it}} = 1, \\ 1, & \text{if } n_{\mathrm{it}} = 2, \end{cases} \quad \text{and} \quad \left(A \neq B,\ \kappa \in l(A),\ \lambda \in l'(B),\ n \in \{1,2\}\right), \tag{D.7}$$

where k_{pol} is a scaling factor that takes large differences in the atomic energy levels H_κ into account, $k_A^{l_a}$ is the so-called Wolfsberg parameter in EHT and $k_{r,A}$ scales an additional interatomic distance dependence R_{AB} of the $\mathbf{H^0}$ elements. The distance dependence is scaled by the global parameter $k_{\mathrm{it},r}$ in the first iteration to account for the slightly changed setup of $\mathbf{H}$ in the second iteration. These contributions are multiplied with overlap matrix elements $S^{S_{H^0}}_{\kappa\lambda}$ calculated with modified basis set exponents, the origin of which is explained in detail in Sec. D.3.3. V^{ECP} is an approximation for the ECP in ωB97X-3c. The **o**ne-**c**enter **o**ff-**d**iagonal elements of H_0 are defined as follows:

$$H^0_{\kappa_A \lambda_A} = S^{S_{H^0}}_{\kappa\lambda} \left(H_\kappa + H_\lambda\right) X + \left(S^{S_{H^0}}_{\kappa\lambda}\right)^2 \left(H_\kappa + H_\lambda\right) k_{\mathrm{ocod},A} X + V^{\mathrm{ECP}}_{\kappa\lambda} + V^{\mathrm{XC}}_{\kappa\lambda}$$

$$\text{with } X = \begin{cases} k^{l_a}_{s,A} \cdot k_{\mathrm{it,ocod}}, & \text{if } n_{\mathrm{it}} = 1, \\ k^{l_a}_{s,A}, & \text{if } n_{\mathrm{it}} = 2. \end{cases} \quad \text{and} \quad \left(\kappa \in l(A), \lambda \in l'(B)\right). \tag{D.8}$$

In both parts, a parameter taking a different behavior for the first iteration into account is incorporated ("X"). The diagonal elements of $\mathbf{H^0}$ are given by

$$H^0_{\kappa_A \kappa_A} = S^{S_{H^0}}_{\kappa\kappa} \left(H_\kappa + H_\kappa\right) + V^{\mathrm{ECP}}_{\kappa\kappa} + V^{\mathrm{XC}}_{\kappa\kappa} \quad \text{with} \quad \left(\kappa \in l(A)\right). \tag{D.9}$$

The atomic energy levels H_κ are based on an empirical parameter H^l_A and are defined as follows

$$H_\kappa = H^l_A + k^l_{H,A} \left(\mathrm{CN}_A + \mathrm{CN}'_A k_{\mathrm{CN}',A}\right) + \mathrm{CN}'_A k_{\mathrm{shift},A} \quad \text{with} \quad \left(\kappa \in l(A)\right), \tag{D.10}$$

including dependencies on two different CN measures which are parameterized by $k^l_{H,A}$, $k_{\mathrm{CN}',A}$ and $k_{\mathrm{shift},A}$.

The CNs[117] CN_A and CN'_A are given by

$$\mathrm{CN}^{(\prime)}_A = \frac{1}{2} \sum_{B \neq A}^{N_{atoms}} \left(1 + erf\left(k^{\mathrm{CN}}_{\mathrm{erf}} \frac{R_{AB} - R^{\mathrm{cov}}_{AB}}{R^{\mathrm{cov}}_{AB}}\right)\right) \quad \text{with} \quad \begin{cases} R^{\mathrm{cov}}_{AB} = r_{f,A} + r_{f,B} & \text{if } \mathrm{CN}_A, \\ R^{\mathrm{cov}}_{AB} = \frac{4}{3}\left(r'_A + r'_B\right) & \text{if } \mathrm{CN}'_A, \end{cases} \tag{D.11}$$

where r' indicates standard covalent radii[117] and r_f are specially fitted element-specific radii. $k^{\mathrm{CN}}_{\mathrm{erf}}$ controls the steepness of the error function that is used for "counting" nearest neighbors. During the development of PTB, it turned out that the use of two different CNs with different (separable) radii includes important information about the second atomic coordination sphere.

Approximation for Fock Exchange

As an efficient approximation to the "exact" Fock exchange included in the RSH functional, a non-local term akin to those taken for simple ECPs (see below) is included

$$V^{\mathrm{XC}}_{\kappa\lambda} = \sum_{\rho} S^{S_{\mathrm{XC}}}_{\kappa\rho} \left(H_{\kappa} \frac{p_{A,l}}{m_{l_a}} a_{\mathrm{XC}} \right) S^{S_{\mathrm{XC}}}_{\rho\lambda}$$
$$\text{with} \begin{cases} a_{\mathrm{XC}} = k_{\mathrm{XC}_1,A} & \text{if } n_{iter} = 1, \\ a_{\mathrm{XC}} = k^{l}_{\mathrm{XC}_2,A} & \text{if } n_{iter} = 2, \end{cases} \quad \text{and} \quad (\kappa \in l(A),\ \lambda \in l'(B),\ \rho \in l''(C)), \tag{D.12}$$

in which m_{l_a} corresponds to the degeneracy $[s \rightarrow 1, p \rightarrow 3, d \rightarrow 5, ...]$ of shell l to normalize the shell occupation $p_{A,l}$. Similar to $\mathbf{H^0}$ itself, a modified overlap matrix is used (see Sec. D.3.3). Shell-resolved occupations in the first iteration are determined according to Eq. D.17. In iteration two, a shell-dependent parameter is utilized as actual shell-resolved charges from the first step are available.

Modified Overlap Integrals

In order to make some terms in the Hamiltonian spatially flexible but still obey the symmetries in the system, modified overlap integrals are introduced. The basic idea is to scale the Gaussian exponents $\alpha_{\kappa,0}$ of the primitive basis function κ by $k^{l}_{\alpha^{H^0},A}$ in the standard calculation of a modified overlap matrix $\mathbf{S}^{\mathbf{S}_{\mathbf{H^0}}}$ used in $\mathbf{H^0}$, and by $k^{l}_{\alpha,\mathrm{XC}}$ for the overlap matrix $\mathbf{S}^{\mathbf{S}_{\mathrm{XC}}}$ used for the approximated exchange potential $\mathbf{V}^{\mathrm{XC}}$. Furthermore, the scaling factor for the overlap matrix employed in the setup of $\mathbf{H^0}$ is modified in dependency of the shell charge $q_{A,l}$ together with an empirical parameter $k^{S^0}_{\mathrm{it},A}$ in the second iteration.

$$\alpha^{H^0}_{\kappa} = \alpha_{\kappa,0} k^{l}_{\alpha^{H^0},A} \cdot X,$$
$$\text{with} \quad X = \begin{cases} 1, & \text{if } n_{\mathrm{it}} = 1, \\ k^{S^0}_{\mathrm{it},A} \cdot q_{A,l}, & \text{if } n_{\mathrm{it}} = 2. \end{cases} \quad \text{and} \quad (\kappa \in l(A)), \tag{D.13}$$

$$\alpha^{V^{\mathrm{XC}}}_{\kappa} = \alpha_{\kappa,0} k^{l}_{\alpha^{\mathrm{XC}},A}, \quad \text{with } (\kappa \in l(A)). \tag{D.14}$$

The basis functions with scaled exponents are normalized as usual.

Approximation for Effective Core Potential

The approximation for the ECP of the inner (not included) electron shells is taken from Ref. [326] (see also Ref. [138, 408]) and reads

$$V^{\mathrm{ECP}}_{\kappa\lambda} = \sum_{\mu}^{l_{\mathrm{core}}} S_{\kappa\mu} \left(\epsilon^{\mathrm{core}}_{C,\mu} k^{\epsilon}_{\mathrm{ECP},C} \right) S_{\mu\lambda} \quad \text{with} \quad \left(\mu \in l_{\mathrm{core}}(C),\ \kappa \in l_{\mathrm{val}}(A),\ \lambda \in l'_{\mathrm{val}}(B) \right) \tag{D.15}$$

where $\epsilon^{\mathrm{core}}_{\mu}$ represents a Hartree-Fock orbital eigenvalue of "core"-shell μ of atom C (here, C denotes a core AO index, not a carbon atom). $k^{e}_{\mathrm{ECP},C}$ scales this orbital energy, respectively, the resulting contribution by $V^{\mathrm{ECP}}_{\kappa\lambda}$. As core orbitals, single-ζ STO-6G expansions from Stewart[145] using Clementi's STO exponents[409,410] are taken.

D.3.4 Electrostatic Potential and Third-Order Contributions

In the following, an empirically weighted average of the Ohno-Klopman- and Mataga-Nishimoto-damped Coulomb interaction $\gamma_{AB,ll'}$ is used, for which only the prefactors a_{OK} and a_{MN} are fitted:

$$\gamma_{AB,ll'} = \frac{a_{\mathrm{OK}}}{\sqrt{{R_{AB}}^2 + \left(\frac{2}{{\eta_A^l}^{-1} + {\eta_B^{l'}}^{-1}}\right)^{-2}}} + \frac{a_{\mathrm{MN}}}{R_{AB} + \left(\frac{2}{{\eta_A^l}^{-1} + {\eta_B^{l'}}^{-1}}\right)^{-1}}$$
$$\text{with } \left(l(A), l'(B) \text{ and } a_{\mathrm{MN}} + a_{\mathrm{OK}} = 1\right). \tag{D.16}$$

Instead of an arithmetic average, harmonically averaged chemical hardness parameters η_A^l are employed. According to our experience (and in agreement with the common view), other than arithmetic atomic averages (so-called combination rules) are almost always preferred and the harmonic one is taken as a usual fall-back when potentially negative quantities are involved. γ enters the electrostatic part of the Hamiltonian as given in detail below.

First Iteration

Initial guess atomic charges (q^{EEQ}) are determined by the semi-classical electronegativity equilibrium charge model (EEQ) as implemented in the D4 dispersion correction method.[120] It employs four parameters per element $\mathrm{EN}_A^{\mathrm{EEQ}}$, η_A^{EEQ}, κ_A^{EEQ} and a_A^{EEQ} (see Ref.[120] for details). Here, they are modified by the scaling parameters denoted $k_{\mathrm{EN},A}^{\mathrm{EEQ}}$, $k_{\eta,A}^{\mathrm{EEQ}}$, $k_{\kappa,A}^{\mathrm{EEQ}}$ and $k_{a,A}^{\mathrm{EEQ}}$. Shell charges result from a projection of the atomic EEQ charges onto shells as follows

$$q_{A,l}^{\mathrm{EEQ}} = p_{A,l}^{\mathrm{ref}} \left(1 - \frac{Z_A^{\mathrm{eff}} - q_A^{\mathrm{EEQ}}}{Z_A^{\mathrm{eff}}}\right). \tag{D.17}$$

In the first iteration, $H_{\kappa\lambda}^{\mathrm{IES}}$ is set up as follows

$$H_{\kappa\lambda}^{\mathrm{IES+TO}} = -S_{\kappa\lambda}\left(\frac{1}{2}\sum_C \sum_{l''} \left(\gamma_{AC,ll''} + \gamma_{BC,l'l''}\right) q_{C,l''}^{\mathrm{EEQ}} + \left(\left(q_A^{\mathrm{EEQ}}\right)^2 k_{\mathrm{TO},A} + \left(q_B^{\mathrm{EEQ}}\right)^2 k_{\mathrm{TO},B}\right)\right)$$
$$\text{with } \left(\kappa \in l(A), \lambda \in l'(B), l''(C)\right). \tag{D.18}$$

where $k_{\mathrm{TO},A}$ is an empirical parameter scaling the third-order contribution. For the damped Coulomb interaction $\gamma_{AB,ll'}$ (see Eq. D.16), the prefactors $a_{\mathrm{OK}} = 1.6$ and $a_{\mathrm{MN}} = -0.6$ are used. The chemical hardness values entering Eq. D.16 are defined as follows:

$$\eta_A^l = \eta_A^{\mathrm{EEQ}}\left(1 + 0.25 \cdot q_A^{\mathrm{EEQ}}\right) k_{1,\mathrm{A}}^l \tag{D.19}$$

η_A^{EEQ} is the element-wise hardness parameter carried over from the original EEQ model that is scaled by $k_{1,\mathrm{A}}^l$ (subscript 1 denotes the first iteration) and additionally depends on the atomic charge on atom A.

Second Iteration

$\mathbf{H}^{\mathrm{IES+TO+U}}$ contains in the second iteration contributions which can be attributed to 1) damped isotropic ES (resolved shell-wise), 2) third-order contributions in which squared atomic charges enter (both analog to

iteration one), and additionally 3) a term related to "DFT+U",[406,407] each of them using atomic and shell charges q and the density matrix $\mathbf{P}$ from the first iteration:

$$\begin{aligned} H_{\kappa\lambda}^{\text{IES+TO+U}} = -S_{\kappa\lambda}\Bigg(\frac{1}{2}\sum_{C}\sum_{l''}\left(\gamma_{AC,ll''}+\gamma_{BC,l'l''}\right)q_{C,l''} \\ +\left(q_A^2 k_{\text{TO},A}+q_B^2 k_{\text{TO},B}\right)\Bigg)+XP_{\kappa\lambda}\left(U_\kappa+U_\lambda\right)U_{AB}^{\text{damp}}, \\ \text{with}\quad X=\begin{cases} c_{U_{\text{d}},\text{A}}, & \text{if } \kappa=\lambda, \\ 1, & \text{else.}\end{cases} \\ \text{and}\ \left(\kappa\in l(A),\lambda\in l'(B),l''(C)\right). \end{aligned} \tag{D.20}$$

The diagonal elements of the "+U" contribution (i.e., $\kappa = \lambda$) are scaled with an additional parameter $c_{U_{\text{d}},\text{A}}$. The calculation of third-order contributions is equal to that described in Eq. D.18 for the first iteration. The damped Coulomb interaction $\gamma_{AB,ll'}$ (see Eq. D.16) is calculated by using a purely Ohno-Klopman formula ($a_{\text{OK}} = 1.0$ and $a_{\text{MN}} = 0.0$) and chemical hardness values scaled with $k_{\eta_2,A}^l$ as follows

$$\eta_A^l = \eta_A^{\text{EEQ}} k_{\eta_2,A}^l. \tag{D.21}$$

U_κ is defined as follows

$$U_\kappa = c_{U_{\text{eff}},A}^l\left(1-\left(c_{U_1,A}\cdot q_A + c_{U_2,A}\cdot\left(q_A\right)^2\right)\right) \quad \text{with} \quad \left(\kappa\in l(A)\right), \tag{D.22}$$

where the element values are scaled with the atomic charge in first and second order with the empirical parameters $c_{U_1,A}$ and $c_{U_2,A}$, respectively. Third-order corrections to U_κ proved to be not beneficial. The "+U" contribution incorporates the following damping function which introduces a different behavior for long-range interactions

$$U_{AB}^{\text{damp}} = \frac{1}{2}\left(1+erf\left(-1.8\frac{R_{AB}-F_{AB}}{F_{AB}}\right)\right). \tag{D.23}$$

F_{AB} corresponds to an empirically derived analog for the sum of the covalent radii of A and B and is given by the following expression

$$F_{AB} = a_{r,A} + \left(\text{CN}_A - \text{avCN}\left[Z\left(A\right)\right]\right)a_{r,A}^{\text{CN}} + a_{r,B} + \left(\text{CN}_B - \text{avCN}\left[Z\left(B\right)\right]\right)a_{r,B}^{\text{CN}}. \tag{D.24}$$

in which "avCN" denotes a "universal" average coordination number determined from the fit set of molecules *a priori*. $a_{r,A}$ and $a_{r,A}^{\text{CN}}$ are an empirically determined atomic radius and a scaling factor for the dependency on the coordination number of atom A, respectively.

Finally, the orbital energies from the second diagonalization are shifted to match the corresponding DFT values according to

$$\epsilon = \epsilon_0\left(1+g_\epsilon\right)+g_{\epsilon_0} \tag{D.25}$$

where g_ϵ and g_{ϵ_0} are empirical parameters. Since the values in the PTB effective Hamiltonian matrix have only a loose relation to the corresponding ones in the converged DFT Fock matrix and are not fit target quantities, the resulting orbital energies show systematic errors that have to be adjusted for further property calculations. Note, that this has no effect on any of our target properties but is useful if the method is used, e.g., for simplified TD-DFT calculations.[411]

D.3.5 Response Approximation

In usual mean-field electronic structure methods that are solved self-consistently, the response of the wave function to an external electric field is determined via solving the coupled-perturbed SCF equations or by finite-field perturbation (static field-including SCF). Since this cannot be done exactly in PTB, an approximate response of the wave function has to be developed. Similar to the procedure for **P**, a two-step finite electric field procedure is used.

After adding the exactly calculated potential of a small finite external electric field to the effective Hamiltonian matrix of the finished PTB calculation, a new density matrix is obtained by diagonalization. Using the updated shell and atomic populations (and ignoring all other changes in the Hamiltonian) and the perturbation by the same external electric field, a second (new) effective Hamiltonian matrix is set up, which is constructed in a similar way as in the second iteration of the usual PTB procedure. The distinct changes to the Hamiltonian are explained in the following.

In the two-center part of the setup of $\mathbf{H^0}$, the parameters k_A^{res} and $k_{\text{pol}}^{\text{res}}$ are used instead of $k_A^{l_a}$ and k_{pol}, respectively, in the original formulation in Eq. D.7. Furthermore, the original overlap matrix **S** instead of the scaled overlap matrix $\mathbf{S}^{\mathbf{S}_{\text{XC}}}$ is used during the setup of $\mathbf{V}^{\text{XC}}$. This avoids the computation of an additional overlap matrix and reduces empiricism in the method. The definition of $\mathbf{H}^{\text{IES+TO+U}}$ contains a mixed field-perturbed (γ) and field-free electrostatic potential (γ^0), that is, the electrostatic potential resulting from the updated atomic and shell charges enters partly, controlled by the pre-factor $c_{V^{\text{ES}}}^{\text{res}}$:

$$\begin{aligned} H_{\kappa\lambda}^{\text{IES+TO+U}} = -S_{\kappa\lambda} \Bigg(& \frac{1}{2} \sum_C \sum_{l''} \Big[c_{V^{\text{ES}}}^{\text{res}} \left(\gamma_{AC,ll''} \right. \\ & \left. + \ \gamma_{BC,l'l''} \right) q_{C,l''} + \left(1 - c_{V^{\text{ES}}}^{\text{res}} \right) \left(\gamma^0_{AC,ll''} + \gamma^0_{BC,l'l''} \right) q^0_{C,l''} \Big] \\ & + \left((q_A)^2 k_{\text{TO},A} + (q_B)^2 k_{\text{TO},B} \right) \Bigg) + P_{\kappa\lambda} \left(U_\kappa^{\text{res}} + U_\lambda^{\text{res}} \right) U_{AB}^{\text{damp}} \\ & \text{with } \left(\kappa \in l(A),\ \lambda \in l'(B),\ l''(C) \right). \end{aligned} \tag{D.26}$$

Note that the superscripts in γ^0 and q^0_C do not denote an exponent. The TO correction employs the updated field-perturbed charges. The diagonal elements of the "+U" part are scaled by an additional parameter $k_{U_{\text{eff}},A}^{\text{res}}$ in the response calculation.

$$U_\kappa^{\text{res}} = k_{U_{\text{eff}},A}^{\text{res}} c_{U_{\text{eff}},A}^{l} \left(1 - \left(c_{U_1,A} \cdot q_A + c_{U_2,A} \cdot {q_A}^2 \right) \right) \quad \text{with} \quad (\kappa \in l(A)). \tag{D.27}$$

Further details of the calculation of the "DFT+U" part are the same as in Eq. D.22 and D.23. The damped Coulomb interaction is set up equally to the second iteration, however, a weighted average of the Ohno-Klopman- and Mataga-Nishimoto damping is applied (similar to the first iteration). Here, the prefactors $a_{\text{OK}} = 0.25$ and $a_{\text{MN}} = 0.75$ are used.

The effective Hamiltonian matrix set up in this way is diagonalized. From the resulting density matrix, the dipole moments are computed as usual and used to derive the static dipole polarizability tensor α_{IJ} by numerical differentiation with respect to the field, where IJ denotes the Cartesian axis. For details on the computation of IR intensities and Raman activities from numerical first derivatives of the dipole moment and the static dipole polarizability, see Sec. III-A of the SI.

D.3.6 Choice and Composition of the Fit Sets

The molecule training sets were derived from the ones used for GFN1-xTB, GFN2-xTB, and GFN-FF but significantly extended in the number of systems and their chemical diversity. Only closed-shell systems were considered in the fit but we note in passing that PTB in its spin-restricted form together with Fermi-smearing is applicable to any open-shell electronic configuration. The molecule size in the set is typically < 30 atoms with only a few containing around 50 atoms. For the most important elements H, C, N, O (HCNO) which were also used to determine the few global parameters, some larger molecules with about 100-150 atoms were included. The number of reference molecules is typically about 100-150 per element leading to about 10000-20000 data points (4300 for He) of the considered fit properties (see below). The corresponding numbers are considerably larger for HCNO (1200 systems, 145000 fit data points). In order to avoid parameter interdependencies as much as possible, the HCNO set is rather "pure" with only a small subset of molecules containing additionally the elements Li, Be, B, and F. For each element *X*, the corresponding fit set molecules contain *X* and mostly HCNO, and from the third row on, also the elements Li-B, F, and Na-Cl. This required at the end of the fitting process a meta-iteration strategy, i.e., the iterative parameter optimization loops successively over all elements.

The reference molecules are mostly neutral (70 %) but highly charged (up to +4/-4) systems are additionally included to put the important description of the electrostatic energy on a solid data basis. They typically refer to equilibrium structures obtained from DFT or GFN2-xTB calculations. However, many distorted (non-equilibrium) geometries are included as well. The actual geometry used is of less importance here because the DFT reference and PTB calculation employ exactly the same input structure, and basically we demand that PTB reproduces the DFT result for any geometry. A special feature of the new element fit sets compared to the ones used previously is that each includes about 5-15 so-called "mindless" molecules (MLMs) which were constructed by optimizing a random ensemble of atoms to the next available energy minimum[301]. The MLM generation procedure employs GFN*n*-xTB/GFN-FF as underlying methods and produces structures with very unusual bonding situations (see Fig. D.7 a) and b) in Sec. D.4.3) making the method robustly working also for difficult electronic structures.

D.3.7 Parameter Fitting

The fit procedure is roughly the same as established previously for GFN1-xTB, GFN2-xTB, and GFN-FF. No special ML techniques were applied. Instead, careful human supervision of the optimization process involving chemical intuition and detailed on-the-fly cross-checks for physical reliability was involved. The model parameters were determined by a minimization of the root mean square deviation (RMSD) between calculated and reference data using three different algorithms depending on the degree of convergence, difficulty of the element, and other details of the parameterization process. Mostly, and for all final parameter optimizations, the Levenberg-Marquardt (LM) algorithm[412] with line-search and numerical parameter gradients was utilized. In the initial stages of the fit, a one-parameter-at-a-time optimization strategy was used to avoid the appearance of non-physical parameter minima which sometimes occur with the LM algorithm in initial optimization stages. As an alternative optimizer not requiring function value derivatives, the Powell[355] algorithm was used which is very robust and also recommended.

The global and element-wise parameters were simultaneously optimized for the elements HCNO. The other parameters were optimized element-wise while keeping already existing parameters fixed. The fitting for the entire periodic table was continued with the halogens and elements Na-S, after which elements were treated in more or less canonical order (low *Z* first) while trying to minimize interdependencies. The *d*-block

Table D.1: Description of the properties for the parameter fit with an approximate weighting of the RMSD for the HCNO set.

property	RMSD weight in %
shell population	15
bond order	3
dipole and second moment	10
polarizability	5
polarizability gradient	5
dipole moment gradient	29
kinetic energy	20
orbital energies	7
$E_{\mathrm{RPBE}}[P_{\mathrm{PTB}}] - E_{\mathrm{RPBE}}[P_{\mathrm{B97X-3c}}]$	6

elements were considered after finalizing the main groups.

The properties from which the fit RMSD is derived and which define the targets of the method are given in Tab. D.1. The column denoted by "weight" gives the percentage contribution to the RMSD for the HCNO molecule set.

The choice of these properties and weights results from the goal of a globally well-balanced performance regarding a large number of possible areas of application. Besides simple properties like shell populations or bond orders which are directly related to **P**, also first-order properties like dipole moment or kinetic energy are considered. The advantage of using these quantities instead of fitting **P** directly is that appropriate weight is automatically placed on practically relevant properties. The inclusion of the derivatives in the fit is mandatory to obtain reasonable IR or Raman spectral intensities. For subsequent use of the PTB density in DFT energy expressions (not discussed here), it turned out to be beneficial to additionally include the total energy difference evaluated with a standard GGA (here the RPBE[316] functional) between the PTB density matrix and one from a full ωB97X-3c treatment ($E_{\mathrm{RPBE}}[P_{\mathrm{PTB}}] - E_{\mathrm{RPBE}}[P_{\mathrm{B97X-3c}}]$) as a target quantity. This quantity of course vanishes for a perfect fit, i.e., for $P_{\mathrm{PTB}} = P_{\mathrm{B97X-3c}}$.

D.3.8 Further Technical Details

All DFT reference calculations were conducted with TURBOMOLE 7.6[294] and applying the semi-numerical exchange approximation[359,360] `$senex` keyword) for the RSH functional. Many test calculations and geometry optimizations were performed on GFN*n*-xTB and GFN-FF levels with the `xtb` code.[259,413] The dipole and polarizability reference data were obtained analytically with TURBOMOLE while they refer to a numerical differentiation for PTB with an atomic displacement of 0.015 Bohr. The finite electric field perturbation step was 0.001 a.u.

D.4 Results and Discussion

In the following, we prove the capability of PTB to describe the electronic and molecular target properties robustly with an accuracy approaching state-of-the-art RSH DFT. The previously developed composite method ωB97X-3c evaluated in exactly the same basis set is considered as the reference throughout. Of particular interest is whether PTB is able to capture important electronic effects visible at the RSH DFT level and avoids typical pitfalls of lower-level methods (i.e., no artificial CT or charge over-delocalization).

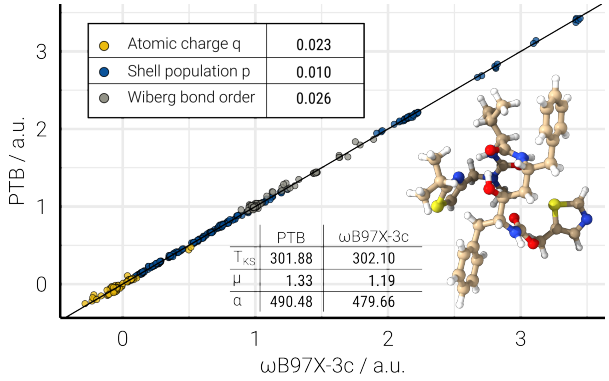

Figure D.3: Electronic and molecular properties for Ritonavir which was not part of the molecule training set. Values (atomic units) with PTB vs. ωB97X-3c are given for the KS kinetic energy T_{KS}, the molecular dipole moment μ, and the static isotropic dipole polarizability α. For the atomic charges q, the shell populations p, and the bond orders, also the RMSDs with respect to the reference are indicated in the legend.

Properties tested in this section are a) atomic charges and shell populations, b) bond orders (Sec. D.4.3), c) KS kinetic energies (Sec. D.4.6), d) orbital energy gaps (Sec. D.4.7), e) dipole moments (Sec. D.4.4), f) polarizabilities (Sec. D.4.5), and derivatives of the g) dipole moment (IR intensities) and h) polarizabilities (Raman activities) with respect to the position of the nuclei (Sec. D.4.8).

Furthermore, we compare the performance on the fit set over the whole periodic table, i.e., the quality of the results for the various elements (Sec. D.4.2) and also check the robustness on difficult and chemically unusual systems to ensure transferability and physically sound behavior (Fig. D.6 and D.7 in Sec. D.4.3).

D.4.1 Molecular Properties of Ritonavir

As a rather typical example for the performance of PTB for organic molecules, we show here results for various properties of ritonavir as a medium-sized drug compound (see Fig. D.3), which was not part of the training set. The correlation plot shows impressively the coherence of atomic charges, shell populations, and bond orders with the corresponding DFT values. Remarkably, no outlier is present, and virtually every data point is on par with the reference. The KS kinetic energy T_{KS}, the molecular dipole moment μ, and the static isotropic polarizability α are similarly well described with deviations of mostly $\ll$ 10 %, which is comparable to the variance between different common density functional results. Data for a broader range of also more difficult systems will be shown in the following subsections.

D.4.2 Variation of PTB Performance Over the Periodic Table

In the spirit of the GFN methods[259] that are parametrized for all elements up to $Z = 86$, we require also PTB to behave robustly over the whole periodic table. Of particular interest is the question if the method can also reproduce the more complicated electronic structure of, e.g., transition metal complexes or alkali oxides, which are challenging with semiempirical methods.[290,368]

We show in Fig. D.4 the final RMSD of the parameter fit for each element in color-coded form relative to the value for fluorine. Since the number of reference data points is similar for all elements except for HCNO, this is a reasonable comparison. The parameters for the elements HCNO were fitted in one common procedure on a noticeably larger data set and served as a starting point for the other elements. Therefore, the

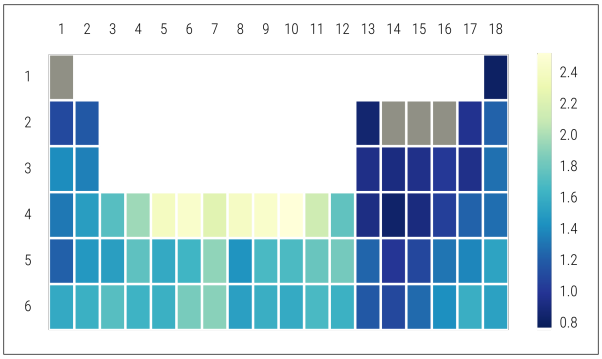

Figure D.4: Relative fit RMSD are given in color-coded form over the whole periodic table excluding lanthanides. The RMSD of fluorine is set to unity. For HCNO, the RMSD is much smaller (about 0.3) and not indicated since the parameters for these elements were obtained in one common initial fit for many more reference data and thus, a comparison on the given scale is not reasonable.

relative RMSD is lower (about 0.3) and not comparable to that of other elements.

Unsurprisingly, light non-metallic main group elements show the lowest RMSD, with helium having the smallest value. Similar low RMSDs were observed for boron and germanium. In groups 13-18, the RMSD increases from the top left (boron, silicon, aluminum) to the bottom right corner (astatine, xenon, radon). On the one hand, the more diffuse and polar electronic nature of heavier atoms in molecules makes it generally more challenging for PTB to reproduce the DFT density matrix. On the other hand, the approximation for ECPs employed in PTB (see Sec. D.3.3) may introduce a larger error than for lighter atoms. In addition, versatile molecular systems as reference data points are more difficult to generate for (heavy) noble gas elements, and hence, these sets contain many difficult cations and/or hyper-valent species. Still, it has to be noted that the RMSD increases only by a factor of about 1.5 in comparison to second and third-row main group elements. Coming from main group elements, a sharp cut is visible when going to groups 1-12. As expected, 3d elements with a partly filled *d*-shell exhibit the largest RMSDs, being 2-3 times larger than that of fluorine, which was set to 1. Alkali and alkaline earth metals rank between transition metals and main group elements. Interestingly, the RMSDs of lithium and beryllium are close to those of main group elements of comparable weight, indicating their partially non-metallic nature. Their stronger tendency to build defined covalent bonds compared to heavier group 1 or 2 elements is probably beneficial for their description in a TB theory.

Since the comprehensive visualization of results for all elements is not convenient, only data for the elements HCNO, S, Pb, and Ti are shown explicitly in the following sections. Together with Fig. D.4, they are representative of the entire periodic table.

D.4.3 Atomic Charges, Shell Populations, and Bond Orders

Since the target of PTB is to provide a single-particle density matrix that approximates the DFT reference as close as possible, the natural choice of property to be checked is the atomic charge q. In Fig. D.5, this is done together with shell populations p and bond orders, as all of these properties are derived directly from the density matrix. For HCNO about 110,000 and for S, Pb, and Ti, about 10,000-20,000 data points from the respective fit sets are evaluated.

Remarkably, despite the sheer size of the data sets, only very few data points deviate from the 1:1 correlation

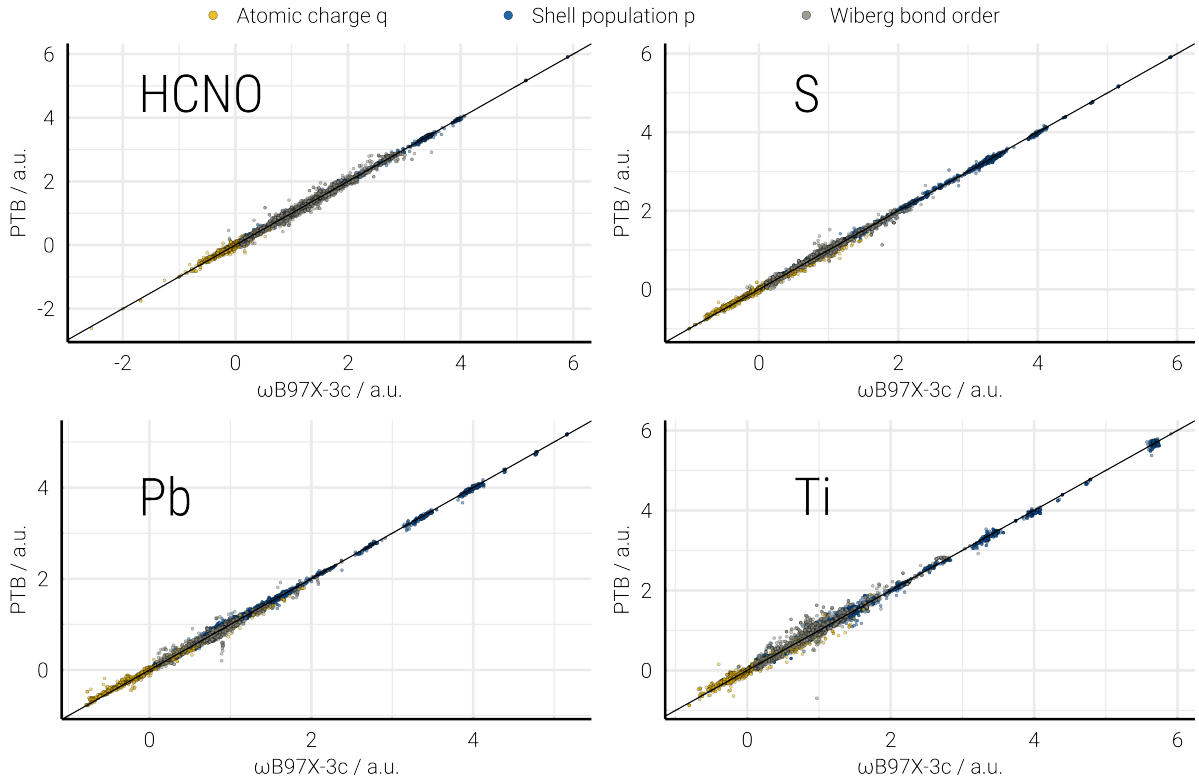

Figure D.5: Atomic charges q, shell charges p, and Wiberg bond orders in atomic units calculated with PTB compared to ωB97X-3c for the examples HCNO, S, Pb, and Ti (see Sec. D.4.2). The data points were taken from all molecules in the full training set for the respective element.

line for HCNO. This also holds true for strongly negative atomic charges close to −2 or very high shell populations. Furthermore, the distributions in Fig. D.5 are consistent with the tendency for larger deviations in the order HCNO < S < Pb < Ti that can be abstracted from Fig. D.4: While PTB yields results for S that are very similar to HCNO, Pb, and in particular the deviations for Ti are more significant. However, it should be noted that, except for rare cases with difficult electronic structures, qualitatively and in most cases even quantitatively correct results are obtained even for Ti as one of the most challenging atom types.

The PTB method attempts to capture CT effects accurately as in RSH DFT and to overcome typical limitations of GGA- or conventional TB-based methods. To assess the amount of artificial CT in PTB, we investigated in Fig. D.6 the atomic charge of a chlorine atom for different distances to an imidazole unit in a zwitterion pair. At very short distances < 2 Bohr, where some covalent bond character is present, the charge should be partially delocalized over the complex, whereas at larger distances > 5 Bohr, it should completely localize at chlorine. This is described correctly by ωB97X-3c. With GGA methods like RPBE (here evaluated in the same basis set), the charge localization in the dissociation limit is, as expected, not fulfilled due to SIE. For larger distances, the corresponding SCF calculation does not even converge. In contrast, with GFN2-xTB the charge is much too localized at intermediate and short distances, which we attribute to the minimal basis set employed. PTB describes the situation correctly along the reaction coordinate and similarly to ωB97X-3c as it is able to capture both the charge delocalization at short distances and the charge localization in the dissociation limit.

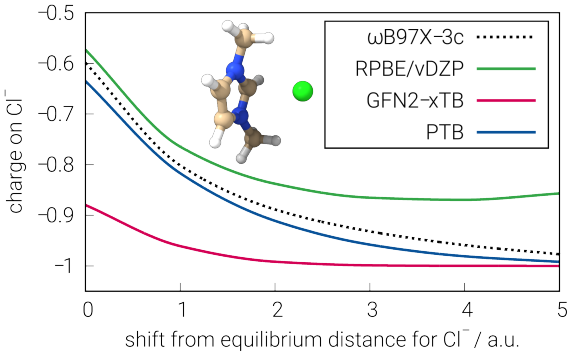

Figure D.6: The calculated atomic charge on the chlorine atom as a function of the distance to the imidazole ring. Besides PTB and ωB97X-3c, also results with RPBE/vDZP and GFN2-xTB are presented.

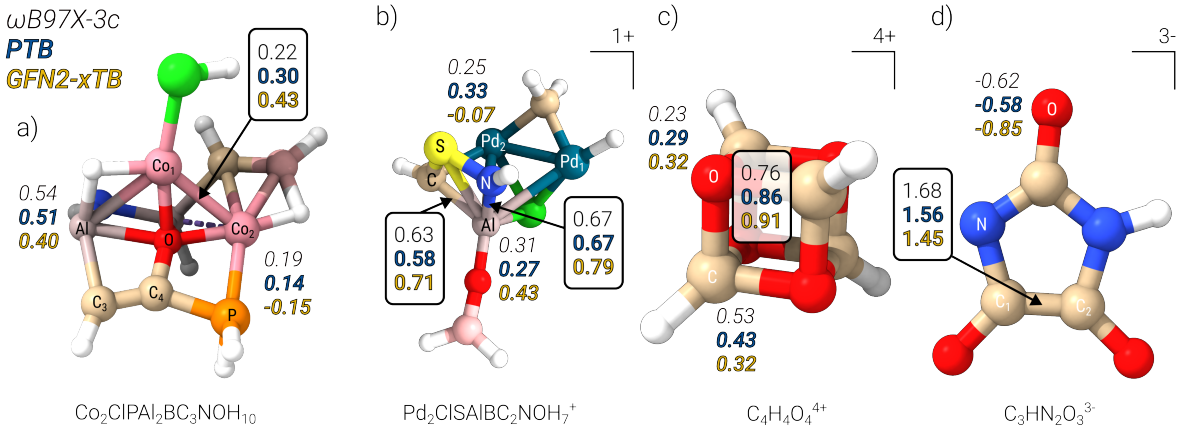

Figure D.7: Atomic charges q (italic letters) and Wiberg bond orders (black boxes) for selected atoms and covalent bonds, respectively, calculated with ωB97X-3c, PTB, and GFN2-xTB. Four molecules (not part of the molecule training set) are investigated, two of which are so-called "mindless" molecules, while the other two are highly charged (anionic and cationic) organic molecules.

Highly parameterized methods belonging to the class of data science- or ML-driven methods, may suffer from over-fitting or missing transferability to structures far away from the learned chemical space.[414] We tried to overcome this issue by a) employing only a limited number of empirical parameters, which always represent physically motivated scaling factors in the Hamiltonian, and b) requiring PTB to give reasonable results even for very unusual or "extreme" structures. The latter is exemplified here for four systems that were not part of the fit set and represent extremely challenging test cases (see Fig. D.7). Two of them are so-called MLMs, constructed semi-automatically (randomly, see Ref. [301]) and which are far from the usual chemical space of known molecules. The only requirement for their use (also in the training set of molecules) is that they have a finite HOMO-LUMO gap and proper SCF convergence with GFN2-xTB. The other two systems are highly charged ionic structures. In Fig. D.7, atomic charges (italic letters) and bond orders (boxes) by PTB are displayed, along with the GFN2-xTB values for comparison and the corresponding ωB97X-3c result for reference.

In system a), the atomic charges for the Al atom and for the Co_2 atom as well as the bond order for the Co_1-Co_2 interaction are given. While GFN2-xTB nevertheless delivers a qualitatively correct result for the atomic charge of the Al atom, it predicts the incorrect charge sign for the Co_2 atom. In both cases, PTB provides results with $\ll 0.01\ e^-$ variances being very close to the reference. The deviations for the bond orders are somewhat larger but the values are still significantly closer to the reference than with GFN2-xTB. For the other MLM b), the observations are similar: PTB yields atomic charges and bond orders very close to the reference with a maximum deviation of $0.08\ e^-$, while GFN2-xTB shows larger variations but is still qualitatively correct in most cases.

For c) ($C_4H_4O_4{}^{4+}$), the deviations with PTB of up to 0.1 a.u. are slightly larger than in the previous examples but still smaller than with GFN2-xTB. In particular, the atomic charge of the carbon atoms is noticeably better reproduced by PTB than with GFN2-xTB. The description of the $C_3HN_2O_3{}^{3-}$ anion in d) profits from the larger vDZP basis set in PTB. Therefore, charges and bond orders are, less surprisingly, much more accurate than with GFN2-xTB. Interestingly and in agreement with the observations for the imidazole-chlorine complex, GFN2-xTB overestimates charge separation as the displayed charge for one of the oxygen atoms is much more negative than with the reference.

In addition to the zwitterion in Fig. D.6 and the very challenging and unusual systems in Fig. D.7, results for a large test set of about 1,000 randomly chosen molecules from the PubChem database,[415] which were also not part of the fit set, are shown in Fig. S3 in the SI. The RMSDs of the discussed properties for this set are below those of the fit set, indicating high transferability. For a more detailed discussion, see Sec. I-C of the SI. Moreover, water hexamer clusters have been investigated in Sec. I-A of the SI.

The presented findings indicate that PTB is not restricted to a specific chemical space but can represent also extremely complex systems that are far beyond typical application domains. Although electronic properties were not the main target during the development of GFN2-xTB and it was not fitted to them, in passing it is noted that the overall results by GFN2-xTB are nevertheless still acceptable.

D.4.4 Dipole Moments

In analogy to the results for charges, populations, and bond orders, the accuracy of dipole moments calculated with PTB is depicted in correlation plots for the elements HCNO, S, Pb, and Ti in Fig. D.8. For HCNO, about 1,700 data points (consisting of x, y, and z components of the dipole moments) from the fit set are displayed, while several hundred data points are contained for the other elements. The conclusion from an analysis of the results follows the key statements of section D.4.3 – almost no outliers are visible and the overall accuracy is very high. Interestingly, the moments for Pb are similarly accurate as those of HCNO and

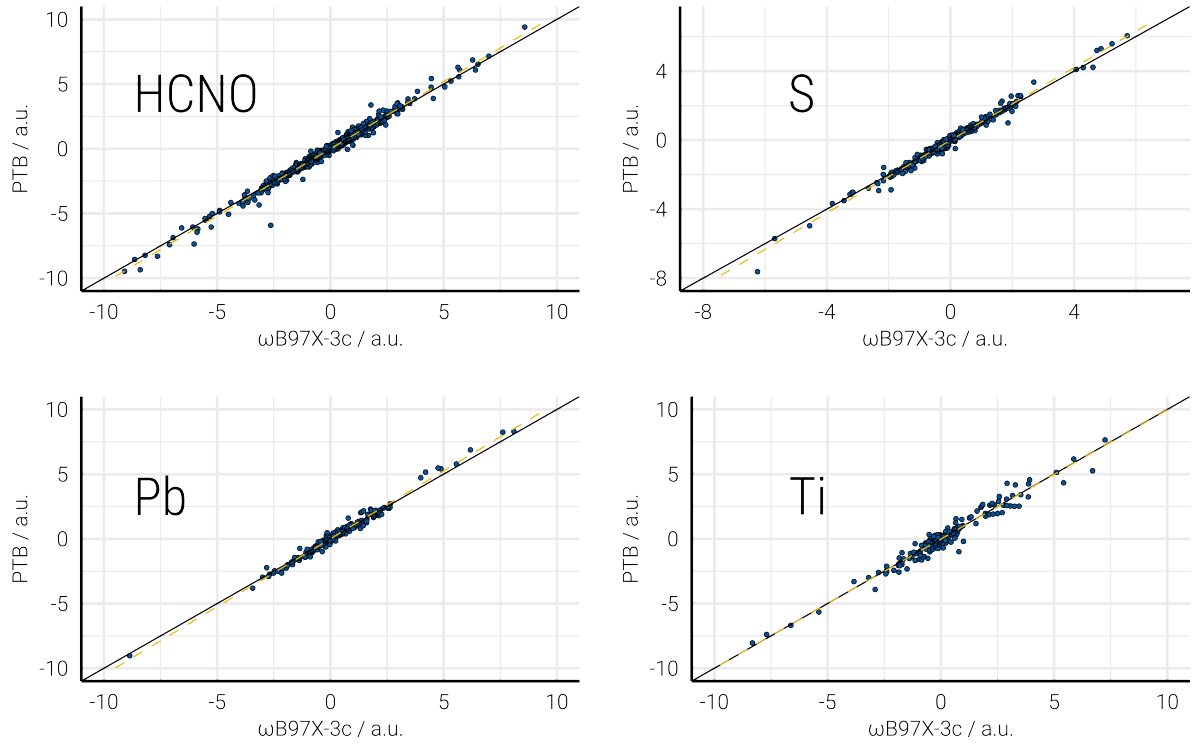

Figure D.8: Dipole moment (x, y, and z component for each molecule) in atomic units calculated with PTB compared to ωB97X-3c. The dashed line corresponds to a linear fit of the data points.

S, although the global fit RMSD was found to be higher. As expected, Ti shows the largest deviations. In general, the discrepancies to the ωB97X-3c reference are larger for dipole moments than for zeroth-order properties like charges and populations, which is plausible since higher-order electric properties probe outer, diffuse parts of the wave function that are particularly difficult to describe for approximated methods not evaluated self-consistently.

Similar to the previous subsection, also for dipole moments, the transferability of the results has been checked for a test set of about 1,000 random molecules from the PubChem database. Even though a very small number of outliers (three, deviations of about 5 a.u.) were detected, the performance was still similar to that for the fit sets. For details, see Sec. I-C of the SI.

D.4.5 Polarizabilities

While minimal AO basis sets (as in GFN2-xTB) enable the robust (though not very accurate) computation of zeroth- or first-order properties (see Sec. D.4.3), they prevent reliable computation of higher-order properties such as electric dipole polarizabilities, for which extended basis sets are required. Opposed to existing SQM methods, PTB employs a significantly larger vDZP basis, which overcomes this limitation. Therefore, we investigated in Fig. D.9 the accuracy of static isotropic polarizabilities α with PTB for 369 small organic molecules (see the SI for details on the test set). Over the entire range of values, the MAE is consistently only about 1.5 a.u. corresponding to small relative deviations of typically 1-5 %. The good performance for polarizabilities is particularly noteworthy because of the approximated wavefunction response in PTB.

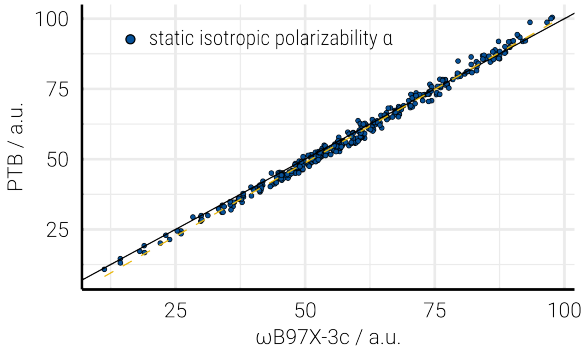

Figure D.9: Static isotropic polarizabilities α calculated with PTB compared to ωB97X-3c for 369 small to medium-sized (< 50 atoms) organic molecules, which were not explicitly part of the fit set. The dashed line corresponds to a linear fit of the data points.

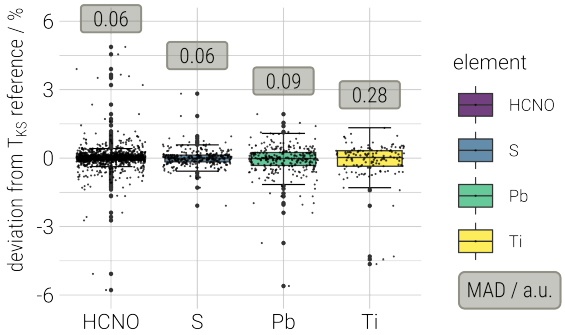

Figure D.10: Relative deviations of the PTB KS kinetic energy T_{KS} compared to ωB97X-3c in %. Results for fit set molecules of the elements HCNO, S, Pb, and Ti are given. Furthermore, the mean absolute deviations (MADs) are indicated in the grey boxes. The relative deviations are represented by standard box plots, in which the box size corresponds to the median 50 % of the data and the distance to the whiskers to 1.5 times the distance between the upper and lower quartiles. Data points outside of the whiskers are labeled as outliers in form of thicker circles. All data points are indicated by small scatter points.

Apparently, the additional two diagonalizations and the use of an adapted Hamiltonian matrix in the response procedure (see Sec. D.3.5) are sufficient for the computation of accurate isotropic polarizabilities.

A further check for the accuracy of isotropic polarizabilities can be found in Sec. II of the SI for conjugated alkenes. For all investigated chain lengths, PTB and reference values agree very well indicating the absence of overpolarization effects in the PTB method. Comparison of ωB97X-3c data to corresponding values for ωB97X-D4 in a large augmented quadruple-ζ basis set are also shown, revealing basis set incompleteness effects for this rather basis set-dependent property on the order of 10 %.

D.4.6 Kohn-Sham Kinetic Energies

Even though not of obvious importance in practical applications, the KS kinetic energy T_{KS} as calculated from the exact kinetic energy integrals and the density matrix, is a valuable descriptor for the quality of the method. Since T_{KS} (and its error) is proportional to the system size, we decided to show relative deviations

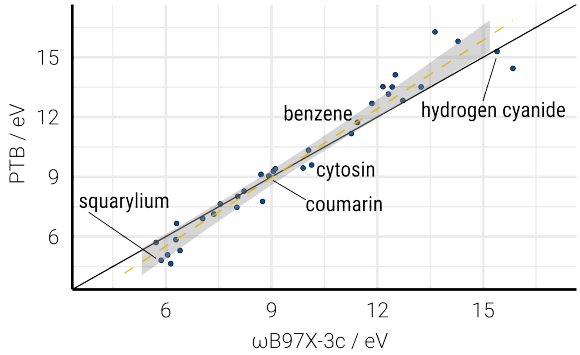

Figure D.11: Orbital HOMO-LUMO gaps of typical small to medium-sized organic dyes with PTB compared to ωB97X-3c in eV. For selected molecules, their (trivial) name is displayed in the plot. The dashed line corresponds to a linear fit of the data points, while the grey-shaded area corresponds to the confidence interval at a level of 0.99.

in Fig. D.10, again for structures in the fit set of HCNO, S, Pb, and Ti. In addition, the absolute MAD with respect to ωB97X-3c data is given.

In general, Fig. D.10 underlines the observations for the previously discussed electronic properties. The MAD increases when moving from HCNO and S to Pb and Ti. However, the relative mean absolute deviation (relMAD) is similarly small for all four elements (0.33 % [HCNO] < relMAD < 0.49 % [Ti]). Thus, the slightly larger absolute errors for Ti are still acceptable and consistent with the overall somewhat more challenging description of the electronic nature of transition metals. Note that the much larger sample size for HCNO (about 1,200) than for the other elements (100-300) and the inclusion also of very small systems (with larger relative errors) lead to a larger number of so-called outliers in the boxplot.

Nevertheless, relMADs consistently below 1 % demonstrate clearly that PTB is capable of describing the electronic structures of the molecules in the fit set accurately, also beyond the properties discussed so far. Another sanity check for the robustness of KS kinetic energies computed with PTB is shown in Fig. S2 in the SI, where T_{KS} as a function of off all bond lengths of the molecule (i.e., the scaled cartesian coordinates) is investigated with respect to the ωB97X-3c reference.

D.4.7 Orbital Energy Gaps

Up to now, we have mostly looked into properties defined by the occupied orbitals in the density matrix **P**. However, PTB aims also at providing robust orbitals and orbital energies in the virtual space for possible application in approximate excited state methods such as sTDA-xTB.[324] To this end, we compared orbital HOMO-LUMO gaps of small to medium-sized organic dyes with PTB to the corresponding ωB97X-3c reference values in Fig. D.11.

The linear fit of the data points (yellow dashed line) reveals a slope that is too steep, i.e., smaller gap values are slightly underestimated and correspondingly, larger gaps are overestimated. This observation may be connected to the non-linear shift of the PTB orbital energies to match DFT orbital energies (see Eq. D.25). Besides that, PTB provides accurate orbital energy gaps with a relMAD of about 7 % and an MAD of only 0.6 eV.

The correct representation of orbital energies and corresponding gaps indicates a physically sound behavior also in the virtual space. The applicability of PTB for computation of, e.g., electronic spectra in the framework of sTDA[411] is part of current research in our lab.

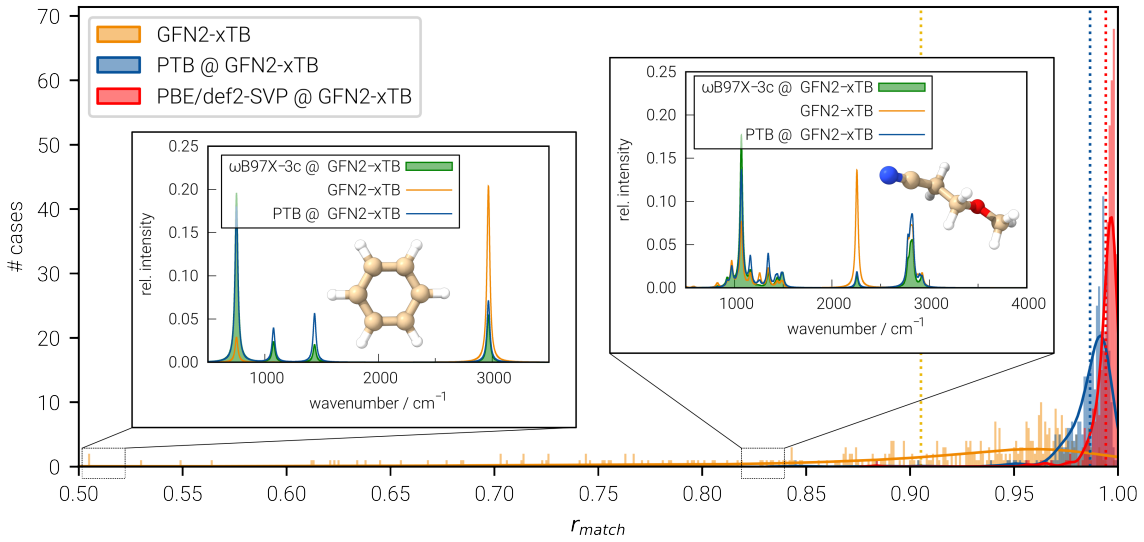

Figure D.12: Histogram with spectral match scores r_{match} for IR spectra of 369 small organic molecules (see the SI for details on the employed similarity measure and on the data set). $r_{\text{match}} = 1$ corresponds to a perfect coherence, whereas $r_{\text{match}} = 0$ indicates no coherence between the spectra. Vibrational frequencies and normal modes, respectively, were taken from GFN2-xTB throughout, while the IR intensities were calculated with ωB97X-3c (reference), GFN2-xTB, PTB, and PBE[81]/def2-SVP[139], denoted by "[method]@GFN2-xTB". The IR spectra of ωB97X-3c@GFN2-xTB type were considered as the reference. Exemplary IR spectra are given for benzene (left, $r_{\text{match}}^{\text{full GFN2-xTB}} = 0.505$) and 3-methoxy-propanenitrile (right, $r_{\text{match}}^{\text{full GFN2-xTB}} = 0.829$). PBE/def2-SVP is omitted in the exemplary spectra since the spectra would look virtually the same as with ωB97X-3c. The vertical colored dashed lines display the respective mean r_{match} score.

D.4.8 Infrared and Raman Spectra

As shown in earlier work, IR spectra computed with (semi)empirical methods such as GFN2-xTB (employing the double harmonic approximation, see Ref. [416]) suffer mainly from errors in the intensities[314]. Since PTB was shown to yield excellent electric dipole moments as well as static dipole polarizabilities (see Sec. D.4.4 and D.4.5), it is obvious to also investigate their derivatives with respect to vibrational normal modes to obtain IR intensities and Raman activities. Given that there is currently no energy expression and hence no Hessian available for PTB, we stick with the GFN2-xTB normal modes, which is also a reasonable approach in practice.

In the following, we combine dipole moment- and dipole polarizability gradients with respect to nuclear coordinates ($\frac{\partial \mu}{\partial R_{xyz}}$ and $\frac{\partial \alpha}{\partial R_{xyz}}$) obtained from PTB with GFN2-xTB vibrational normal modes. The same is done with corresponding gradients calculated with PBE[81]/def2-SVP[139] and ωB97X-3c to obtain spectra based on DFT intensities for comparison and reference, respectively. In this way, we are only evaluating IR intensities and Raman activities, but not the influence of the vibrational frequencies or normal modes. The use of GFN2-xTB vibrational normal modes together with *external* intensities or activities is indicated via the notation "[method]@GFN2-xTB". Consequently, a simple "GFN2-xTB" notation indicates the use of GFN2-xTB frequencies *and* intensities. The vibrational spectra derived in this way are quantitatively compared by introducing a finite linewidth and applying a match score r_{match} that is a similarity measure

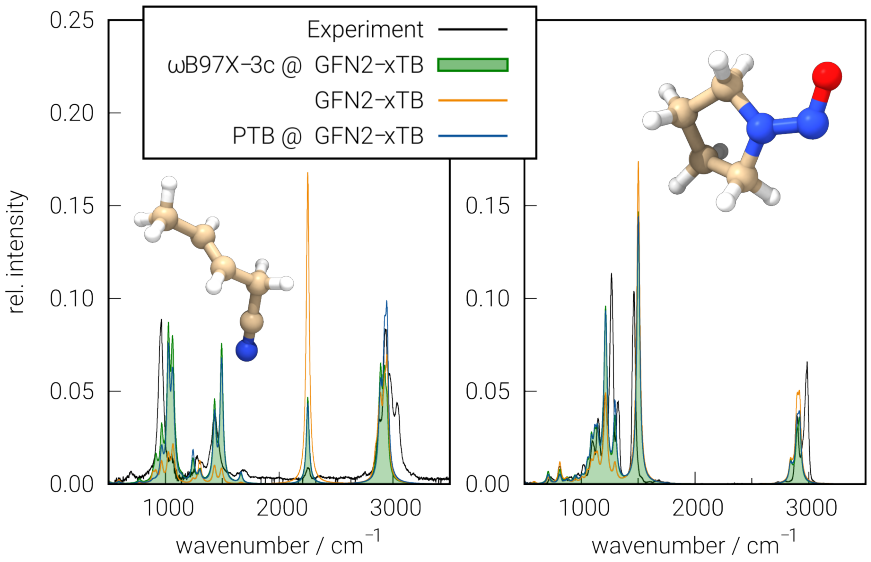

Figure D.13: IR spectra of 3-pentenenitrile and 1-nitrosopyrrolidine. The experimental gas phase spectra and spectra calculated with ωB97X-3c@GFN2-xTB, full GFN2-xTB, and PTB@GFN2-xTB are shown.

for two spectra (see also Ref. [314]), where ωB97X-3c@GFN2-xTB spectra are considered as the reference. Further details on the technical implementation, the generation of spectra with a finite linewidth, and the employed similarity measure are provided in the SI.

For Fig. D.12, IR spectra of 369 organic molecules (see the SI for details on this test set) were investigated in this way. The results are visualized in form of a histogram, in which the r_{match} scores for the whole test set are plotted for all three methods. ωB97X-3c@GFN2-xTB spectra consequently correspond to $r_{\text{match}} = 1$. Moreover, two exemplary spectra are presented.

The histogram shows a fairly broad distribution of r_{match} values when employing GFN2-xTB intensities, which is in agreement with the mentioned observations in Ref. [314]. The example spectrum on the left, which led to an r_{match} score of about 0.5 for GFN2-xTB, illustrates this nicely: The relative intensities of a low-lying mode below 1000 cm^{-1} and the C-H stretching vibration at $\approx$ 3000 cm^{-1} are far away from the reference. PTB in contrast provides relative intensities that are very close to the DFT intensities and this is furthermore indicated by the statistical distribution in the histogram. The lowest histogram entry is at about $r_{\text{match}} = 0.84$, and a large majority of data points is $\gg 0.95$. For comparison, we included also the respective distribution for PBE/def2-SVP@GFN2-xTB, which is only marginally better than that of PTB@GFN2-xTB. Thus, PTB yields accurate IR intensities with deviations from the ωB97X-3c reference in the same order as the difference between different DFT functionals. The second example spectrum with an r_{match} value for GFN2-xTB close to its average score also underlines this observation. The relative intensities of many vibrational modes are well described by GFN2-xTB, but the deviation for two of the modes is still significant. PTB, on the other hand, achieves overall quite accurate intensities also for this example.

In Fig. D.13, we also compare spectra calculated with the combined methods as explained above with experimental gas-phase spectra to highlight the importance of the improved intensities by PTB compared to GFN2-xTB for practical applications. The spectrum of 3-pentenenitrile on the left shows clearly that the relative IR intensities with GFN2-xTB are qualitatively wrong for lower-lying scaffold vibrational modes around 1000 to 1500 cm^{-1} and for the strongly overestimated intensity of the C≡N stretching mode at about

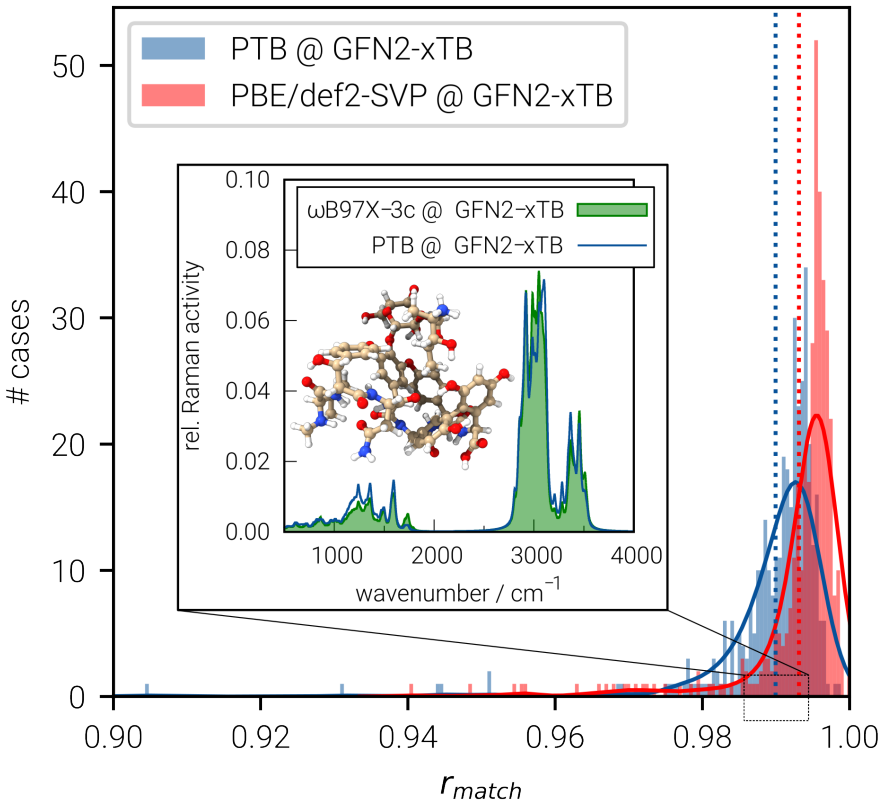

Figure D.14: Histogram with match scores r_{match} for Raman spectra (Raman activity per normal mode) of 369 small organic molecules. See Fig. D.12 and the SI for details on the method notation, the employed similarity measure, and the data set. The inlay shows the Raman spectrum of vancomycin (176 atoms, $r_{\text{match}}^{\text{PTB@GFN2-xTB}} = 0.988$), which was added to the set of 369 small organic molecules used for the histogram. Values below 0.9 were not present in the data set. The vertical-colored dashed lines display the respective mean r_{match} score.

2300 cm^{-1}. The intensities provided by ωB97X-3c are much closer to the experimental spectrum. This also applies to PTB, which in most cases yields relative intensities that are virtually indistinguishable from the DFT reference. A similar analysis results for the spectrum of 1-nitrosopyrrolidine, for which the intensities of many lower-lying modes are distinctly too low with GFN2-xTB as well, while PTB provides intensities very close to the DFT reference. The spectra show that the remaining deviations in the intensities compared to experimental spectra are due to a) errors in the vibrational normal modes computed with GFN2-xTB that also affect the intensities, b) errors resulting from the double harmonic approximation, or c) artifacts in the experimental data, which cannot be accounted for with the applied theoretical approach (for a more in-depth discussion on this, see Ref. [314]). Importantly, the insufficient accuracy for dipole moment gradients as in GFN2-xTB is overcome with PTB.

In contrast to IR intensities that can be routinely calculated also with SQM methods (or even FF methods like GFN-FF[189,314]), the calculation of Raman activities has so far been almost exclusively the domain of DFT or WFT methods. As already hinted in Sec. D.4.5, PTB overcomes this limitation by utilizing for the first time a significantly larger basis set, which enables the reliable computation of dipole polarizabilities. The accuracy of their derivative with respect to nuclear coordinates is investigated in the following. Analogously

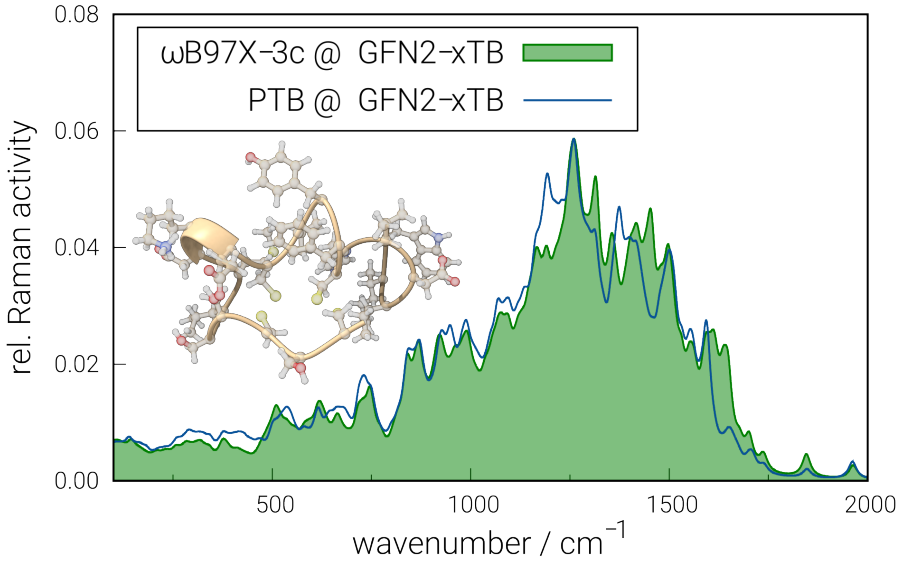

Figure D.15: Raman activities of the polypeptide endothelin (327 atoms) computed by ωB97X-3c@GFN2-xTB and PTB@GFN2-xTB. The high relative intensity of bands at around 3000 cm^{-1} (regardless of the method) would prevent a reasonable comparison, which is why only the range from 100 to 2000 cm^{-1}, being more relevant for the protein structure, is shown. See Fig. D.12 and the SI for details on the method notation.

to the evaluation of IR intensities, we compare Raman activities from PTB with ωB97X-3c reference data and for better perspective, additionally with PBE/def2-SVP. As for IR intensities, GFN2-xTB frequencies are used throughout. The notation, the utilized test set (except for vancomycin, see below), and the employed spectral similarity measure are also identical to the analysis of the IR intensities.

The histogram in Fig. D.14 shows that the results for Raman activities by PTB are even slightly better than for IR intensities. No spectrum with $r_{\text{match}} < 0.9$ is found in the test set (see the SI for tabulated data in text file format) and the average match score with PTB exceeds that for IR intensities. Similar to the previous analysis for IR spectra, PTB is very close to PBE/def2-SVP and the deviations are only marginal. Besides the small molecules in the aforementioned test set, we also included vancomycin as a typical drug molecule consisting of 176 atoms, shown in the inlay in Fig. D.14. As displayed in the figure, its spectral match score corresponds roughly to the average value (dashed line) very close to unity indicating only tiny differences between the two spectra.

To further illustrate the capabilities of PTB, we computed Raman activities for the polypeptide endothelin[417]. The most relevant part of the spectrum with respect to the protein structure ranging from 100-2000 cm^{-1}[418] is shown in Fig. D.15. Except for small details in the fine structure of the spectrum at around 1200 to 1400 cm^{-1}, the Raman activities of most vibrational modes are reproduced correctly by PTB, even though their absolute values (in $[\frac{\text{Å}^4}{\text{u}}]$) are distinctly smaller compared to the previous examples. This PTB@GFN2-xTB calculation (employing a preliminary PTB implementation) of endothelin consisting of 327 atoms took about two hours, whereas the reference DFT treatment lasted about ten days with equivalent computational resources.

D.5 Conclusions and Outlook

We presented in this work new ideas and strategies to develop semiempirical electronic structure theory further and finally presented a usable method termed PTB, which will be made available free-of-charge in the `xtb` quantum chemical program[413]. A preliminary implementation of PTB is available upon request.

It is out-of-the-box applicable to various electronic structure properties for molecules composed of all *spd*-block elements up to $Z = 86$. The main point is to substitute the so far commonly used minimal atomic orbital basis sets with an extended and properly polarized basis set taken from previous DFT work (vDZP). This choice together with one of the most accurate range-separated hybrid density functionals (ωB97X-V) as reference forms the basis for the development of the new PTB method. Its main purpose is to provide extremely fast the one-particle electronic density matrix **P** and derived properties like atomic charges, shell populations, dipole moments, and geometric derivatives as well as static dipole polarizabilities. The effective Hamiltonian matrix elements, which approximate those from a ωB97X-V SCF converged Fock matrix, are derived in analogy to DFTB3/GFN2-xTB theory but contain additionally new non-local terms to account for the RSH character of the reference method and the presence of spatially different atomic shells of the same angular momentum.

For a wide range of molecules from various chemical compound classes (organic, inorganic, transition metal complexes, clusters, and aggregates) including multiple anions/cations, excellent agreement between PTB and ωB97X-3c computed properties is obtained. Typical RMSDs are 0.02 e^- for atomic charges or shell populations and about 5-10% for dipole moments and dipole polarizabilities. Artificial charge transfer in, e.g., zwitterions or SCF convergence problems in electronically difficult small gap systems is largely avoided, which also results from the applied non-iterative, two-step diagonalization scheme.

As example applications, we show the efficient computation of IR absorption intensities in combination with GFN2-xTB computed normal modes and for the first time, semi-empirical results for the Raman spectrum of an entire protein. This is made possible by the achieved speed-up of typically about three orders of magnitude compared to the corresponding reference DFT calculation (see Sec. V in the SI for details). Compared to GFN2-xTB, PTB requires only 3-7 times more computational effort. For systems converging slowly in GFN2-xTB, this ratio is even more advantageous.

No energy expression is available at this point and the entire development was focused to reproduce the DFT density matrix and some response features of it as accurately as possible. The electronic energy as a functional of **P** and corresponding derivatives (nuclear forces) shall be developed in future works. To this end, strategies either based on DFT and (DF)TB theory or even ML (or combinations thereof) seem to be promising and are currently being pursued in our laboratory.

D.6 Data Availability

The data that supports the findings of this study are available within the article and its supplementary material. Any further information as well as a preliminary implementation of PTB is available upon request from the authors. *Additional note in the reprint: PTB is fully implemented in the `xtb` program package.*[413]

D.7 Supplementary Material

See the supplementary material for the following data, which are available free of charge.

- Additional investigations of electronic and molecular properties with PTB, including a cross-check on a set of about 1,000 random molecules
- Additional theory on IR intensities and Raman activities, and the quantitative comparison of vibrational spectra
- Investigation of the computational timings in comparison to the reference DFT calculation for a large supramolecular system
- Table of all empirical parameters utilized in PTB, including a text file with each parameter given for all parameterized elements
- Statistical error measures
- Raw data required for reproducing each plot and statistical analysis in the main text
- Test set for static isotropic polarizabilities and for IR and Raman spectra with 369 small organic molecules and corresponding match scores

D.8 Acknowledgements

The authors thank all members of the Grimme work group for helpful comments and discussions during the long PTB development period. M.M. acknowledges the Fonds der Chemischen Industrie (FCI) for funding under a Kekulé scholarship and thanks M. Stahn for technical support. This work was supported by the Deutsche Forschungsgemeinschaft in the framework of the priority program SPP 2363 ”Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning”, project number 497190956.

D.9 Conflict of Interest

The authors have no conflicts to disclose.

APPENDIX E

Appendix: An Atom-in-Molecule Adaptive Polarized Valence Single-ζ Atomic Orbital Basis for Electronic Structure Calculations

Marcel Müller,[†] Andreas Hansen,[†] Stefan Grimme[†]

Received: August 15, 2023
Published online: October 25, 2023

Reprinted in Appendix E with permission[§] from
M. Müller, A. Hansen, and S. Grimme, *An atom-in-molecule adaptive polarized valence single-ζ atomic orbital basis for electronic structure calculations*, The Journal of Chemical Physics **159** *16* (2023) 164108,
DOI: 10.1063/5.0172373

Own contributions

- Conceptualization of the minimal basis set approach
- Development of an `ORCA` input generator for q-vSZP
- Initial implementation of the CEH charge model into `tblite`
- Performing all calculations and generation of the results
- Interpretation of the results
- Writing of the manuscript

[†] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany
[§] Permission requests to reuse material from this chapter should be directed to AIP Publishing.

E.1 Abstract

Many low-cost or semiempirical quantum mechanical-based electronic structure methods suffer from the use of unpolarized minimal atomic orbital (AO) basis sets. In this work, we overcome this limitation by a fully DFT variationally optimized, adaptive minimal basis set consistently available for the elements up to radon (Z = 86). The new key feature is to make the linear coefficients of the primitive Gaussians in a contracted AO dependent on the effective atomic charge of the atom in the molecule, i.e., each symmetry-unique atom obtains its "own" specifically adapted basis functions. In this way, the physically important "breathing" of the AOs in a molecule with a) atomic charge (expansion/contraction for anionic/cationic states) and b) the number of close-lying bonded neighbor atoms is accounted for. The required atomic charges are obtained from a specially developed extended Hückel type Hamiltonian and the coordination numbers from the molecule geometry. Proper analytical derivatives of the resulting adaptive basis functions can easily be derived. Moreover, the basis functions are electric field-dependent, thus improving the description of, e.g., dipole moments and polarizabilities. The new basis set termed q-vSZP (charge dependent valence single-ζ, polarized) is thoroughly benchmarked for atomic/molecular and thermochemical properties compared to standard minimal and double-ζ basis sets at the DFT level with the accurate ωB97X-D4 functional. It is shown that q-vSZP is clearly superior to existing minimal basis sets, often reaching double-ζ quality or even better results. We expect it to be the optimal choice in future semiempirical quantum mechanical methods.

E.2 Introduction

Quantum chemistry has revolutionized our understanding of molecular properties and chemical reactivity by providing a powerful framework for predicting and interpreting chemical phenomena at the electronic level. At the heart of electronic structure theory are standardized AO basis sets. These sets, comprised of atom-centered Gaussian or Slater functions, serve as fundamental tools to approximate the electronic wavefunction of molecules accurately. Comprehensive surveys of state-of-the-art basis set families used in DFT or WFT based calculations are given in Refs. [132, 419].

The choice of an appropriate AO basis set plays a pivotal role in the accuracy and reliability of quantum chemical calculations and is usually based on a compromise between target accuracy and accessible computational resources. While this is intensively discussed in WFT/DFT,[70,420] basis set issues are less noticed or even ignored in SQM methods. Standard approaches like PM6,[238] DFTB,[267,268] or xTB[259,277,278] mostly employ MB sets without polarization functions, where the exponent of the (usually applied) Slater functions is considered as an empirical rather than a variational parameter.

Recently, we presented the PTB semiempirical model,[2] which for the first time employs a consistent vDZP AO basis set[1] for all chemical elements up to Z = 86. A vital feature of the variational optimization procedure for the primitive Gaussian basis function exponents and contraction coefficients in vDZP was to simultaneously minimize the sum of the total energy of various atomic states, molecules, and charged species [reference molecules (RMs)]. This, together with a relatively deep contraction (a large number of Gaussian primitives), leads to chemically flexible AOs reaching the accuracy of standard (atom optimized) triple- or sometimes even quadruple-ζ quality sets.

In this work, we apply this basis set construction concept to the polarized valence single-ζ (vSZP) level and extend it (*vide infra*). The main reason for this truncation is that the vDZP basis set is *i)* too large for typical SQM methods and *ii)* that non-vanishing SQM Hamiltonian matrix elements between functions of the same angular momentum on the same atom (e.g., s and s', p and p') potentially lead to unstable results in

conventional SCF approaches (see Ref. [2] for a detailed discussion).

A severe limitation of any MB set (each orbital shell occupied in the free atom is described by only one radial AO basis function) is that the spatial expansion or contraction of the atomic orbital in the molecular environment with the effective nuclear charge (screening) cannot be described. In the following, this effect will also be referred to as "breathing" of the orbitals, a concept well known in valence bond theory[421] and also mentioned in a recent publication by Neese *et al.*[243] (see below). In fact, the whole idea of molecular basis optimization more or less breaks down in the MB case because the resulting single functions as an average over the RMs would resemble the AOs of fixed average molecular states, but would never be able to describe distinctly different situations properly. The only way to overcome this fundamental problem is to make the basis set adaptive, i.e., individual for every symmetry-unique atom in the system, depending on the local geometric and electronic structure.

Here, to our knowledge, we describe for the first time a generally applicable, analytically differentiable procedure for an atom-in-molecule adaptive AO basis set. Based on this, we present a corresponding fully optimized basis set (for existing ECPs) for all elements up to $Z = 86$, except the lanthanides. In order to indicate the dependence of the AOs on the effective atomic charge q as a critical ingredient (see below), the new basis set is dubbed q-vSZP (q-dependent vSZP). A principal use is in typical SQM methods, where the contraction depth does almost not affect the computation time. A version with fewer primitive functions to speed up 'low-cost' DFT calculations is currently developed along the same lines in our lab, and will be described separately.

Some special, non-automatic approaches to modify orbital exponents in SQM[245,422,423] methods and HF theory[424] depending on the atomic charge state were already proposed decades ago. Very recently, Wang and Neese, independently from our work, proposed a related idea to make the contraction coefficients in a minimal ANO-type basis set dependent on the bond length for diatomic molecules.[243] While this accounts for the important expansion/compression of AOs with increasing/decreasing bond distance, in the mentioned work, the similarly important change of AOs with atomic charge was not considered explicitly.

The ideas presented here account in a general way for both effects, change of q and the CN[117,118] of an atom in any polyatomic molecule. The critical point is to pre-compute these descriptors by simple and robust theoretical models and only make the contraction coefficients (not the Gaussian exponents) linearly dependent on them. This enables convenient basis set optimization on the one hand and easy differentiation of integrals to obtain, e.g., nuclear forces on the other hand. The newly introduced contraction coefficient dependency is variationally optimized as before in DFT calculations for RMs. Therefore, this part of the approach is considered non-empirical, while the theoretical model to obtain the atomic charges q is empirically adjusted to reproduce DFT data.

The primary purpose of the q-vSZP basis is to be used in next-generation SQM methods developed and published separately (working title: GP3-xTB). It may also be employed in other low-cost or machine-learning theoretical models as a physically relatively complete yet efficient and compact representation of electronic wavefunctions. To highlight the importance of the basis set and exclude other influences, we evaluate q-vSZP compared to standard basis sets in this work consistently with the same state-of-the-art RSH-DFT functional (ωB97X-D4,[103]) which has been used in our lab also in previous works on this topic.[1] However, we note that the observations described in the following are relatively independent of the choice of the DFT functional.

After an outline of the theory and some technical details of the optimization procedure, exemplary results for electron densities, dipole moments, and static polarizabilities are given to illustrate the built-in flexibility of the new basis set. The main results section describes thorough benchmarking on thermochemical properties and molecular structures, for which we employ standard databases such as GMTKN55.[93]

E.3 Theory

E.3.1 Environment dependency of the basis functions

Within the q-vSZP basis set, the basis functions ϕ are no longer identical per atom type or element, as in standard Gaussian-type AO basis sets. Instead, each of the basis functions ϕ_{κ,A_i} depends on the effective charge $q^{\text{eff}}_{A_i}$ of the atom in the molecule. In the following, A_i denotes a distinct atom of element A. The effective charge is determined from the atomic charge $q^{\text{CEH}}_{A_i}$ (obtained from a new CEH model, see Sec. E.3.3) and the coordination number CN_{A_i} of the respective atom A_i (see below). The detailed relations are explained in the following.

The usual linear combination of primitive Gaussian functions to one contracted Gaussian-type AO (i.e., a basis function) is given in Eq. E.1. Therein, ϕ_{κ,A_i} denotes a basis function of atom A_i consisting of N_{pr} primitive functions. c_{λ,A_i} and $\chi_{\lambda,A}$ are the effective contraction coefficient and the primitive Gauss function λ, respectively.

$$\phi_{\kappa,A_i}(\zeta_\kappa, q^{\text{eff}}_{A_i}) = \sum_{\lambda\in\kappa}^{N_{\text{pr}}} c_{\lambda,A_i}(q^{\text{eff}}_{A_i})\,\chi_{\lambda,A}(\zeta_{\lambda,A}) \quad \text{with} \quad (A_i \in A)\,. \tag{E.1}$$

To avoid non-linear parameterization (i.e., the Gaussian exponents), we decided to make only the contraction coefficients atom-in-molecule specific. In deeply contracted basis functions (AOs), the effect of changed exponents can be emulated by linear variation of the primitive functions. Consequently, the primitive Gauss functions χ and the associated exponents ζ do not differ between atoms of the same type, as noted in Eq. E.1, and are therefore optimized element-wise.

The dependency of the contraction coefficients c_{λ,A_i} of atom A_i on its effective atomic charge $q^{\text{eff}}_{A_i}$ is defined in Eq. E.2

$$c_{\lambda,A_i}(q^{\text{eff}}_{A_i}) = c_{0,\lambda,A} + c_{1,\lambda,A}\cdot q^{\text{eff}}_{A_i}, \tag{E.2}$$

where $c_{0,\lambda,A}$ is the base (free atom) contraction coefficient of the primitive function λ for element A. The coefficients $c_{1,\lambda,A}$ to be determined linearly change the primitive contribution in dependence of the chemical environment, i.e., reshaping the molecule's AOs. The effective charge descriptor consists of the most important atomic charge (linearly and quadratically) and CN-dependent parts as well as an additional cross-term according to

$$q^{\text{eff}}_{A_i}(q_{A_i}, \text{CN}_{A_i}) = q^{\text{CEH}}_{A_i} + k_{1,A}\cdot\left(q^{\text{CEH}}_{A_i}\right)^2 + k_{2,A}\cdot\sqrt{\text{CN}_{A_i}} + k_{3,A}\cdot\text{CN}_{A_i}\cdot q^{\text{CEH}}_{A_i}. \tag{E.3}$$

The square root of the CN is taken in the third term so that large values, which typically appear in dense metal-containing systems, are effectively damped. $k_{x,A}$ denote element-specific prefactors that are variationally optimized together with the coefficients $c_{0,\lambda,A}$ and $c_{1,\lambda,A}$. The atomic charges $q^{\text{CEH}}_{A_i}$ are obtained from the newly developed CEH model described in Sec. E.3.3 and CN is the error function-modified coordination number from Ref. [2]. For a neutral atom ($q^{\text{CEH}}_{A_i} = 0$, $\text{CN}_{A_i} = 0$), $c_{\lambda,A_i} = c_{0,\lambda,A}$ holds.

Because the analytical derivatives of the q^{CEH} and CN with respect to nuclear displacement are readily available, the corresponding AO derivatives can be derived, thus enabling efficient computation of analytical forces. In passing, we note that the $q^{\text{CEH}}_{A_i}$ may depend on an applied external electric field, making the basis functions also field-dependent, which will be explored in Sec. E.5.3 for dipole moments and static dipole polarizabilities.

E.3.2 Basis set construction

To approach the target accuracy as close as possible (realistically about vDZP level) while preserving the MB character (i.e., only one basis function per occupied AO shell), we decided to use a relatively deep contraction (a large number of primitive functions) for each AO to reduce incompleteness errors. This was inspired by the recently developed vDZP[1] basis set, for which a similar treatment (but with two valence shells) was successfully applied.

The second feature taken over from the vDZP basis set is the optimization of the exponents and coefficients in atomic *and* molecular calculations. Thereby, polarization functions, which are not occupied in the ground state of the free atom, are optimized to describe typical bonding situations best. Moreover, as a consequence, deeply contracted basis functions better resemble the actual molecular orbital with larger basis sets, partially circumventing an extensive LCAO. In the previous optimization of the vDZP basis, the free atom and its positive and negative ions and hydrides of different compositions were utilized. Here, for q-vSZP, which features a larger number of variational parameters, additional ions and small molecules containing different first-row elements were included. See Sec. E.4.3 for details on the optimization process of the basis set and Ref. [1] for a more in-depth discussion of molecular basis set optimization.

The third similarity to the vDZP basis set is the use of large-core ECPs in the q-vSZP basis set to make it usable for semiempirical methods, which generally only consider the valence electrons explicitly. As mentioned above, this choice avoids problems in SQM theory with atomic functions of the same angular momentum. However, in contrast to the ECPs used in the vDZP basis set, alkaline, and alkaline earth metals, as well as transition metals (groups 1-12), are now treated as valence only, also due to the beforementioned reason. Moreover, for group 12 elements, only two electrons are explicitly treated, since the *d*-electrons are considered by matching ECPs. This is realized by using the `ECP-<ncore>-SDF` ECPs instead of the `CRENBL` ECPs for groups 1 and 2 and the `CRENBS` ECPs instead of the `ECP-<ncore>-MWB` ECPs for groups 3–11. For group 12, the `ECP-<ncore>-SDF` ECPs with two valence electrons are used. See Tab. E.1 for details on the employed ECP for each element. Most importantly, all first- and second-row elements contain (contracted) polarization functions, which are generally absent in (standard) SQM methods. We think that in this way, "the right result for the right reason" can be obtained for electronically more complex cases or important standard thermochemistry problems (e.g., conformational energies). However, this substantially increases the computational cost because the basis set size, e.g., for carbon, is roughly doubled due to the additional d-shell. In passing, we note that for the crucial elements H, C, N, and O, the polarization functions are also relatively deeply contracted with three primitives to enable proper "breathing" of these orbitals as well. Here, we only consider the full basis with polarization functions. Still, for specific SQM methods, it might be beneficial (for efficiency reasons or to reduce the number of required empirical parameters) to discard them on, e.g., first- and second-row elements.

In summary, chemical elements in q-vSZP are described consistently by a [1s1p1d] basis ([1s1p] for H and He). Because we are not aware of any consistent, valence-only ECP for the lanthanide elements, they are not supported yet in the q-vSZP basis set (details are given in Tab. E.1). In contrast, the new CEH model described in the following Sec. E.3.3 is also available for the lanthanides.

E.3.3 Charge Extended Hückel Model for Atomic Charges

Initially, we tested the well-established, semiclassical EEQ model, which has been successfully applied for some years in our D4[119,120] and GFN-FF[189] methods, to obtain robust atomic charges efficiently. However, in certain situations (mainly noncovalent complexes composed of elements with considerable electronegativity

Table E.1: Primitive contraction scheme and ECPs of the q-vSZP basis set for all elements up to radon (except for the lanthanides). The contraction scheme corresponds always to (xsypzd) → [1s1p1d] (for H and He: [1s1p]). Columns three and four indicate the type of effective core potential used and the corresponding number of core electrons.

element	prim. functs.	ECP type	# core el.
H–He	(8s3p)	-	-
Li	(5s4p2d)	ECP-2-SDF[425]	2
Be	(6s4p2d)	ECP-2-SDF[425]	2
B	(6s5p2d)	ECP-2-MWB[141†]	2
C–O	(6s6p3d)	ECP-2-MWB[141]	2
F–Ne	(6s6p2d)	ECP-2-MWB[141†]	2
Na	(4s4p2d)	ECP-10-SDF[425]	10
Mg	(4s3p2d)	ECP-10-SDF[425]	10
Al	(5s4p2d)	ECP-10-MWB[141]	10
Si	(6s4p2d)	ECP-10-MWB[141]	10
P–Ar	(6s6p2d)	ECP-10-MWB[141]	10
K,Ca	(4s3p2d)	ECP-18-SDF[425]	18
Sc–Cu	(5s3p6d)	CRENBS[348]	18
Zn	(5s3p2d)	ECP-28-SDF[426]	28
Ga,Br,Kr	(6s5p2d)	ECP-28-MWB[141,349]	28
Ge–Se	(6s6p2d)	ECP-28-MWB[141,349]	28
Rb	(4s3p2d)	ECP-36-SDF[425,427]	36
Sr	(4s4p2d)	ECP-36-SDF[425,427]	36
Y–Mo,Ru–Ag	(5s3p6d)	CRENBS[144]	36
Tc	(5s2p6d)	CRENBS[144]	36
Cd	(5s3p2d)	ECP-46-SDF[426]	46
In	(5s4p2d)	ECP-46-MWB[141,349]	46
Sn–Xe	(6s5p2d)	ECP-46-MWB[141,349]	46
Cs–Ba	(4s3p2d)	ECP-54-SDF[427]	54
La	(4s2p5d)	CRENBS[351]	54
Hf	(4s3p5d)	CRENBS[351]	68
Ta–Au	(5s3p5d)	CRENBS[351]	68
Hg	(5s3p2d)	ECP-78-SDF[428]	78
Tl,Bi–Rn	(6s5p2d)	ECP-78-MWB[354]	78
Pb	(5s5p2d)	ECP-78-MWB[354]	78

differences), the EEQ model produces artificial CT, sometimes on the order of about 0.1-0.2 electrons. This, in turn, leads to an incorrect AO expansion/contraction and, as a result, unacceptable interaction energy errors of several kcal·mol^{-1}.

In recent years, we have had a good experience with EHT-type SQM Hamiltonians for the determination of atomic charges by a standard Mulliken population analysis.[324] Here, we introduce a stand-alone method termed Charge Extended Hückel (CEH) to obtain atomic charges for given cartesian coordinates and element ordinal numbers.

For the CEH method, we solve the general eigenvalue problem

$$\mathbf{H}^{\mathrm{CEH}}\mathbf{C} = \mathbf{S}\mathbf{C}\boldsymbol{\epsilon} \tag{E.4}$$

once, where $\mathbf{H}^{\mathrm{CEH}}$ denotes the effective Hamiltonian matrix in the special CEH minimal AO basis, $\mathbf{S}$ is the corresponding AO overlap matrix, and $\mathbf{C}$ are the MO coefficients with eigenvalues $\boldsymbol{\epsilon}$. A restricted (Fermi-smearing) procedure as in the GFNn-xTB methods[277,278] is applied. The valence MOs are expanded as usual in atom centered STOs, which are approximated by Stewart's Gaussian expansions[145] with usually four or six primitives per AO similar to GFNn-xTB.[259] The corresponding Slater exponents ζ^l are treated in this part as empirical parameters.

In the following description, A refers to an atom, l to an AO shell (1s, 2s, 2p, …), l_a to the angular momentum (s, p, d, …), L stands for the interaction type (see below), and greek indices denote AOs. The matrix elements of $\mathbf{H}^{\mathrm{CEH}}$ are given by

$$H^{\mathrm{CEH}}_{\kappa\lambda} = \frac{1}{2}\left(k^{l_a(l)} + k^{l_a(l')}\right)_{\kappa\lambda} S^{\mathrm{sc}}_{\kappa\lambda}\left(H^{l} + H^{l'}\right)_{\kappa\lambda} \quad \text{with} \quad (\kappa \in l(A), \lambda \in l'(B))\,, \tag{E.5}$$

where k^{l_a} are three (s, p, d) global (angular momentum-dependent) Wolfsberg parameters, H^l are atomic shell energy levels (see Eq. E.7 below). $\mathbf{S}^{\mathrm{sc}}$ is an overlap matrix that is scaled in the diatomic frame and then transformed back to the Cartesian coordinate system:[247,325,326]

$$\begin{aligned} S^{\mathrm{sc}}_{\kappa\lambda} &= \sum_{mn} O_{\kappa m}\tilde{S}^{\mathrm{sc}}_{mn}O^{T}_{n\lambda} \\ &= \sum_{mn} O_{\kappa m}\left(\left(\sum_{\kappa\lambda} O^{T}_{m\kappa}S_{\kappa\lambda}O_{\lambda n}\right)\cdot\frac{2}{\left(\frac{1}{k^{L}_{A}}+\frac{1}{k^{L}_{B}}\right)}\right)O^{T}_{n\lambda}, \end{aligned} \tag{E.6}$$

with $\mathbf{O}$ as the unitary transformation matrix for transforming an overlap submatrix between A and B into the diatomic frame. The exact diatomic overlap integrals $\tilde{S}_{kl}$ are scaled depending on whether the interaction between κ and λ is of σ ($L = 0$), π ($L = 1$), or δ ($L = 2$). The k^{L}_{A} are additional element-specific parameters for the three different interaction types. For further details on the overlap transformation into the diatomic frame, see Ref. [247, 325, 326].

The atomic shell energy levels H^l are based on the free atom level energies H^l_A modified by CN-dependent terms

$$H^{l} = H^{l}_{A} + k^{l}_{\mathrm{CN},A}\mathrm{CN}_{A} + k_{\mathrm{CN},A}\mathrm{CN}^{\chi}_{A}, \tag{E.7}$$

with l being an AO shell of atom A. Here, CN_A is the standard CN obtained with empirically fitted atomic radii $R_{\mathrm{CN},A}$, parameterized shell-wise by $k^{l}_{\mathrm{CN},A}$. In order to include some electrostatic screening effects, CN^{χ}_{A}

† modified d-f projector for B and F, see Ref. [1] for details.

is a second, distinct coordination number with electronegativity weighting, i.e.,

$$\mathrm{CN}_A^{\chi} = \frac{1}{2}\sum_{B\neq A}^{N_{\mathrm{atoms}}} (\mathrm{EN}_A - \mathrm{EN}_B)\left(1 + \mathrm{erf}\left(k_{\mathrm{erf}}^{\mathrm{CN}}\frac{R_{AB} - R_{AB}^{\mathrm{cov}}}{R_{AB}^{\mathrm{cov}}}\right)\right), \tag{E.8}$$

where R_{AB} is the interactomic distance, $R_{AB}^{cov} = R_{\mathrm{CN},A} + R_{\mathrm{CN},B}$ is the covalent distance, EN_A is the empirical electronegativity, and $k_{\mathrm{erf}}^{\mathrm{CN}}$ is a global parameter. The last term in Eq. E.7 shifts the levels depending on the number of spatially close electropositive or -negative neighbors. Atomic charges are determined from the core charges Z_A and a standard Mulliken population analysis based on the unscaled, exact overlap matrix $\mathbf{S}$ and the final density matrix $\mathbf{P}$:

$$q_A = Z_A - \sum_{\kappa\in A}^{\mathrm{AO}} (\mathbf{PS})_{\kappa\kappa}\,. \tag{E.9}$$

For optimization of the eight (*sp*-block element) or 14 (*spd*-block element), respectively, empirical parameters employed in the CEH model, refer to Sec. E.4.2.

The accuracy of the atomic charges obtained with the CEH model and electric dipole moments as an additional check quantity is evaluated and discussed in detail in Sec. I B of the Supporting Information (SI). In general, the CEH model provides accurate atomic charges and dipole moments that are only slightly worse than those by more sophisticated methods like GFN2-xTB[278] or PTB.[2] Furthermore, the deviations are reasonably consistent over the element groups in the periodic table. Severe outliers, even for chemically unusual, electronically complicated structures, have not been observed yet. We think that this efficient atomic charge model can be of general use in quantum chemistry beyond the q-vSZP basis set, e.g., for atomic charges in semi-empirical dispersion corrections[119,120] or for SCC guesses in SQM methods.[267,278]

E.4 Technical Details

E.4.1 General

All DFT calculations for this work were conducted with the ORCA quantum chemistry package in version 5.0.4.[309,357] The DFT functional is fixed to ωB97X-D4[1,103,307] to evaluate the performance of various basis sets. The resolution of the identity (RI) approximation for the electronic Coulomb energy with the large universal auxiliary basis sets by Weigend[358] as well as seminumerical treatment of Fock exchange[178] were used throughout. To exclude possible numerical noise (especially for numerical derivatives), high convergence thresholds (keyword: `VeryTightSCF`) and grid settings (keyword: `DefGrid3`) were applied. Because analytical gradients of the adaptive basis set have not been implemented yet, geometry optimizations are conducted with numerical energy derivatives, exploiting the adapted basis set for distorted geometries. An analogous procedure is used for numerical derivatives with respect to an external electric field to obtain either dipole moments ($\mu = -\frac{dE}{dF}$) or static dipole polarizabilities ($\alpha = \frac{d\mu}{dF}$).

All numerical derivatives with respect to nuclear coordinates or external electric fields were conducted with the help of an in-house Python code, which is publicly available on our GitHub channel (github.com/grimme-lab).[429] Geometry optimizations were done in the external driver mode with `xtb`.[259,278] `ORCA` input files taking the atom-in-molecule adaptive basis set into account were automatically generated via a program incorporating the CEH charge and CN calculation. This program is also publicly accessible on GitHub[430] and can easily be modified to generate input files for other quantum chemistry packages as well. A complete sample q-vSZP input for `ORCA 5.0.4`, including the atom-in-molecule adaptive basis set, is given in Sec. II D of the SI for the H_2O molecule. The CEH charge model is implemented in the `tblite` program package,

accessible via the corresponding GitHub repository (github.com/tblite/tblite).

E.4.2 Optimization of the CEH Model Parameters

The element-specific parameters appearing in the CEH Hamiltonian, i.e., k^{l_a}, H^l_A, $k^l_{\mathrm{CN},A}$, k^L_A, $k_{\mathrm{CN},A}$, $R_{\mathrm{CN},A}$, EN_A, the Slater orbital exponents ζ^l (in total eight for an *sp*-block element, 14 for an *spd*-block element), and $k^{\mathrm{CN}}_{\mathrm{erf}}$ have been determined by a least-squares fit of computed Mulliken atomic charges (based on the unscaled exact overlap integrals) to DFT-based Hirshfeld charges (ωB97X-D4/TZVP level). In order to ensure robust and physically reasonable behavior, DFT-computed dipole moments and bond orders have been included with an overall weight of about 35 % in the minimized loss function. On average, about 50 small- to medium-sized molecules (up to about 50 atoms) per element were included in the training set, leading to roughly 1000-2000 training data points per element. The model is consistently available for all elements up to radon, including the lanthanides. All required empirical parameters mentioned above are provided in an additional text file `parameter_CEH.dat`. See Sec. II A of the SI for details and further explanations.

E.4.3 Basis Set Optimization

The optimization procedure for the q-vSZP basis set is very similar to the one applied recently for the vDZP valence basis set used in the ωB97X-3c composite DFT[1] and the PTB[2] methods. With the help of the Powell optimization algorithm,[431] which does not require any objective function value derivative, the sum of atomic and molecular ωB97X-D3[356] energies has been minimized. As for the vDZP basis functions, the free atom and its positive and negative ions and hydrides of XH_n and $\mathrm{X_2H}_m$ type were used. In addition, small molecules/ions containing Li, Be, C, N, O, F, and Cl were included to account for more variational parameters in the present approach. The atoms in the RMs different from the target element were described with the vDZP basis set to avoid basis set inter-dependencies in the optimization procedure. Because of the additional $c_{1,\lambda,A}$ coefficients and three $k_{x,A}$ parameters (per element), the overall number of degrees of freedom is somewhat higher than in a conventional basis set optimization but still manageable. The optimization procedure showed proper (but slow) convergence for all elements. In about 2-3 % of all cases, two primitives received very similar Gaussian exponents during optimization, so one of them had to be eliminated to avoid linear dependencies. The $c_{1,\lambda,A}$ were initialized with small positive values for all primitives except the most diffuse one to which a larger negative initial value was assigned. This corresponds to the physically expected situation where in an anion with $q < 0$ compared to neutral systems, the smallest exponent primitive is increased in weight while the contribution of the steeper functions decreases.

The new basis set together with the used ECPs is provided in the SI (as additional text files and in Sec. II B and C) and within the program described in Sec. E.4.1, enabling single-point energy calculations with standard quantum chemistry programs. Implementation of the new basis function derivations for analytical DFT gradient calculations is currently underway in our laboratory and will be described separately.

E.4.4 D4 Parameter Optimization

The D4 parameters for ωB97X-D4/q-vSZP were optimized for the S66x8[376,377] benchmark set to reduce the remaining BSSE. The same procedure was applied to all competing basis sets (including the versions with added polarization functions, see first paragraph of Sec. E.5), which should ensure a comparison on the same footing. To avoid problems with finding physically reasonable parameter minima in typical optimization algorithms for MBs, scans were conducted for a considerable space of D4 parameter combinations. For vDZP

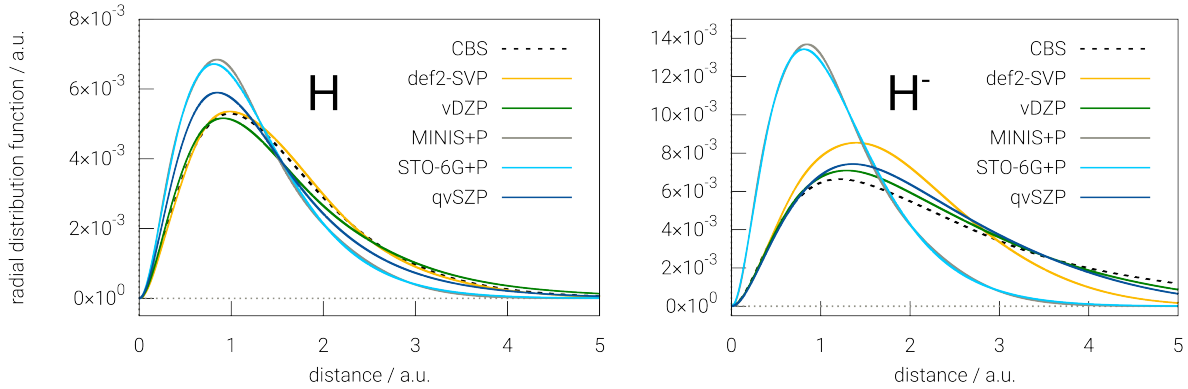

Figure E.1: Radial distribution function (RDF) of the electron density of the hydrogen atom (left) and hydride (right) by employing ωB97X-D4[1,103] in combination with different basis sets. The annotation "CBS" is used for an uncontracted 14s basis set from the TURBOMOLE[294,295] basis set library.

and def2-QZVP, the optimized D4 parameters presented in Ref. [1] were taken. The applied D4 parameters are listed in Sec. II E of the SI.

E.5 Results and Discussion

The goal of the presented atom-in-molecule adaptive q-vSZP basis set is to overcome the inherent accuracy limitations of typical MB sets without adding additional orbital shells (i.e., avoiding a higher cardinal number). Therefore, common MBs, such as STO-6G[293] or MINIS[296] are the natural choice as comparators on the lower end. Since q-vSZP includes polarization functions for all elements, while the aforementioned MBs are not polarized per default, standard polarization functions taken from the def2-SVP basis set[139] were added to the MBs, which should ensure a fair comparison. The resulting basis sets are termed "<MB>+P" in the following.

The (static) DZ basis set,[1] which is optimized similarly and contains roughly the same number of total primitive functions split into two AOs (i.e., vDZP), is taken as an upper limit for the achievable accuracy with the q-vSZP basis. Additionally, we compare against def2-SVP[139] as a very commonly used DZ basis set.

E.5.1 Electron Density of Neutral and Anionic Atoms

The simplest example showing the necessity for radial expansion or contraction of an AO is atomic hydrogen. In larger-than-minimal basis sets, the SCF determines the contribution of each basis function in the AOs individually for each calculation by exploiting the LCAO approach. In the example of H^-, more diffuse *s* functions have a greater contribution compared to the neutral atom – the AO expands. In contrast, when using standard MBs, the AOs are fixed to an average atom-in-molecule state.

This effect is illustrated in Fig. E.1 for different basis sets, using ωB97X-D4 throughout. As a reference, an uncontracted 14*s* basis set is used. The shapes of the radial distribution functions (RDFs) of the electron density do not change between H and H^- when employing the polarized MBs STO-6G+P and MINIS+P. In fact, they even over-localize the electron density in both cases, which is probably owed to the fact that hydrogen carries a positive partial charge in most common bonding situations, for which these sets have been developed.

Table E.2: BB CP corrections evaluated at the ωB97X-D4 level for different basis sets. Relative values for the BB-CP correction in % are determined from the absolute values in kcal·mol^{-1} with respect to the reference dimerization energies. The complexes are taken from the S22 benchmark set.[432,433]

	rel. CP corr. (%)	abs. CP corr. (kcal·mol^{-1})
STO-6G+P	40.4	4.05
STO-6G	33.5	3.29
MINIS+P	32.1	2.22
MINIS	27.0	2.00
q-vSZP	15.2	0.86
def2-SVP	34.4	2.20
vDZP	13.6	0.64
def-TZVP	8.1	0.49

In contrast, q-vSZP behaves similarly to the DZ bases def2-SVP and vDZP. For the anionic case, the electron density by q-vSZP is even closer to the converged result than def2-SVP. Since no coordinating atoms are present, this is purely attributed to the charge dependency in Eq. E.2 and E.3, increasing the contraction coefficients of the diffuse compared to the compact primitive functions. The vDZP basis set is remarkably close to the CBS limit for both H and H$^-$.

In Sec. III A of the SI, similarly good results are shown for the neutral and anionic carbon atom.

E.5.2 Basis Set Superposition Error

Inspired by the approach for generating the vDZP basis set, the q-vSZP basis set presented here has been optimized for atoms and molecules. This leads to generally better fitting AOs in molecules and complexes and, consequently, to a diminished amount of BSSE since the unphysical use of basis functions of other atoms is reduced, as discussed in Ref. [1]. The residual amount of BSSE is estimated by calculating BB CP corrections for the noncovalent complexes in the S22[432,433] benchmark set and given in Tab. E.2.

As expected, the additional polarization functions added to STO-6G and MINIS increase the BSSE slightly by about 5-7 % in comparison to the pure MBs. However, even without the additional polarization functions, the BSSE is at a level of about 30 %, which is similar to usual DZ basis sets such as def2-SVP. The molecular optimized q-vSZP and vDZP bases show distinctly smaller relative BSSEs of about 15 %, which is close to that of def-TZVP.[342] Importantly, the surprisingly small absolute BSSE of q-vSZP (<1 kcal/mol) enables its absorbance in the anyway required D3[117,118] or D4[119,120] dispersion damping parameters, thereby avoiding further empirical corrections like, e.g., gCP or DFT-C.[153,195]

E.5.3 Molecular Properties

In typical QM molecular property calculations, the dipole moment and the static dipole polarizability are evaluated as expectation values from dipole integrals or as analytical derivatives of the electronic energy with respect to an external electric field via the coupled-perturbed self-consistent field (CP-SCF) equations, respectively, while leaving the basis functions fixed. In contrast, in the presented q-vSZP approach, the basis set is no longer fixed and depends on the CEH atomic charges, which themselves depend on the electric field. Considering the external electric field within the CEH model, the basis set can adapt to it. This behavior is numerically investigated for the dipole moment and dipole polarizability by finite-field

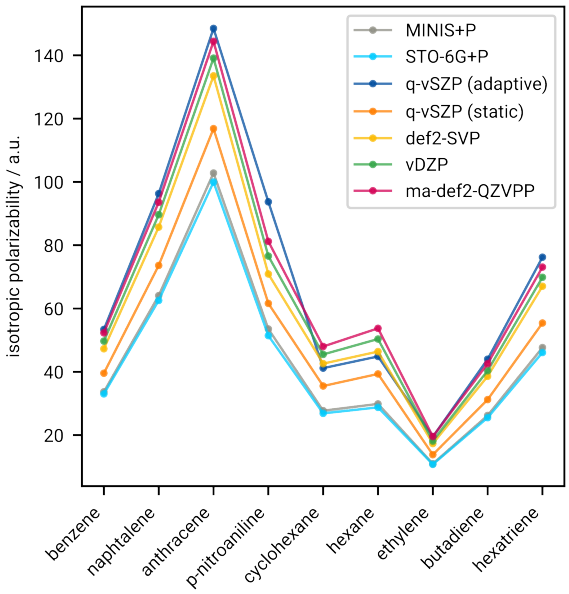

Figure E.2: Static isotropic dipole polarizabilities for nine different organic molecules calculated with ωB97X-D4 using different basis sets. For "q-vSZP (adaptive)", the electric field dependency of the basis set is exploited by calculating the polarizability numerically from dipole moment gradients and turned-on electric field. "q-vSZP (static)" indicates the analytical polarizability calculation with a zero-field basis. The minimally augmented def2-QZVPP basis set[434] is considered as the reference.

perturbation calculations.

In Fig. E.2, this is explored for static dipole polarizabilities, of which the isotropic value is discussed (termed *polarizability* in the following, for the sake of simplicity). For all competing basis sets and "q-vSZP (static)", the polarizability is calculated analytically via the CP-SCF method. In contrast, for "q-vSZP (adaptive)", it is determined via numerical differentiation of the dipole moment with respect to an external electric field. As expected, the polarizability increases significantly with basis set size, i.e., MINIS+P ≈ STO-6G+P < q-vSZP (static) < def2-SVP < vDZP < ma-def2-QZVPP.[434] This trend is relatively insensitive to the type of (organic) compound. On the other side, the adaptive q-vSZP basis yields significantly larger polarizabilities than the conventional MBs. This is particularly true for unsaturated and conjugated molecules such as aromatic systems (*f.l.t.r.*: 1-4) and a series of conjugated alkenes (7-9). For saturated alkanes (5-6), polarizabilities with the adaptive q-vSZP basis are still larger than with the static approach, but the increase is minor. This effect can probably be attributed to the CEH model, which allows for the physically correct field-induced CT. As these CT effects are more dominant in conjugated systems than in saturated alkanes, the charges and, consequently, the basis set changes more strongly for the former systems. The adaptive q-vSZP basis set yields polarizabilities that are very close to the ma-def2-QZVPP values. This could provide new opportunities for the calculation of higher-order molecular properties using MBs (e.g., in future SQM methods).

The analog comparison for dipole moments (expectation value with a fixed basis set vs. field-dependent numerical derivative) is shown in Fig. 3 in Sec. III B of the SI. The MADs (with respect to ma-def2-QZVPP dipole moments) amount to 0.39 a.u. for the adaptive approach and 0.08 a.u. for the static approach.

E.5.4 Main-Group Thermochemistry

A primary target of the presented q-vSZP basis set is the replacement of standard MBs (e.g., STO-6G) in SQM and potentially also in simplified DFT methods for thermochemical applications. We thoroughly evaluated the q-vSZP performance at the DFT level together with competing basis sets on the well-known GMTKN55[93] benchmark set. GMTKN55 is a comprehensive benchmark database for main group (thermo-)chemistry. All data points correspond to relative (chemical) energies, avoiding the meaningless comparison of absolute energies between different basis sets with different ECPs. The database is grouped into five categories, ranging from basic properties and reactions of small and large molecules over barrier heights to inter- and intramolecular NCIs. Over 1500 individual reactions with high-level CCSD(T)/CBS reference data are included. However, a small number of the molecules in the database contain atoms heavier than calcium, for which the MINIS basis set is unavailable. Therefore, we set up a reduced variant of GMTKN55 containing only 1414 data points. In this modification, the subsets HAL59, HEAVYSB11, and RG18 contain fewer data points, and the subset HEAVY28 is omitted. To level considerable differences in reaction energies throughout the whole benchmark set, the GMTKN55 introduces a weighted error measure for MADs of the distinct subsets, termed WTMAD-2, which is used in the following throughout to discuss the performance of the different basis sets. The mean reaction energies, average reaction energies, and number of data points per subset that enter the WTMAD-2 are modified accordingly in this reduced variant of the GMTKN55.

WTMAD-2 values for each subcategory are shown in Fig. E.3 (top) for ωB97X-D4 employing the q-vSZP and the competing basis sets vDZP, def2-SVP, STO-6+P, and MINIS+P. The overall WTMAD-2 values are visualized in the bottom part of Fig. E.3. Results for the original MBs STO-6G and MINIS, as well as for the q-vSZP basis after truncation of all polarization functions ("q-vSZ"), are given in Sec. I C of the SI. STO-6G+P shows enormous errors for all subclasses of GMTKN55, indicating the (expected) severe amount of BSIE. Interestingly, this is significantly less pronounced for MINIS+P, but the errors are still larger than that for def2-SVP by at least a factor of two. The unpolarized pure STO-6G and MINIS basis sets exhibit even slightly larger errors. The commonly employed DZ basis set def2-SVP suffers mainly from errors in the NCIs, which is related to a large BSSE, as shown in Sec. E.5.2. The resulting WTMAD-2 of 13.0 kcal·mol^{-1} is still far away from the performance of the functional in a close-to-converged def2-QZVP basis set (3.7 kcal·mol^{-1}, only shown in the bottom part of Fig. E.3). The vDZP basis set (corresponding to ωB97X-3c from Ref. [1]) is close to the def2-QZVP result, with a WTMAD-2 of 5.5 kcal·mol^{-1}.

The here presented q-vSZP basis set lies between def2-SVP and vDZP with a WTMAD-2 of 12.6 kcal·mol^{-1}, outperforming STO-6G+P and MINIS+P and their unpolarized counterparts significantly. Especially for intra- and even more so for intermolecular NCIs, q-vSZP is more accurate than def2-SVP, which is in agreement with the above-mentioned BSSEs for the different methods, shown in Fig. E.2. def2-SVP introduces slightly smaller errors than q-vSZP for basic properties, reactions, and barrier heights. This indicates that atom-in-molecule-specific contraction coefficients cannot fully compensate for the missing second AO per orbital shell. However, even for reactions of small molecules, the improvement by q-vSZP is still of about 8 to 13 kcal·mol^{-1} in WTMAD-2 compared to MINIS+P and STO-6G+P. These results underline the importance of both (*i*) the molecular optimization of the basis set and (*ii*) the atom-in-molecule-adaptive contraction coefficients. In any case, we emphasize that even a MB can keep up with the performance of standard DZ basis sets such as def2-SVP concerning thermochemical properties.

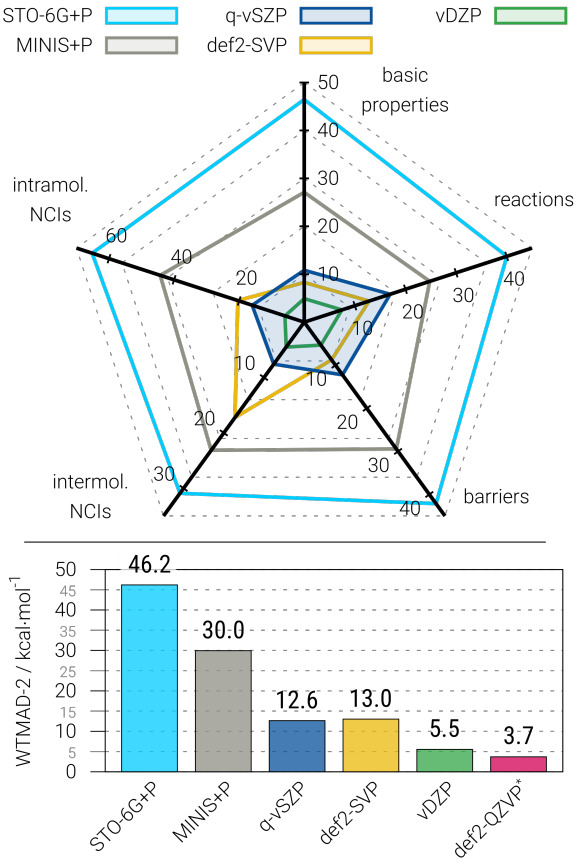

Figure E.3: **Top:** WTMAD-2 values for the subclasses of the GMTKN55 benchmark set in kcal·mol^{-1}. **Bottom:** Overall WTMAD-2 values for the whole GMTKN55 benchmark set in kcal·mol^{-1}. The exact numbers for each basis set are shown on top of each bar. The asterisk at "def2-QZVP" indicates that for the sets WATER27, G21EA, AHB21, and IL16, additional diffuse functions were added (see Ref. [93] for details). Raw data required for the reproduction of the figures can be found in the archive `gmtkn55_data.zip` in the SI.

E.5.5 Molecular Structures

Very efficient SQM and simplified composite DFT methods are heavily used for optimizing molecular structures or reaction path searches.[171,289,435] Therefore, the presented q-vSZP basis set, designed to provide the basis for future methods of the types above, should yield accurate equilibrium geometries. This is checked on the LMGB35 (experimental bond lengths for small, light-main group molecules)[115,373] and CCse21 (high-level semi-experimental bond lengths and angles)[375] benchmark sets.

For the q-vSZP basis, the correct derivative of the electronic energy with respect to the nuclear coordinates involves a derivative with respect to the basis set itself (cf. Sec. E.4.1 and E.5.3). Since this has not yet been implemented analytically, the gradient is obtained numerically via a finite-step size approach. For all other basis sets, the analytical gradient is used.

The boxplots in Fig. E.4 display the statistical results, with **a)** showing deviations from reference bond lengths for the combined CCse21 and LMGB35 benchmark sets. Refer to Sec. I D in the SI for the individual benchmark sets. Furthermore, **b)** visualizes deviations from reference bond angles in the CCse21 benchmark

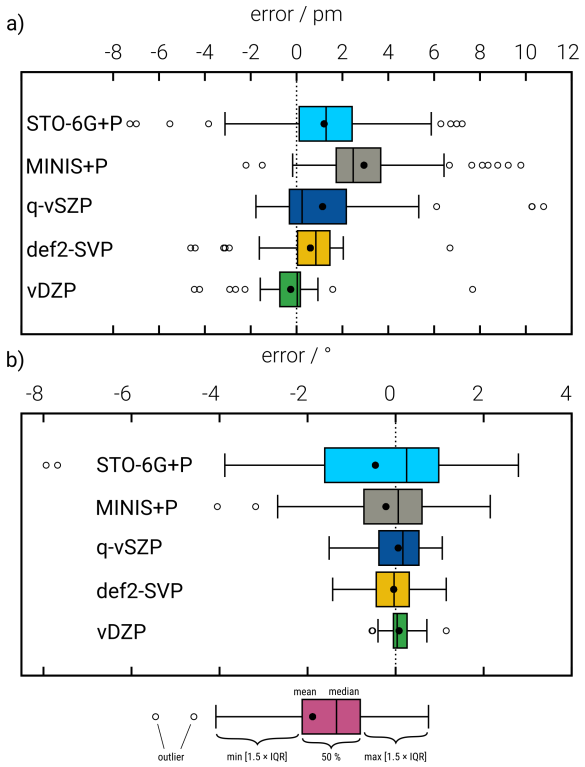

Figure E.4: **a)** Bond length deviations from reference values in the combined LMGB35[373] and CCse21[375] geometry benchmark sets. **b)** Deviations from reference angles in the CCse21[375] benchmark set. The sketched boxplot under Figure *b)* displays the utilized statistical boxplot definition. "IQR" corresponds to the interquartile range, the distance between the upper and lower quartiles. All results refer to ωB97X-D4 in combination with the respective basis set. Raw data required for the reproduction of the figures can be found in the SI files `ccse21.ods` and `lmgb35.ods`.

set. The comparison of bond lengths in **a)** reveals that the vDZP basis set performs best, followed by def2-SVP. All MBs exhibit more significant errors, but q-vSZP yields the lowest MAD compared to STO-6+P and MINIS+P in both benchmark sets. For CCse21, q-vSZP even outperforms def2-SVP concerning the signed ME, the MAE, and the ErrR. It is noteworthy that in contrast to the observations made for thermochemical accuracy in Sec. E.5.4, STO-6+P produces lower (systematic) errors than MINIS+P, also indicated by lower MEs. The error distributions for bond angles in **b)** follow the same trends as for the bond lengths with STO-6G+P and MINIS+P exhibiting more significant deviations compared to def2-SVP and, more particularly, to vDZP. However, similar to the CCse21 bond lengths, q-vSZP reaches the accuracy of def2-SVP, revealing almost equal error statistics. The observations made for bond lengths and angles demonstrate the excellent applicability of the q-vSZP basis set for molecular structure determination.

E.6 Summary and Conclusions

This work introduces an unprecedented general way of setting up an atom-in-molecule adaptive AO minimal basis set. By adjusting the coefficients of the deeply contracted AOs based on the CN and pre-calculated atomic charges, the AOs can "breathe", i.e., expand or contract depending on the chemical environment. The atomic charges required for this ansatz are efficiently calculated via a robust, novel EHT method termed

"CEH". Both the CEH method and the q-vSZP basis set are available for the whole periodic table up to radon (currently excluding the lanthanides for the basis set because of the lack of appropriate ECPs). A key point of the CEH method is the account of challenging CT effects exclusive to a QM treatment and not accessible with, e.g., classical charge equilibration methods.

Besides the atom-in-molecule adaptive contraction coefficients, the q-vSZP basis set profits immensely from molecular optimization of the entire basis set, the influence of which was proven earlier in Ref. [1] at the vDZP level. In particular, the low amounts of BSSE of the q-vSZP basis set can be attributed to the molecular optimization of the basis. The importance of the above-mentioned "breathing" effects depending on the atomic partial charge was revealed by RDFs of the electron density for neutral and anionic single atoms. This advantage was further investigated for numerical derivatives of static dipole polarizabilities, for which q-vSZP by its field dependence circumvents the systematic underestimation of polarizabilities with usual MBs.

Thermochemical and molecular structure benchmark sets convincingly demonstrate the practical advantage of combining atom-in-molecule adaptive AOs with molecular optimization of the basis set. The q-vSZP basis was applied with the state-of-the-art ωB97X-D4 functional. For comparison, we set up polarized versions of the standard MBs STO-6G and MINIS ("+P"). Additionally, we compared to the DZ basis sets vDZP and def2-SVP. In all benchmark sets, the single-ζ (minimal) q-vSZP basis can compete with def2-SVP as a standard DZ basis set. Thus, it is to some extent possible to emulate the flexibility of two AOs per shell with a single AO that adapts to its environment. For the established GMTKN55 benchmark set, the q-vSZP basis performs immensely better than STO-6G+P and MINIS+P. For molecular structures, q-vSZP provided equally good and, in most cases, more accurate geometries than standard MBs.

We are currently working on implementing a modified (less polarized) q-vSZP basis in upcoming SQM tight-binding methods. Additionally, we are exploring its usage in simplified DFT techniques, including analytical basis set gradients. We plan to incorporate these highly efficient methods for large-scale QM investigations into freely available QM packages in the near future.

E.7 Data Availability

The data supporting the findings of this study are available within the article and its supplementary material. Any further information is available upon request from the authors.

E.8 Supplementary Material

In the supplementary material, we provide further statistics on the GMTKN55 benchmark results and on training data for the CEH model, additional boxplot analyses for the distinct molecular structure benchmark sets, the q-vSZP basis set and matching ECPs in tabulated form, a reference input and total energy output (`ORCA 5.0.4`) for a molecular example, the optimized D4 parameters for all discussed small basis sets, an analysis of the computational effort associated with the q-vSZP basis set, tests for the robustness of numerical derivatives with the q-vSZP basis, and further investigations on the q-vSZP electron density and dipole moments. The q-vSZP basis set (including the corresponding ECPs) is provided additionally in text file format within the `q-vSZP_basis.zip` archive. All empirical parameters for the CEH method are provided within the text file `parameter_CEH.dat`. The appropriate legend for the parameter file is also given in the SI. Raw data for the GMTKN55 and molecular structure benchmark sets (LMGB35 and CCse21) are provided as additional files.

E.9 Acknowledgements

The German Science Foundation (DFG) is gratefully acknowledged for financial support through the SPP 2363 “Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning”. M. M. thanks the Fonds der Chemischen Industrie (FCI) for funding via a Kekulé scholarship. We would also like to express our appreciation to T. Froitzheim for helpful discussions on the technical implementation.

APPENDIX F

Appendix: Advanced Charge Extended Hückel (CEH) Model and a Consistent Adaptive Minimal Basis Set for the Elements Z=1–103

Marcel Müller,[†] Thomas Froitzheim,[†] Andreas Hansen,[†] Stefan Grimme[†]

Received: October 15, 2024
Published online: November 25, 2024

Reprinted in Appendix F with permission[§] from
M. Müller, T. Froitzheim, A. Hansen, and S. Grimme, *Advanced Charge Extended Hückel (CEH) Model and a Consistent Adaptive Minimal Basis Set for the Elements Z = 1–103*, The Journal of Physical Chemistry A **128** *49* (2024), Publisher: American Chemical Society 10723, DOI: 10.1021/acs.jpca.4c06989

Own contributions

- Conceptualization of improvements
- Performing all calculations and generation of the results
- Interpretation of the results
- Writing of the manuscript

[†] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, D-53115 Bonn, Germany
[§] Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

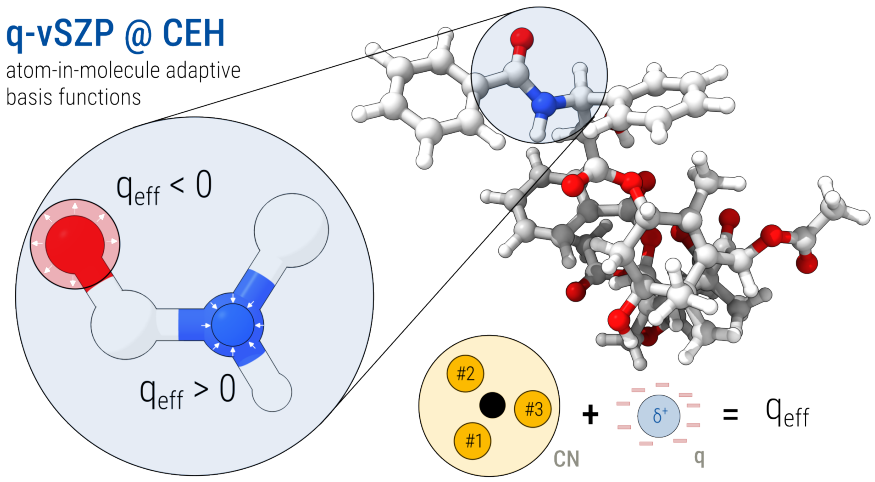

Figure F.1: Associated Table of Contents graphic for publication in The Journal of Physical Chemistry A.

F.1 Abstract

The Charge Extended Hückel (CEH) model, initially introduced for adaptive atomic orbital (AO) basis set construction (*J. Chem. Phys.* **2023**, *159*, 164108), has been significantly revised to enhance accuracy and robustness, particularly in challenging electronic situations. This revision includes an extension towards *f*-elements, covering actinoids with their *f*-electrons in the valence space. We present a novel non-iterative approximation for the electrostatic contribution to the effective Fock matrix, which substantially improves performance in polar or charged systems. Additionally, the training dataset for elements $Z = 1 - 103$ has been expanded to encompass even more chemically diverse reference molecules as well as dipole moments and shell populations in addition to atomic charges. It includes a greater variety of "mindless" molecules (MLMs) as well as more complex electronic structures through open-shell and highly charged species. The revised method achieves mean absolute errors for atomic charges q of approximately 0.02 e^- for randomly selected (mostly organic) molecules and 0.09 e^- for MLMs, outperforming both classical charge models and established tight-binding methods. Furthermore, the revised CEH model has been validated through density functional theory calculations with the updated adaptive q-vSZP AO basis set on common thermochemical databases. Consistent with the extension of the CEH model, q-vSZP has also been variationally optimized and tested for elements $Z = 58 - 71$ and $87 - 103$. The original versions of both CEH and q-vSZP are now considered deprecated.

F.2 Introduction

Semiempirical electronic structure methods have been experiencing a resurgence, as highlighted by a recent special issue on this topic.[215] Modern advancements aim to enhance accuracy, broaden applicability to more elements and applications, and further improve efficiency for large systems.[2,243,259,277,278,436] These developments are significantly influenced by the rise of ML techniques, leading to potential mutual enhancements.[191,437–439]

Many low-cost or SQM electronic structure methods are limited by their use of unpolarized minimal AO basis sets.[277,278] To address this issue, we recently introduced a DFT variationally optimized, atom-in-molecule

adaptive MB set, which has been made available consistently for elements up to radon ($Z = 86$, excluding the lanthanides (Lns)) and termed q-vSZP.[3,430] The key innovation is the introduction of a linear dependence for the primitive Gaussian coefficients in a contracted AO on the effective atomic charge of the atom within the molecule. Each symmetry-unique atom thus acquires specifically adapted basis functions. This approach enables the "breathing" of the AOs with both the atomic charge (expansion/contraction for anionic/cationic states) and the number of nearby bonded atoms within a molecule ("Pauli"-type contraction/expansion with more/less neighbors). The q-vSZP basis underpins our current development of an improved TB SQM method, intended as a successor to the established GFN2-xTB model[278] (working title: g-xTB).

The atomic charges required for the q-vSZP basis setup are determined by the specially developed CEH model, while coordination numbers are derived from the molecular geometry. During the g-xTB development, we identified problematic cases for the CEH model addressed here through specific enhancements of the extended Hückel-type Hamiltonian. Additionally, we extend the model to include the previously missing elements $Z = 87 - 103$ (francium, radium, and the actinides), providing a consistent atomic charge model and basis set for the first 103 elements. This extension is particularly notable, as (semi)empirical methods and consistent Gaussian AO basis sets for the 7*s* and 5*f* elements are rare.[440–448] Recent exceptions for the former include the parameterization of DFTB as well as GFN-FF for 5*f* elements.[449,450]

Initially developed as an auxiliary tool for the charge-dependence of the q-vSZP basis set, the CEH model has evolved into a stand-alone method for obtaining atomic charges from given Cartesian coordinates, element ordinal numbers, total charge, and multiplicity of a molecular system. With this, CEH expands on a broad set of charge models developed in the past based both on SQM[404,451–453] and ML[454–457] models. Very recently, Pracht *et al.* have tested its performance for dipole moments and infrared intensities.[458] Investigations of the accuracy of atomic partial charges with more elaborated DFT and WFT methods have been conducted for a long time[319,459–462] and show that RSH-DFT in a large basis set is a suitable reference for (semi)empirical approaches[463]. However, a plethora of charge partitioning schemes exist that could in principle be used to generate reference charges from DFT calculations.[453,460,464–468] In the current work, we have chosen the Hirshfeld population analysis to assign atomic partial charges, because of its robust definition for large and diffuse basis sets.[469] In contrast, Mulliken population analysis[282] exhibits strong basis set dependence, limiting its application for larger than minimal basis sets. Further tests with the Minimal Basis Iterative Stockholder (MBIS) method[470], a variant of Hirshfeld analysis, yielded excessive polarization in certain (rather common) cases, such as in zwitterionic aminobutyric acid (see SI for further details). Such strong polarizations and in general missing robustness are unsuitable for the atom-in-molecule parameterization of basis functions pursued here. Moreover, the moderate magnitude of the Hirshfeld charges compared to those of other methods[469] is not crucial here, as their contribution to the change of the AOs is scaled linearly on an element-wise basis.

CEH atomic charges can be used in various applications, such as refining molecular cavities for solvation free energy calculations[471], enhancing the electrostatic energy in force fields[189] or non-SCF tight-binding schemes[279], and accounting for the charge dependence of the dispersion coefficients.[120,320] Moreover, CEH charges could serve as drop-in replacements in 4$^{\text{th}}$ generation neural network potentials, which often rely on classical charge equilibration schemes.[404,454,472] A focus of this work shall be the general-purpose application of CEH charges, independent of the q-vSZP basis set.

The paper is structured as follows: After revisiting the theory and detailing the Hamiltonian improvements, we describe the changes in both the CEH parameter fit and the q-vSZP optimization, as well as the extensions to newly considered elements for both parts. The quality of the CEH charges is assessed by comparison to corresponding DFT data across a large and chemically diverse set of molecules, including MLMs and actinide complexes. Lastly, an established range-separated density functional (ωB97X-D4[1,103,307,473]) is used

to benchmark the q-vSZP basis set on standard thermochemical data and realistic *f*-element applications.

F.3 Theory

F.3.1 Charge Extended Hückel Model

The CEH charge of atom A is defined by the standard Mulliken population analysis[282]

$$q_A^{\mathrm{CEH}} = Z_A - \sum_{\kappa \in A}^{\mathrm{AO}} (\mathbf{PS})_{\kappa\kappa}\,, \tag{F.1}$$

where Z_A denotes its effective core charge, $\mathbf{S}$ the overlap matrix, and $\mathbf{P}$ the density matrix. $\mathbf{P}$ is constructed from the MO coefficient matrix $\mathbf{C}$ obtained by a single solution of the Roothaan-Hall-type equation[127,128]

$$\mathbf{H}^{\mathrm{CEH}}\mathbf{C} = \mathbf{SC}\boldsymbol{\epsilon}. \tag{F.2}$$

$\mathbf{H}^{\mathrm{CEH}}$ is an extended Hückel-type[221] Hamiltonian, and $\boldsymbol{\epsilon}$ are its eigenvalues. The valence MOs are occupied based on the (restricted) Fermi-Dirac distribution at an elevated temperature to mimic the effect of static correlation in systems with a small fundamental gap. Atom-centered Slater-type orbitals with empirical exponents ζ^l approximated by Stewart's Gaussian expansions[145] are used as the valence minimal AO basis set. All elements use a consistent contraction depth of six primitives per AO. The Gaussian expansions for Fr, Ra, and the actinides (principal quantum number (PQN) 7) are unavailable and thus substituted by the corresponding PQN=6 AOs. For the Lns ($Z = 57 - 71$), *f*-electrons are not explicitly treated in the valence space (*f*-in-core approximation).

In a major improvement over the previous CEH version, we introduce a pseudo-charge-dependent (first-order) Hamiltonian contribution $\mathbf{H}^{1^*}$, which accounts explicitly for long-range electrostatic interactions:

$$\mathbf{H}^{\mathrm{CEH}} = \mathbf{H}^0 + \mathbf{H}^{1^*} \tag{F.3}$$

Long-range interactions are missing in the extended Hückel-type (zeroth-order) Hamiltonian $\mathbf{H}^0$ due to its exponential decay with the (diatomic frame scaled) atomic orbital overlap $\mathbf{S}^{\mathrm{sc}}$ (see the SI for details on the construction of $\mathbf{H}^0$). To avoid the self-consistent solution of equation F.2 necessary with an explicitly charge-dependent SQM Hamiltonian, $\mathbf{H}^{1^*}$ uses a newly defined purely geometry-dependent approximation for the local atomic charges q^{loc}. For this, we distribute the total molecular charge Q_{tot} equally to all atoms in the molecule and further assign partial charges based on the relative electronegativities χ. As a measure, we use the χ-weighted coordination number CN^{χ}, which is already part of the previous CEH Hamiltonian:[3]

$$q_A^{\mathrm{loc}} = k_{q,A} \cdot \mathrm{CN}^{\chi} + k_{\mathrm{tot}} \cdot \frac{Q_{\mathrm{tot}}}{n_{\mathrm{at}}} \tag{F.4}$$

Here, $k_{q,A}$ and k_{tot} are element-wise and global parameters, respectively, and n_{at} is the number of atoms. Note that the sum of the local charges does not necessarily conserve the overall molecular charge. Instead, it aims to reproduce only the relative magnitude of atomic charges in a local environment. While this is a crude approximation, its local nature has the benefit that no artificial charge transfer can occur, which plagues more sophisticated electronegativity-based charge equilibration models.[3,120] With q^{loc}, the standard expressions for

the atom-resolved second- and onsite third-order tight-binding define the matrix elements of $\mathbf{H}^{1^*}$:

$$H^{1^*}_{\kappa\lambda} = -\frac{1}{2}\left(\sum_{C}^{n_{\text{at}}} q_C^{\text{loc}}\left(\gamma_{AC}+\gamma_{BC}\right) + \Gamma_A (q_A^{\text{loc}})^2 + \Gamma_B (q_B^{\text{loc}})^2\right) S_{\kappa\lambda} \qquad \forall \kappa \in A, \lambda \in B \tag{F.5}$$

The distance dependence of the second-order term is described by the standard, screened Coulomb-type Mataga-Nishimoto-Ohno-Klopman formula[226,280,474]

$$\gamma_{AB} = \frac{1}{R_{AB} + \eta^{-1}} \tag{F.6}$$

with the interatomic distance R_{AB} and the arithmetic average η of the Hubbard parameters U_A and U_B. U_A and its derivative Γ_A are treated as element-wise fit parameters. For optimization of the 15 (*sp*-block element), 19 (*spd*-block element), or 23 (*spdf*-block element) element-wise empirical parameters and 7 global parameters employed in the CEH model, see the following chapters.

F.3.2 The q-vSZP Basis

Environment Dependency of the q-vSZP Basis Functions

The q-vSZP basis set, as described in Ref. [3], operates on the simple principle that the contraction coefficients of the primitive Gaussian functions depend on the chemical environment of each atom. Thus, every symmetry-unique atom in a molecule obtains an individual basis, enabling the AOs to "breathe". Since this fundamental concept remains unchanged in the revised version presented here, only a brief overview of the theory will be provided (for a comprehensive overview, refer to Ref. [3]).

Eq. F.7 defines a contracted Gaussian-type AO ϕ_{κ,A_i} as a linear combination of N_{pr} primitive Gaussian functions $\chi_{\lambda,A}$ with the contraction coefficients c_{λ,A_i}. Here, A_i is a symmetry-unique atom of element A.

$$\phi_{\kappa,A_i}(\zeta_\kappa, q^{\text{eff}}_{A_i}) = \sum_{\lambda\in\kappa}^{N_{\text{pr}}} c_{\lambda,A_i}(q^{\text{eff}}_{A_i})\,\chi_{\lambda,A}(\zeta_{\lambda,A}) \quad \text{with} \quad (A_i \in A)\,. \tag{F.7}$$

The radial extent of the AO can be changed by linear variation of the contribution of each primitive function in the deep contraction. Note that the exponents $\zeta_{\lambda,A}$ of each primitive Gauss function $\chi_{\lambda,A}$ remain unchanged between atoms of the same element. The dependency of the contraction coefficients c_{λ,A_i} of atom A_i on its effective atomic charge $q^{\text{eff}}_{A_i}$ is linear:

$$c_{\lambda,A_i}(q^{\text{eff}}_{A_i}) = c_{0,\lambda,A} + c_{1,\lambda,A}\cdot q^{\text{eff}}_{A_i}, \tag{F.8}$$

where $c_{0,\lambda,A}$ is the base (free atom) contraction coefficient and $c_{1,\lambda,A}$ is the linear coefficient for the primitive function $\chi_{\lambda,A}$. The effective charge descriptor $q^{\text{eff}}_{A_i}$ consists of atomic charge (linearly and quadratically) and CN-dependent parts as well as an additional cross-term according to

$$q^{\text{eff}}_{A_i}(q_{A_i}, \text{CN}_{A_i}) = q^{\text{CEH}}_{A_i} + k_{1,A}\cdot\left(q^{\text{CEH}}_{A_i}\right)^2 + k_{2,A}\cdot\sqrt{\text{CN}_{A_i}} + k_{3,A}\cdot\text{CN}_{A_i}\cdot q^{\text{CEH}}_{A_i}. \tag{F.9}$$

The square root of the CN is taken in the third term to damp large values appearing in dense metal-containing systems. $k_{x,A}$ denote three element-specific prefactors that were variationally optimized together with the coefficients $c_{0,\lambda,A}$ and $c_{1,\lambda,A}$. The atomic charges $q^{\text{CEH}}_{A_i}$ are obtained from the CEH model, and CN is the

error function-modified coordination number from Ref. [2]. Thus, the effective contraction coefficient c_{λ,A_i} reduces to $c_{0,\lambda,A}$ for a neutral atom ($q_{A_i}^{\mathrm{CEH}} = 0$, $\mathrm{CN}_{A_i} = 0$).

The development of the analytical gradient in the context of DFT and SQM calculations for the now implicitly nuclear position-dependent basis set contraction (including the necessary $\frac{\partial q^{\mathrm{CEH}}}{\partial R}$ analytical derivative) is currently underway in our laboratory and will be described separately.

Basis Set Construction and Extension for Lanthanides, Francium, Radium, and the Actinides

The main design principles of the first q-vSZP version were maintained:

- Deep contraction (a large number of primitive functions) of each AO. This generally reduces BSIEs despite having only one basis function per AO shell, i.e., a MB. Additionally, a large number of primitive functions allows for greater flexibility within the concept of adaptive primitive coefficients.
- Variational optimization of all basis set parameters in Eqs. F.7, F.8, and F.9 in both atomic *and* molecular calculations. See the Technical Details section below for an in-depth explanation of the optimization process, and Refs. [1, 3] for a more extensive perspective on the background of molecule-optimized basis sets.
- Consistent large-core ECPs leading to a valence-only basis set. This makes the q-vSZP basis compatible with typical SQM methods that usually consider valence electrons only.

While the fundamental construction principle of deeply contracted basis functions remains the same, the specific contraction scheme varies for some elements in the revised version (see Tab. F.1, changes in **bold** letters). Most importantly, this work presents an extension for the lanthanides (Z=58–71), francium, radium, and the actinides (89–103) (see Tab. F.2), which have gained increasing importance also in theoretical studies in recent years.[477] The third design principle of a valence-only basis set is mostly maintained; however, the q-vSZP basis for the added elements relies on standard small-core ECPs instead of large-core ECPs. This decision was made because robust large-core ECPs for these elements were unavailable, according to some initial tests. Instead, we chose standard Stuttgart-Cologne small-core ECPs[140,142,475,476] and added semi-core basis functions for the remaining electrons aside from the valence shells. These semi-core functions emulate a larger-core ECP but are not adaptive to the environment, i.e., the second set of coefficients $c_{1,\lambda,A}$ is zero for all non-valence functions. This allows for the optimization of a valence basis set without the need for matching ECPs and reduces the number of parameters in the basis set optimization. In novel SQM methods, the semi-core basis functions can be discarded, resulting in a consistent adaptive q-vSZP basis set for the valence region. Similar to the first q-vSZP version, polarization functions on first- and second-row elements might also be dropped for specific SQM methods to reduce the number of parameters and the computational effort.

F.4 Technical Details

F.4.1 General

† Modified d-f projector for B and F, see Ref. [1] for details.

‡ The ECP was taken from the TURBOMOLE[294,295] basis set library, with the *h* projector being removed from the published ECP.

Table F.1: Primitive contraction schemes and ECPs of the q-vSZP basis set for all elements up to radon (except for the Lns). The contraction scheme corresponds always to (*x*s*y*p*z*d) → [1s1p1d] (for H and He: [1s1p]). Columns three and four indicate the type of effective core potential used and the corresponding number of core electrons. Changes with respect to the first q-vSZP version are denoted in bold letters.

element	prim. functs.	ECP type	# core el.
H	(8s3p)	-	-
He	**(8s2p)**	-	-
Li	**(5s5p2d)**	ECP-2-SDF[425]	2
Be	**(6s4p2d)**	ECP-2-SDF[425]	2
B	(6s5p2d)	ECP-2-MWB[141†]	2
C–O	(6s6p3d)	ECP-2-MWB[141]	2
F,Ne	(6s6p2d)	ECP-2-MWB[141†]	2
Na	(4s4p2d)	ECP-10-SDF[425]	10
Mg	(4s3p2d)	ECP-10-SDF[425]	10
Al	(5s4p2d)	ECP-10-MWB[141]	10
Si	**(5s4p2d)**	ECP-10-MWB[141]	10
P–Ar	**(5s5p2d)**	ECP-10-MWB[141]	10
K	(4s3p2d)	ECP-18-SDF[425]	18
Ca	**(4s3p3d)**	ECP-18-SDF[425]	18
Sc	**(5s2p6d)**	CRENBS[348]	18
Ti-Cu	(5s2p6d)	CRENBS[348]	18
Zn	(5s3p2d)	ECP-28-SDF[426]	28
Ga	**(5s4p2d)**	ECP-28-MWB[141,349]	28
Ge-Kr	**(6s5p2d)**	ECP-28-MWB[141,349]	28
Rb	(4s3p2d)	ECP-36-SDF[425,427]	36
Sr	**(4s3p3d)**	ECP-36-SDF[425,427]	36
Y	**(5s2p6d)**	CRENBS[144]	36
Zr–Mo,Ru–Ag	(5s3p6d)	CRENBS[144]	36
Tc	(5s2p6d)	CRENBS[144]	36
Cd	(5s3p2d)	ECP-46-SDF[426]	46
In	(5s4p2d)	ECP-46-MWB[141,349]	46
Sn–Xe	(6s5p2d)	ECP-46-MWB[141,349]	46
Cs	(4s3p2d)	ECP-54-SDF[427]	54
Ba	**(4s3p3d)**	ECP-54-SDF[427]	54
La	(4s2p5d)	CRENBS[351]	54
Hf	(4s3p5d)	CRENBS[351]	68
Ta–Au	(5s3p5d)	CRENBS[351]	68
Hg	(5s3p2d)	ECP-78-SDF[428]	78
Tl	**(5s4p2d)**	ECP-78-MWB[354]	78
Pb	(5s5p2d)	ECP-78-MWB[354]	78
Bi–Rn	(6s5p2d)	ECP-78-MWB[354]	78

Table F.2: Contraction schemes and ECPs of the q-vSZP basis set for the Lns, Fr, Ra, and the actinides (Ans). Columns five and six indicate the type of effective core potential used and the corresponding number of core electrons.

element	contraction scheme			ECP type	# core el.
Ce,Pm	(23s19p14d7f)	[3s3p2d1f]	{8/8 8/7/7 7/4/7/7}	ECP-28-MWB[140]	28
Pr,Nd,Dy,Er	(23s19p13d7f)	[3s3p2d1f]	{8/8 8/7/7 7/4/6/7}	ECP-28-MWB[140]	28
Sm,Eu	(23s18p12d7f)	[3s3p2d1f]	{8/7 8/7/6 7/4/6/7}	ECP-28-MWB[140]	28
Gd,Ho,Yb	(23s18p13d7f)	[3s3p2d1f]	{8/7 8/7/7 7/4/6/7}	ECP-28-MWB[140]	28
Tb	(23s18p14d7f)	[3s3p2d1f]	{8/7 8/7/7 7/4/7/7}	ECP-28-MWB[140]	28
Tm	(23s19p12d7f)	[3s3p2d1f]	{8/8 8/7/6 7/4/6/7}	ECP-28-MWB[140]	28
Lu	(23s18p13d7f)	[3s3p2d1f]	{8/8 8/6/7 7/4/6/7}	ECP-28-MWB[140]	28
Fr	(18s18p4d)	[2s2p1d]	{9/9 9/9/4}	ECP-78-MDF[475‡]	78
Ra	(18s18p4d)	[2s2p1d]	{9/9 9/9/4}	ECP-78-MDF[476‡]	78
Ac,Pa	(22s19p12d7f)	[3s3p2d1f]	{8/8 7/7/6 7/4/6/7}	ECP-60-MWB[142]	60
Th	(23s19p12d7f)	[3s3p2d1f]	{8/8 8/7/6 7/4/6/7}	ECP-60-MWB[142]	60
U–Am	(23s18p13d7f)	[3s3p2d1f]	{8/7 8/7/7 7/4/6/7}	ECP-60-MWB[142]	60
Cm–Es,Md,No	(23s17p13d7f)	[3s3p2d1f]	{8/7 8/6/7 7/4/6/7}	ECP-60-MWB[142]	60
Fm,Lr	(22s17p13d7f)	[3s3p2d1f]	{8/7 7/6/7 7/4/6/7}	ECP-60-MWB[142]	60

The CEH charge model is implemented in the `tblite` program package, accessible via the corresponding GitHub repository (github.com/tblite/tblite).

All DFT calculations for this work were conducted with the ORCA quantum chemistry package in version 5.0.4 or 6.0.0.[309,357] If not stated otherwise, the DFT functional for all q-vSZP comparisons is ωB97X-D4[1,103,307,320]. The RI approximation for the electronic Coulomb energy with the large universal auxiliary basis sets by Weigend[358] as well as the seminumerical COSX[178,179] were used throughout. For numerical derivatives, high convergence thresholds (keyword: `VeryTightSCF`) and grid settings (keyword: `DefGrid3`) were applied to exclude numerical noise to the extent possible. The D4[119,120,320] parameters employed with the (revised) q-vSZP basis set were taken from the original publication in Ref.[3].

Like in the first version, geometry optimizations are conducted with numerical energy derivatives, fully adapting the basis set for each distorted geometry. All numerical derivatives with respect to nuclear coordinates or external electric fields were conducted with the help of an in-house Python code, which is publicly available within the corresponding GitHub repository[429] (github.com/grimme-lab/NumgradPy). Geometry optimizations were conducted with the ANCopt optimizer in `xtb`[259,278,478], driving the single-point and gradient calculations by `ORCA`. The program for automatic generation of the `ORCA` input files with the atom-in-molecule adaptive basis set is publicly accessible on GitHub (github.com/grimme-lab/qvSZP).[430] It can easily be modified to generate input files for other quantum chemistry packages as well.

F.4.2 Optimization of CEH Parameters

The parameters of the revised CEH model presented in this work were determined by a least-squares fit to Hirshfeld[317] charges at the ωB97M-V[104]/def2-TZVPPD[149] (def-TZVP[479] for actinides) level of theory. Like in the first version, dipole moments and bond orders were additionally included in the parameter fit to ensure robustness and stability. A new ingredient to the fit is the size of the fundamental gap, which is crucial in light of the substantially increased electronic temperature (4000 K compared to 300 K in the CEH-v1[3]). While the increased temperature improves electronically complicated systems, a small fundamental gap can result in artificial CT. Since CEH contains no non-local exchange to open the HOMO-LUMO gap naturally, a dynamic penalty is applied to the gap. For complicated systems (e.g., open-shell radicals), where the reference method has a small gap, a low weight in the fit-loss function is placed on the gap, while for simple systems, the large

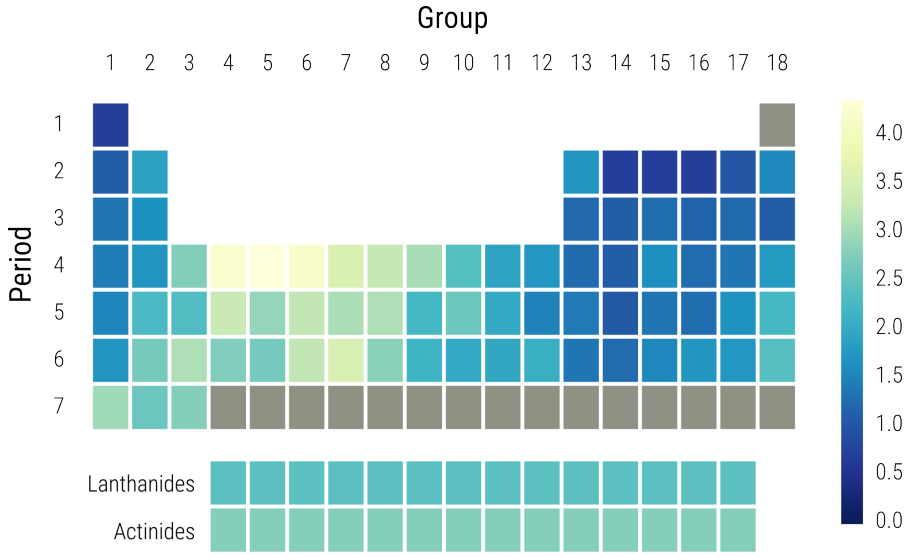

Figure F.2: Relative final RMSEs of the parameter fit in color-coded form over the whole periodic table. The RMSEs have been evaluated against Hirshfeld[317] charges, molecular dipole moments, Wiberg bond orders[480,481], and fundamental gaps (see main text) at the ωB97M-V[104]/def2-TZVPPD[149] (def-TZVP[479] for actinides) level of theory. The parameters for the elements "HCNO" as well as for the Lns and Ans were obtained in one common parameter fit.

gap has to be reproduced more closely. Beyond atomic charges q, which remain the primary target property, shell populations and dipole moments were also utilized as training data in order to obtain a physically well founded model. For the parameter fit of the revised CEH version presented here, a stronger focus has been put on electronically complex situations: "Mindless" molecules (see the corresponding Results section and Ref. [301] and [2] for details), systems with high absolute charges of up to +4/-4, and open-shell structures with unusual spin and oxidation states.

To determine the relative accuracy between different elements, Fig. F.2 compares the final RMSEs of the parameter fit throughout the whole periodic table. Since the parameters for the elements H, C, N, and O ("HCNO"), the lanthanides, and the actinides were obtained in one common parameter fit each, they appear as a single color in the figure. The relative accuracy of CEH with respect to the DFT reference corresponds to chemical intuition: Non-metallic elements that form defined covalent bonds are easier to describe than metals and heavier atoms. Still, for early *3d* transition metals, the fit RMSE is of the same order of magnitude, being only 4-5 times higher than for HCNO. This is in line with the findings during the development of our latest tight-binding method, PTB,[2] and thus underlines the robustness and consistency of the CEH model.

F.4.3 Optimization of the q-vSZP Basis Set

The parameters $\zeta_{\lambda,A}$, $c_{0,\lambda,A}$, $c_{1,\lambda,A}$, $k_{1,A}$, $k_{2,A}$, and $k_{3,A}$ were optimized variationally, similarly to the first version of q-vSZP. See Ref. [3] for details of the optimization process. An important difference from the first version is the use of reference DFT charges at the ωB97M-V[104]/def2-TZVPPD[149] Hirshfeld[317] level for optimizing the basis set instead of the CEH charges used during application. This has two advantages: First, the optimization process becomes independent of changes in the CEH model, e.g., avoiding re-parameterizations due to changed CEH parameters. Second, since the basis set parameters were optimized with reference DFT charges, rare cases where the CEH charges are qualitatively incorrect do not affect the basis set optimization.

Table F.3: Statistical metrics for the evaluation of atomic partial charges of randomly selected molecules from the PubChem database with respect to DFT reference charges (see the main text for technical details). Outliers ($|q_{i,j} - q_{i,\mathrm{DFT}}| > 1.0$) were not taken into account. Table entries that correspond to the least accurate result are given in italic letters, bold letters denote the most accurate results.

Method	PCC	SRCC	MAE	RMSE	# outliers
EEQ	0.879	0.873	0.089	0.109	1
EspalomaCharge	*0.788*	*0.743*	*0.194*	*0.251*	*92*
CEH-v1	0.961	0.922	0.024	0.036	1
CEH-v2	0.963	0.895	**0.019**	**0.027**	1
GFN1-xTB	0.947	0.919	0.063	0.110	1
GFN2-xTB	**0.971**	**0.945**	0.037	0.063	1

The optimization of the q-vSZP for the new elements (Lns, Fr, Ra, and Ans) requires a good choice of the initial starting point for the basis set optimization. Since small-core ECPs are used, a set of orthogonal coefficients is constructed from an initial fully decontracted calculation of the atom. During the optimization, the orthogonality between the semi-core (later excluded in the SQM development) and valence basis functions is enforced by a penalty calculated from their overlap. The new basis set, along with the ECPs used, is provided in the SI and within the program in Ref. [430], enabling single-point energy calculations with ORCA[309,357].

F.5 Results and Discussion

The main goal of this work is to provide a consistent and reliable charge model as well as an adaptive minimal basis set for the whole periodic table ($Z \leq 103$). A particular focus lies on the behavior of the newly included *f*-elements. In the following, CEH-v2 and q-vSZP-v2 denote the updated version of the original CEH and q-vSZP (denoted CEH-v1 throughout the manuscript, respectively).

F.5.1 Atomic Partial Charges

Randomly Selected Molecules from the PubChem Database

Realistic test cases outside the fitting regime are essential to validate the robustness and accuracy of novel methods. This applies not only to modern ML-FFs[192] but also to SQM methods, such as the CEH model. Here, we chose to fetch a test set of random molecules from the PubChem database[415]; see also Ref. [2] and [482] for further examples. The random selection of the molecules and preparation of their 3D structure is conducted via two Python scripts that are publicly available[483,484]: First, random numbers between 1 and 10^7 serving as PubChem Compound Identification integers for unique chemical structures are generated. After downloading the structures, they are converted from 2D to 3D, if necessary. Lastly, the geometries are optimized on the GFN2-xTB level of theory.[278] The final set of 916 random molecules and a detailed register of its composition is included in the SI.

For comparison, alternative methods for generating atomic partial charges were tested on the same dataset. In addition to the classical EEQ[120,402–404] and SQM methods – GFN1-xTB[277] and GFN2-xTB[278] – a hybrid physical/artificial neural network (ANN)-based approximation to AM1-BCC[451] charges, called EspaloMaCharge[455], was evaluated. For this method, only neutral molecules were included, as specifying the

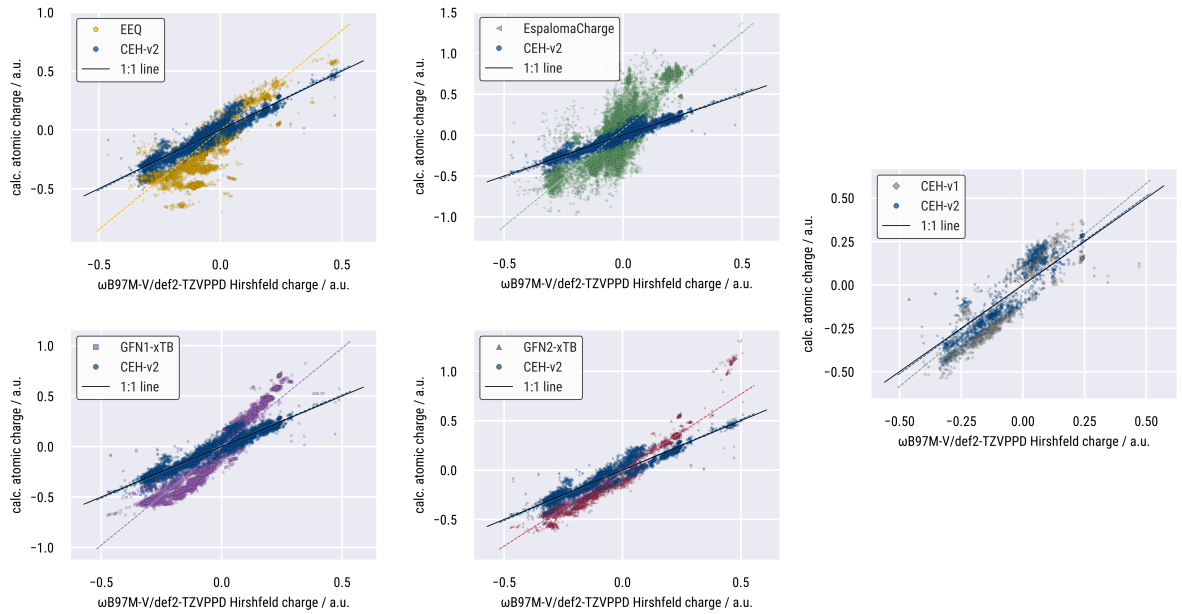

Figure F.3: Comparison of the correlation of atomic partial charges calculated with EEQ[120] (yellow), EspalomaCharge[455] (green), GFN1-xTB[277] (purple), and GFN2-xTB[278] (red) versus CEH-v2 (blue) with DFT Hirshfeld[317] charges at the ωB97M-V[104]/def2-TZVPPD[149] level of theory. Additionally, the first CEH version ("CEH-v1") is also included for comparison. The raw data set comprises 916 molecules with 43240 atomic charge data points i. Outliers ($|q_{i,j} - q_{i,\mathrm{DFT}}| > 1.0$) as well as hydrogen charges were omitted. The dashed lines correspond to a linear regression of the data. For the scatter plot, data points for which all compared methods provided charges within a range of 0.1 were not included graphically to improve readability, by avoiding a large number of points very close to the 1:1 line. Thus, a data point is only plotted if the following condition holds: $\max |q_{i,j} - q_{i,k}| \geq 0.1$, with j and k being the different charge predictions in the plot.

formal charge was not possible. Additionally, molecules that encountered technical errors with EspalomaCharge were excluded from the comparison. Throughout the study, Hirshfeld[317] charges obtained from ωB97M-V[104]/def2-TZVPPD[149] calculations were used as the reference.

The correlation of atomic partial charges with the reference DFT charges is visualized in Fig. F.3. To simplify the visualization, each figure contains the CEH-v2 partial charges and those of one of the competing methods listed above. In addition, statistical metrics for the accuracy of the partial charges are provided in Tab. F.3. The CEH-v2 model (blue) clearly outperforms the EEQ model (yellow) on the test data. Since both models (EEQ as implemented in Ref. [120]) were parameterized on DFT/triple-ζ Hirshfeld charges, the better performance of CEH-v2 can be purely assigned to its more sophisticated construction. However, it should be noted that EEQ is faster by a factor of about ten in medium-sized systems as it requires only the solution of a linear system of equations in the number of atoms compared to the diagonalization in the number of AOs for CEH. The preceding version, CEH-v1, is only moderately less precise, which is unsurprising given that the objective of the revised version was to enhance robustness for challenging cases and expand applicability rather than improving the accuracy.

The hybrid physical/ANN approach "EspalomaCharge" yields only mediocre accuracy on the test data with a significantly lower Pearson correlation coefficient (PCC) of about 0.79 compared to 0.96 with both versions of the CEH model. In part, this can be attributed to the different targets of emulating AM1-BCC

and not Hirshfeld[317] charges. However, Spearman's rank correlation coefficient (SRCC), which should be less susceptible to absolute errors or shifts, is also lower by a similar degree (0.74 vs 0.90). Furthermore, EspalomaCharge yields 87 outliers excluded from the statistical analysis and the plot, while CEH-v2 led to only one outlier. In contrast, the SQM methods GFN1- (purple) and GFN2-xTB (red) exhibit results comparable to CEH-v2. While the CEH-v2 MAE and RMSE are lower by a factor of about two to three compared to the xTB methods, the SRCC is slightly closer to unity for both GFN1- and GFN2-xTB. This aligns with the difference in construction between xTB and CEH: The self-consistent xTB charges lead to more consistent trends, while CEH yields lower absolute errors due to the direct fit on Hirshfeld charges. An observation that holds for all compared methods (yet less pronounced for GFN2-xTB) is the overestimation of absolute charges (for both positive and negative charges) compared to the DFT Hirshfeld charges, which is largely the reason for the higher absolute errors. However, this might also be due to the differences in charge partitioning, as xTB provides Mulliken charges which are generally larger than Hirshfeld charges.

Actinoid Complexes

While for organic molecules consisting of elements $Z \leq 18$, many different methods for atomic partial charge assignment are available, this is not the case for the An series of elements, i.e., for $89 \leq Z \leq 103$. Especially empirical approaches such as charge equilibration, FF, and SQM methods are rare due to the complex electronic structure and high effort of individually parameterizing each element. Recently, the EEQ model used for the D4 dispersion correction has been extended to the An series based on a large set of automatically generated An complexes ("AcQM" data set, see Ref. [320]), coming with ωB97M-V[104]/ma-def-TZVP[434] (def2-TZVPP[139] for non-An atoms) Hirshfeld[317] charges. Out of the 2531 structures, 1282 are realistic An complexes, and 1249 are small so-called "mindless" complexes. The latter were generated by random placement of 1-8 atoms around a single An atom and subsequent geometry optimization (see Ref. [320] for a more detailed explanation). Since the CEH-v2 model is also available for all elements $Z \leq 103$, the accuracy of atomic charges of both methods in comparison to reference DFT values is assessed on AcQM in Fig. F.4, with the statistical evaluation in Tab. F.4. For all structures, only the charge of the An atom was taken into account.

Both methods agree well with the reference charges, yielding almost identical MAEs and RMSEs of 0.2 and 0.3 e^-, respectively. This is particularly noteworthy as the classical EEQ model does not treat the electronic structure but only distributes charges over the molecule based on electronegativities and three further empirical parameters. In contrast, the CEH-v2 model explicitly treats the valence electrons (including the *f*-shell), which is expected to improve the consistency of the atomic charges. This is also observed for the PCCs and SRCCs, which significantly improve with CEH-v2. Visual inspection of the regression shows that CEH-v2 provides almost no systematic error, whereas EEQ underestimates the absolute value of the charges. The latter is not in agreement with the previous analysis on randomly selected organic molecules, for which EEQ overestimated the magnitude of the partial charges. The number of outliers with $|q_{i,j} - q_{i,\mathrm{DFT}}| > 1.0$ is slightly higher for CEH-v2 compared to the EEQ model. The reason for the distinctly higher CEH-v2 error measures for the actinides in comparison to the randomly selected molecules might be both *i)* the fact that only charges of the electronically more complex actinide atoms were considered and *ii)* the unusual and partially artificial multiplicity and oxidation state assignments in the AcQM data set. The performance for the lanthanide series has been assessed in the original publication of the CEH model and is similar to the actinides.

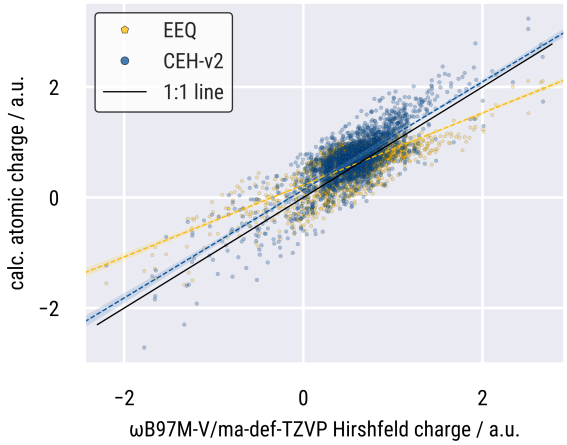

Figure F.4: Comparison of the correlation of atomic partial charges calculated with EEQ[120,320] (yellow) versus CEH-v2 (blue) with DFT Hirshfeld charges at the ωB97M-V[104]/ma-def-TZVP level of theory. Only the actinide charges i (i.e., charges for atoms with $Z > 88$) were taken into account. Outliers ($|q_{i,j} - q_{i,\mathrm{DFT}}| > 1.0$) were omitted. The dashed lines correspond to a linear regression of the data. For the scatter plot, data points for which all compared methods provided charges within a range of 0.1 were not included graphically to improve readability, by avoiding a large number of points very close to the 1:1 line. Thus, a data point is only plotted if the following condition holds: $\max |q_{i,j} - q_{i,k}| \geq 0.1$, with j and k being the different charge predictions in the plot.

Table F.4: Statistical metrics for the evaluation of atomic partial charges of actinide complexes with respect to DFT reference charges (see the main text for technical details). Outliers ($|q_{i,j} - q_{i,\mathrm{DFT}}| > 1.0$) were not taken into account.

Method	PCC	SRCC	MAE	RMSE	# outliers
EEQ	0.764	0.637	0.231	0.289	32
CEH-v2	0.848	0.738	0.241	0.308	48

"Mindless" Molecules

A recently revived approach for evaluating the performance for previously unseen systems involves the use of so-called "mindless" molecules (MLMs).[2,301,485] To facilitate their routine generation, we developed a new Python package, "MindlessGen," which enables the efficient and highly customizable creation of MLMs.[486] Further details on the software and in-depth analyses of these molecules will be presented elsewhere.

We generated a small set of five MLMs with diverse atomic compositions and compared the atomic partial charges of selected atoms derived from EEQ and CEH-v2 with reference DFT charges at the ωB97M-V/def2-TZVPPD level. All molecules are of closed-shell character and carry a net charge ranging from -2 to 2. Fig. F.5 illustrates the distinct charges for each molecule and selected atom. Consistent with the previous analyses on randomly selected mostly organic molecules, the EEQ model overestimates the charges. However, it is noteworthy that the classical charge equilibration method successfully captures basic trends and the correct sign of the atomic partial charge. In the examined cases, CEH-v2 yields absolute errors ranging from 0.02 to 0.17 e^-, which exceed the statistical error measures for the chemically more meaningful, randomly selected molecules but remain within a comparable range. This underscores the value of using a chemically diverse training set to improve the method's robustness.

The trend observed in Fig. F.5 is further supported by the statistical comparison of 100 distinct MLMs in

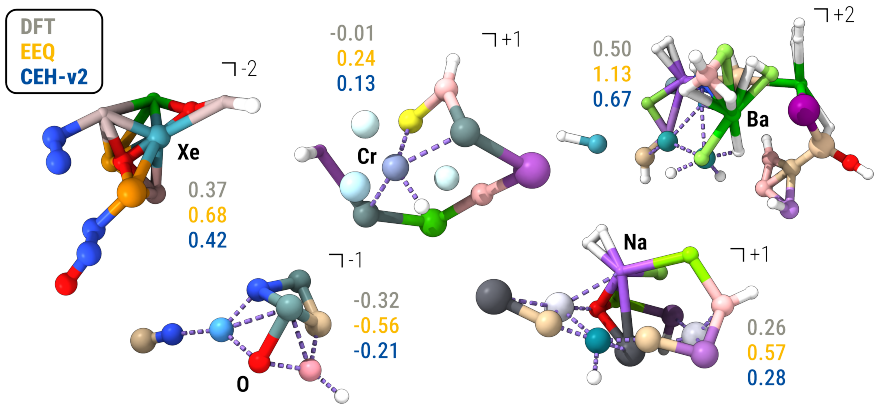

Figure F.5: "Mindless" molecules with atomic partial charges for selected atoms, generated with ωB97M-V/def2-TZVPPD ("DFT"), EEQ, and CEH-v2. The black element symbol indicates the atom for which the partial charge is examined. Molecular structures and atomic partial charges for all molecules are contained in the SI.

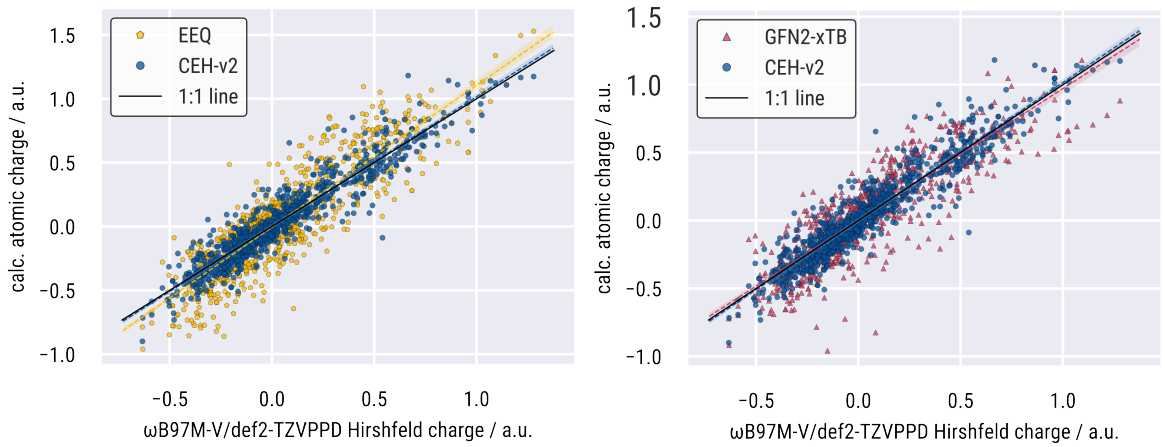

Figure F.6: Comparison of the correlation of atomic partial charges calculated with EEQ[120] (yellow) and GFN2-xTB[278] (red) versus CEH-v2 (blue) with DFT Hirshfeld[317] charges at the ωB97M-V[104]/def2-TZVPPD[149] level of theory. The dashed lines correspond to a linear regression of the data. The raw data set comprises 100 MLMs with 944 atomic charge data points. For a single molecule, the GFN2-xTB SCF did not converge and the corresponding GFN2-xTB charge entries were excluded. Except for these charges, no data points were omitted. The `mindlessgen.toml` file defining the generated molecules is provided in the SI.

Table F.5: Statistical metrics for atomic partial charges of 100 MLMs with respect to DFT reference charges (see the main text and the SI for technical details).

Method	PCC	SRCC	MAE	RMSE
EEQ	0.876	0.857	0.150	0.192
GFN2-xTB	0.874	0.881	0.124	0.167
CEH-v2	0.946	0.942	0.077	0.108

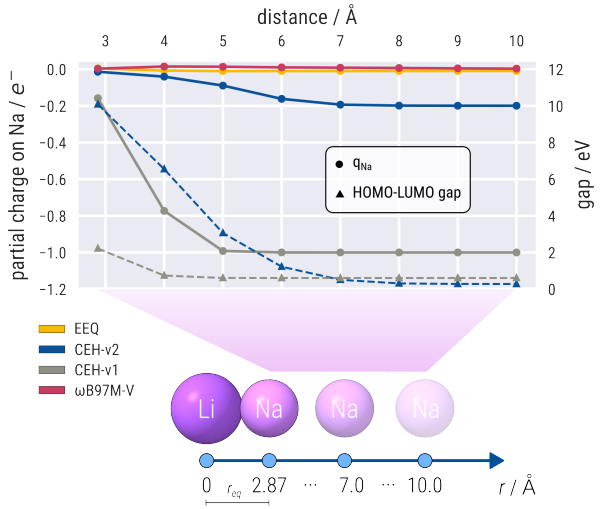

Figure F.7: Atomic partial charge of sodium in LiNa depending on the inter-atomic distance. The equilibrium distance of 2.87 Å was determined by optimization with ωB97X-3c[1]. For ωB97M-V[104], the def2-TZVPPD basis set[149] was employed.

Fig. F.6 (technical details are given in the SI). In addition to the EEQ model, we compared the results with GFN2-xTB. The CEH-v2 model produces significantly fewer data points with large deviations from the 1:1 line than both EEQ and GFN2-xTB. This is reflected in the correlation measures with respect to the reference charges listed in Tab. F.5: EEQ and GFN2-xTB exhibit almost identical SRCC and PCC values, which are distinctly lower than with CEH-v2. Notably, the simpler EEQ model comes close to the performance of the self-consistent GFN2-xTB. The same trend is observed for the absolute deviations measured by the MAE which is almost halved compared to the other methods. Further statistical measures for GFN1-xTB and CEH-v1 and technical details are given in the SI.

Systems Susceptible to Artificial Charge Transfer

Artificial CT is a general problem of empirical charge models and even low-level DFT calculations.[70,487] The development of the CEH model was even prompted initially by artificial CT observed with the EEQ model.[3] Here, the dissociation of non-covalently bound subsystems with a large difference in electronegativity would lead to significant CT (i.e., CH_4 and O_2 transfer $0.17e^-$ if calculated as a complex). Analogously, ammonium fluoride does not dissociate heterolytically with EEQ ($q_F = -0.33$ at a distance of 6 Å). CEH alleviates this problem by introducing the energy levels of a Hamiltonian populated by the Aufbau principle. Thus, both CEH and DFT provide an almost integer negative charge on the fluorine atom in NH_4F at the same distance. However, the explicit electronic structure poses a new challenge if the electronic levels are energetically close, which is common during homolytic bond cleavage. In QM methods, this situation can lead to convergence failures in SCF-based approaches and artificial CT to the lower energy level in single-diagonalization approaches.

During the development of a new TB method (working title: g-xTB), we observed the latter issue in CEH-v1: For a diatomic system like LiNa, artificial CT occurs from the slightly higher lying Li valence 2s level to the sodium, leading to the unphysical heterolytic dissociation into Li^+ and Na^- above 5.0 Å. As apparent from the gray lines in Fig. F.7, the increasing CT (solid line) runs in parallel with the quickly

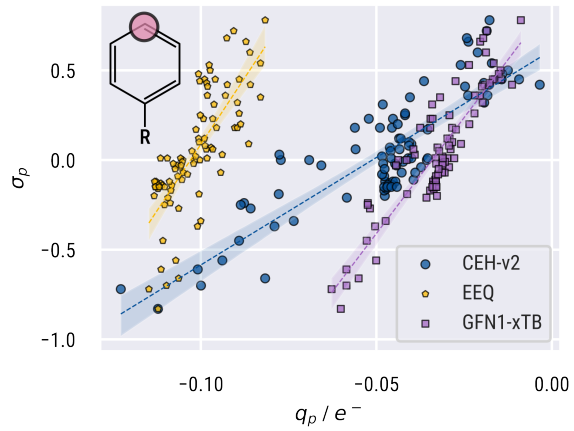

Figure F.8: Correlation of experimental Hammett parameters for para-substitution (σ_p) from Ref. [488] and [489] with CEH-v2, EEQ, and GFN1-xTB atomic charges of the carbon atoms (q_p) in *para*-position. The dashed line represents a linear regression fit with a confidence interval of 95 % in light gray.

diminishing fundamental gap (dashed line) at larger distances (0.02 eV above 5 Å). One approach to mitigate this issue in a single-diagonalization approach is an increased electronic temperature, which "smears out" the occupation over the energy levels. For close-lying levels, Fermi statistics at a sufficiently high electronic temperature (4000 K for CEH-v2) yields an almost equal population of both levels. The increased electronic temperature in CEH-v2 (blue lines) limits the CT to $< 0.2e^-$ in the dissociation limit, even though the HOMO-LUMO gap is even smaller than in CEH-v1 (0.01 eV at 10 Å). Beyond the electronic temperature, CEH-v2 uses an explicit electrostatics term and a slightly more diffuse basis set to delay the decay of the off-diagonal Hamiltonian matrix elements. This leads to a larger and more slowly decaying fundamental gap and also a limitation of artificial CT at shorter distances ($< 0.1e^-$ below 5Å). As expected, both EEQ and ωB97M-V correctly dissociate also with an equal population at lithium and sodium.

Prediction of Experimental Hammett Parameters with CEH Charges

Linear free energy relationships (LFERs) have long been a cornerstone of physical organic chemistry, providing a quantitative framework to relate substituent electronic effects with chemical reactivity. These relationships, particularly in the context of the Hammett equation, allow for the correlation of reaction rates and equilibria with substituent constants, providing insights into mechanistic pathways. Traditionally, these substituent constants (Hammett parameters such as σ_p) have been derived from empirical data, yet modern computational methods offer an alternative path to capture these effects. In recent work, Luchini and Paton benchmarked the correlation of various computational charge schemes and other computationally derived properties with experimental Hammett parameters.[488] In Fig. F.8, we employ CEH-v2 atomic charges to demonstrate its capability to predict substituent effects on reactivity in aromatic systems using the same experimental data.

The squared PCC R^2 amounts to 0.73, which is very close to the published results for DFT Hirshfeld[317] and GFN1-xTB[277] Mulliken[282] charges of 0.92 and 0.79, respectively. EEQ[120,404] yields a distinctly lower R^2 of 0.57. The robust performance of CEH-v2 compared to related yet computationally more involved methods is particularly remarkable since very small charge deviations of $< 0.05\ e^-$ determine the quality of the correlation. These subtle relative differences are generally easier to capture with self-consistent methods

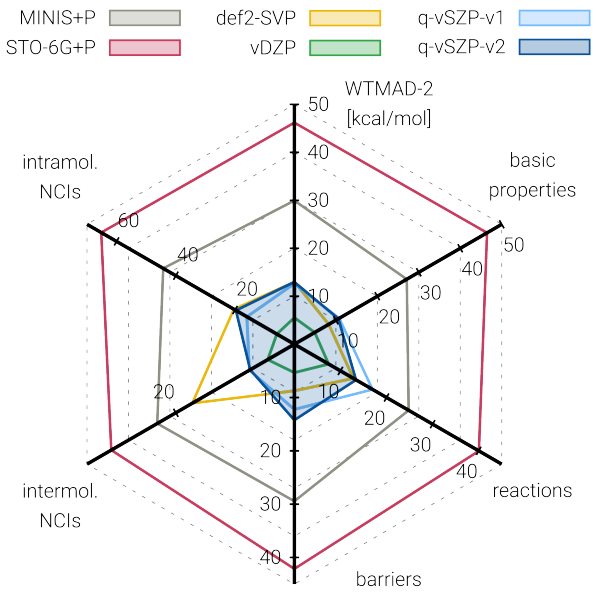

Figure F.9: WTMAD-2 values for the subclasses of the GMTKN55 benchmark set in kcal·mol^{-1}. Raw data required for the reproduction of the figure is provided in the SI.

since initial errors can be corrected during later iterations.

F.5.2 DFT Calculations Using q-vSZP

Main Group Thermochemistry

Since the revised version of the q-vSZP basis is supposed to be the foundation of novel minimal basis (MB) (semiempirical) QM methods, its accuracy shall be tested along similar lines as the first version. Thus, the following analysis highlights the differences to the first version in Ref. [3]. As with the first version, a reduced version of the GMTKN55 database is used, which does not contain elements heavier than calcium to ensure comparability to other MBs.

Fig. F.9 shows that the overall performance of the q-vSZP basis is maintained, with slight differences from the first version. The WTMAD-2 changes from 12.6 to 13.0 kcal·mol^{-1}. In particular, q-vSZP-v1 (light blue) performed slightly better on intramolecular NCIs and reaction barrier heights. Meanwhile, q-vSZP-v2 (dark blue) is more accurate for reaction energies for large systems and isomerizations. Overall, the differences between the two versions are negligible, especially when compared with conventional non-adaptive MBs. The electron density on charged species, the amount of BSSE of the q-vSZP basis set, and electric field-dependent molecular properties are virtually unaffected by the presented update and will thus not be discussed again.

Diuranium Carbide Cluster Stabilized Inside C_{80}

Smaller and more efficient yet still accurate basis sets are particularly rare for heavier elements like lanthanides and actinides. To the best of our knowledge, q-vSZP is the first minimal basis set to cover these elements consistently. It is crucial to ensure that both qualitative and, where possible, quantitative accuracy are preserved since SQM methods using q-vSZP otherwise have to compensate for potential basis set limitations.

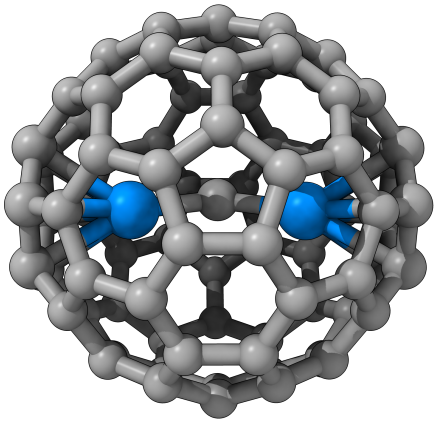

Figure F.10: $U_2C@C_{80}$, a diuranium cluster stabilized inside a C_{80} fullerene cage. The depicted geometry has been taken over from Ref. [320], and was optimized at the PBE0-D4/def2-SVP level of theory.

A relevant test case is U2C@C80 (cf. to Fig. F.10), a diuranium cluster stabilized inside a C_{80} fullerene cage. By assessing its interaction energy, which involves a mixture of covalent and non-covalent contributions[320,490], also BSSE and effects dominated by polarization and more diffuse functions are probed.

Calculating the interaction energy with ωB97X-D4[103] using a standard triple-ζ basis set such as def2-TZVP[139] (def-TZVP[479] for uranium) yields an interaction energy of -268.7 kcal·mol^{-1}. This result closely aligns with the PBE0-D4[491,492] interaction energy of -277.1 kcal·mol^{-1} reported in Ref. [320] with the same basis set. In comparison, ωB97X-D4/q-vSZP gives an interaction energy of -287.3 kcal·mol^{-1}, representing a deviation of -19 kcal·mol^{-1} or 7 %, which might be partly attributed to a small BSSE. All calculations used fixed geometries based on the PBE0-D4/def2-SVP[139] level of theory from Ref.[320]. For the def2-TZVP calculations, the optimized D4 parameters from Ref.[1] were employed, while the q-vSZP-v2 calculation utilized the D4 parameters from q-vSZP-v1.[3]

In addition to reduced computational demands per SCF step, smaller basis sets notably alleviate the issue of tedious SCF convergence with extended Gaussian basis sets. In this example, q-vSZP required only 20 % of the computation time, completing 188 SCF cycles compared to 392 cycles with def2-TZVP, albeit reduced convergence and numerical grid settings for the latter.

Molecular Structures of Lanthanide Complexes

Lanthanide complexes from the LnQM data set[493] were taken as exemplary compounds to ensure the accuracy of molecular geometries optimized with q-vSZP-v2. While the analytical nuclear gradient (current work in our lab) of the q-vSZP basis set is indispensable for effective application in computational chemistry tasks, the geometries for small systems can still be optimized with the numerical gradient.

Fig. F.11 depicts structure overlays for two different octahedral Ln complexes with common ligands such as ammonia, halides, cyanide, and water. Both the visual inspection and the RMSE of the Cartesian coordinates suggest a qualitatively good agreement with the def2-TZVPP reference structure. As a single exception, the nitrogen atom of the thiocyanate group in complex *a)* binds almost at a 180° angle to the central ytterbium atom in the reference structure, whereas the q-vSZP-v2 structure is kinked. A similar albeit less pronounced situation is the case for the cyanide group in complex *b)*. Both observations can be attributed to the minimal basis set character of q-vSZP, potentially preventing some important hybridizations.

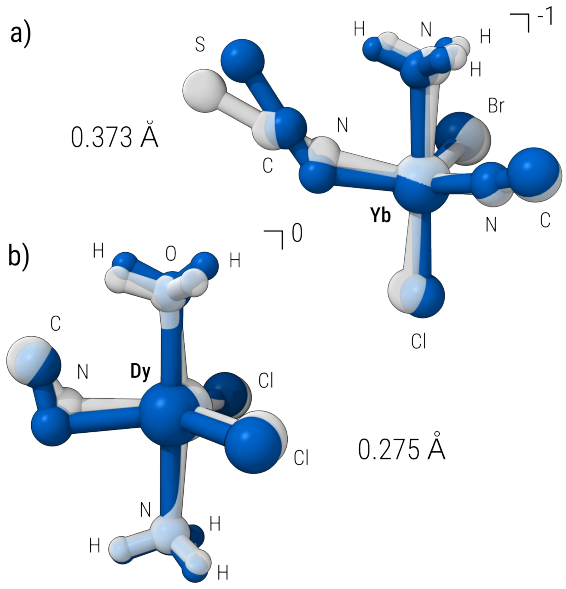

Figure F.11: Overlay of molecular geometries optimized with ωB97X-D4[1,103,120], using the q-vSZP-v2 (blue) and def2-TZVPP[139] (light gray) basis set. The RMSE on the Cartesian q-vSZP-v2 coordinates with respect to the def2-TZVPP geometry is given in Ångström. For q-vSZP, numerical gradients generated by infinitesimal deflection of the Cartesian coordinates via a Python package written in-house[429] were employed. Single-point energies were obtained via `ORCA-6.0.0`.[309] The actual optimization was conducted with `xtb`[259] as a driver. The def2-TZVPP geometry optimization was conducted using analytical gradients in `ORCA-6.0.0`.

F.6 Conclusions

The CEH model and the adaptive minimal q-vSZP basis set have been revised and extended towards the often underrepresented lanthanides and actinides. The revised versions (denoted CEH-v2 and q-vSZP-v2 throughout the manuscript) replace the original versions from Ref. [3].

The extension of the CEH model to accommodate *f*-elements, alongside the introduction of a novel non-iterative approximation for electrostatic interactions and further parameterization enhancements, leads to improved applicability and robustness, particularly in challenging electronic environments. For actinides, *f*-electrons are even explicitly treated in the valence space. The ability of CEH to provide accurate atomic partial charges has been validated against DFT reference data for a wide range of molecular systems, from randomly selected PubChem molecules to actinide complexes and "mindless" molecule. The CEH model consistently surpasses traditional charge equilibration schemes and hybrid physical/ANN-based methods, delivering lower absolute errors and stronger correlations with DFT-calculated charges. Moreover, it is approximately ten times faster than GFN*n*-xTB for typical applications. Its success in reproducing correlations with experimental Hammett parameters further demonstrates its potential as a versatile partial charge model in computational chemistry.

The updated adaptive q-vSZP basis set benefits from a more consistent optimization using reference charges and extends the variationally optimized minimal basis to cover the entire periodic table up to lawrencium ($Z = 103$). With its combination of generality, consistency, and the minimal basis approach to alleviate the diagonalization bottleneck, q-vSZP stands as an optimal basis set choice for future SQM methods.

These advancements provide a robust foundation for the development of future (semiempirical) methods,

enabling the efficient treatment of large, chemically diverse systems while maintaining the accuracy of more computationally intensive QM approaches.

F.7 Acknowledgement

The German Science Foundation (DFG) is gratefully acknowledged for financial support through the SPP 2363 "Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning." M. M. and T. F. thank the Fonds der Chemischen Industrie (FCI) for funding via a Kekulé scholarship.

F.8 Supplementary Material

The Supporting Information is available free of charge at pubs.acs.org/doi/10.1021/acs.jpca.4c06989: Theoretical details of the CEH model, additional information on the random molecule selection and analysis, the q-vSZP basis set and corresponding ECPs in a machine-readable format, and details on the "mindless" molecule generation and benchmark data are provided in the electronic SI. Additionally, molecular structures and corresponding charges for randomly selected molecules, actinide complexes, and "mindless" molecules; CEH-v2 and EEQ charges for the LFER correlation; tabulated results for the U2C@C80 association energy; detailed GMTKN55 benchmark data; and relevant scripts for generation of the most important figures are contained in `assets/SI-qvSZP-v2` within the q-vSZP GitHub repository (github.com/grimme-lab/qvSZP).[430]

Bibliography

[1] M. Müller, A. Hansen, and S. Grimme, *ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set*,
The Journal of Chemical Physics **158** *1* (2023) 014103, DOI: 10.1063/5.0133026.

[2] S. Grimme, M. Müller, and A. Hansen, *A non-self-consistent tight-binding electronic structure potential in a polarized double-ζ basis set for all spd-block elements up to Z = 86*,
The Journal of Chemical Physics **158** *12* (2023) 124111, DOI: 10.1063/5.0137838.

[3] M. Müller, A. Hansen, and S. Grimme, *An atom-in-molecule adaptive polarized valence single-ζ atomic orbital basis for electronic structure calculations*,
The Journal of Chemical Physics **159** *16* (2023) 164108, DOI: 10.1063/5.0172373.

[4] M. Müller, T. Froitzheim, A. Hansen, and S. Grimme, *Advanced Charge Extended Hückel (CEH) Model and a Consistent Adaptive Minimal Basis Set for the Elements Z = 1–103*,
The Journal of Physical Chemistry A **128** *49* (2024), Publisher: American Chemical Society 10723,
DOI: 10.1021/acs.jpca.4c06989.

[5] K. Škoch, C. G. Daniliuc, G. Kehr, S. Ehlert, M. Müller, S. Grimme, and G. Erker,
Frustrated Lewis-Pair Neighbors at the Xanthene Framework: Epimerization at Phosphorus and Cooperative Formation of Macrocyclic Adduct Structures,
Chemistry - A European Journal **27** *47* (2021), Publisher: John Wiley & Sons, Ltd 12104,
DOI: 10.1002/chem.202100835.

[6] N. Fleck, R. M. Thomas, M. Müller, S. Grimme, and B. H. Lipshutz,
An environmentally responsible route to tezacaftor, a drug for treatment of cystic fibrosis prepared in water via ppm Au catalysis as entry to 2-substituted indoles,
Green Chemistry **24** *17* (2022), Publisher: The Royal Society of Chemistry 6517,
DOI: 10.1039/D2GC01828D.

[7] K. Škoch, C. G. Daniliuc, M. Müller, S. Grimme, G. Kehr, and G. Erker, *Stereochemical Behavior of Pairs of P-stereogenic Phosphanyl Groups at the Dimethylxanthene Backbone*,
Chemistry - A European Journal **28** *20* (2022), Publisher: John Wiley & Sons, Ltd e202200248,
DOI: 10.1002/chem.202200248.

[8] P. Pracht et al., *CREST – A program for the exploration of low-energy molecular chemical space*,
The Journal of Chemical Physics **160** *11* (2024) 114110, DOI: 10.1063/5.0197592.

[9] P. A. M. Dirac and R. H. Fowler, *Quantum mechanics of many-electron systems*,
Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **123** *792* (1929), Publisher: Royal Society 714, DOI: 10.1098/rspa.1929.0094.

[10] A. L. Lavoisier, *Traité élémentaire de chimie*, Paris: Chez Cuchet, 1789.

[11] J. Dalton, *A New System of Chemical Philosophy*, vol. 1,
Cambridge Library Collection - Physical Sciences, Cambridge: Cambridge University Press, 1808,
DOI: 10.1017/CBO9780511736391.

[12] A. Kekulé, *Sur la constitution des substances aromatiques*,
Bulletin de la Société Chimique de Paris **3** (1865) 98.

[13] A. Kekulé, *Untersuchungen über aromatische Verbindungen*,
Annalen der Chemie und Pharmacie **137** *2* (1866) 129.

[14] A. Kekulé, *Lehrbuch der Organischen Chemie oder der Chemie der Kohlenstoffverbindungen*,
Enke, 1867, 784 pp.

[15] D. Mendeleev, *The relation between the properties and atomic weights of the elements*,
Journal of the Russian Chemical Society **1** (1869) 60.

[16] J. J. Thomson, *XL. Cathode Rays*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **44** *269* (1897), Publisher: Taylor & Francis 293,
DOI: 10.1080/14786449708621070.

[17] M. Planck, *Ueber das Gesetz der Energieverteilung im Normalspectrum*,
Annalen der Physik **309** *3* (1901) 553, DOI: 10.1002/andp.19013090310.

[18] N. Bohr, *I. On the constitution of atoms and molecules*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **26** *151* (1913), Publisher: Taylor & Francis 1,
DOI: 10.1080/14786441308634955.

[19] L. D. Broglie, *Recherches sur la théorie des Quanta*,
Annales de Physique **10** *3* (1925), Number: 3 Publisher: EDP Sciences 22,
DOI: 10.1051/anphys/192510030022.

[20] W. Heisenberg,
Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen.,
Zeitschrift für Physik **33** *1* (1925) 879, DOI: 10.1007/BF01328377.

[21] W. Pauli, *Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren*, Zeitschrift für Physik **31** *1* (1925) 765, DOI: 10.1007/BF02980631.

[22] E. Schrödinger, *Quantisierung als Eigenwertproblem*, Annalen der Physik **384** *4* (1926) 361,
DOI: 10.1002/andp.19263840404.

[23] M. Barbatti,
When Theory Came First: A Review of Theoretical Chemical Predictions Ahead of Experiments,
2025, DOI: 10.26434/chemrxiv-2025-gn7lm.

[24] J. J. P. Eckert and J. W. Mauchly, "Electronic numerical integrator and computer,"
U.S. pat. 3120606A, S. R. Corp, 1964.

[25] P. A. Freiberger and M. R. Swaine, *ENIAC*, *Encyclopedia Britannica*, 2024.

[26] P. Pulay, *Analytical derivatives, forces, force constants, molecular geometries, and related response properties in electronic structure theory*,
WIREs Computational Molecular Science **4** *3* (2014) 169, DOI: 10.1002/wcms.1171.

[27] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton,
Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), JOM **65** *11* (2013) 1501,
DOI: 10.1007/s11837-013-0755-4.

[28] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio,
From DFT to machine learning: recent approaches to materials science–a review,
Journal of Physics: Materials **2** *3* (2019), Publisher: IOP Publishing 032001,
DOI: 10.1088/2515-7639/ab084b.

[29] A. Jain, Y. Shin, and K. A. Persson,
Computational predictions of energy materials using density functional theory,
Nature Reviews Materials **1** *1* (2016), Publisher: Nature Publishing Group 1,
DOI: 10.1038/natrevmats.2015.4.

[30] F. Strieth-Kalthoff et al.,
Delocalized, asynchronous, closed-loop discovery of organic laser emitters, Science **384** *6697* (2024), Publisher: American Association for the Advancement of Science eadk9227,
DOI: 10.1126/science.adk9227.

[31] R. Pollice, B. Ding, and A. Aspuru-Guzik, *Rational design of organic molecules with inverted gaps between the first excited singlet and triplet*, Matter **7** *3* (2024), Publisher: Elsevier 1161,
DOI: 10.1016/j.matt.2024.01.002.

[32] C. Gorgulla et al., *An open-source drug discovery platform enables ultra-large virtual screens*, Nature **580** *7805* (2020), Publisher: Nature Publishing Group 663,
DOI: 10.1038/s41586-020-2117-z.

[33] J. (Zeng, Y. (Tao, T. J. Giese, and D. M. York,
Modern semiempirical electronic structure methods and machine learning potentials for drug discovery: Conformers, tautomers, and protonation states,
The Journal of Chemical Physics **158** *12* (2023) 124110, DOI: 10.1063/5.0139281.

[34] M. Iff, K. Atz, C. Isert, I. Pachon-Angona, L. Cotos, M. Hilleke, J. A. Hiss, and G. Schneider,
Combining de novo molecular design with semiempirical protein–ligand binding free energy calculation, RSC Advances **14** *50* (2024), Publisher: The Royal Society of Chemistry 37035,
DOI: 10.1039/D4RA05422A.

[35] Z. Zhang, J. B. Stückrath, S. Grimme, and A. Gansäuer,
Titanocene-Catalyzed [2+2] Cycloaddition of Bisenones and Comparison with Photoredox Catalysis and Established Methods, Angewandte Chemie International Edition **60** *26* (2021) 14339,
DOI: 10.1002/anie.202102739.

[36] B. W. J. Chen, L. Xu, and M. Mavrikakis, *Computational Methods in Heterogeneous Catalysis*, Chemical Reviews **121** *2* (2021), Publisher: American Chemical Society 1007,
DOI: 10.1021/acs.chemrev.0c01060.

[37] P. Zhang, Y. Yang, X. Duan, Y. Liu, and S. Wang,
Density Functional Theory Calculations for Insight into the Heterocatalyst Reactivity and Mechanism in Persulfate-Based Advanced Oxidation Reactions,
ACS Catalysis **11** *17* (2021), Publisher: American Chemical Society 11129,
DOI: 10.1021/acscatal.1c03099.

[38] A. N. Morozov and A. M. Mebel,
Theoretical study of the reaction mechanism and kinetics of the phenyl + propargyl association,
Physical Chemistry Chemical Physics **22** *13* (2020), Publisher: The Royal Society of Chemistry 6868,
DOI: 10.1039/D0CP00306A.

[39] M. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln*,
Annalen der Physik **389** *20* (1927) 457, DOI: 10.1002/andp.19273892002.

[40] D. R. Hartree,
The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods,
Mathematical Proceedings of the Cambridge Philosophical Society **24** *1* (1928) 89,
DOI: 10.1017/S0305004100011919.

[41] P. Weiss, *L'hypothèse du champ moléculaire et la propriété ferromagnétique*, Journal de Physique Théorique et Appliquée **6** *1* (1907), Publisher: Société Française de Physique 661,
DOI: 10.1051/jphystap:019070060066100.

[42] J. C. Slater, *The Theory of Complex Spectra*,
Physical Review **34** *10* (1929), Publisher: American Physical Society 1293,
DOI: 10.1103/PhysRev.34.1293.

[43] V. Fock, *Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems*,
Zeitschrift für Physik **61** *1* (1930) 126, DOI: 10.1007/BF01340294.

[44] W. Ritz,
Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik,
Journal für die reine und angewandte Mathematik **135** *1* (1909) 1.

[45] J. K. L. MacDonald, *Successive Approximations by the Rayleigh-Ritz Variation Method*,
Physical Review **43** *10* (1933), Publisher: American Physical Society 830,
DOI: 10.1103/PhysRev.43.830.

[46] P.-O. Löwdin, *Quantum Theory of Many-Particle Systems. I. Physical Interpretations by Means of Density Matrices, Natural Spin-Orbitals, and Convergence Problems in the Method of Configurational Interaction*,
Physical Review **97** *6* (1955), Publisher: American Physical Society 1474,
DOI: 10.1103/PhysRev.97.1474.

[47] B. Roos, *A new method for large-scale Cl calculations*, Chemical Physics Letters **15** *2* (1972) 153,
DOI: 10.1016/0009-2614(72)80140-4.

[48] F. Coester and H. Kümmel, *Short-range correlations in nuclear wave functions*,
Nuclear Physics **17** (1960) 477, DOI: 10.1016/0029-5582(60)90140-1.

[49] O. Sinanoğlu, *Many-Electron Theory of Atoms and Molecules. I. Shells, Electron Pairs vs Many-Electron Correlations*, The Journal of Chemical Physics **36** *3* (1962) 706,
DOI: 10.1063/1.1732596.

[50] J. Čížek, *On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods*,
The Journal of Chemical Physics **45** *11* (1966) 4256, DOI: 10.1063/1.1727484.

[51] G. D. Purvis III and R. J. Bartlett,
A full coupled-cluster singles and doubles model: The inclusion of disconnected triples,
The Journal of Chemical Physics **76** *4* (1982) 1910, DOI: 10.1063/1.443164.

[52] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon,
A fifth-order perturbation comparison of electron correlation theories,
Chemical Physics Letters **157** *6* (1989) 479, DOI: 10.1016/S0009-2614(89)87395-6.

[53] C. Møller and M. S. Plesset, *Note on an Approximation Treatment for Many-Electron Systems*,
Physical Review **46** *7* (1934), Publisher: American Physical Society 618,
DOI: 10.1103/PhysRev.46.618.

[54] P. Siegbahn, A. Heiberg, B. Roos, and B. Levy, *A Comparison of the Super-CI and the Newton-Raphson Scheme in the Complete Active Space SCF Method*,
Physica Scripta **21** *3* (1980) 323, DOI: 10.1088/0031-8949/21/3-4/014.

[55] B. O. Roos, P. R. Taylor, and P. E. M. Sigbahn, *A complete active space SCF method (CASSCF) using a density matrix formulated super-CI approach*, Chemical Physics **48** *2* (1980) 157,
DOI: 10.1016/0301-0104(80)80045-0.

[56] P. E. M. Siegbahn, J. Almlöf, A. Heiberg, and B. O. Roos, *The complete active space SCF (CASSCF) method in a Newton–Raphson formulation with application to the HNO molecule*,
The Journal of Chemical Physics **74** *4* (1981) 2384, DOI: 10.1063/1.441359.

[57] R. J. Buenker and S. D. Peyerimhoff, *CI method for the study of general molecular potentials*,
Theoretica chimica acta **12** *3* (1968) 183, DOI: 10.1007/BF00528266.

[58] H.-J. Werner and P. J. Knowles,
An efficient internally contracted multiconfiguration–reference configuration interaction method,
The Journal of Chemical Physics **89** *9* (1988) 5803, DOI: 10.1063/1.455556.

[59] P. Hohenberg and W. Kohn, *Inhomogeneous Electron Gas*,
Physical Review **136** *3* (1964), Publisher: American Physical Society B864,
DOI: 10.1103/PhysRev.136.B864.

[60] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*,
John Wiley & Sons, 2015, 378 pp.

[61] L. H. Thomas, *The calculation of atomic fields*,
Mathematical Proceedings of the Cambridge Philosophical Society **23** *5* (1927) 542,
DOI: 10.1017/S0305004100011683.

[62] E. Fermi, *Eine statistische Methode zur Bestimmung einiger Eigenschaften des Atoms und ihre Anwendung auf die Theorie des periodischen Systems der Elemente*,
Zeitschrift für Physik **48** *1* (1928) 73, DOI: 10.1007/BF01351576.

[63] J. C. Slater, *A Simplification of the Hartree-Fock Method*,
Physical Review **81** *3* (1951), Publisher: American Physical Society 385,
DOI: 10.1103/PhysRev.81.385.

[64] W. Kohn and L. J. Sham, *Self-Consistent Equations Including Exchange and Correlation Effects*,
Physical Review **140** *4* (1965), Publisher: American Physical Society A1133,
DOI: 10.1103/PhysRev.140.A1133.

[65] A. D. Kaplan, M. Levy, and J. P. Perdew,
The Predictive Power of Exact Constraints and Appropriate Norms in Density Functional Theory,
Annual Review of Physical Chemistry **74** (Volume 74, 2023 2023), Publisher: Annual Reviews 193,
DOI: 10.1146/annurev-physchem-062422-013259.

[66] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun,
Accurate and Numerically Efficient r2SCAN Meta-Generalized Gradient Approximation, The
Journal of Physical Chemistry Letters **11** *19* (2020), Publisher: American Chemical Society 8208,
DOI: 10.1021/acs.jpclett.0c02405.

[67] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, *Correction to "Accurate and
Numerically Efficient r2SCAN Meta-Generalized Gradient Approximation"*, The Journal of
Physical Chemistry Letters **11** *21* (2020), Publisher: American Chemical Society 9248,
DOI: 10.1021/acs.jpclett.0c03077.

[68] R. G. Parr and Y. Weitao, *Density-Functional Theory of Atoms and Molecules*,
Oxford University Press, 1994, 344 pp.

[69] F. Jensen, *Introduction to Computational Chemistry*, Wiley-VCH Verlag, 2017, 457 pp.

[70] M. Bursch, J.-M. Mewes, A. Hansen, and S. Grimme,
*Best-Practice DFT Protocols for Basic Molecular Computational Chemistry***,
Angewandte Chemie International Edition **61** *42* (2022) e202205735,
DOI: 10.1002/anie.202205735.

[71] Y. Zhao and D. G. Truhlar, *Density Functionals with Broad Applicability in Chemistry*,
Accounts of Chemical Research **41** *2* (2008), Publisher: American Chemical Society 157,
DOI: 10.1021/ar700111a.

[72] W. Thiel, *Semiempirical quantum-chemical methods*,
Wiley Interdisciplinary Reviews: Computational Molecular Science **4** *2* (2014) 145,
DOI: 10.1002/wcms.1161.

[73] S. Kurth, J. P. Perdew, and P. Blaha, *Molecular and solid-state tests of density functional
approximations: LSD, GGAs, and meta-GGAs*,
International Journal of Quantum Chemistry **75** *4* (1999) 889,
DOI: 10.1002/(SICI)1097-461X(1999)75:4/5<889::AID-QUA54>3.0.CO;2-8.

[74] J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka,
*Prescription for the design and selection of density functional approximations: More constraint
satisfaction with fewer fits*, The Journal of Chemical Physics **123** *6* (2005) 062201,
DOI: 10.1063/1.1904565.

[75] S. H. Vosko, L. Wilk, and M. Nusair, *Accurate spin-dependent electron liquid correlation energies
for local spin density calculations: a critical analysis*,
Canadian Journal of Physics **58** *8* (1980), Publisher: NRC Research Press 1200,
DOI: 10.1139/p80-159.

[76] J. P. Perdew and Y. Wang,
Accurate and simple analytic representation of the electron-gas correlation energy,
Physical Review B **45** *23* (1992), Publisher: American Physical Society 13244,
DOI: 10.1103/PhysRevB.45.13244.

[77] A. D. Becke,
Density-functional exchange-energy approximation with correct asymptotic behavior,
Physical Review A **38** *6* (1988), Publisher: American Physical Society 3098,
DOI: 10.1103/PhysRevA.38.3098.

[78] C. Lee, W. Yang, and R. G. Parr, *Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density*, Physical Review B **37** *2* (1988) 785,
DOI: 10.1103/PhysRevB.37.785.

[79] B. Miehlich, A. Savin, H. Stoll, and H. Preuss,
Results obtained with the correlation energy density functionals of becke and Lee, Yang and Parr,
Chemical Physics Letters **157** *3* (1989) 200, DOI: 10.1016/0009-2614(89)87234-3.

[80] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais,
Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation,
Physical Review B **46** *11* (1992), Publisher: American Physical Society 6671,
DOI: 10.1103/PhysRevB.46.6671.

[81] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized gradient approximation made simple*,
Physical Review Letters **77** *18* (1996) 3865, DOI: 10.1103/PhysRevLett.77.3865.

[82] J. P. Perdew, S. Kurth, A. Zupan, and P. Blaha, *Accurate Density Functional with Correct Formal Properties: A Step Beyond the Generalized Gradient Approximation*,
Physical Review Letters **82** *12* (1999), Publisher: American Physical Society 2544,
DOI: 10.1103/PhysRevLett.82.2544.

[83] J. P. Perdew, S. Kurth, A. Zupan, and P. Blaha,
Erratum: Accurate Density Functional with Correct Formal Properties: A Step Beyond the Generalized Gradient Approximation [Phys. Rev. Lett. 82, 2544 (1999)],
Physical Review Letters **82** *25* (1999), Publisher: American Physical Society 5179,
DOI: 10.1103/PhysRevLett.82.5179.

[84] J. P. Perdew, J. Tao, V. N. Staroverov, and G. E. Scuseria, *Meta-generalized gradient approximation: Explanation of a realistic nonempirical density functional*,
The Journal of Chemical Physics **120** *15* (2004) 6898, DOI: 10.1063/1.1665298.

[85] N. Mardirossian and M. Head-Gordon, *Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V*,
Journal of Chemical Physics **142** *7* (2015), Publisher: AIP Publishing LLCAIP Publishing 074111,
DOI: 10.1063/1.4907719.

[86] J. Harris, *Adiabatic-connection approach to Kohn-Sham theory*,
Physical Review A **29** *4* (1984), Publisher: American Physical Society 1648,
DOI: 10.1103/PhysRevA.29.1648.

[87] A. D. Becke, *A new mixing of Hartree–Fock and local density-functional theories*,
The Journal of Chemical Physics **98** *2* (1993) 1372, DOI: 10.1063/1.464304.

[88] A. D. Becke, *Density-functional thermochemistry. III. The role of exact exchange*,
The Journal of Chemical Physics **98** *7* (1993), Publisher: American Institute of PhysicsAIP 5648,
DOI: 10.1063/1.464913.

[89] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, *Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields*, The Journal of Physical Chemistry **98** *45* (1994), Publisher: American Chemical Society 11623, DOI: 10.1021/j100096a001.

[90] M. Ernzerhof and G. E. Scuseria, *Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional*, The Journal of Chemical Physics **110** *11* (1999) 5029, DOI: 10.1063/1.478401.

[91] V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, *Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes*, The Journal of Chemical Physics **119** *23* (2003) 12129, DOI: 10.1063/1.1626543.

[92] V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, *Erratum: "Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes" [J. Chem. Phys. 119, 12129 (2003)]*, The Journal of Chemical Physics **121** *22* (2004) 11507, DOI: 10.1063/1.1795692.

[93] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, *A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions*, Physical Chemistry Chemical Physics **19** *48* (2017) 32184, DOI: 10.1039/c7cp04913g.

[94] S. Grimme, *Semiempirical hybrid density functional with perturbative second-order correlation*, The Journal of Chemical Physics **124** *3* (2006) 034108, DOI: 10.1063/1.2148954.

[95] A. Savin and H.-J. Flad, *Density functionals for the Yukawa electron-electron interaction*, International Journal of Quantum Chemistry **56** *4* (1995) 327, DOI: 10.1002/qua.560560417.

[96] T. Leininger, H. Stoll, H.-J. Werner, and A. Savin, *Combining long-range configuration interaction with short-range density functionals*, Chemical Physics Letters **275** *3* (1997) 151, DOI: 10.1016/S0009-2614(97)00758-6.

[97] H. Iikura, T. Tsuneda, T. Yanai, and K. Hirao, *A long-range correction scheme for generalized-gradient-approximation exchange functionals*, The Journal of Chemical Physics **115** *8* (2001) 3540, DOI: 10.1063/1.1383587.

[98] T. Tsuneda and Kimihiko Hirao, *Long-range correction for density functional theory*, WIREs Computational Molecular Science **4** *4* (2014) 375, DOI: 10.1002/wcms.1178.

[99] T. Stein, L. Kronik, and R. Baer, *Prediction of charge-transfer excitations in coumarin-based dyes using a range-separated functional tuned from first principles*, The Journal of Chemical Physics **131** *24* (2009) 244119, DOI: 10.1063/1.3269029.

[100] T. Yanai, D. P. Tew, and N. C. Handy, *A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP)*, Chemical Physics Letters **393** *1* (2004) 51, DOI: 10.1016/j.cplett.2004.06.011.

[101] J.-D. Chai and M. Head-Gordon, *Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections*, Physical Chemistry Chemical Physics **10** *44* (2008), Publisher: The Royal Society of Chemistry 6615, DOI: 10.1039/B810189B.

[102] J.-D. Chai and M. Head-Gordon,
Systematic optimization of long-range corrected hybrid density functionals,
The Journal of Chemical Physics **128** *8* (2008) 084106, DOI: 10.1063/1.2834918.

[103] N. Mardirossian and M. Head-Gordon,
ωb97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy,
Physical Chemistry Chemical Physics **16** *21* (2014), Publisher: The Royal Society of Chemistry 9904,
DOI: 10.1039/c3cp54374a.

[104] N. Mardirossian and M. Head-Gordon, *ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation*,
Journal of Chemical Physics **144** *21* (2016), Publisher: American Institute of Physics Inc.,
DOI: 10.1063/1.4952647.

[105] J. Heyd, G. E. Scuseria, and M. Ernzerhof,
Hybrid functionals based on a screened Coulomb potential,
The Journal of Chemical Physics **118** *18* (2003) 8207, DOI: 10.1063/1.1564060.

[106] J. Heyd, G. E. Scuseria, and M. Ernzerhof, *Erratum: "Hybrid functionals based on a screened Coulomb potential" [J. Chem. Phys. 118, 8207 (2003)]*,
The Journal of Chemical Physics **124** *21* (2006) 219906, DOI: 10.1063/1.2204597.

[107] S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth,
Dispersion-Corrected Mean-Field Electronic Structure Methods,
Chemical Reviews **116** *9* (2016) 5105, DOI: 10.1021/acs.chemrev.5b00533.

[108] L. Wittmann, H. Neugebauer, S. Grimme, and M. Bursch,
Dispersion-corrected r2SCAN based double-hybrid functionals,
The Journal of Chemical Physics **159** *22* (2023) 224103, DOI: 10.1063/5.0174988.

[109] S. Kristyán and P. Pulay,
Can (semi)local density functional theory account for the London dispersion forces?
Chemical Physics Letters **229** *3* (1994) 175, DOI: 10.1016/0009-2614(94)01027-7.

[110] D. J. Lacks and R. G. Gordon, *Pair interactions of rare-gas atoms as a test of exchange-energy-density functionals in regions of large density gradients*,
Physical Review A **47** *6* (1993), Publisher: American Physical Society 4681,
DOI: 10.1103/PhysRevA.47.4681.

[111] P. Hobza, J. Šponer, and T. Reschel, *Density functional theory and molecular clusters*,
Journal of Computational Chemistry **16** *11* (1995) 1315, DOI: 10.1002/jcc.540161102.

[112] J. Seponer, J. Leszczynski, and P. Hobza,
Base stacking in cytosine dimer. A comparison of correlated ab initio calculations with three empirical potential models and density functional theory calculations,
Journal of Computational Chemistry **17** *7* (1996) 841,
DOI: 10.1002/(SICI)1096-987X(199605)17:7<841::AID-JCC8>3.0.CO;2-S.

[113] S. Grimme, W. Hujo, and B. Kirchner, *Performance of dispersion-corrected density functional theory for the interactions in ionic liquids*,
Physical Chemistry Chemical Physics **14** *14* (2012), Publisher: The Royal Society of Chemistry 4875,
DOI: 10.1039/c2cp24096c.

[114] J. G. Brandenburg and S. Grimme, "Dispersion Corrected Hartree–Fock and Density Functional Theory for Organic Crystal Structure Prediction,"
Prediction and Calculation of Crystal Structures: Methods and Applications,
ed. by S. Atahan-Evrenk and A. Aspuru-Guzik, Cham: Springer International Publishing, 2014 1, DOI: 10.1007/128_2013_488.

[115] S. Grimme, J. G. Brandenburg, C. Bannwarth, and A. Hansen, *Consistent structures and interactions by density functional theory with small atomic orbital basis sets*,
Journal of Chemical Physics **143** *5* (2015) 54107, DOI: 10.1063/1.4927476.

[116] J. Gorges, B. Bädorf, S. Grimme, and A. Hansen, *Efficient Computation of the Interaction Energies of Very Large Non-covalently Bound Complexes*,
Synlett **34** *10* (2023), Publisher: Georg Thieme Verlag KG 1135, DOI: 10.1055/s-0042-1753141.

[117] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, *A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu*,
Journal of Chemical Physics **132** *15* (2010), Publisher: American Institute of PhysicsAIP 154104,
DOI: 10.1063/1.3382344.

[118] S. Grimme, S. Ehrlich, and L. Goerigk,
Effect of the damping function in dispersion corrected density functional theory, Journal of Computational Chemistry **32** *7* (2011), Publisher: John Wiley & Sons, Ltd ISBN: 1096-987X 1456,
DOI: 10.1002/jcc.21759.

[119] E. Caldeweyher, C. Bannwarth, and S. Grimme, *Extension of the D3 dispersion coefficient model*,
Journal of Chemical Physics **147** *3* (2017) 34112, DOI: 10.1063/1.4993215.

[120] E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme,
A generally applicable atomic-charge dependent London dispersion correction,
Journal of Chemical Physics **150** *15* (2019) 154122, DOI: 10.1063/1.5090222.

[121] A. Tkatchenko and M. Scheffler, *Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data*,
Physical Review Letters **102** *7* (2009), Publisher: American Physical Society 073005,
DOI: 10.1103/PhysRevLett.102.073005.

[122] A. Tkatchenko, R. A. DiStasio, R. Car, and M. Scheffler,
Accurate and Efficient Method for Many-Body van der Waals Interactions,
Physical Review Letters **108** *23* (2012), Publisher: American Physical Society 236402,
DOI: 10.1103/PhysRevLett.108.236402.

[123] A. D. Becke and E. R. Johnson, *A density-functional model of the dispersion interaction*,
The Journal of Chemical Physics **123** *15* (2005) 154101, DOI: 10.1063/1.2065267.

[124] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist,
Van der Waals Density Functional for General Geometries,
Physical Review Letters **92** *24* (2004), Publisher: American Physical Society 246401,
DOI: 10.1103/PhysRevLett.92.246401.

[125] O. A. Vydrov and T. Van Voorhis,
Nonlocal van der Waals density functional: The simpler the better,
Journal of Chemical Physics **133** *24* (2010), Publisher: American Institute of PhysicsAIP 244103,
DOI: 10.1063/1.3521275.

[126] A. M. Teale et al., *DFT exchange: sharing perspectives on the workhorse of quantum chemistry and materials science*, Physical Chemistry Chemical Physics **24** *47* (2022), Publisher: The Royal Society of Chemistry 28700, DOI: 10.1039/D2CP02827A.

[127] C. C. J. Roothaan, *New Developments in Molecular Orbital Theory*, Reviews of Modern Physics **23** *2* (1951), Publisher: American Physical Society 69, DOI: 10.1103/RevModPhys.23.69.

[128] G. G. Hall and J. E. Lennard-Jones, *The molecular orbital theory of chemical valency VIII. A method of calculating ionization potentials*, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **205** *1083* (1951), Publisher: Royal Society 541, DOI: 10.1098/rspa.1951.0048.

[129] J. C. Slater, *Atomic Shielding Constants*, Physical Review **36** *1* (1930), Publisher: American Physical Society 57, DOI: 10.1103/PhysRev.36.57.

[130] S. F. Boys and A. C. Egerton, *Electronic wave functions - I. A general method of calculation for the stationary states of any molecular system*, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **200** *1063* (1950), Publisher: Royal Society 542, DOI: 10.1098/rspa.1950.0036.

[131] P. M. W. Gill, "Molecular integrals Over Gaussian Basis Functions," *Advances in Quantum Chemistry*, ed. by J. R. Sabin and M. C. Zerner, vol. 25, Academic Press, 1994 141, DOI: 10.1016/S0065-3276(08)60019-2.

[132] E. Perlt, ed., *Basis Sets in Computational Chemistry*, vol. 107, Lecture Notes in Chemistry, Cham: Springer International Publishing, 2021, DOI: 10.1007/978-3-030-67262-1.

[133] P. Pyykko, *Relativistic effects in structural chemistry*, Chemical Reviews **88** *3* (1988), Publisher: American Chemical Society 563, DOI: 10.1021/cr00085a006.

[134] H. Hellmann, *A New Approximation Method in the Problem of Many Electrons*, The Journal of Chemical Physics **3** *1* (1935) 61, DOI: 10.1063/1.1749559.

[135] J. C. Phillips and L. Kleinman, *New Method for Calculating Wave Functions in Crystals and Molecules*, Physical Review **116** *2* (1959), Publisher: American Physical Society 287, DOI: 10.1103/PhysRev.116.287.

[136] L. R. Kahn, P. Baybutt, and D. G. Truhlar, *Ab initio effective core potentials: Reduction of all-electron molecular structure calculations to calculations involving only valence electrons*, The Journal of Chemical Physics **65** *10* (1976) 3826, DOI: 10.1063/1.432900.

[137] M. Dolg, U. Wedig, H. Stoll, and H. Preuss, *Energy-adjusted ab initio pseudopotentials for the first row transition elements*, The Journal of Chemical Physics **86** *2* (1986) 866, DOI: 10.1063/1.452288.

[138] L. R. Kahn and W. A. Goddard, *Ab initio effective potentials for use in molecular calculations*, The Journal of Chemical Physics **56** *6* (1972) 2702, DOI: 10.1063/1.1677597.

[139] F. Weigend and R. Ahlrichs, *Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy*, Physical Chemistry Chemical Physics **7** *18* (2005), Publisher: The Royal Society of Chemistry 3297, DOI: 10.1039/b508541a.

[140] M. Dolg, H. Stoll, and H. Preuss, *Energy-adjusted ab initio pseudopotentials for the rare earth elements*, The Journal of Chemical Physics **90** *3* (1989), Publisher: Springer-Ver!ag 1730, DOI: 10.1063/1.456066.

[141] A. Bergner, M. Dolg, W. Küchle, H. Stoll, and H. Preuß, *Ab initio energy-adjusted pseudopotentials for elements of groups 13-17*, Molecular Physics **80** *6* (1993), Publisher: Taylor & Francis Group 1431, DOI: 10.1080/00268979300103121.

[142] W. Küchle, M. Dolg, H. Stoll, and H. Preuss, *Energy-adjusted pseudopotentials for the actinides. Parameter sets and test calculations for thorium and thorium monoxide*, The Journal of Chemical Physics **100** *10* (1994) 7535, DOI: 10.1063/1.466847.

[143] P. J. Hay and W. R. Wadt, *Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg*, The Journal of Chemical Physics **82** *1* (1985) 270, DOI: 10.1063/1.448799.

[144] L. A. LaJohn, P. A. Christiansen, R. B. Ross, T. Atashroo, and W. C. Ermler, *Ab initio relativistic effective potentials with spin–orbit operators. III. Rb through Xe*, The Journal of Chemical Physics **87** *5* (1987) 2812, DOI: 10.1063/1.453069.

[145] R. F. Stewart, *Small Gaussian Expansions of Slater-Type Orbitals*, The Journal of Chemical Physics **52** *1* (1970) 431, DOI: 10.1063/1.1672702.

[146] W. J. Hehre, R. Ditchfield, and J. A. Pople, *Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules*, The Journal of Chemical Physics **56** *5* (1972) 2257, DOI: 10.1063/1.1677527.

[147] R. T. Ireland and L. K. McKemmish, *On the specialization of Gaussian basis sets for core-dependent properties*, The Journal of Chemical Physics **159** *6* (2023) 064102, DOI: 10.1063/5.0159119.

[148] B. O. Roos, R. Lindh, P.-Å. Malmqvist, V. Veryazov, and P.-O. Widmark, *Main Group Atoms and Dimers Studied with a New Relativistic ANO Basis Set*, The Journal of Physical Chemistry A **108** *15* (2004), Publisher: American Chemical Society 2851, DOI: 10.1021/jp031064+.

[149] D. Rappoport and F. Furche, *Property-optimized Gaussian basis sets for molecular response calculations*, The Journal of Chemical Physics **133** *13* (2010) 134105, DOI: 10.1063/1.3484283.

[150] T. H. Dunning, *Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen*, The Journal of Chemical Physics **90** *2* (1989) 1007, DOI: 10.1063/1.456153.

[151] K. L. Schuchardt, B. T. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li, and T. L. Windus, *Basis Set Exchange: A Community Database for Computational Sciences*, Journal of Chemical Information and Modeling **47** *3* (2007), Publisher: American Chemical Society 1045, DOI: 10.1021/ci600510j.

[152] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson, and T. L. Windus, *New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community*, Journal of Chemical Information and Modeling **59** *11* (2019), Publisher: American Chemical Society 4814, DOI: 10.1021/acs.jcim.9b00725.

[153] J. Witte, J. B. Neaton, and M. Head-Gordon, *Effective empirical corrections for basis set superposition error in the def2-SVPD basis: GCP and DFT-C*, Journal of Chemical Physics **146** *23* (2017) 234105, DOI: 10.1063/1.4986962.

[154] H. B. Jansen and P. Ros, *Non-empirical molecular orbital calculations on the protonation of carbon monoxide*, Chemical Physics Letters **3** *3* (1969) 140, DOI: 10.1016/0009-2614(69)80118-1.

[155] B. Liu and A. D. McLean, *Accurate calculation of the attractive interaction of two ground state helium atoms*, The Journal of Chemical Physics **59** *8* (1973) 4557, DOI: 10.1063/1.1680654.

[156] F. Jensen, *Polarization consistent basis sets: II. Estimating the Kohn-Sham basis set limit*, Journal of Chemical Physics **116** *17* (2002), Publisher: American Institute of PhysicsAIP 7372, DOI: 10.1063/1.1465405.

[157] L. K. McKemmish and P. M. W. Gill, *Gaussian Expansions of Orbitals*, Journal of Chemical Theory and Computation **8** *12* (2012), Publisher: American Chemical Society 4891, DOI: 10.1021/ct300559t.

[158] T. Helgaker, W. Klopper, H. Koch, and J. Noga, *Basis-set convergence of correlated calculations on water*, Journal of Chemical Physics **106** *23* (1997), Publisher: American Institute of Physics Inc. 9639, DOI: 10.1063/1.473863.

[159] F. Jensen, *Polarization consistent basis sets: Principles*, Journal of Chemical Physics **115** *20* (2001), Publisher: American Institute of PhysicsAIP 9113, DOI: 10.1063/1.1413524.

[160] J. VandeVondele and J. Hutter, *Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases*, Journal of Chemical Physics **127** *11* (2007), DOI: 10.1063/1.2770708.

[161] R. A. Shaw and J. G. Hill, *BasisOpt: A Python package for quantum chemistry basis set optimization*, The Journal of Chemical Physics **159** *4* (2023) 044802, DOI: 10.1063/5.0157878.

[162] J. A. Calvin, C. Peng, V. Rishi, A. Kumar, and E. F. Valeev, *Many-Body Quantum Chemistry on Massively Parallel Computers*, Chemical Reviews **121** *3* (2021), Publisher: American Chemical Society 1203, DOI: 10.1021/acs.chemrev.0c00006.

[163] S. M. Woodley and C. R. A. Catlow, *High performance computing in the chemistry of materials*, Physical Chemistry Chemical Physics **16** *39* (2014), Publisher: The Royal Society of Chemistry 21001, DOI: 10.1039/C4CP90126F.

[164] S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher, and M. Stahn, *Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules*, Journal of Physical Chemistry A **125** *19* (2021), Publisher: American Chemical Society 4039, DOI: 10.1021/acs.jpca.1c00971.

[165] P. Pracht, F. Bohle, and S. Grimme, *Automated exploration of the low-energy chemical space with fast quantum chemical methods*, Physical Chemistry Chemical Physics **22** *14* (2020), Publisher: Royal Society of Chemistry 7169, DOI: 10.1039/c9cp06869d.

[166] C. Plett and S. Grimme, *Automated and Efficient Generation of General Molecular Aggregate Structures*, Angewandte Chemie International Edition **62** *4* (2023) e202214477, DOI: 10.1002/anie.202214477.

[167] S. Spicher, C. Plett, P. Pracht, A. Hansen, and S. Grimme, *Automated Molecular Cluster Growing for Explicit Solvation by Efficient Force Field and Tight Binding Methods*, Journal of Chemical Theory and Computation **18** *5* (2022), Publisher: American Chemical Society 3174, DOI: 10.1021/acs.jctc.2c00239.

[168] S. Grimme, *Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations*, Journal of Chemical Theory and Computation **15** *5* (2019), Publisher: American Chemical Society 2847, DOI: 10.1021/acs.jctc.9b00143.

[169] B. Kirchner, P. J. di Dio, and J. Hutter, "Real-World Predictions from Ab Initio Molecular Dynamics Simulations," *Multiscale Molecular Methods in Applied Chemistry*, ed. by B. Kirchner and J. Vrabec, Berlin, Heidelberg: Springer, 2012 109, DOI: 10.1007/128_2011_195.

[170] A. Nigam, R. Pollice, G. Tom, K. Jorner, J. Willes, L. Thiede, A. Kundaje, and A. Aspuru-Guzik, *Tartarus: A Benchmarking Platform for Realistic And Practical Inverse Molecular Design*, Advances in Neural Information Processing Systems **36** (2023) 3263.

[171] J. P. Unsleber, S. A. Grimmel, and M. Reiher, *Chemoton 2.0: Autonomous Exploration of Chemical Reaction Networks*, Journal of Chemical Theory and Computation **18** *9* (2022), Publisher: American Chemical Society 5393, DOI: 10.1021/acs.jctc.2c00193.

[172] M. Häser and R. Ahlrichs, *Improvements on the direct SCF method*, Journal of Computational Chemistry **10** *1* (1989) 104, DOI: 10.1002/jcc.540100111.

[173] J. Kussmann, M. Beer, and C. Ochsenfeld, *Linear-scaling self-consistent field methods for large molecules*, WIREs Computational Molecular Science **3** *6* (2013) 614, DOI: 10.1002/wcms.1138.

[174] J. L. Whitten, *Coulombic potential energy integrals and approximations*, The Journal of Chemical Physics **58** *10* (1973) 4496, DOI: 10.1063/1.1679012.

[175] K. Eichkorn, O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs,
Auxiliary basis sets to approximate Coulomb potentials,
Chemical Physics Letters **240** *4* (1995), Publisher: North-Holland 283,
DOI: 10.1016/0009-2614(95)00621-A.

[176] K. Eichkorn, F. Weigend, O. Treutler, and R. Ahlrichs, *Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials*,
Theoretical Chemistry Accounts **97** *1* (1997), Publisher: Springer New York 119,
DOI: 10.1007/s002140050244.

[177] F. Weigend, *A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency*,
Physical Chemistry Chemical Physics **4** *18* (2002), Publisher: The Royal Society of Chemistry 4285,
DOI: 10.1039/B204199P.

[178] F. Neese, F. Wennmohs, A. Hansen, and U. Becker,
Efficient, approximate and parallel Hartree-Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree-Fock exchange,
Chemical Physics **356** *1* (2009), Publisher: North-Holland 98,
DOI: 10.1016/j.chemphys.2008.10.036.

[179] B. Helmich-Paris, B. de Souza, F. Neese, and R. Izsák,
An improved chain of spheres for exchange algorithm,
The Journal of Chemical Physics **155** *10* (2021) 104109, DOI: 10.1063/5.0058766.

[180] C. Ochsenfeld, C. A. White, and M. Head-Gordon,
Linear and sublinear scaling formation of Hartree–Fock-type exchange matrices,
The Journal of Chemical Physics **109** *5* (1998) 1663, DOI: 10.1063/1.476741.

[181] A. Sodt and M. Head-Gordon,
Hartree-Fock exchange computed using the atomic resolution of the identity approximation,
The Journal of Chemical Physics **128** *10* (2008) 104106, DOI: 10.1063/1.2828533.

[182] J. Almlöf, K. Faegri Jr., and K. Korsell,
Principles for a direct SCF approach to LICAO–MOab-initio calculations,
Journal of Computational Chemistry **3** *3* (1982) 385, DOI: 10.1002/jcc.540030314.

[183] A. D. Daniels and G. E. Scuseria, *What is the best alternative to diagonalization of the Hamiltonian in large scale semiempirical calculations?*
The Journal of Chemical Physics **110** *3* (1999) 1321, DOI: 10.1063/1.478008.

[184] S. Goedecker, *Linear scaling electronic structure methods*,
Reviews of Modern Physics **71** *4* (1999), Publisher: American Physical Society 1085,
DOI: 10.1103/RevModPhys.71.1085.

[185] D. Richters and T. D. Kühne,
Self-consistent field theory based molecular dynamics with linear system-size scaling,
The Journal of Chemical Physics **140** *13* (2014) 134109, DOI: 10.1063/1.4869865.

[186] C. Ochsenfeld, J. Kussmann, and D. S. Lambrecht,
"Linear-Scaling Methods in Quantum Chemistry," *Reviews in Computational Chemistry*, Section: 1,
John Wiley & Sons, Ltd, 2007 1, DOI: 10.1002/9780470116449.ch1.

[187] Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal, and P. Ren, *Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications*, Annual Review of Biophysics **48** (Volume 48, 2019 2019), Publisher: Annual Reviews 371, DOI: 10.1146/annurev-biophys-070317-033349.

[188] A. Croitoru, S.-J. Park, A. Kumar, J. Lee, W. Im, A. D. J. MacKerell, and A. Aleksandrov, *Additive CHARMM36 Force Field for Nonstandard Amino Acids*, Journal of Chemical Theory and Computation **17** *6* (2021), Publisher: American Chemical Society 3554, DOI: 10.1021/acs.jctc.1c00254.

[189] S. Spicher and S. Grimme, *Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems*, Angewandte Chemie **132** *36* (2020) 15795, DOI: 10.1002/ange.202004239.

[190] D. Anstine, R. Zubatyuk, and O. Isayev, *AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs*, 2024, DOI: 10.26434/chemrxiv-2023-296ch-v3.

[191] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, *MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields*, Advances in Neural Information Processing Systems **35** (2022) 11423.

[192] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, *Machine Learning Force Fields*, Chemical Reviews **121** *16* (2021), Publisher: American Chemical Society 10142, DOI: 10.1021/acs.chemrev.0c01111.

[193] Y. Yang, S. Zhang, K. D. Ranasinghe, O. Isayev, and A. E. Roitberg, *Machine Learning of Reactive Potentials*, Annual Review of Physical Chemistry **75** (Volume 75, 2024 2024), Publisher: Annual Reviews 371, DOI: 10.1146/annurev-physchem-062123-024417.

[194] S. Boys and F. Bernardi, *The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors*, Molecular Physics **19** *4* (1970), Publisher: Taylor & Francis 553, DOI: 10.1080/00268977000101561.

[195] H. Kruse and S. Grimme, *A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems*, Journal of Chemical Physics **136** *15* (2012), Publisher: American Institute of PhysicsAIP 154101, DOI: 10.1063/1.3700154.

[196] E. Torres and G. A. Dilabio, *A (nearly) universally applicable method for modeling noncovalent interactions using B3LYP*, Journal of Physical Chemistry Letters **3** *13* (2012), Publisher: American Chemical Society 1738, DOI: 10.1021/jz300554y.

[197] J. A. Van Santen and G. A. Dilabio, *Dispersion Corrections Improve the Accuracy of Both Noncovalent and Covalent Interactions Energies Predicted by a Density-Functional Theory Approximation*, Journal of Physical Chemistry A **119** *25* (2015), Publisher: American Chemical Society 6703, DOI: 10.1021/acs.jpca.5b02809.

[198] R. Sure and S. Grimme, *Corrected small basis set Hartree-Fock method for large systems*, Journal of Computational Chemistry **34** *19* (2013), Publisher: John Wiley & Sons, Ltd 1672, DOI: 10.1002/jcc.23317.

[199] P.-F. Loos, B. Pradines, A. Scemama, J. Toulouse, and E. Giner, *A Density-Based Basis-Set Correction for Wave Function Theory*, The Journal of Physical Chemistry Letters **10** *11* (2019), Publisher: American Chemical Society 2931, DOI: 10.1021/acs.jpclett.9b01176.

[200] D. Mester and M. Kállay, *Near-Basis-Set-Limit Double-Hybrid DFT Energies with Exceptionally Low Computational Costs*, The Journal of Physical Chemistry Letters (2025), Publisher: American Chemical Society 2136, DOI: 10.1021/acs.jpclett.5c00122.

[201] W. Klopper and W. Kutzelnigg, *Møller-plesset calculations taking care of the correlation CUSP*, Chemical Physics Letters **134** *1* (1987) 17, DOI: 10.1016/0009-2614(87)80005-2.

[202] D. P. Tew and W. Klopper, *New correlation factors for explicitly correlated electronic wave functions*, The Journal of Chemical Physics **123** *7* (2005) 074101, DOI: 10.1063/1.1999632.

[203] J. A. Pople, M. Head-Gordon, D. J. Fox, K. Raghavachari, and L. A. Curtiss, *Gaussian-1 theory: A general procedure for prediction of molecular energies*, The Journal of Chemical Physics **90** *10* (1989), Publisher: American Institute of PhysicsAIP 5622, DOI: 10.1063/1.456415.

[204] L. A. Curtiss, K. Raghavachari, G. W. Trucks, and J. A. Pople, *Gaussian-2 theory for molecular energies of first- and second-row compounds*, The Journal of Chemical Physics **94** *11* (1991), Publisher: American Institute of PhysicsAIP ISBN: 10.1063/1.477422 7221, DOI: 10.1063/1.460205.

[205] L. A. Curtiss, K. Raghavach Ari, P. C. Redfern, V. Rassolov, and J. A. Pople, *Gaussian-3 (G3) theory for molecules containing first and second-row atoms*, Journal of Chemical Physics **109** *18* (1998), Publisher: American Institute of PhysicsAIP ISBN: 10.1063/1.477422 7764, DOI: 10.1063/1.477422.

[206] L. A. Curtiss, P. C. Redfern, and K. Raghavachari, *Gaussian-4 theory*, Journal of Chemical Physics **126** *8* (2007), DOI: 10.1063/1.2436888.

[207] J. M. Martin and G. De Oliveira, *Towards standard methods for benchmark quality ab initio thermochemistry - W1 and W2 theory*, Journal of Chemical Physics **111** *5* (1999), Publisher: American Institute of Physics Inc. 1843, DOI: 10.1063/1.479454.

[208] A. Daniel Boese, M. Oren, O. Atasoylu, J. M. Martin, M. Kállay, and J. Gauss, *W3 theory: Robust computational thermochemistry in the kJ/mol accuracy range*, Journal of Chemical Physics **120** *9* (2004) 4129, DOI: 10.1063/1.1638736.

[209] A. Karton, E. Rabinovich, J. M. Martin, and B. Ruscic, *W4 theory for computational thermochemistry: In pursuit of confident sub-kJ/mol predictions*, Journal of Chemical Physics **125** *14* (2006), Publisher: American Institute of PhysicsAIP 144108, DOI: 10.1063/1.2348881.

[210] J. M. Martin, *Ab initio total atomization energies of small molecules - Towards the basis set limit*, Chemical Physics Letters **259** *5* (1996) 669, DOI: 10.1016/0009-2614(96)00898-6.

[211] A. Schäfer, H. Horn, and R. Ahlrichs,
Fully optimized contracted Gaussian basis sets for atoms Li to Kr,
The Journal of Chemical Physics **97** *4* (1992) 2571, DOI: 10.1063/1.463096.

[212] J. G. Brandenburg, E. Caldeweyher, and S. Grimme, *Screened exchange hybrid density functional for accurate and efficient structures and interaction energies*, Physical Chemistry Chemical Physics **18** *23* (2016), Publisher: The Royal Society of Chemistry 15519,
DOI: 10.1039/c6cp01697a.

[213] J. G. Brandenburg, C. Bannwarth, A. Hansen, and S. Grimme,
B97-3c: A revised low-cost variant of the B97-D density functional method,
Journal of Chemical Physics **148** *6* (2018), Publisher: American Institute of Physics Inc. 64104,
DOI: 10.1063/1.5012601.

[214] S. Grimme, A. Hansen, S. Ehlert, and J. M. Mewes,
R2SCAN-3c: A "swiss army knife" composite electronic-structure method,
Journal of Chemical Physics **154** *6* (2021) 64103, DOI: 10.1063/5.0040021.

[215] P. O. Dral, B. Hourahine, and S. Grimme, *Modern semiempirical electronic structure methods*, The Journal of Chemical Physics **160** *4* (2024) 040401, DOI: 10.1063/5.0196138.

[216] G. Seifert and J.-O. Joswig,
Density-functional tight binding—an approximate density-functional theory method,
WIREs Computational Molecular Science **2** *3* (2012) 456, DOI: 10.1002/wcms.1094.

[217] E. Hückel, *Quantentheoretische Beiträge zum Benzolproblem*,
Zeitschrift für Physik **70** *3* (1931) 204, DOI: 10.1007/BF01339530.

[218] R. Pariser and R. G. Parr, *A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. I.*, The Journal of Chemical Physics **21** *3* (1953) 466, DOI: 10.1063/1.1698929.

[219] R. Pariser and R. G. Parr, *A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. II*,
The Journal of Chemical Physics **21** *5* (1953) 767, DOI: 10.1063/1.1699030.

[220] J. A. Pople, *Electron interaction in unsaturated hydrocarbons*,
Transactions of the Faraday Society **49** *0* (1953), Publisher: The Royal Society of Chemistry 1375,
DOI: 10.1039/TF9534901375.

[221] R. Hoffmann, *An Extended Hückel Theory. I. Hydrocarbons*,
The Journal of Chemical Physics **39** *6* (1963) 1397, DOI: 10.1063/1.1734456.

[222] S. L. Altmann and C. A. Coulson,
π-σ Electronic states in molecules. I. The Hückel approximation,
Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **210** *1102* (1952), Publisher: Royal Society 327, DOI: 10.1098/rspa.1952.0004.

[223] P. Bultinck, H. D. Winter, W. Langenaeker, and J. P. Tollenare, eds.,
Computational Medicinal Chemistry for Drug Discovery, Boca Raton: CRC Press, 2003, 1169 pp.

[224] *Zero differential overlap*, *Wikipedia*, Page Version ID: 1030852890, 2021.

[225] K. Jorner, R. Pollice, C. Lavigne, and A. Aspuru-Guzik,
Ultrafast Computational Screening of Molecules with Inverted Singlet–Triplet Energy Gaps Using the Pariser–Parr–Pople Semiempirical Quantum Chemistry Method,
The Journal of Physical Chemistry A **128** *12* (2024), Publisher: American Chemical Society 2445, DOI: 10.1021/acs.jpca.3c06357.

[226] K. Ohno, *Some remarks on the Pariser-Parr-Pople method*, Theoretica chimica acta **2** *3* (1964) 219, DOI: 10.1007/BF00528281.

[227] R. Podeszwa, S. A. Kucharski, and L. Z. Stolarczyk, *Electronic correlation in cyclic polyenes. Performance of coupled-cluster methods with higher excitations*,
The Journal of Chemical Physics **116** *2* (2002) 480, DOI: 10.1063/1.1425408.

[228] M. Bedogni, D. Giavazzi, F. Di Maiolo, and A. Painelli,
Shining Light on Inverted Singlet–Triplet Emitters, Journal of Chemical Theory and Computation **20** *2* (2024), Publisher: American Chemical Society 902, DOI: 10.1021/acs.jctc.3c01112.

[229] M. Lee, K. Leiter, C. Eisner, J. Crone, and J. Knap, *Extended Hückel and Slater's rule initial guess for real space grid-based density functional theory*,
Computational and Theoretical Chemistry **1062** (2015) 24, DOI: 10.1016/j.comptc.2015.03.011.

[230] S. Lehtola, *Assessment of Initial Guesses for Self-Consistent Field Calculations. Superposition of Atomic Potentials: Simple yet Efficient*, Journal of Chemical Theory and Computation **15** *3* (2019), Publisher: American Chemical Society 1593, DOI: 10.1021/acs.jctc.8b01089.

[231] J. A. Pople and G. A. Segal, *Approximate Self-Consistent Molecular Orbital Theory. III. CNDO Results for AB2 and AB3 Systems*, The Journal of Chemical Physics **44** *9* (1966) 3289,
DOI: 10.1063/1.1727227.

[232] J. A. Pople, D. L. Beveridge, and P. A. Dobosh, *Approximate Self-Consistent Molecular-Orbital Theory. V. Intermediate Neglect of Differential Overlap*,
The Journal of Chemical Physics **47** *6* (1967) 2026, DOI: 10.1063/1.1712233.

[233] J. A. Pople, D. P. Santry, and G. A. Segal,
Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures,
The Journal of Chemical Physics **43** *10* (1965) S129, DOI: 10.1063/1.1701475.

[234] J. Pople and D. Beveridge, *Approximate molecular orbital theory*,
McGraw-Hill series in advanced chemistry, McGraw-Hill, 1970, 214 pp.

[235] M. J. S. Dewar and W. Thiel,
Ground states of molecules. 38. The MNDO method. Approximations and parameters, Journal of the American Chemical Society **99** *15* (1977), Publisher: American Chemical Society 4899,
DOI: 10.1021/ja00457a004.

[236] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart,
Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model, Journal of the American Chemical Society **107** *13* (1985), Publisher: American Chemical Society 3902, DOI: 10.1021/ja00299a024.

[237] J. J. P. Stewart, *Optimization of parameters for semiempirical methods II. Applications*,
Journal of Computational Chemistry **10** *2* (1989) 221, DOI: 10.1002/jcc.540100209.

[238] J. J. P. Stewart, *Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements*, Journal of Molecular Modeling **13** *12* (2007) 1173, DOI: 10.1007/s00894-007-0233-4.

[239] M. Kolb and W. Thiel,
Beyond the MNDO model: Methodical considerations and numerical results,
Journal of Computational Chemistry **14** *7* (1993) 775, DOI: 10.1002/jcc.540140704.

[240] W. Weber and W. Thiel, *Orthogonalization corrections for semiempirical methods*,
Theoretical Chemistry Accounts **103** *6* (2000) 495, DOI: 10.1007/s002149900083.

[241] M. Scholten, *Semiempirische Verfahren mit Orthogonalisierungskorrekturen: Die OM3 Methode*,
PhD thesis: Universität Düsseldorf, 2003.

[242] P. O. Dral, X. Wu, L. Spörkel, A. Koslowski, W. Weber, R. Steiger, M. Scholten, and W. Thiel,
Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Theory, Implementation, and Parameters, Journal of Chemical Theory and Computation **12** *3* (2016),
Publisher: American Chemical Society 1082, DOI: 10.1021/acs.jctc.5b01046.

[243] Z. Wang and F. Neese, *Development of NOTCH, an all-electron, beyond-NDDO semiempirical method: Application to diatomic molecules*,
The Journal of Chemical Physics **158** *18* (2023) 184102, DOI: 10.1063/5.0141686.

[244] R. Dixon, *Approximate self-consistent field molecular orbital calculations for valence shell electronic states*, Molecular Physics **12** *1* (1967) 83, DOI: 10.1080/00268976700100091.

[245] A. D. Bacon and M. C. Zerner, *An intermediate neglect of differential overlap theory for transition metal complexes: Fe, Co and Cu chlorides*, Theoretica chimica acta **53** *1* (1979) 21,
DOI: 10.1007/BF00547605.

[246] D. N. Nanda and K. Jug, *SINDO1. A semiempirical SCF MO method for molecular binding energy and geometry I. Approximations and parametrization*,
Theoretica chimica acta **57** *2* (1980) 95, DOI: 10.1007/BF00574898.

[247] B. Ahlswede and K. Jug, *Consistent modifications of SINDO1: I. Approximations and parameters*,
Journal of Computational Chemistry **20** *6* (1999) 563,
DOI: 10.1002/(SICI)1096-987X(19990430)20:6<563::AID-JCC1>3.0.CO;2-2.

[248] K. Jug, G. Geudtner, and T. Homann,
MSINDO parameterization for third-row main group elements,
Journal of Computational Chemistry **21** *11* (2000) 974,
DOI: 10.1002/1096-987X(200008)21:11<974::AID-JCC7>3.0.CO;2-X.

[249] T. Bredow, G. Geudtner, and K. Jug, *MSINDO parameterization for third-row transition metals*,
Journal of Computational Chemistry **22** *8* (2001) 861, DOI: 10.1002/jcc.1051.

[250] J. C. Slater and G. F. Koster, *Simplified LCAO Method for the Periodic Potential Problem*,
Physical Review **94** *6* (1954), Publisher: American Physical Society 1498,
DOI: 10.1103/PhysRev.94.1498.

[251] K. Laasonen and R. M. Nieminen, *Molecular dynamics using the tight-binding approximation*,
Journal of Physics: Condensed Matter **2** *6* (1990) 1509, DOI: 10.1088/0953-8984/2/6/010.

[252] C. H. Xu, C. Z. Wang, C. T. Chan, and K. M. Ho, *A transferable tight-binding potential for carbon*,
Journal of Physics: Condensed Matter **4** *28* (1992) 6047, DOI: 10.1088/0953-8984/4/28/006.

[253] K. Ishida, H. Aoki, and T. Chikyu,
One-dimensional exciton in a two-band tight-binding model with long-range interactions,
Physical Review B **47** *12* (1993), Publisher: American Physical Society 7594,
DOI: 10.1103/PhysRevB.47.7594.

[254] M. Sugawara, *Tight-binding model including cation d orbitals to predict valence-band offset in zinc-blende semiconductor heterojunctions*,
Physical Review B **47** *12* (1993), Publisher: American Physical Society 7588,
DOI: 10.1103/PhysRevB.47.7588.

[255] M. Menon and K. R. Subbaswamy,
Nonorthogonal tight-binding molecular-dynamics study of silicon clusters,
Physical Review B **47** *19* (1993), Publisher: American Physical Society 12754,
DOI: 10.1103/PhysRevB.47.12754.

[256] Neil W. Ashcroft, *Solid state physics*, in collab. with Internet Archive,
Orlando: Holt, Rinehart and Winston, 1976, 858 pp.

[257] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, *Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon*,
Physical Review B **51** *19* (1995), Publisher: American Physical Society 12947,
DOI: 10.1103/PhysRevB.51.12947.

[258] G. Seifert and H. Eschrig, *LCAO-Xα Calculations of Transition Metal Clusters*,
physica status solidi (b) **127** *2* (1985) 573, DOI: 10.1002/pssb.2221270218.

[259] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, *Extended tight-binding quantum chemistry methods*,
Wiley Interdisciplinary Reviews: Computational Molecular Science **11** *2* (2021) 1,
DOI: 10.1002/wcms.1493.

[260] O. K. Andersen and O. Jepsen, *Explicit, First-Principles Tight-Binding Theory*,
Physical Review Letters **53** *27* (1984), Publisher: American Physical Society 2571,
DOI: 10.1103/PhysRevLett.53.2571.

[261] E. Artacho and F. Ynduráin, *Nonparametrized tight-binding method for local and extended defects in homopolar semiconductors*,
Physical Review B **44** *12* (1991), Publisher: American Physical Society 6169,
DOI: 10.1103/PhysRevB.44.6169.

[262] R. E. Cohen, M. J. Mehl, and D. A. Papaconstantopoulos,
Tight-binding total-energy method for transition and noble metals,
Physical Review B **50** *19* (1994), Publisher: American Physical Society 14694,
DOI: 10.1103/PhysRevB.50.14694.

[263] G. Seifert, D. Porezag, and T. Frauenheim,
Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme,
International Journal of Quantum Chemistry **58** *2* (1996) 185,
DOI: 10.1002/(SICI)1097-461X(1996)58:2<185::AID-QUA7>3.0.CO;2-U.

[264] K. Takegahara, Y. Aoki, and A. Yanase, *Slater-Koster tables for f electrons*,
Journal of Physics C: Solid State Physics **13** *4* (1980) 583, DOI: 10.1088/0022-3719/13/4/016.

[265] J. D. Shore and D. A. Papaconstantopoulos,
Slater-Koster parametrization of the band structure of TiNi,
Journal of Physics and Chemistry of Solids **45** *4* (1984) 439,
DOI: 10.1016/0022-3697(84)90152-5.

[266] R. Kashikar, M. Gupta, and B. R. K. Nanda, *A generic Slater–Koster description of the electronic structure of centrosymmetric halide perovskites*,
The Journal of Chemical Physics **154** *10* (2021) 104706, DOI: 10.1063/5.0044338.

[267] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, *Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties*, Physical Review B **58** *11* (1998), Publisher: American Physical Society 7260, DOI: 10.1103/PhysRevB.58.7260.

[268] M. Gaus, Q. Cui, and M. Elstner, *DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)*, Journal of Chemical Theory and Computation **7** *4* (2011), Publisher: American Chemical Society 931, DOI: 10.1021/ct100684s.

[269] Q. Cui and M. Elstner,
Density functional tight binding: Values of semi-empirical methods in an ab initio era,
Physical Chemistry Chemical Physics **16** *28* (2014) 14368, DOI: 10.1039/c4cp00908h.

[270] R. G. Parr and R. G. Pearson,
Absolute hardness: companion parameter to absolute electronegativity, Journal of the American Chemical Society **105** *26* (1983), Publisher: American Chemical Society 7512,
DOI: 10.1021/ja00364a005.

[271] M. Gaus, H. Jin, D. Demapan, A. S. Christensen, P. Goyal, M. Elstner, and Q. Cui,
DFTB3 Parametrization for Copper: The Importance of Orbital Angular Momentum Dependence of Hubbard Parameters, Journal of Chemical Theory and Computation **11** *9* (2015), Publisher: American Chemical Society 4205, DOI: 10.1021/acs.jctc.5b00600.

[272] S. Sanna, B. Hourahine, T. Frauenheim, and U. Gerstmann,
Theoretical study of rare earth point defects in GaN, Physica Status Solidi C **5** *6* (2008) 2358,
DOI: 10.1002/pssc.200778667.

[273] V. Q. Vuong, J. M. L. Madridejos, B. Aradi, B. G. Sumpter, G. F. Metha, and S. Irle,
Density-functional tight-binding for phosphine-stabilized nanoscale gold clusters,
Chemical Science **11** *48* (2020), Publisher: The Royal Society of Chemistry 13113,
DOI: 10.1039/D0SC04514D.

[274] J. Frenzel, A. F. Oliveira, H. A. Duarte, T. Heine, and G. Seifert,
Structural and Electronic Properties of Bulk Gibbsite and Gibbsite Surfaces,
Zeitschrift für anorganische und allgemeine Chemie **631** *6* (2005) 1267,
DOI: 10.1002/zaac.200500051.

[275] M. Wahiduzzaman, A. F. Oliveira, P. Philipsen, L. Zhechkov, E. van Lenthe, H. A. Witek, and T. Heine, *DFTB Parameters for the Periodic Table: Part 1, Electronic Structure*, Journal of Chemical Theory and Computation **9** *9* (2013), Publisher: American Chemical Society 4006,
DOI: 10.1021/ct4004959.

[276] A. F. Oliveira, P. Philipsen, and T. Heine, *DFTB Parameters for the Periodic Table, Part 2: Energies and Energy Gradients from Hydrogen to Calcium*, Journal of Chemical Theory and Computation **11** *11* (2015), Publisher: American Chemical Society 5209, DOI: 10.1021/acs.jctc.5b00702.

[277] S. Grimme, C. Bannwarth, and P. Shushkov, *A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86)*, Journal of Chemical Theory and Computation **13** *5* (2017), Publisher: American Chemical Society 1989, DOI: 10.1021/acs.jctc.7b00118.

[278] C. Bannwarth, S. Ehlert, and S. Grimme, *GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions*, Journal of Chemical Theory and Computation **15** *3* (2019) 1652, DOI: 10.1021/acs.jctc.8b01176.

[279] P. Pracht, E. Caldeweyher, S. Ehlert, and S. Grimme, *A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules*, 2019, DOI: 10.26434/chemrxiv.8326202.v1.

[280] G. Klopman, *A Semiempirical Treatment of molecular Structures. II. Molecular Terms and Application to diatomic Molecules*, Journal of the American Chemical Society **86** *21* (1964), Publisher: American Chemical Society 4550, DOI: 10.1021/ja01075a008.

[281] J. Řezáč and P. Hobza, *Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods*, Journal of Chemical Theory and Computation **8** *1* (2012), Publisher: American Chemical Society 141, DOI: 10.1021/ct200751e.

[282] R. S. Mulliken, *Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I*, The Journal of Chemical Physics **23** *10* (1955) 1833, DOI: 10.1063/1.1740588.

[283] H. Neugebauer, B. Bädorf, S. Ehlert, A. Hansen, and S. Grimme, *High-throughput screening of spin states for transition metal complexes with spin-polarized extended tight-binding methods*, Journal of Computational Chemistry **44** *27* (2023) 2120, DOI: 10.1002/jcc.27185.

[284] J. Gorges and S. Grimme, *QCxMS2 - a program for the calculation of electron ion- ization mass spectra via automated reaction network dis- covery*, 2025, DOI: 10.26434/chemrxiv-2025-277zm.

[285] S. Spicher, E. Caldeweyher, A. Hansen, and S. Grimme, *Benchmarking London dispersion corrected density functional theory for noncovalent ion-π interactions*, Physical Chemistry Chemical Physics **23** *20* (2021), Publisher: The Royal Society of Chemistry 11635, DOI: 10.1039/d1cp01333e.

[286] J. G. Brandenburg, T. Maas, and S. Grimme, *Benchmarking DFT and semiempirical methods on structures and lattice energies for ten ice polymorphs*, The Journal of Chemical Physics **142** *12* (2015) 124104, DOI: 10.1063/1.4916070.

[287] J. G. Brandenburg, M. Hochheim, T. Bredow, and S. Grimme, *Low-Cost Quantum Chemical Methods for Noncovalent Interactions*, The Journal of Physical Chemistry Letters **5** *24* (2014), Publisher: American Chemical Society 4275, DOI: 10.1021/jz5021313.

[288] K. Kříž and J. Řezáč, *Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design*,
Journal of Chemical Information and Modeling **60** *3* (2020), Publisher: UTC 1453,
DOI: 10.1021/acs.jcim.9b01171.

[289] S. Spicher, M. Bursch, and S. Grimme, *Efficient Calculation of Small Molecule Binding in Metal–Organic Frameworks and Porous Organic Cages*,
The Journal of Physical Chemistry C **124** *50* (2020), Publisher: American Chemical Society 27529,
DOI: 10.1021/acs.jpcc.0c08617.

[290] L. R. Maurer, M. Bursch, S. Grimme, and A. Hansen, *Assessing Density Functional Theory for Chemically Relevant Open-Shell Transition Metal Reactions*, Journal of Chemical Theory and Computation **17** *10* (2021), Publisher: American Chemical Society 6134,
DOI: 10.1021/acs.jctc.1c00659.

[291] C. Plett, S. Grimme, and A. Hansen, *Conformational energies of biomolecules in solution: Extending the MPCONF196 benchmark with explicit water molecules*,
Journal of Computational Chemistry **45** *7* (2024) 419, DOI: 10.1002/jcc.27248.

[292] S. Ehlert, S. Grimme, and A. Hansen,
Conformational Energy Benchmark for Longer n-Alkane Chains,
The Journal of Physical Chemistry A **126** *22* (2022), Publisher: American Chemical Society 3521,
DOI: 10.1021/acs.jpca.2c02439.

[293] W. J. Hehre, R. F. Stewart, and J. A. Pople, *Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals*,
The Journal of Chemical Physics **51** *6* (1969) 2657, DOI: 10.1063/1.1672392.

[294] S. G. Balasubramani et al., *TURBOMOLE: Modular program suite for ab initio quantum-chemical and condensed-matter simulations*,
Journal of Chemical Physics **152** *18* (2020), Publisher: American Institute of Physics Inc.,
DOI: 10.1063/5.0004635.

[295] *TURBOMOLE V7.7.1, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2023, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.* Karlsruhe, 2023.

[296] H. Tatewaki and S. Huzinaga, *A systematic preparation of new contracted Gaussian-type orbital sets. III. Second-row atoms from Li through ne*,
Journal of Computational Chemistry **1** *3* (1980) 205, DOI: 10.1002/jcc.540010302.

[297] A. Karton and J. M. L. Martin, *Explicitly correlated Wn theory: W1-F12 and W2-F12*,
The Journal of Chemical Physics **136** *12* (2012) 124114, DOI: 10.1063/1.3697678.

[298] Y. Zhao, N. González-García, and D. G. Truhlar,
Benchmark Database of Barrier Heights for Heavy Atom Transfer, Nucleophilic Substitution, Association, and Unimolecular Reactions and Its Use to Test Theoretical Methods,
The Journal of Physical Chemistry A **109** *9* (2005), Publisher: American Chemical Society 2012,
DOI: 10.1021/jp045141s.

[299] S. Spicher and S. Grimme, *Single-Point Hessian Calculations for Improved Vibrational Frequencies and Rigid-Rotor-Harmonic-Oscillator Thermodynamics*, Journal of Chemical Theory and Computation **17** *3* (2021), Publisher: American Chemical Society 1701, DOI: 10.1021/acs.jctc.0c01306.

[300] S. Ehlert, M. Stahn, S. Spicher, and S. Grimme, *Robust and efficient implicit solvation model for fast semiempirical methods*, Journal of Chemical Theory and Computation **17** *7* (2021), Publisher: American Chemical Society 4250, DOI: 10.1021/acs.jctc.1c00471.

[301] M. Korth and S. Grimme, *Mindless DFT benchmarking*, Journal of Chemical Theory and Computation **5** *4* (2009), Publisher: American Chemical Society 993, DOI: 10.1021/ct800511q.

[302] D. E. Taylor et al., *Blind test of density-functional-based methods on intermolecular interaction energies*, Journal of Chemical Physics **145** *12* (2016), Publisher: AIP Publishing LLCAIP Publishing 124105, DOI: 10.1063/1.4961095.

[303] R. Sedlak, T. Janowski, M. Pitoňák, J. Řezáč, P. Pulay, and P. Hobza, *Accuracy of quantum chemical methods for large noncovalent complexes*, Journal of Chemical Theory and Computation **9** *8* (2013), Publisher: American Chemical Society 3364, DOI: 10.1021/ct400036b.

[304] D. I. Sharapa, A. Genaev, L. Cavallo, and Y. Minenkov, *A Robust and Cost-Efficient Scheme for Accurate Conformational Energies of Organic Molecules*, ChemPhysChem **20** *1* (2019), Publisher: Wiley-VCH Verlag 92, DOI: 10.1002/cphc.201801063.

[305] J. Řezáč, D. Bím, O. Gutten, and L. Rulíšek, *Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set*, Journal of Chemical Theory and Computation **14** *3* (2018), Publisher: American Chemical Society 1254, DOI: 10.1021/acs.jctc.7b01074.

[306] R. Sure and S. Grimme, *Comprehensive Benchmark of Association (Free) Energies of Realistic Host-Guest Complexes*, Journal of Chemical Theory and Computation **11** *8* (2015), Publisher: American Chemical Society 3785, DOI: 10.1021/acs.jctc.5b00296.

[307] A. Najibi and L. Goerigk, *DFT-D4 counterparts of leading meta-generalized-gradient approximation and hybrid density functionals for energetics and geometries*, Journal of Computational Chemistry **41** *30* (2020), Publisher: John Wiley and Sons Inc 2562, DOI: 10.1002/jcc.26411.

[308] A. Najibi and L. Goerigk, *The Nonlocal Kernel in van der Waals Density Functionals as an Additive Correction: An Extensive Analysis with Special Emphasis on the B97M-V and ωb97M-V Approaches*, Journal of Chemical Theory and Computation **14** *11* (2018) 5725, DOI: 10.1021/acs.jctc.8b00842.

[309] F. Neese, *Software update: The ORCA program system—Version 5.0*, Wiley Interdisciplinary Reviews: Computational Molecular Science **12** *5* (2022), Publisher: John Wiley and Sons Inc e1606, DOI: 10.1002/wcms.1606.

[310] D. G. A. Smith et al., *PSI4 1.4: Open-source software for high-throughput quantum chemistry*, The Journal of Chemical Physics **152** *18* (2020) 184108, DOI: 10.1063/5.0006002.

[311] C. T. Ser, H. Hao, S. Pablo-García, K. Jorner, S. Li, R. Pollice, and A. Aspuru-Guzik, *Bulky phosphine ligands promote palladium-catalysed protodeboronation*, 2024, DOI: 10.26434/chemrxiv-2024-cw8cs.

[312] A. A. Schoepfer, R. Laplaza, M. D. Wodrich, J. Waser, and C. Corminboeuf, *Reaction-Agnostic Featurization of Bidentate Ligands for Bayesian Ridge Regression of Enantioselectivity*, ACS Catalysis **14** *12* (2024), Publisher: American Chemical Society 9302, DOI: 10.1021/acscatal.4c02452.

[313] D. Abdullin, N. Fleck, C. Klein, P. Brehm, S. Spicher, A. Lützen, S. Grimme, and O. Schiemann, *Synthesis of μ 2 -Oxo-Bridged Iron(III) Tetraphenylporphyrin–Spacer–Nitroxide Dimers and their Structural and Dynamics Characterization by using EPR and MD Simulations*, Chemistry - A European Journal **25** *10* (2019) 2586, DOI: 10.1002/chem.201805016.

[314] P. Pracht, D. F. Grant, and S. Grimme, *Comprehensive Assessment of GFN Tight-Binding and Composite Density Functional Theory Methods for Calculating Gas-Phase Infrared Spectra*, Journal of Chemical Theory and Computation **16** *11* (2020) 7044, DOI: 10.1021/acs.jctc.0c00877.

[315] S. Vuckovic, S. Song, J. Kozlowski, E. Sim, and K. Burke, *Density Functional Analysis: The Theory of Density-Corrected DFT*, Journal of Chemical Theory and Computation **15** *12* (2019), Publisher: American Chemical Society 6636, DOI: 10.1021/acs.jctc.9b00826.

[316] B. Hammer, L. B. Hansen, and J. K. Nørskov, *Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals*, Physical Review B - Condensed Matter and Materials Physics **59** *11* (1999) 7413, DOI: 10.1103/PhysRevB.59.7413.

[317] F. L. Hirshfeld, *Bonded-atom fragments for describing molecular charge densities*, Theoretica chimica acta **44** *2* (1977) 129, DOI: 10.1007/BF00549096.

[318] A. E. Reed, R. B. Weinstock, and F. Weinhold, *Natural population analysis*, The Journal of Chemical Physics **83** *2* (1985) 735, DOI: 10.1063/1.449486.

[319] K. B. Wiberg and P. R. Rablen, *Comparison of atomic charges derived via different procedures*, Journal of Computational Chemistry **14** *12* (1993) 1504, DOI: 10.1002/jcc.540141213.

[320] L. Wittmann, I. Gordiy, M. Friede, B. Helmich-Paris, S. Grimme, A. Hansen, and M. Bursch, *Extension of the D3 and D4 London dispersion corrections to the full actinides series*, Physical Chemistry Chemical Physics **26** *32* (2024) 21379, DOI: 10.1039/D4CP01514B.

[321] J. W. Ponder et al., *Current Status of the AMOEBA Polarizable Force Field*, The Journal of Physical Chemistry B **114** *8* (2010), Publisher: American Chemical Society 2549, DOI: 10.1021/jp910674d.

[322] K. Vanommeslaeghe, E. P. Raman, and A. D. J. MacKerell, *Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges*, Journal of Chemical Information and Modeling **52** *12* (2012), Publisher: American Chemical Society 3155, DOI: 10.1021/ci3003649.

[323] T. Gensch et al.,
A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis, Journal of the American Chemical Society **144** *3* (2022), Publisher: American Chemical Society 1205, DOI: 10.1021/jacs.1c09718.

[324] S. Grimme and C. Bannwarth, *Ultra-fast computation of electronic spectra for large systems by tight-binding based simplified Tamm-Dancoff approximation (sTDA-xTB)*,
Journal of Chemical Physics **145** *5* (2016), Publisher: AIP Publishing LLCAIP Publishing 054103, DOI: 10.1063/1.4959605.

[325] R. L. Ellis, G. Kuehnlenz, and H. H. Jaffé, *The use of the CNDO method in spectroscopy*, Theoretica chimica acta **26** *2* (1972) 131, DOI: 10.1007/BF00526540.

[326] M. Scholz and H.-J. Koehler, *Band 3: Quantenchemische Näherungsverfahren und ihre Anwendung in der organischen Chemie*, Hüthig, 1981, 557 pp.

[327] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons, 2001, 540 pp.

[328] A. Klamt and G. Schüürmann, *COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient*, Journal of the Chemical Society, Perkin Transactions 2 *5* (1993), Publisher: The Royal Society of Chemistry 799, DOI: 10.1039/P29930000799.

[329] A. Klamt, *The COSMO and COSMO-RS solvation models*,
WIREs Computational Molecular Science **1** *5* (2011) 699, DOI: 10.1002/wcms.56.

[330] J. M. Mewes, A. Hansen, and S. Grimme, *Comment on "The Nature of Chalcogen-Bonding-Type Tellurium–Nitrogen Interactions": Fixing the Description of Finite-Temperature Effects Restores the Agreement Between Experiment and Theory*,
Angewandte Chemie - International Edition **60** *24* (2021) 13144, DOI: 10.1002/anie.202102679.

[331] H. M. Senn, J. Kästner, J. Breidung, and W. Thiel, *Finite-temperature effects in enzymatic reactions — Insights from QM/MM free-energy simulations*,
Canadian Journal of Chemistry **87** *10* (2009), Publisher: NRC Research Press 1322, DOI: 10.1139/V09-092.

[332] X. Li, J. C. Tully, H. B. Schlegel, and M. J. Frisch, *Ab initio Ehrenfest dynamics*,
The Journal of Chemical Physics **123** *8* (2005) 084106, DOI: 10.1063/1.2008258.

[333] S. Mai, P. Marquetand, and L. González, *Nonadiabatic dynamics: The SHARC approach*,
WIREs Computational Molecular Science **8** *6* (2018) e1370, DOI: 10.1002/wcms.1370.

[334] J. C. Tully, *Perspective: Nonadiabatic dynamics theory*,
The Journal of Chemical Physics **137** *22* (2012) 22A301, DOI: 10.1063/1.4757762.

[335] M. Musiał, A. Perera, and R. J. Bartlett, *Multireference coupled-cluster theory: The easy way*,
The Journal of Chemical Physics **134** *11* (2011) 114108, DOI: 10.1063/1.3567115.

[336] S. R. White, *Density matrix formulation for quantum renormalization groups*,
Physical Review Letters **69** *19* (1992), Publisher: American Physical Society 2863, DOI: 10.1103/PhysRevLett.69.2863.

[337] "Atomic Radii of the Elements," *CRC Handbook of Chemistry and Physics*, ed. by W. M. Haynes, 92nd ed., Num Pages: 2, CRC Press, 2011 2.

[338] N. Mardirossian and M. Head-Gordon, *Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals*, Molecular Physics **115** *19* (2017) 2315, DOI: 10.1080/00268976.2017.1333644.

[339] E. Caldeweyher and J. G. Brandenburg, *Simplified DFT methods for consistent structures and energies of large systems*, Journal of Physics Condensed Matter **30** *21* (2018), Publisher: IOP Publishing 213001, DOI: 10.1088/1361-648X/aabcfb.

[340] J. Kohn, M. Bursch, A. Hansen, and S. Grimme, *Computational study of ground-state properties of μ2-bridged group 14 porphyrinic sandwich complexes*, Journal of Computational Chemistry (2022), Publisher: John Wiley and Sons Inc, DOI: 10.1002/jcc.26870.

[341] J. Jayapaul, S. Komulainen, V. V. Zhivonitko, J. Mareš, C. Giri, K. Rissanen, P. Lantto, V. V. Telkki, and L. Schröder, *Hyper-CEST NMR of metal organic polyhedral cages reveals hidden diastereomers with diverse guest exchange kinetics*, Nature Communications **13** *1* (2022), Publisher: Nature Publishing Group 1, DOI: 10.1038/s41467-022-29249-w.

[342] A. Schäfer, C. Huber, and R. Ahlrichs, *Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr*, The Journal of Chemical Physics **100** *8* (1994), Publisher: American Institute of PhysicsAIP 5829, DOI: 10.1063/1.467146.

[343] T. Gasevic, J. B. Stückrath, S. Grimme, and M. Bursch, *Optimization of the r2SCAN-3c Composite Electronic-Structure Method for Use with Slater-Type Orbital Basis Sets*, The Journal of Physical Chemistry A (2022), Publisher: American Chemical Society, DOI: 10.1021/ACS.JPCA.2C02951.

[344] T. D. Kühne et al., *CP2K: An electronic structure and molecular dynamics software package -Quickstep: Efficient and accurate electronic structure calculations*, Journal of Chemical Physics **152** *19* (2020), Publisher: {AIP} Publishing 194103, DOI: 10.1063/5.0007045.

[345] E. S. Zijlstra, N. Huntemann, A. Kalitsov, M. E. Garcia, and U. Von Barth, *Optimized Gaussian basis sets for Goedecker-Teter-Hutter pseudopotentials*, Modelling and Simulation in Materials Science and Engineering **17** *1* (2009), DOI: 10.1088/0965-0393/17/1/015009.

[346] G. A. Petersson, S. Zhong, J. A. Montgomery, and M. J. Frisch, *On the optimization of gaussian basis sets*, Journal of Chemical Physics **118** *3* (2003) 1101, DOI: 10.1063/1.1516801.

[347] L. F. Pacios and P. A. Christiansen, *Ab initio relativistic effective potentials with spin-orbit operators. I. Li through Ar*, The Journal of Chemical Physics **82** *6* (1985) 2664, DOI: 10.1063/1.448263.

[348] M. M. Hurley, L. F. Pacios, P. A. Christiansen, R. B. Ross, and W. C. Ermler, *Ab initio relativistic effective potentials with spin-orbit operators. II. K through Kr*, The Journal of Chemical Physics **84** *12* (1986) 6840, DOI: 10.1063/1.450689.

[349] T. Leininger, A. Berning, A. Nicklass, H. Stoll, H. J. Werner, and H. J. Flad,
Spin-orbit interaction in heavy group 13 atoms and TlAr,
Chemical Physics **217** *1* (1997), ISBN: 15.81178166 19, DOI: 10.1016/S0301-0104(97)00043-8.

[350] D. Andrae, U. Häußermann, M. Dolg, H. Stoll, and H. Preuß,
Energy-adjusted ab initio pseudopotentials for the second and third row transition elements,
Theoretica Chimica Acta **77** *2* (1990) 123, DOI: 10.1007/BF01114537.

[351] R. B. Ross, J. M. Powers, T. Atashroo, W. C. Ermler, L. A. LaJohn, and P. A. Christiansen,
Ab initio relativistic effective potentials with spin-orbit operators. IV. Cs through Rn,
The Journal of Chemical Physics **93** *9* (1990) 6654, DOI: 10.1063/1.458934.

[352] M. Dolg, H. Stoll, and H. Preuss, *A combination of quasirelativistic pseudopotential and ligand field calculations for lanthanoid compounds*, Theoretica Chimica Acta **85** *6* (1993) 441,
DOI: 10.1007/BF01112983.

[353] T. R. Cundari and W. J. Stevens, *Effective core potential methods for the lanthanides*,
The Journal of Chemical Physics **98** *7* (1993), Publisher: American Institute of PhysicsAIP 5555,
DOI: 10.1063/1.464902.

[354] W. Küchle, M. Dolg, H. Stoll, and H. Preuss,
Ab initio pseudopotentials for hg through rn i. parameter sets and atomic calculations,
Molecular Physics **74** *6* (1991), Publisher: Taylor & Francis Group 1245,
DOI: 10.1080/00268979100102941.

[355] M. J. D. Powell, *An efficient method for finding the minimum of a function of several variables without calculating derivatives*, The Computer Journal **7** *2* (1964) 155,
DOI: 10.1093/comjnl/7.2.155.

[356] Y. S. Lin, G. D. Li, S. P. Mao, and J. D. Chai,
Long-range corrected hybrid density functionals with improved dispersion corrections, Journal of Chemical Theory and Computation **9** *1* (2013), Publisher: American Chemical Society 263,
DOI: 10.1021/ct300715s.

[357] F. Neese, F. Wennmohs, U. Becker, and C. Riplinger,
The ORCA quantum chemistry program package,
Journal of Chemical Physics **152** *22* (2020), Publisher: AIP Publishing LLCAIP Publishing 224108,
DOI: 10.1063/5.0004608.

[358] F. Weigend, *Accurate Coulomb-fitting basis sets for H to Rn*,
Physical Chemistry Chemical Physics **8** *9* (2006) 1057, DOI: 10.1039/b515623h.

[359] P. Plessow and F. Weigend,
Seminumerical calculation of the Hartree-Fock exchange matrix: Application to two-component procedures and efficient evaluation of local hybrid density functionals,
Journal of Computational Chemistry **33** *7* (2012) 810, DOI: 10.1002/jcc.22901.

[360] C. Holzer, *An improved seminumerical Coulomb and exchange algorithm for properties and excited states in modern density functional theory*,
Journal of Chemical Physics **153** *18* (2020), Publisher: AIP Publishing LLCAIP Publishing 184115,
DOI: 10.1063/5.0022755.

[361] R. Ditchfield, W. J. Hehre, and J. A. Pople, *Self-consistent molecular-orbital methods. IX. An extended gaussian-type basis for molecular-orbital studies of organic molecules*, The Journal of Chemical Physics **54** *2* (1971) 720, DOI: 10.1063/1.1674902.

[362] Z. Ni, Y. Guo, F. Neese, W. Li, and S. Li, *Cluster-in-Molecule Local Correlation Method with an Accurate Distant Pair Correction for Large Systems*, Journal of Chemical Theory and Computation **17** *2* (2021), Publisher: American Chemical Society 756, DOI: 10.1021/acs.jctc.0c00831.

[363] V. K. Prasad, Z. Pei, S. Edelmann, A. Otero-De-La-Roza, and G. A. Dilabio, *BH9, a New Comprehensive Benchmark Data Set for Barrier Heights and Reaction Energies: Assessment of Density Functional Approximations and Basis Set Incompleteness Potentials*, Journal of Chemical Theory and Computation **18** *1* (2022), Publisher: American Chemical Society 151, DOI: 10.1021/acs.jctc.1c00694.

[364] J. Řezáč, *Non-Covalent Interactions Atlas Benchmark Data Sets 2: Hydrogen Bonding in an Extended Chemical Space*, Journal of Chemical Theory and Computation **16** *10* (2020), Publisher: American Chemical Society 6305, DOI: 10.1021/acs.jctc.0c00715.

[365] V. M. Miriyala and J. Řezáč, *Testing Semiempirical Quantum Mechanical Methods on a Data Set of Interaction Energies Mapping Repulsive Contacts in Organic Molecules*, Journal of Physical Chemistry A **122** *10* (2018), Publisher: American Chemical Society 2801, DOI: 10.1021/acs.jpca.8b00260.

[366] N. Mehta, T. Fellowes, J. M. White, and L. Goerigk, *CHAL336 Benchmark Set: How Well Do Quantum-Chemical Methods Describe Chalcogen-Bonding Interactions?* Journal of Chemical Theory and Computation **17** *5* (2021) 2783, DOI: 10.1021/acs.jctc.1c00006.

[367] Y. S. Al-Hamdani, P. R. Nagy, A. Zen, D. Barton, M. Kállay, J. G. Brandenburg, and A. Tkatchenko, *Interactions between large molecules pose a puzzle for reference quantum mechanical methods*, Nature Communications **12** *1* (2021), Publisher: Nature Publishing Group 1, DOI: 10.1038/s41467-021-24119-3.

[368] M. Bursch, A. Hansen, P. Pracht, J. T. Kohn, and S. Grimme, *Theoretical study on conformational energies of transition metal complexes*, Physical Chemistry Chemical Physics **23** *1* (2021), Publisher: The Royal Society of Chemistry 287, DOI: 10.1039/d0cp04696e.

[369] S. Dohm, A. Hansen, M. Steinmetz, S. Grimme, and M. P. Checinski, *Comprehensive Thermochemical Benchmark Set of Realistic Closed-Shell Metal Organic Reactions*, Journal of Chemical Theory and Computation **14** *5* (2018), Publisher: American Chemical Society 2596, DOI: 10.1021/acs.jctc.7b01183.

[370] M. Blaško, L. F. Pašteka, and M. Urban, *DFT Functionals for Modeling of Polyethylene Chains Cross-Linked by Metal Atoms. DLPNO-CCSD(T) Benchmark Calculations*, Journal of Physical Chemistry A **125** *34* (2021), Publisher: American Chemical Society 7382, DOI: 10.1021/acs.jpca.1c04793.

[371] M. A. Iron and T. Janes, *Evaluating Transition Metal Barrier Heights with the Latest Density Functional Theory Exchange-Correlation Functionals: The MOBH35 Benchmark Database*, Journal of Physical Chemistry A **123** *17* (2019), Publisher: American Chemical Society 3761, DOI: 10.1021/acs.jpca.9b01546.

[372] E. Semidalas and J. M. Martin,
The MOBH35 Metal-Organic Barrier Heights Reconsidered: Performance of Local-Orbital Coupled Cluster Approaches in Different Static Correlation Regimes, Journal of Chemical Theory and Computation **18** *2* (2022), Publisher: American Chemical Society 883,
DOI: 10.1021/acs.jctc.1c01126.

[373] K. P. Huber and G. Herzberg, *Molecular Spectra and Molecular Structure*,
Molecular Spectra and Molecular Structure (1979), Publisher: Springer US,
DOI: 10.1007/978-1-4757-0961-2.

[374] M. Bühl, C. Reimann, D. A. Pantazis, T. Bredow, and F. Neese,
Geometries of third-row transition-metal complexes from density-functional theory, Journal of Chemical Theory and Computation **4** *9* (2008), Publisher: American Chemical Society 1449,
DOI: 10.1021/ct800172j.

[375] M. Piccardo, E. Penocchio, C. Puzzarini, M. Biczysko, and V. Barone,
Semi-Experimental Equilibrium Structure Determinations by Employing B3LYP/SNSD Anharmonic Force Fields: Validation and Application to Semirigid Organic Molecules,
Journal of Physical Chemistry A **119** *10* (2015), Publisher: UTC 2058, DOI: 10.1021/jp511432m.

[376] J. Řezáč, K. E. Riley, and P. Hobza, *S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures*, Journal of Chemical Theory and Computation **7** *8* (2011), Publisher: American Chemical Society 2427, DOI: 10.1021/ct2002946.

[377] B. Brauer, M. K. Kesharwani, S. Kozuch, and J. M. Martin, *The S66x8 benchmark for noncovalent interactions revisited: Explicitly correlated: Ab initio methods and density functional theory*,
Physical Chemistry Chemical Physics **18** *31* (2016) 20905, DOI: 10.1039/c6cp00688d.

[378] Y. Q. Chen, Y. J. Sheng, Y. Q. Ma, and H. M. Ding, *Efficient calculation of protein-ligand binding free energy using GFN methods: the power of the cluster model*,
Physical Chemistry Chemical Physics **24** *23* (2022) 14339, DOI: 10.1039/d2cp00161f.

[379] N. D. Yilmazer and M. Korth,
Enhanced semiempirical QM methods for biomolecular interactions,
Computational and Structural Biotechnology Journal **13** (2015) 169,
DOI: 10.1016/j.csbj.2015.02.004.

[380] A. S. Christensen, T. Kubař, Q. Cui, and M. Elstner, *Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications*,
Chemical Reviews **116** *9* (2016) 5301, DOI: 10.1021/acs.chemrev.5b00584.

[381] P. Zheng, R. Zubatyuk, W. Wu, O. Isayev, and P. O. Dral,
Artificial intelligence-enhanced quantum chemical method with broad applicability,
Nature Communications **12** *1* (2021), Publisher: Nature Publishing Group 1,
DOI: 10.1038/s41467-021-27340-2.

[382] F. P. Bonafé, B. Aradi, B. Hourahine, C. R. Medrano, F. J. Hernández, T. Frauenheim, and C. G. Sánchez, *A Real-Time Time-Dependent Density Functional Tight-Binding Implementation for Semiclassical Excited State Electron-Nuclear Dynamics and Pump-Probe Spectroscopy Simulations*, Journal of Chemical Theory and Computation **16** *7* (2020) 4454,
DOI: 10.1021/acs.jctc.9b01217.

[383] J. C. Kromann, J. H. Jensen, M. Kruszyk, M. Jessing, and M. Jørgensen, *Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions*, Chemical Science **9** *3* (2018) 660, DOI: 10.1039/c7sc04156j.

[384] J. Murillo-López et al., *Studying the phosphoryl transfer mechanism of the: E. coli phosphofructokinase-2: From X-ray structure to quantum mechanics/molecular mechanics simulations*, Chemical Science **10** *10* (2019), Publisher: The Royal Society of Chemistry 2882, DOI: 10.1039/c9sc00094a.

[385] Y. Yang, H. Yu, D. York, Q. Cui, and M. Elstner, *Extension of the self-consistent-charge density-functional tight-binding method: Third-order expansion of the density functional theory total energy and introduction of a modified effective coulomb interaction*, Journal of Physical Chemistry A **111** *42* (2007), Publisher: American Chemical Society 10861, DOI: 10.1021/jp074167r.

[386] T. Frauenheim, G. Seifert, M. Elstner, T. Niehaus, C. Köhler, M. Amkreutz, M. Sternberg, Z. Hajnal, A. Di Carlo, and S. Suhai, *Atomistic simulations of complex materials: Ground-state and excited-state properties*, Journal of Physics Condensed Matter **14** *11* (2002), Publisher: IOP Publishing 3015, DOI: 10.1088/0953-8984/14/11/313.

[387] Q. Cui, M. Elstner, E. Kaxiras, T. Frauenheim, and M. Karplus, *A QM/MM implementation of the self-consistent charge density functional tight binding (SCC-DFTB) method*, Journal of Physical Chemistry B **105** *2* (2001), Publisher: American Chemical Society 569, DOI: 10.1021/jp0029109.

[388] T. A. Niehaus, S. Suhai, F. Della Sala, M. Elstner, G. Seifert, and T. Frauenheim, *Tight-binding approach to time-dependent density-functional response theory*, Physical Review B - Condensed Matter and Materials Physics **63** *8* (2001), DOI: 10.1103/PhysRevB.63.085108.

[389] A. H. Abazid and B. J. Nachtsheim, *A Triazole-Substituted Aryl Iodide with Omnipotent Reactivity in Enantioselective Oxidations*, Angewandte Chemie - International Edition **59** *4* (2020), Publisher: John Wiley & Sons, Ltd 1479, DOI: 10.1002/anie.201912023.

[390] D. Lemm, G. F. von Rudorff, and O. A. von Lilienfeld, *Machine learning based energy-free structure predictions of molecules, transition states, and solids*, Nature Communications **12** *1* (2021), Publisher: Nature Publishing Group 1, DOI: 10.1038/s41467-021-24525-7.

[391] J. Zhang and T. Lu, *Efficient evaluation of electrostatic potential with computerized optimized code*, Physical Chemistry Chemical Physics **23** *36* (2021), Publisher: The Royal Society of Chemistry 20323, DOI: 10.1039/d1cp02805g.

[392] R. Schade et al., *Towards electronic structure-based ab-initio molecular dynamics simulations with hundreds of millions of atoms*, Parallel Computing **111** (2022), Publisher: Elsevier B.V., DOI: 10.1016/j.parco.2022.102920.

[393] D. B. Diaz, S. D. Appavoo, A. F. Bogdanchikova, Y. Lebedev, T. J. McTiernan, G. dos Passos Gomes, and A. K. Yudin, *Illuminating the dark conformational space of macrocycles using dominant rotors*, Nature Chemistry **13** *3* (2021), Publisher: Nature Publishing Group 218, DOI: 10.1038/s41557-020-00620-y.

[394] P. Zheng, W. Yang, W. Wu, O. Isayev, and P. O. Dral, *Toward Chemical Accuracy in Predicting Enthalpies of Formation with General-Purpose Data-Driven Methods*, Journal of Physical Chemistry Letters **13** *15* (2022), Publisher: American Chemical Society 3479, DOI: 10.1021/acs.jpclett.2c00734.

[395] Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, and T. F. Miller, *OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features*, Journal of Chemical Physics **153** *12* (2020), Publisher: AIP Publishing LLCAIP Publishing 124111, DOI: 10.1063/5.0021955.

[396] V. Sinha, J. J. Laan, and E. A. Pidko, *Accurate and rapid prediction of pKaof transition metal complexes: semiempirical quantum chemistry with a data-augmented approach*, Physical Chemistry Chemical Physics **23** *4* (2021), Publisher: The Royal Society of Chemistry 2557, DOI: 10.1039/d0cp05281g.

[397] O. D. Abarbanel and G. R. Hutchison, *Machine learning to accelerate screening for Marcus reorganization energies*, Journal of Chemical Physics **155** *5* (2021) 54106, DOI: 10.1063/5.0059682.

[398] S. Dohm, M. Bursch, A. Hansen, and S. Grimme, *Semiautomated Transition State Localization for Organometallic Complexes with Semiempirical Quantum Chemical Methods*, Journal of Chemical Theory and Computation **16** *3* (2020), Publisher: American Chemical Society 2002, DOI: 10.1021/acs.jctc.9b01266.

[399] M. Fanti, G. Orlandi, and F. Zerbetto, *The Raman activity of C60 and C70: A computational semiempirical study*, Journal of Physics B: Atomic, Molecular and Optical Physics **29** *21* (1996), Publisher: IOP Publishing 5065, DOI: 10.1088/0953-4075/29/21/020.

[400] S. Kaminski, T. J. Giese, M. Gaus, D. M. York, and M. Elstner, *Extended polarization in third-order SCC-DFTB from chemical-potential equalization*, Journal of Physical Chemistry A **116** *36* (2012) 9131, DOI: 10.1021/jp306239c.

[401] R. C. Bingham, M. J. S. Dewar, and D. H. Lo, *Ground states of molecules. XXV. MINDO/3. Improved version of the MINDO semiempirical SCF-MO method*, Journal of the American Chemical Society **97** *6* (1975), Publisher: American Chemical Society 1285, DOI: 10.1021/ja00839a001.

[402] W. J. Mortier, S. K. Ghosh, and S. Shankar, *Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules*, Journal of the American Chemical Society **108** *15* (1986) 4315, DOI: 10.1021/ja00275a013.

[403] A. K. Rappé and W. A. Goddard, *Charge equilibration for molecular dynamics simulations*, Journal of Physical Chemistry **95** *8* (1991) 3358, DOI: 10.1021/j100161a070.

[404] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, *Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network*, Physical Review B - Condensed Matter and Materials Physics **92** *4* (2015) 1, DOI: 10.1103/PhysRevB.92.045131.

[405] V. I. Anisimov, J. Zaanen, and O. K. Andersen, *Band theory and Mott insulators: Hubbard U instead of Stoner I*, Publication Title: Physical Review B Volume: 44 Issue: 3 ISSN: 01631829, 1991 943, DOI: 10.1103/PhysRevB.44.943.

[406] B. Himmetoglu, A. Floris, S. De Gironcoli, and M. Cococcioni, *Hubbard-corrected DFT energy functionals: The LDA+U description of correlated systems*, International Journal of Quantum Chemistry **114** *1* (2014), Publisher: John Wiley & Sons, Ltd 14, DOI: 10.1002/qua.24521.

[407] H. J. Kulik, *Perspective: Treating electron over-delocalization with the DFT+U method*, Journal of Chemical Physics **142** *24* (2015), Publisher: AIP Publishing LLCAIP Publishing 240901, DOI: 10.1063/1.4922693.

[408] P. Durand and J. C. Barthelat, *A theoretical method to determine atomic pseudopotentials for electronic structure calculations of molecules and solids*, Theoretica Chimica Acta **38** *4* (1975), Publisher: Springer 283, DOI: 10.1007/BF00963468.

[409] E. Clementi and D. L. Raimondi, *Atomic screening constants from SCF functions*, The Journal of Chemical Physics **38** *11* (1963) 2686, DOI: 10.1063/1.1733573.

[410] E. Clementi, D. L. Raimondi, and W. P. Reinhardt, *Atomic screening constants from SCF functions. II. Atoms with 37 to 86 electrons*, The Journal of Chemical Physics **47** *4* (1967), Publisher: American Institute of PhysicsAIP ISBN: 1.18422.8094 1300, DOI: 10.1063/1.1712084.

[411] S. Grimme, *A simplified Tamm-Dancoff density functional approach for the electronic excitation spectra of very large molecules*, Journal of Chemical Physics **138** *24* (2013), Publisher: American Institute of PhysicsAIP 244104, DOI: 10.1063/1.4811331.

[412] D. W. Marquardt, *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*, Journal of the Society for Industrial and Applied Mathematics **11** *2* (1963) 431, DOI: 10.1137/0111030.

[413] *Semiempirical extended tight-binding program package xtb*, Bonn, 2022.

[414] J. Westermayr, M. Gastegger, K. T. Schütt, and R. J. Maurer, *Perspective on integrating machine learning into computational chemistry and materials science*, Journal of Chemical Physics **154** *23* (2021) 230903, DOI: 10.1063/5.0047760.

[415] *PubChem database*, 2023.

[416] J. Neugebauer, M. Reiher, C. Kind, and B. A. Hess, *Quantum chemical calculation of vibrational spectra of large molecules - Raman and IR spectra for Buckminsterfullerene*, Journal of Computational Chemistry **23** *9* (2002) 895, DOI: 10.1002/jcc.10089.

[417] R. W. Janes, D. H. Peapus, and B. A. Wallace, *The crystal structure of human endothelin*, Nature Structural Biology **1** *5* (1994), Publisher: Nature Publishing Group 311, DOI: 10.1038/nsb0594-311.

[418] J. M. Benevides, S. A. Overman, and G. J. Thomas, *Raman Spectroscopy of Proteins*, Current Protocols in Protein Science **33** *1* (2003), Publisher: Wiley, DOI: 10.1002/0471140864.ps1708s33.

[419] F. Jensen, *Atomic orbital basis sets*, WIREs Computational Molecular Science **3** *3* (2013) 273, DOI: 10.1002/wcms.1123.

[420] B. Chan, *Optimal Small Basis Set and Geometric Counterpoise Correction for DFT Computations*, Journal of Chemical Theory and Computation (2023), Publisher: American Chemical Society, DOI: 10.1021/acs.jctc.3c00298.

[421] P. C. Hiberty, "The breathing orbital valence bond method," *Modern Electronic Structure Theory and Applications in Organic Chemistry*, WORLD SCIENTIFIC, 1997 289, DOI: 10.1142/9789812839756_0007.

[422] R. Iffert and K. Jug, *Treatment of anions in SINDO1*, Theoretica Chimica Acta **72** *5* (1987) 373, DOI: 10.1007/BF01192229.

[423] J. D. Baker and M. C. Zerner, *Charge-dependent Hamiltonian for first- and second-row atomic properties*, The Journal of Physical Chemistry **94** *7* (1990), Publisher: American Chemical Society 2866, DOI: 10.1021/j100370a026.

[424] R. L. Longo, *Charge-dependent basis sets. I. First row elements*, International Journal of Quantum Chemistry **75** *4* (1999) 585, DOI: 10.1002/(SICI)1097-461X(1999)75:4/5<585::AID-QUA23>3.0.CO;2-L.

[425] P. Fuentealba, H. Preuss, H. Stoll, and L. Von Szentpály, *A proper account of core-polarization with pseudopotentials: single valence-electron alkali compounds*, Chemical Physics Letters **89** *5* (1982) 418, DOI: 10.1016/0009-2614(82)80012-2.

[426] G. Igel-Mann, *Doktorarbeit Gudrun Igel-Mann (unpublished)*, PhD thesis: Stuttgart University, 1987.

[427] L. von Szentpály, P. Fuentealba, H. Preuss, and H. Stoll, *Pseudopotential calculations on Rb+2, Cs+2, RbH+, CsH+ and the mixed alkali dimer ions*, Publication Title: Chemical Physics Letters Volume: 93 Issue: 6 ISSN: 00092614, 1982 555, DOI: 10.1016/0009-2614(82)83728-7.

[428] P. Schwerdtfeger, *Doktorarbeit Peter Schwerdtfeger (unpublished)*, PhD thesis: Stuttgart University, 1986.

[429] M. Müller, *NumgradPy, github.com/grimme-lab/NumgradPy*, version 1.0, Bonn, 2023.

[430] M. Müller, *q-vSZP input generator*, version 2.2.0, Bonn, 2024.

[431] M. J. D. Powell, "The NEWUOA software for unconstrained optimization without derivatives," *Large-Scale Nonlinear Optimization*, ed. by G. Di Pillo and M. Roma, Nonconvex Optimization and Its Applications, Boston, MA: Springer US, 2006 255, DOI: 10.1007/0-387-30065-1_16.

[432] P. Jurečka, J. Šponer, J. Černý, and P. Hobza,
Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs,
Physical Chemistry Chemical Physics **8** *17* (2006), Publisher: The Royal Society of Chemistry 1985,
DOI: 10.1039/B600027D.

[433] M. S. Marshall, L. A. Burns, and C. D. Sherrill, *Basis set convergence of the coupled-cluster correction, $\delta _{\text{MP2}}^{\text{CCSD(T)}}$δMP2CCSD(T): Best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases*, The Journal of Chemical Physics **135** *19* (2011) 194102,
DOI: 10.1063/1.3659142.

[434] J. Zheng, X. Xu, and D. G. Truhlar, *Minimally augmented Karlsruhe basis sets*,
Theoretical Chemistry Accounts **128** *3* (2011) 295, DOI: 10.1007/s00214-010-0846-z.

[435] R. Sure and S. Grimme, *Halogen bonded supramolecular capsules: a challenging test case for quantum chemical methods*,
Chemical Communications **52** *64* (2016), Publisher: The Royal Society of Chemistry 9893,
DOI: 10.1039/C6CC03664C.

[436] B. Hourahine et al., *DFTB+, a software package for efficient approximate density functional theory based atomistic simulations*, The Journal of Chemical Physics **152** *12* (2020) 124101,
DOI: 10.1063/1.5143190.

[437] D. Anstine, R. Zubatyuk, and O. Isayev, *AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs*, 2024,
DOI: 10.26434/chemrxiv-2023-296ch-v2.

[438] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller,
"SchNet: a continuous-filter convolutional neural network for modeling quantum interactions,"
Proceedings of the 31st International Conference on Neural Information Processing Systems,
NIPS'17, Red Hook, NY, USA: Curran Associates Inc., 2017 992.

[439] J. Behler, *Perspective: Machine learning potentials for atomistic simulations*,
The Journal of Chemical Physics **145** *17* (2016) 170901, DOI: 10.1063/1.4966192.

[440] S. Lehtola, *Polarized Gaussian basis sets from one-electron ions*,
The Journal of Chemical Physics **152** *13* (2020) 134108, DOI: 10.1063/1.5144964.

[441] X. Cao and M. Dolg,
Segmented contraction scheme for small-core actinide pseudopotential basis sets,
Journal of Molecular Structure: THEOCHEM **673** *1* (2004) 203,
DOI: 10.1016/j.theochem.2003.12.015.

[442] K. A. Peterson, *Correlation consistent basis sets for actinides. I. The Th and U atoms*,
The Journal of Chemical Physics **142** *7* (2015) 074105, DOI: 10.1063/1.4907596.

[443] R. Feng and K. A. Peterson,
Correlation consistent basis sets for actinides. II. The atoms Ac and Np–Lr,
The Journal of Chemical Physics **147** *8* (2017) 084108, DOI: 10.1063/1.4994725.

[444] K. A. Peterson and K. G. Dyall, "Gaussian Basis Sets for Lanthanide and Actinide Elements," *Computational Methods in Lanthanide and Actinide Chemistry*, Section: 8, John Wiley & Sons, Ltd, 2015 195, DOI: 10.1002/9781118688304.ch8.

[445] D. A. Pantazis and F. Neese, *All-Electron Scalar Relativistic Basis Sets for the Actinides*, Journal of Chemical Theory and Computation **7** *3* (2011), Publisher: American Chemical Society 677, DOI: 10.1021/ct100736b.

[446] J.-B. Lu, D. C. Cantu, C.-Q. Xu, M.-T. Nguyen, H.-S. Hu, V.-A. Glezakou, R. Rousseau, and J. Li, *Norm-Conserving Pseudopotentials and Basis Sets to Explore Actinide Chemistry in Complex Environments*, Journal of Chemical Theory and Computation **17** *6* (2021), Publisher: American Chemical Society 3360, DOI: 10.1021/acs.jctc.1c00026.

[447] K. G. Dyall, *Relativistic Double-Zeta, Triple-Zeta, and Quadruple-Zeta Basis Sets for the 4s, 5s, 6s, and 7s Elements*, The Journal of Physical Chemistry A **113** *45* (2009), Publisher: American Chemical Society 12638, DOI: 10.1021/jp905057q.

[448] K. G. Dyall, P. Tecmer, and A. Sunaga, *Diffuse Basis Functions for Relativistic s and d Block Gaussian Basis Sets*, Journal of Chemical Theory and Computation **19** *1* (2023), Publisher: American Chemical Society 198, DOI: 10.1021/acs.jctc.2c01050.

[449] C. Liu, N. F. Aguirre, M. J. Cawkwell, E. R. Batista, and P. Yang, *Efficient Parameterization of Density Functional Tight-Binding for 5f-Elements: A Th–O Case Study*, Journal of Chemical Theory and Computation **20** *14* (2024), Publisher: American Chemical Society 5923, DOI: 10.1021/acs.jctc.4c00145.

[450] T. Rose, M. Bursch, J.-M. Mewes, and S. Grimme, *Fast and Robust Modeling of Lanthanide and Actinide Complexes, Biomolecules, and Molecular Crystals with the Extended GFN-FF Model*, Inorganic Chemistry (2024), Publisher: American Chemical Society, DOI: 10.1021/acs.inorgchem.4c03215.

[451] A. Jakalian, D. B. Jack, and C. I. Bayly, *Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation*, Journal of Computational Chemistry **23** *16* (2002) 1623, DOI: 10.1002/jcc.10128.

[452] W. Thiel, "Perspectives on Semiempirical Molecular Orbital Theory," *Advances in Chemical Physics*, Advances in Chemical Physics, 1996 703, DOI: 10.1002/9780470141526.ch10.

[453] A. V. Marenich, S. V. Jerome, C. J. Cramer, and D. G. Truhlar, *Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases*, Journal of Chemical Theory and Computation **8** *2* (2012), Publisher: American Chemical Society 527, DOI: 10.1021/ct200866d.

[454] R. Zubatyuk, J. S. Smith, B. T. Nebgen, S. Tretiak, and O. Isayev, *Teaching a neural network to attach and detach electrons from molecules*, Nature Communications **12** *1* (2021), Publisher: Nature Publishing Group 4870, DOI: 10.1038/s41467-021-24904-0.

[455] Y. Wang, I. Pulido, K. Takaba, B. Kaminow, J. Scheen, L. Wang, and J. D. Chodera, *EspalomaCharge: Machine Learning-Enabled Ultrafast Partial Charge Assignment*, The Journal of Physical Chemistry A **128** *20* (2024), Publisher: American Chemical Society 4160, DOI: 10.1021/acs.jpca.4c01287.

[456] Q. Xie and A. P. Horsfield, *Coordinate-Free and Low-Order Scaling Machine Learning Model for Atomic Partial Charge Prediction for Any Size of Molecules*, Journal of Chemical Information and Modeling **64** *11* (2024), Publisher: American Chemical Society 4419, DOI: 10.1021/acs.jcim.4c00376.

[457] Z. Song, J. Han, G. Henkelman, and L. Li, *Charge-Optimized Electrostatic Interaction Atom-Centered Neural Network Algorithm*, Journal of Chemical Theory and Computation **20** *5* (2024), Publisher: American Chemical Society 2088, DOI: 10.1021/acs.jctc.3c01254.

[458] P. Pracht, Y. Pillai, V. Kapil, G. Csányi, N. Gönnheimer, M. Vondrák, J. T. Margraf, and D. J. Wales, *Efficient Composite Infrared Spectroscopy: Combining the Double-Harmonic Approximation with Machine Learning Potentials*, Journal of Chemical Theory and Computation **20** *24* (2024), Publisher: American Chemical Society 10986, DOI: 10.1021/acs.jctc.4c01157.

[459] K. B. Wiberg and P. R. Rablen, *Atomic Charges*, The Journal of Organic Chemistry **83** *24* (2018), Publisher: American Chemical Society 15463, DOI: 10.1021/acs.joc.8b02740.

[460] M. Cho, N. Sylvetsky, S. Eshafi, G. Santra, I. Efremenko, and J. M. L. Martin, *The Atomic Partial Charges Arboretum: Trying to See the Forest for the Trees*, ChemPhysChem **21** *8* (2020) 688, DOI: 10.1002/cphc.202000040.

[461] M. Veit, D. M. Wilkins, Y. Yang, R. A. DiStasio Jr., and M. Ceriotti, *Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles*, The Journal of Chemical Physics **153** *2* (2020) 024113, DOI: 10.1063/5.0009106.

[462] J. Rigby and E. I. Izgorodina, *Assessment of atomic partial charge schemes for polarisation and charge transfer effects in ionic liquids*, Physical Chemistry Chemical Physics **15** *5* (2013), Publisher: The Royal Society of Chemistry 1632, DOI: 10.1039/C2CP42934A.

[463] N. Mehta and J. M. L. Martin, *On the sensitivity of computed partial charges toward basis set and (exchange-)correlation treatment*, Journal of Computational Chemistry **45** *13* (2024) 1017, DOI: 10.1002/jcc.27294.

[464] J. Cioslowski, *A new population analysis based on atomic polar tensors*, Journal of the American Chemical Society **111** *22* (1989), Publisher: American Chemical Society 8333, DOI: 10.1021/ja00204a001.

[465] C. Fonseca Guerra, J.-W. Handgraaf, E. J. Baerends, and F. M. Bickelhaupt, *Voronoi deformation density (VDD) charges: Assessment of the Mulliken, Bader, Hirshfeld, Weinhold, and VDD methods for charge analysis*, Journal of Computational Chemistry **25** *2* (2004) 189, DOI: 10.1002/jcc.10351.

[466] G. Knizia, *Intrinsic Atomic Orbitals: An Unbiased Bridge between Quantum Theory and Chemical Concepts*, Journal of Chemical Theory and Computation **9** *11* (2013), Publisher: American Chemical Society 4834, DOI: 10.1021/ct400687b.

[467] F. Weinhold and C. R. Landis,
Valency and Bonding: A Natural Bond Orbital Donor-Acceptor Perspective,
Cambridge: Cambridge University Press, 2005, DOI: 10.1017/CBO9780511614569.

[468] F. M. Bickelhaupt, N. J. R. van Eikema Hommes, C. Fonseca Guerra, and E. J. Baerends,
The Carbon−Lithium Electron Pair Bond in (CH3Li)n (n = 1, 2, 4),
Organometallics **15** *13* (1996), Publisher: American Chemical Society 2923,
DOI: 10.1021/om950966x.

[469] E. R. Davidson and S. Chakravorty,
A test of the Hirshfeld definition of atomic charges and moments,
Theoretica chimica acta **83** *5* (1992) 319, DOI: 10.1007/BF01113058.

[470] T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. Van Speybroeck,
M. Waroquier, and P. W. Ayers,
Minimal Basis Iterative Stockholder: Atoms in Molecules for Force-Field Development, Journal of Chemical Theory and Computation **12** *8* (2016), Publisher: American Chemical Society 3894,
DOI: 10.1021/acs.jctc.6b00456.

[471] C. Plett, M. Stahn, M. Bursch, J.-M. Mewes, and S. Grimme, *Improving Quantum Chemical Solvation Models by Dynamic Radii Adjustment for Continuum Solvation (DRACO)*,
The Journal of Physical Chemistry Letters **15** *9* (2024), Publisher: American Chemical Society 2462,
DOI: 10.1021/acs.jpclett.3c03551.

[472] M. Gubler, J. A. Finkler, M. R. Schäfer, J. Behler, and S. Goedecker,
Accelerating Fourth-Generation Machine Learning Potentials Using Quasi-Linear Scaling Particle Mesh Charge Equilibration, Journal of Chemical Theory and Computation **20** *16* (2024), Publisher: American Chemical Society 7264, DOI: 10.1021/acs.jctc.4c00334.

[473] M. Friede, S. Ehlert, S. Grimme, and J.-M. Mewes, *Do Optimally Tuned Range-Separated Hybrid Functionals Require a Reparametrization of the Dispersion Correction? It Depends*, Journal of Chemical Theory and Computation **19** *22* (2023), Publisher: American Chemical Society 8097,
DOI: 10.1021/acs.jctc.3c00717.

[474] K. Nishimoto and N. Mataga, *Electronic Structure and Spectra of Some Nitrogen Heterocycles*,
Zeitschrift für Physikalische Chemie **12** *5* (1957), Publisher: De Gruyter (O) 335,
DOI: 10.1524/zpch.1957.12.5_6.335.

[475] I. S. Lim, P. Schwerdtfeger, B. Metz, and H. Stoll, *All-electron and relativistic pseudopotential studies for the group 1 element polarizabilities from K to element 119*,
The Journal of Chemical Physics **122** *10* (2005) 104103, DOI: 10.1063/1.1856451.

[476] I. S. Lim, H. Stoll, and P. Schwerdtfeger, *Relativistic small-core energy-consistent pseudopotentials for the alkaline-earth elements from Ca to Ra*,
The Journal of Chemical Physics **124** *3* (2006) 034107, DOI: 10.1063/1.2148945.

[477] N.-X. Zhang, C.-Z. Wang, J.-H. Lan, Q.-Y. Wu, and W.-Q. Shi,
Actinide endohedral inter-metalloid clusters of the group 15 elements, Physical Chemistry Chemical Physics **26** *38* (2024), Publisher: The Royal Society of Chemistry 25069,
DOI: 10.1039/D4CP02546F.

[478] S. Grimme, *ANCOPT: Approximate normal coordinate rational function optimization program*,
Universität Bonn (2013).

[479] X. Cao, M. Dolg, and H. Stoll,
Valence basis sets for relativistic energy-consistent small-core actinide pseudopotentials,
The Journal of Chemical Physics **118** *2* (2003) 487, DOI: 10.1063/1.1521431.

[480] I. Mayer, *Bond orders and valences in the SCF theory: a comment*,
Theoretica chimica acta **67** *4* (1985) 315, DOI: 10.1007/BF00529303.

[481] K. B. Wiberg, *Application of the pople-santry-segal CNDO method to the cyclopropylcarbinyl and cyclobutyl cation and to bicyclobutane*, Tetrahedron **24** *3* (1968) 1083,
DOI: 10.1016/0040-4020(68)88057-3.

[482] T. Salthammer, U. Hohm, M. Stahn, and S. Grimme, *Proton-transfer rate constants for the determination of organic indoor air pollutants by online mass spectrometry*,
RSC Advances **13** *26* (2023), Publisher: The Royal Society of Chemistry 17856,
DOI: 10.1039/D3RA01705B.

[483] M. Müller and M. Stahn, *PubGrep*, version 0.4.0, 2024.

[484] M. Müller, *randommolecules*, version 0.1.0, 2024.

[485] T. Gould, B. Chang, S. Dale, and S. Vuckovic,
Transferable diversity – a data-driven representation of chemical space, 2024,
DOI: 10.26434/chemrxiv-2023-5075x-v3.

[486] M. Müller, J. Schoeps, and S. Grimme, *MindlessGen*, version v0.4.0, 2024.

[487] J.-B. Lu, Y.-Y. Zhang, X.-L. Jiang, L.-W. Ye, and J. Li, *Improved Gaussian basis sets for norm-conserving 4f-in-core pseudopotentials of trivalent lanthanides (Ln = Ce–Lu)*,
The Journal of Chemical Physics **161** *13* (2024) 134115, DOI: 10.1063/5.0228388.

[488] G. Luchini and R. S. Paton, *Bottom-Up Atomistic Descriptions of Top-Down Macroscopic Measurements: Computational Benchmarks for Hammett Electronic Parameters*,
ACS Physical Chemistry Au **4** *3* (2024), Publisher: American Chemical Society 259,
DOI: 10.1021/acsphyschemau.3c00045.

[489] P. Ertl,
A Web Tool for Calculating Substituent Descriptors Compatible with Hammett Sigma Constants,
Chemistry–Methods **2** *12* (2022) e202200041, DOI: 10.1002/cmtd.202200041.

[490] X. Zhang et al., *A diuranium carbide cluster stabilized inside a C80 fullerene cage*,
Nature Communications **9** *1* (2018), Publisher: Nature Publishing Group 2753,
DOI: 10.1038/s41467-018-05210-8.

[491] J. P. Perdew, M. Ernzerhof, and K. Burke,
Rationale for mixing exact exchange with density functional approximations,
The Journal of Chemical Physics **105** *22* (1996) 9982, DOI: 10.1063/1.472933.

[492] C. Adamo and V. Barone,
Toward reliable density functional methods without adjustable parameters: The PBE0 model,
The Journal of Chemical Physics **110** *13* (1999) 6158, DOI: 10.1063/1.478522.

[493] C. Hölzer, I. Gordiy, S. Grimme, and M. Bursch,
Hybrid DFT Geometries and Properties for 17k Lanthanoid Complexes The LnQM Data Set,
Journal of Chemical Information and Modeling **64** *3* (2024), Publisher: American Chemical Society 825, DOI: 10.1021/acs.jcim.3c01832.

List of Figures

List of Tables

List of Acronyms

A

ACF adiabatic connection formula
An actinide
ANN artificial neural network
AO atomic orbital
ATM Axilrod–Teller–Muto

B

BB Boys-Bernardi
BJ Becke-Johnson
BSE basis set error
BSIE basis set incompleteness error
BSSE basis set superposition error

C

CASSCF complete active space self-consistent field
CBS complete basis set
CC coupled-cluster
CEH Charge Extended Hückel
CI configuration interaction
CN coordination number
CNDO complete neglect of differential overlap
COSX chain-of-spheres algorithm for exchange matrix computations
CP counterpoise
CP-SCF coupled-perturbed self-consistent field
CT charge transfer

D

DFA density functional approximation
DFT density functional theory
DFTB density functional tight-binding
DZ double-ζ

E

ECP effective core potential
EEQ electronegativity equilibration
EHT extended Hückel theory
ErrR error range
ES electrostatics

F

FF force field

G

gCP geometrical counterpoise
GGA generalized gradient approximation
GMTKN general main-group thermochemistry, kinetics, and noncovalent interactions
GTO Gaussian-type orbital

H

HF Hartree-Fock
HK Hohenberg and Kohn
HPC high-performance computing
HTS high-troughput screening

I

INDO intermediate neglect of differential overlap
IR infrared

K

KS Kohn-Sham

L

LCAO linear combination of atomic orbitals
LDA local density approximation
LFER linear free energy relationship
LM Levenberg-Marquardt
Ln lanthanide

M

m meta
MAD mean absolute deviation
MAE mean absolute error
MB minimal basis
MBD many-body dispersion
MBIS Minimal Basis Iterative Stockholder
MD molecular dynamics
ME mean error
ML machine learning
MLM "mindless" molecule
MLP machine learning potential
MNDO modified neglect of diatomic overlap
MO molecular orbital
MTD metadynamics

N

NCI non-covalent interaction
NDDO neglect of diatomic differential overlap
NPA natural population analysis

P

PCC Pearson correlation coefficient
PPP Pariser–Parr–Pople
PQN principal quantum number

Q

q-vSZP charge-dependent (polarized) valence single-ζ
QC quantum chemistry
QEq charge equilibration
QM quantum mechanical
QZ quadruple-ζ

R

RDF radial distribution function
relMAD relative mean absolute deviation
RI resolution of the identity
RM reference molecule
RMSD root mean square deviation
RMSE root mean square error
RSH range-separated hybrid

S

SCC self-consistent charge
SCF self-consistent field
SI Supporting Information
SIE self-interaction error
SQM semiempirical quantum mechanical
SRB short-ranged basis
SRCC Spearman's rank correlation coefficient
StdDev standard deviation
STO Slater-type orbital

T

TB tight-binding
TZ triple-ζ

V

vDZP polarized valence double-ζ
vSZP polarized valence single-ζ

W

WFT wave function theory
WTMAD-2 weighted total mean absolute deviation

X

XC exchange-correlation
xTB eXtended Tight-Binding

Z

ZDO zero differential overlap

Diese Dissertation ist im Zeitraum 06/2021 – 03/2025
entstanden am

Mulliken Center for Theoretical Chemistry
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn