From Density Functional Theory to Tight Binding

Development of Robust and Efficient Quantum Chemistry Methods

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

von
Sebastian Ehlert
aus
Magdeburg

Bonn, 2022

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

1. Gutachter:	Prof. Dr. Stefan Grimme
2. Gutachter:	Prof. Dr. Thomas Bredow

Tag der Promotion:	14.10.2022
Erscheinungsjahr:	2022

I want to predict bonding in molecules and solids, not to fit it.

— John P. Perdew —

Statement of Authorship

I, Sebastian Ehlert, hereby declare that I am the sole author of this thesis. The ideas and work of others, whether published or unpublished, have been fully acknowledged and referenced in my thesis.

Publications

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

1. C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, *Extended tight-binding quantum chemistry methods*, WIREs Comput. Mol. Sci. **11** (2021) e1493, DOI: 10.1002/wcms.1493.

2. S. Ehlert, U. Huniar, J. Ning, J. W. Furness, J. Sun, A. D. Kaplan, J. P. Perdew, and J. G. Brandenburg, *r^2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications*, J. Chem. Phys. **154** (2021) 061101, DOI: 10.1063/5.0041008.

3. S. Ehlert, S. Grimme, and A. Hansen, *Conformational Energy Benchmark for Longer n-Alkane Chains*, J. Phys. Chem. A **126** (2022) 3521, DOI: 10.1021/acs.jpca.2c02439.

4. S. Ehlert, M. Stahn, S. Spicher, and S. Grimme, *Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods*, J. Chem. Theory Comput. **17** (2021) 4250, DOI: 10.1021/acs.jctc.1c00471.

For the following articles significant contributions have been made.

5. M. Bursch, H. Neugebauer, S. Ehlert, and S. Grimme, *Dispersion corrected r^2SCAN based global hybrid functionals: r^2SCANh, r^2SCAN0, and r^2SCAN50*, J. Chem. Phys. **156** (2022) 134105, DOI: 10.1063/5.0086040.

6. P. Zaby, J. Ingenmey, B. Kirchner, S. Grimme, and S. Ehlert, *Calculation of improved enthalpy and entropy of vaporization by a modified partition function in quantum cluster equilibrium theory*, J. Chem. Phys. **155** (2021) 104101, DOI: 10.1063/5.0061187.

7. E. Caldeweyher, J.-M. Mewes, S. Ehlert, and S. Grimme, *Extension and evaluation of the D4 London-dispersion model for periodic systems*, Phys. Chem. Chem. Phys. **22** (2020) 8499, DOI: 10.1039/D0CP00502A.

8. 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*, J. Chem. Theory Comput. **15** (2019) 1652, DOI: 10.1021/acs.jctc.8b01176.

9. E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme, *A generally applicable atomic-charge dependent London dispersion correction*, J. Chem. Phys. **150** (2019) 154122, DOI: 10.1063/1.5090222.

The development of software projects as part of this thesis has been published in the following articles.

10. L. Kedward, B. Aradi, O. Certik, M. Curcic, S. Ehlert, P. Engel, R. Goswami, M. Hirsch, A. Lozada-Blanco, V. Magnin, A. Markus, E. Pagone, I. Pribec, B. Richardson, H. Snyder, J. Urban, and J. Vandenplas, *The State of Fortran*, Comput. Sci. Eng. **24** (2022) 63, doi: 10.1109/MCSE.2022.3159862.

11. D. G. A. Smith, A. T. Lolinco, Z. L. Glick, J. Lee, A. Alenaizan, T. A. Barnes, C. H. Borca, R. Di Remigio, D. L. Dotson, S. Ehlert, A. G. Heide, M. F. Herbst, J. Hermann, C. B. Hicks, J. T. Horton, A. G. Hurtado, P. Kraus, H. Kruse, S. J. R. Lee, J. P. Misiewicz, L. N. Naden, F. Ramezanghorbani, M. Scheurer, J. B. Schriber, A. C. Simmonett, J. Steinmetzer, J. R. Wagner, L. Ward, M. Welborn, D. Altarawy, J. Anwar, J. D. Chodera, A. Dreuw, H. J. Kulik, F. Liu, T. J. Martínez, D. A. Matthews, H. F. Schaefer, J. Šponer, J. M. Turney, L.-P. Wang, N. De Silva, R. A. King, J. F. Stanton, M. S. Gordon, T. L. Windus, C. D. Sherrill, and L. A. Burns, *Quantum Chemistry Common Driver and Databases (QCDB) and Quantum Chemistry Engine (QCEngine): Automation and interoperability among computational chemistry programs*, J. Chem. Phys. **155** (2021) 204801, doi: 10.1063/5.0059356.

12. E. Epifanovsky et al., *Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package*, J. Chem. Phys. **155** (2021) 084801, doi: 10.1063/5.0055522.

13. B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu, and T. Frauenheim, *DFTB+, a software package for efficient approximate density functional theory based atomistic simulations*, J. Chem. Phys. **152** (2020) 124101, doi: 10.1063/1.5143190.

Further articles published in the course of this thesis are listed below.

14. S. Grimme, A. Hansen, S. Ehlert, and J.-M. Mewes, *r^2SCAN-3c: A "Swiss army knife" composite electronic-structure method*, J. Chem. Phys. **154** (2021) 064103, doi: 10.1063/5.0040021.

15. 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*, Chem. Eur. J. **27** (2021) 12104, doi: 10.1002/chem.202100835.

16. X. Jie, C. G. Daniliuc, R. Knitsch, M. R. Hansen, H. Eckert, S. Ehlert, S. Grimme, G. Kehr, and G. Erker, *Aggregation Behavior of a Six-Membered Cyclic Frustrated Phosphane/Borane Lewis Pair: Formation of a Supramolecular Cyclooctameric Macrocyclic Ring System*, Angew. Chem. Int. Ed. **58** (2019) 882, doi: 10.1002/anie.201811873.

17. M. Bursch, E. Caldeweyher, A. Hansen, H. Neugebauer, S. Ehlert, and S. Grimme, *Understanding and Quantifying London Dispersion Effects in Organometallic Complexes*, Acc. Chem. Res. **52** (2019) 258, doi: 10.1021/acs.accounts.8b00505.

18. L. Trombach, S. Ehlert, S. Grimme, P. Schwerdtfeger, and J.-M. Mewes, *Exploring the chemical nature of super-heavy main-group elements by means of efficient plane-wave density-functional theory*, Phys. Chem. Chem. Phys. **21** (2019) 18048, DOI: 10.1039/c9cp02455g.

Presentations and posters on conferences and workshops are listed below.

1. Invited talk on "Extended tight-binding quantum chemistry," *Daresbury DFTB+ School*, **June 2022**, Daresbury, UK.

2. Invited talk on "Using objects across language boundaries," *ESL Fortran Object Oriented Programming Seminar Series*, **Feb 2022**, Online Seminar.

3. Contributed talk on "Fortran package manager: Toward a rich ecosystem of Fortran packages," *1st International Packaging Conference*, **Nov 2021**, Online conference.

4. Invited talk on "Creating a Reusable Software Ecosystem," *Extended Software Development Workshop: high performance computing for simulation of complex phenomena*, **Oct 2021**, Online conference.

5. Invited talk on "Fortran package manager: Toward a rich ecosystem of Fortran packages," *2nd International Fortran Conference*, **Sep 2021**, Online conference.

6. Poster on "r^2SCAN-D4: Dispersion corrected meta-GGA for general chemical applications," *57th Symposium on Theoretical Chemistry*, **Sep 2021**, Online conference.

7. Contributed talk on "r^2SCAN-3c: A Swiss army knife electronic structure method," *CECAM flagship workshop: Non-Covalent Interactions in Large Molecules*, **Aug 2021**, Lausanne, Switzerland.

8. Poster on "A robust and broadly parametrized non-selfconsistent tight-binding quantum chemistry method," *CECAM workshop: Beyond machine learning for quantum chemistry*, **Oct 2019**, Bremen, Germany.

9. Poster on "D4 – A Generally applicable charge-dependent London-dispersion correction," *9th Molecular Quantum Mechanics Conference*, **June 2019**, Heidelberg, Germany.

10. Poster on "D4 – A Generally applicable charge-dependent London-dispersion correction," *Workshop: Developing High-Dimensional Potential Energy Surfaces*, **April 2019**, Göttingen, Germany.

11. Poster on "Application and benchmarking of the DFT-D4 method," *69th Sanibel Symposium*, **Feb 2019**, St. Simons Island, GA, USA.

12. Poster on "Application and benchmarking of the DFT-D4 method," *54th Symposium on Theoretical Chemistry*, **Sep 2018**, Halle, Germany.

Abstract

The main topics of this thesis are efficient quantum chemical methods, their development, and the verification using existing and newly devised benchmark sets. The spectrum of quantum chemical methods discussed here includes wavefunction theory (WFT), (dispersion corrected) density functional theory (DFT), and semiempirical quantum mechanics (SQM), like density functional tight binding (DFTB) and extended tight binding (xTB). Special focus is set on the development and testing of dispersion corrected density functional theory to create widely applicable and robust methods for a broad range of chemical applications. Recent advances in modern density functional theory provide room for further improvements in combination with accurate dispersion corrections. For example, the latest generation of semi-classical D4 dispersion correction provides a more accurate model for describing long-range correlation effects compared to alternative methods. By extensively testing dispersion corrected density functionals, insights into the capabilities of the best available methods for computational chemistry can be obtained. With the general improvements available in density functional theory, the demand for more diverse and challenging benchmarks is increased to allow for meaningful comparisons between available methods. Especially, for non-covalent interactions obtaining accurate references is computationally demanding since the benchmarked energy differences are usually small. Large basis sets and converged numerical settings for correlated methods are needed to distinguish and rank well-performing methods.

On the other hand, with the increasing capabilities of computers and computational chemistry, real world applications become more important, including larger systems and longer time scales. This makes the development and advancement of approximate electronic structure methods another important aspect investigated here. Composite density functional methods of the “3c” scheme employing tailored corrections and optimized basis sets allow devising computationally efficient yet accurate methods. While these methods proved to be a good compromise between accuracy and efficiency, the computational efficiency needed for even larger applications is only reached with further approximations to the underlying theory in the context of SQM-based methods.

The first chapter of this thesis deals with the development of dispersion corrections for a recently devised density functional and its validation on a large collection of diverse benchmark sets for thermochemistry, reaction barriers, general properties, non-covalent interactions, and structural properties. Furthermore, state-of-the-art benchmark sets for organo-metallic reactions and lattice energies are employed to check the transferability of the performance observed for molecular main group chemistry. The comprehensive validation of density functional methods over a large chemical space is crucial to allow an informed assessment of the expected quality for a specific chemical application. The devised dispersion corrected functional, r^{2}SCAN-D4, shows excellent performance over a wide range of the conducted tests. Compared to other similarly constructed functionals the numerical stability is also greatly enhanced making it a robust choice for computational chemistry and material science. The individual steps of the functional construction from the original SCAN-D4 over the regularized variant, rSCAN-D4, to the regularized and restored one, r^{2}SCAN-D4, provide insight into the effect of the respective changes. While the

regularization reduces spurious midrange correlation and significantly improves non-covalent interaction, the violation of the exact constraints negatively impacts thermochemistry and kinetics. The restoration of the exact constraints retains the improved non-covalent interactions and further improves upon the thermochemistry and kinetics compared to the original SCAN-D4 functional.

In the second chapter the development and application of the extended tight binding methods is discussed. The xTB Hamiltonian has recently emerged as a widely applicable solution for approximate quantum mechanical calculations, with a good cost–accuracy ratio in their geometry, frequency, and non-covalent interaction (GFN) parametrizations for target properties. In particular, the systematic extension of tight binding based SQM methods to a wider range of chemical applications in the condensed phase, like in solvation, is a topic discussed in this work. The wide range of investigated models provides important tools for computational simulations from exploratory work over large-scale screening to accurate thermochemistry calculations.

In the third chapter, a new benchmark set is developed to investigate the challenging problem of computing conformational energies for flexible molecules, for which linear *n*-alkane chains are chosen as the most prototypical system. An in-detail investigation is provided about the quality of the available localized wavefunction theory (WFT) methods to allow an accurate assessment of the small energy differences between the respective conformations. It must be stressed that the development of benchmark sets containing larger chemical systems is important to detect deficiencies in the tested methods, which are not present or have only a minor impact on the predominantly small systems in most of the common training sets. While for most density functionals with exception of one empirical class of functional an excellent agreement with WFT can be found, many WFT methods of the Møller–Plesset family (MP*n*) show remarkably worse results. This unexpected result can be explained by only few benchmarks probing MP*n* for larger systems, leaving their deficiencies for conformational energies largely unexplored. Besides testing WFT and DFT methods, the comparison of different SQM methods and also force fields is conducted to evaluate their reliability for describing the conformational ensemble of flexible *n*-alkane chains. Careful analysis of the tested methods allows providing insights into potential shortcomings of the assessed methods.

Finally, the fourth chapter focuses on computing solvation contributions to free energies using SQM methods. Special focus is put on devising a computationally efficient scheme to not hamper the performance of the SQM methods while exploiting theoretical and technical advancements to create the best tailored implicit solvation model for SQM methods as well as general force fields. The proposed solvation models are extensively validated against experimental values and theoretical methods for conformational energies, large supramolecular associations of charged complex, or organometallic compounds. For the solvation models, analytical derivatives with respect to the atom positions were implemented to allow for efficient geometry optimizations, molecular dynamics, and vibrational frequency calculations.

While not discussed extensively in this thesis, the implementation, distribution, and integration of computational chemistry methods in existing and new software packages has been an integral part of this work. All methods developed in the course of this work were implemented in open-source software packages to ensure they are widely accessible to the computational chemistry community. To summarize, this work establishes standards for testing and validating of new methods against extensive benchmark collections as well as for the provision of packages for the application of those methods.

Kurzzusammenfassung

Die Hauptthemen dieser Arbeit sind effiziente quantenchemische Methoden, ihre Entwicklung und Verifikation anhand bestehender und neu entwickelter Benchmark-Sets. Das Spektrum der hier diskutierten quantenchemischen Methoden umfasst Wellenfunktionstheorie (WFT), (dispersionskorrigierte) Dichtefunktionaltheorie (DFT), und semiempirische Quantenmechanik (SQM) wie Dichtefunktionale Tight-Binding (DFTB) und extended Tight-Binding (xTB). Ein besonderer Schwerpunkt liegt auf der Entwicklung und Erprobung von dispersionskorrigierter Dichtefunktionaltheorie, um allgemein nutzbare und robuste Methoden für ein breites Spektrum chemischer Anwendungen zu schaffen. Die aktuellen Fortschritte in der modernen Dichtefunktionaltheorie bieten Raum für weitere Verbesserungen in Kombination mit genauen Dispersionskorrekturen. Hier bietet die semiklassische D4-Dispersionskorrektur der letzten Generation ein genaueres Modell für die Beschreibung von langreichweitigen Korrelationseffekten als vergleichbare Alternative. Umfassende Tests moderner Funktionale mit den neuesten Dispersionskorrekturen bieten Einblicke in die Möglichkeiten, die mit den besten Methoden der computergestützten Chemie zur Verfügung stehen. Mit den allgemeinen Verbesserungen in der Dichtefunktionaltheorie steigt der Bedarf an vielfältigeren und anspruchsvolleren Benchmarks, um aussagekräftige Vergleiche zwischen den verfügbaren Methoden zu ermöglichen. Insbesondere bei nicht-kovalenten Wechselwirkungen ist die Ermittlung genauer Referenzen rechnerisch anspruchsvoll, da die Energieunterschiede bei den Benchmarks in der Regel gering sind. Große Basissätze und konvergierte numerische Einstellungen für korrelierte Ansätze sind erforderlich, um gut funktionierende Methoden zu unterscheiden und zu bewerten.

Auf der anderen Seite, ist mit dem zunehmenden Interesse an rechnerischen Simulationen größerer Systeme oder längerer Zeitskalen die Entwicklung und Verbesserung approximativer elektronischer Strukturmethoden ein weiterer wichtiger Aspekt, der hier untersucht wird. Dichtefunktionalmethoden des „3c“-Schemas, die maßgeschneiderte Korrekturen und optimierte Basissätze verwenden, ermöglichen die Entwicklung rechnerisch effizienter und dennoch genauer Methoden. Trotz der Genauigkeit dieser, wird die für noch umfangreichere Anwendungen erforderliche Recheneffizienz nur durch weitere Näherungen an der zugrundeliegenden Theorie im Rahmen von SQM-basierten Methoden erreicht.

Das erste Kapitel dieser Arbeit befasst sich mit der Entwicklung von Dispersionskorrekturen für ein kürzlich entwickeltes Dichtefunktional und dessen Validierung anhand einer großen Sammlung verschiedener Benchmark-Sätze für Thermochemie, Reaktionsbarrieren, allgemeine Eigenschaften, nicht-kovalente Wechselwirkungen und strukturelle Eigenschaften. Darüber hinaus werden Benchmark-Sets für metallorganische Reaktionen und Gitterenergien verwendet, um die Übertragbarkeit der für die molekulare Hauptgruppenchemie beobachteten Leistung zu überprüfen. Die umfassende Validierung von Dichtefunktionalmethoden über einen großen chemischen Raum ist von entscheidender Bedeutung, um eine fundierte Bewertung der erwarteten Qualität für eine bestimmte chemische Anwendung zu ermöglichen. Das entwickelte dispersionskorrigierte Funktional, r^{2}SCAN-D4, zeigt ausgezeichnete Leistung über einen weiten Bereich der durchgeführten Tests. Im Vergleich zu anderen, ähnlich konstruierten

Funktionalen ist auch die numerische Stabilität stark verbessert, was es zu einer robusten Wahl für die Computer-Chemie und Materialwissenschaft macht. Die einzelnen Schritte der Funktionskonstruktion vom ursprünglichen SCAN-D4 über die regularisierte Variante rSCAN-D4 bis hin zum regularisierten und wiederhergestellten r^{2}SCAN-D4 geben Aufschluss über die Wirkung der jeweiligen Änderungen. Während die Regularisierung die unerwünschte Korrelation im mittelreichweitigen Bereich reduziert und die nicht-kovalente Wechselwirkung deutlich verbessert, wirkt sich die Verletzung der exakten Beschränkungen negativ auf die Thermochemie und Kinetik aus. Durch die Wiederherstellung der exakten Beschränkungen bleiben die verbesserten nicht-kovalenten Wechselwirkungen erhalten und die Thermochemie und Kinetik werden im Vergleich zum ursprünglichen SCAN-D4-Funktional weiter verbessert.

Im zweiten Kaptel wird die Entwicklung und Anwendung der extended tight binding Methoden diskutiert. Der xTB-Hamiltonian hat sich als anwendbare Lösung für approximative quantenmechanische Berechnungen erwiesen, wobei die Parametrisierung für Geometrien, Frequenzen und nicht-kovalenten Wechselwirkungen (GFN) ein gutes Kosten-Genauigkeits-Verhältnis zeigt. Insbesondere die systematische Erweiterung von SQM-Methoden auf der Grundlage von Tight-Binding auf eine breitere Palette chemischer Anwendungen in der kondensierten Phase, wie z. B. bei der Solvatation, ist ein Thema dieser Arbeit. Das weite Spektrum der untersuchten Modelle bietet wichtige Werkzeuge für Computersimulationen, von der Erkundung über groß angelegte Screenings bis hin zu genauen thermochemischen Berechnungen.

Im dritten Kapitel wird ein neuer Benchmark-Satz entwickelt, um das schwierige Problem der Berechnung von Konformationsenergien für flexible Moleküle zu untersuchen, für die lineare *n*-Alkan-Ketten als das prototypischste System ausgewählt wurden. Es wird eine detaillierte Untersuchung der Qualität der verfügbaren Methoden der lokalisierten Wellenfunktionstheorie (WFT) durchgeführt, um eine genaue Bewertung der geringen Energieunterschiede zwischen den jeweiligen Konformationen zu ermöglichen. Es muss betont werden, dass die Entwicklung von Benchmark-Sets, die größere chemische Systeme enthalten, wichtig ist, um Defizite in den getesteten Methoden zu erkennen, die bei den überwiegend kleinen Systemen in den meisten gängigen Benchmark-Sets nicht vorhanden sind oder nur eine geringe Auswirkung haben. Während für die meisten Dichtefunktionale mit Ausnahme einer empirischen Klasse von Funktionalen eine ausgezeichnete Übereinstimmung mit der WFT gefunden werden kann, zeigen viele WFT-Methoden der Møller-Plesset-Familie (MP*n*) deutlich schlechtere Ergebnisse. Neben dem Testen von WFT- und DFT-Methoden werden auch verschiedene SQM-Methoden und Kraftfelder verglichen, um ihre Zuverlässigkeit bei der Beschreibung des Konformationsensembles von flexiblen *n*-Alkan-Ketten zu bewerten. Eine sorgfältige Analyse der getesteten Methoden ermöglicht es, Einblicke in mögliche Schwächen diser zu erhalten.

Das vierte Kapitel schließlich konzentriert sich auf die Berechnung von Solvatationsbeiträgen zu freien Energien mit SQM-Methoden. Besonderes Augenmerk wird auf die Entwicklung eines rechnerisch effizienten Schemas gelegt, um die Leistung der SQM-Methoden nicht zu beeinträchtigen und gleichzeitig theoretische und technische Fortschritte zu nutzen, um das beste maßgeschneiderte implizite Solvatationsmodell für SQM-Methoden sowie allgemeine Kraftfelder zu erstellen. Die vorgeschlagenen Solvatationsmodelle werden anhand von experimentellen Werten und theoretischen Methoden für Konformationsenergien, große supramolekulare Assoziationen von geladenen Komplexen oder metallorganischen Verbindungen umfassend validiert. Für die Solvatationsmodelle wurden analytische Ableitungen in Bezug auf die Atompositionen implementiert, um effiziente Geometrieoptimierungen, Molekulardynamik- und Schwingungsfrequenzberechnungen zu ermöglichen.

Die Implementierung, Verbreitung und Integration von Methoden der computergestützten Chemie in bestehende und neue Softwarepakete war ein wesentlicher Bestandteil dieser Arbeit, auch wenn sie in dieser Arbeit nicht ausführlich behandelt wird. Alle Methoden, die im Rahmen dieser Arbeit entwickelt wurden, wurden in Open-Source-Softwarepaketen implementiert, um sicherzustellen, dass sie für die Gemeinschaft der computergestützten Chemie allgemein zugänglich sind. Zusammenfassend lässt sich sagen, dass diese Arbeit Standards für die Prüfung und Validierung neuer Methoden anhand umfangreicher Benchmark-Sammlungen sowie für die Bereitstellung von Softwarepaketen, die die Anwendung dieser Methoden ermöglichen, geschaffen hat.

Contents

CHAPTER 1

Introduction and theoretical background

1.1 Introduction

For the theoretical description of chemical processes and systems, the application of computational simulations has become an indispensable tool. Computational chemistry offers an approach to the fundamental understanding where experimental realizations are difficult due to toxicological hazards, resource constraints, or time limitations. The advent of computational simulations in chemistry, like in other fields of modern science, was transformative in the way chemical and physical properties are validated and even predicted. The application of computational chemistry in drug discovery, catalyst design, and material science has significantly advanced productivity in the respective research fields.

Especially chemical reactivity is of central interest for computational modeling, which requires an accurate description of the free energy of the chemical system. Free energy accounts for both the energy of the system itself and its degrees of freedom. Experimental measurements of this quantity are challenging, even more for reactions with multiple or short-lived intermediates. To answer the question of whether a chemical reaction takes place in an experiment or a new drug is pharmaceutical active, a computational model for assessing the difference in free energy of the reaction is required. The first task for the model is to select the relevant structures for the reactants and the products as well as the relevant experimental conditions, like solvents or temperature. Especially for flexible molecules acquiring an input structure for the model can be challenging. While experimental input for the three-dimensional structure can be provided from X-ray crystallography, the measurement conditions are usually vastly different in terms of temperature and environment, furthermore the resolution for light atoms like hydrogens can add additional uncertainty. Sampling the conformational degrees of freedom for the reactant and product structures is therefore usually the first step of creating a computational model.[19,20] In this process several thousand candidate structures for a single flexible compound must be considered to allow a reliable sampling and discovery of the relevant conformations. Using a computational model to describe the potential energy surface of the reaction allows finding the stationary points, local minima and first-order saddle points, to compute the thermodynamic and kinetic properties of the reaction and provide insight into the mechanism unavailable by experimental means. Developing methods for computing the potential energy surface consistently is a key challenge for computational chemistry. While computing the full energy of the system is a challenging task, computing the relative differences between structures or along trajectories is more important than predicting the total energy to its full accuracy.

Quantum chemical methods provide approximate yet consistent solutions for computing relative

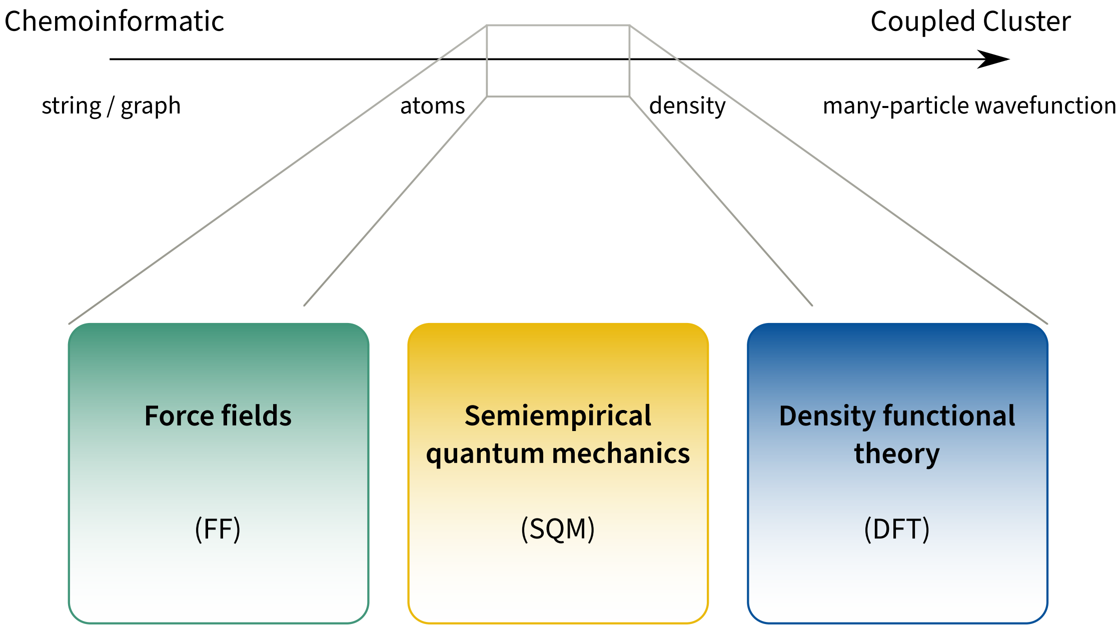

Chemoinformatic
Coupled Cluster
string / graph
atoms
density
many-particle wavefunction
Force fields
(FF)
Semiempirical quantum mechanics
(SQM)
Density functional theory
(DFT)

Figure 1.1: Detailed view on chemical system representations ranging from coarse and incomplete representations like two-dimensional graphs to the complete many-particle wavefunction. Graph or string based low dimensional representations find usage in chemoinformatics while sophisticated methods like coupled cluster theory employ the high dimensional many-particle wavefunction to represent a chemical system. Quantum mechanical methods use representations accounting at least for the electrons, like the electron density.

energy changes in chemical systems. The resolution used to describe the chemical system largely varies in the spectrum of available quantum chemical methods as sketched in Fig. 1.1. With the many-particle wavefunction, on the one hand, a complete description of the electronic structure is available for a given three-dimensional molecular geometry. However, the computational effort to obtain and work with the complete wavefunction makes it feasible for small systems of only a few atoms. While systematic approximations are available in the context of wavefunction theory (WFT), like coupled cluster (CC) theory, a more feasible representation is available with the electron density. Using the electron density to find a mapping toward other quantities is the central objective of density functional theory (DFT), which has become the most common approach to computational modeling in chemistry. With no exact functional known describing this mapping, the development of density functional approximations is an active field of research. The development of new approximate functionals, as well as their validation on different chemical problems, is a central research topic in quantum chemistry. Different strategies in the development, like satisfying exact constraints on the functional[21] or data-driven approaches to optimize functional parameters,[22] give rise to several possible functional choices for computational applications. Testing and validating functional approximations for different chemical problems motivated the creation and curation of diverse benchmark collections[23,24] and extensive comparison and ranking of functionals. Problems like the absence of London-dispersion in semi-local functionals have been well understood[25] and mended either by additive correction schemes[9,26,27] or non-local functional components.[28,29] On the other hand the self-interaction error in semi-local functionals is still an unsolved issue and actively investigated in the context of density-corrected functionals,[30] multiconfigurational DFT,[31] and local hybrid functionals.[32,33]

With computational simulations targeting larger chemical systems or screening a multitude of compounds, the computational effort for the quantum chemical evaluation becomes a significant and prohibitive factor for the effectiveness of the simulation or workflow. Computationally more efficient models, which retain most of the favorable properties of density functionals, are required to support contemporary applications in chemistry. Optimization and acceleration of existing methods, by using highly optimized performance libraries or dedicated hardware acceleration provide a possible avenue. However, making use of highly optimized programs on dedicated hardware becomes technically involved both in the deployment and usage, leaving such programs only accessible to technical experts. On the other hand, simplified theories can provide an intrinsic advantage in computational efficiency. A successful example of such are composite density functional methods[14,34] targeted at describing a wide range of general thermochemistry and reaction barriers well and providing an efficient approach to explore the potential energy surface for structural properties or molecular dynamics.

Another alternative to full quantum mechanical calculations is semiempirical molecular orbital theories, which were predominant in the advent of quantum chemistry[35] and now see renewed interest due to their favorable computational cost.[36] The two main approaches are based on the neglect of diatomic differential overlap (NDDO)[37] or density functional tight binding (DFTB).[38] Semiempirical methods can yield a speed-up of two to three orders of magnitude compared to their parent methods, like density functional theory. Furthermore, most techniques for optimization like linear scaling methods[39] or acceleration by using dedicated hardware[40] can be used for semiempirical methods as well. The drawback of the approximations used for semiempirical methods is that they usually cannot provide the same global accuracy compared to density functional theory. Rather, the parametrization of the semiempirical methods is necessarily limited to a set of target properties that were part of their optimization or training set. Larger errors can be expected for off-target properties or systems which are chemically different from the ones used for parametrization. However, a well-informed choice taking into account those shortcomings can yield a significant speed-up for computational simulations enabling to investigate larger problems or to screen more candidates.

A promising candidate for future applications of semiempirical quantum mechanic (SQM) methods were recently introduced with the extended tight binding (xTB) Hamiltonian.[1] While DFTB methods have been present for more than two decades, they found only slow adoption due to the involved computation of the reference wavefunctions as well as the difficult parametrization of the pairwise repulsive potentials. In contrast to this, the xTB methods follow an element-specific parametrization strategy without the need to precalculate wavefunctions and integrals as in DFTB. Rather than aiming for a general method, the xTB methods focus on a limited scope of target properties including geometries, frequencies, and non-covalent interactions, dubbed GFN parametrization. While unintuitive, the approximations made in the derivation of DFTB and xTB impact the generality of the resulting methods, which are difficult to mend by parametrization, instead selecting a limited scope allows for creating a special purpose method describing the selected target properties well. The resulting methods, GFN1-xTB[41] and GFN2-xTB,[8] have found wide adoption in the computational chemistry community[42,43] and are part of several screening workflows.[19,44] Generally, SQM methods require extensive testing and validation to ensure that the proposed parametrization is robust. This holds true for all semiempirical methods, independent of how rigorous they were derived in the first place, as the introduced approximations can have a severe impact on the overall accuracy and robustness. The development of new SQM methods has spawned an interest in developing more diverse benchmark sets to provide the necessary coverage of chemical diverse systems.[17]

This work aims to provide improved methods in the spectrum from DFT, SQM, and also FF methods. For this purpose a dispersion corrected density functional is introduced in Chapter 2, based on the recently proposed r^{2}SCAN functional.[21] Together with the development of composite electronic structure method of the "3c" kind,[14,34] and its hybrid variants,[5] the r^{2}SCAN family of functionals provide reliable performance at the DFT level of theory. The xTB Hamiltonian and the methods of the GFN family are discussed in Chapter 3. Especially the GFN methods, including the parametrizations of the xTB Hamiltonian and the GFN-FF, provide a computationally efficient yet accurate way for simulating structural properties in large scale screening applications. To take upon the challenging problem of conformational energies, the seemingly easiest system of long, unbranched alkane chains is investigated in Chapter 4. Unbranched alkanes provide flexible molecules with a large conformer ensemble while having a simple electronic structure, making them a prime example to investigate intramolecular non-covalent interactions. The r^{2}SCAN family of functionals provides excellent performance for conformational energies compared to several established methods. The study is further extended to include SQM and FF methods, which are commonly used for generating conformer ensembles. Methods of the GFN family like GFN-FF[45] and GFN2-xTB, are identified for providing the best cost–accuracy ratio for these flexible conformer ensembles. Furthermore, this work investigates the computation of solvation free energies in Chapter 5 with SQM and FF methods. The description of chemical systems in the condensed phase, like solvation, provides a cornerstone for computational simulation. Especially, the impact of the proposed solvation model on the structural properties and conformational energies is part of the validation and testing. A new implicit solvation model is proposed leveraging developments in the field of generalized Born (GB) theory including effects of finite dielectric constants or a revised interaction kernel for the Coulombic screening. Finally, the impact of this thesis is summarized in Chapter 6 and put into the perspective with the field of computational chemistry.

1.2 Theoretical background

In this chapter the fundamental theories used or developed in the course of this work will be introduced. Atomic units will be used throughout the thesis for clarity.

1.2.1 Electronic structure theory

The foundation for electronic structure methods in quantum chemistry is the time-independent, non-relativistic Schrödinger equation[46] given as

$$\widehat{H}\Psi = E\Psi \tag{1.1}$$

where $\widehat{H}$ is the Hamiltonian operator, Ψ is the wavefunction and E is the energy. The Hamiltonian operator $\widehat{H}$ for a Coulombic system is formed from the kinetic energy of the nuclei $\widehat{T}_n$ and electrons $\widehat{T}_e$ as well as the potential energy between the nuclei $\widehat{V}_{nn}$, the electrons $\widehat{V}_{ee}$, and the electron and nuclei $\widehat{V}_{ne}$ (Eq. 1.2)

$$\widehat{H} = \widehat{T}_n + \widehat{T}_e + \widehat{V}_{nn} + \widehat{V}_{ee} + \widehat{V}_{ne} \tag{1.2}$$

The Born–Oppenheimer approximation[47] is commonly employed to simplify the Hamiltonian operator, by accounting for the relative time-scales on which the heavier nuclei move compared to the lighter electrons. Treating the nuclei as classical particles allows to neglect their kinetic energy $\widehat{T}_n$ and the potential energy between the nuclei $\widehat{V}_{nn}$ becomes a constant, the resulting electronic Hamiltonian $\widehat{H}_e$ is given as

$$\widehat{H}_e = \widehat{T}_e + \widehat{V}_{ee} + \widehat{V}_{ne} + V_{nn} \tag{1.3}$$

The electronic kinetic energy operator for an N-electron system is given as

$$\widehat{T}_e = -\frac{1}{2}\sum_{i}^{N} \widehat{\Delta}_i \tag{1.4}$$

where i is the index referring to the individual particle in the system and $\widehat{\Delta}_i$ is the Laplace operator $\widehat{\nabla}_i^2$. The electron repulsion energy $\widehat{V}_{ee}$ and nuclear–electron attraction energy $\widehat{V}_{ne}$ are given as the Coulombic interaction between all particles.

$$\widehat{V}_{ee} = \sum_{i}^{N}\sum_{\substack{j \\ j<i}}^{N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{1.5}$$

$$\widehat{V}_{ne} = -\sum_{i}^{N}\sum_{A}^{K} \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} \tag{1.6}$$

here $\mathbf{r}_i$ are the coordinates of the electron i, $\mathbf{R}_A$ are the coordinates of the nucleus A of the K nuclei and Z_A are their nuclear charge. The operators can be grouped in one-electron and two-electron operators, depending on the number of electron indices the Hamiltonian operator is rewritten as

$$\widehat{H}_e = \widehat{h} + \widehat{V}_{ee} + V_{nn} \qquad \text{with} \quad \widehat{h} = \widehat{T}_e + \widehat{V}_{ne} \tag{1.7}$$

For a given wavefunction the expectation value of any operator $\widehat{O}$ can be evaluated by the integral

$$O = \int \Psi^* \widehat{O}\, \Psi \, d\mathbf{r} = \langle \Psi | \widehat{O} | \Psi \rangle \tag{1.8}$$

over the electron coordinate **r**. For example the energy E can be obtained from the Hamiltonian operator $\widehat{H}_e$ (cf. Eq. 1.1).

The many-particle wavefunction Ψ can be parametrized using simpler mathematical functions, which retain the antisymmetry of the wavefunction and the indistinguishability of the electrons. The simplest of such parametrizations is given with a Slater-determinant Φ of orthonormal one-particle wavefunctions $\psi_i(j)$

$$\Psi \approx \Phi(1,2,\ldots,N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(1) & \psi_2(1) & \cdots & \psi_N(1) \\ \psi_1(2) & \psi_2(2) & \cdots & \psi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(N) & \psi_2(N) & \cdots & \psi_N(N) \end{vmatrix} \tag{1.9}$$

abbreviated as $|\psi_1\psi_2\cdots\psi_N\rangle$. Evaluating the expectation value of the Slater determinant for the Hamiltonian operator results in the Hartree–Fock energy expression[48,49]

$$E = \sum_i^N \underbrace{\langle\psi_i|\widehat{h}|\psi_i\rangle}_{h_i} + \frac{1}{2}\sum_i^N\sum_j^N \Big(\underbrace{\langle\psi_i\psi_j|r_{ij}^{-1}|\psi_i\psi_j\rangle}_{J_{ij}} - \underbrace{\langle\psi_i\psi_j|r_{ij}^{-1}|\psi_j\psi_i\rangle}_{K_{ij}} \Big) + V_{nn} \tag{1.10}$$

with the one-particle energies h_i and the two-particle energies from Coulomb J_{ij} and exchange interactions K_{ij}, and the nuclear repulsion energy V_{nn}. Since the Hartree–Fock energy expression is a variational functional of the Slater determinant, the wavefunction can be obtained by minimization. To preserve the orthonormality of the one-particle wavefunctions a Lagrangian constraint is included as shown in the following equation.

$$L = \frac{\langle\psi_1\cdots\psi_N|\widehat{H}_e|\psi_1\cdots\psi_N\rangle}{\langle\psi_1\cdots\psi_N|\psi_1\cdots\psi_N\rangle} + \sum_i^N\sum_j^N \varepsilon_{ij}\left(\langle\psi_i|\psi_j\rangle - \delta_{ij}\right) \tag{1.11}$$

By introducing a variation to the Lagrangian a set of N coupled integro-differential equations is obtained

$$\Big(\widehat{h}_i + \sum_j^N (\widehat{J}_j - \widehat{K}_j)\Big)\psi_i = \sum_j^N \varepsilon_{ij}\psi_i \tag{1.12}$$

where $\widehat{J}_j$ and $\widehat{K}_j$ are the effective one-particle operators for Coulomb and exchange interactions with the mean-field of the electrons, respectively. For a diagonal ε_{ij} matrix the solution are the canonical orbital energies ε_i. Due to the mean-field treatment of the electrons their instantaneous Coulomb correlation, also known as dynamic correlation, is not captured with this wavefunction parametrization. The missing energy between Hartree–Fock and the exact solution is termed therefore correlation energy.

A systematic approach to include correlation is to augment the Hartree–Fock solution with additional determinants generated by excitions of particles from the occupied one-particle wavefunctions to virtual ones. The main contribution to the correlation energy is given by the doubly excited determinants

$$E_{\text{corr}} = \frac{1}{4}\sum_{i}^{N}\sum_{j}^{N}\sum_{a}^{N_{\text{virt}}}\sum_{b}^{N_{\text{virt}}}\Big(\langle\psi_i\psi_j|\psi_a\psi_b\rangle - \langle\psi_i\psi_j|\psi_b\psi_a\rangle\Big)\,t_{ij}^{ab} \tag{1.13}$$

where t_{ij}^{ab} are the amplitudes of the doubly excited determinants. An approximation to the double amplitudes provided by second order Møller–Plesset (MP2)[50] perturbation theory is

$$t_{ij}^{ab} = \frac{\langle\psi_a\psi_b|\psi_i\psi_j\rangle - \langle\psi_a\psi_b|\psi_j\psi_i\rangle}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{1.14}$$

The resulting MP2 energy is given by

$$E_{\text{MP2}} = E_{\text{HF}} + \frac{1}{4}\sum_{i}^{N}\sum_{j}^{N}\sum_{a}^{N_{\text{virt}}}\sum_{b}^{N_{\text{virt}}}\frac{\Big(\langle\psi_i\psi_j|\psi_a\psi_b\rangle - \langle\psi_i\psi_j|\psi_b\psi_a\rangle\Big)^2}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}\,. \tag{1.15}$$

The correlation energy as defined by MP2 is commonly applied in the context of double hybrid density functionals as discussed in the next section. Further discussion of correlation methods is beyond the scope of this thesis.

1.2.2 Density functional theory

Another approach for including electron correlation is to modify the Hamiltonian rather than improving the Hartree–Fock wavefunction.[51] Density functional theory (DFT) is based on a functional mapping the ground-state density to the external potential[52–54] and describing the interactions between electrons by an observable in 3D space. Starting from the density ρ given as

$$\rho = \sum_{i}^{N} n_i|\psi_i|^2 \tag{1.16}$$

where n_i is the occupation number of the molecular orbital i and an external potential $\nu(\mathbf{r}) = -\sum_A(Z_A/|\mathbf{r} - \mathbf{R}_A|)$, the ground state energy expression of Kohn–Sham DFT can be expressed as a functional of the electron density by

$$E[\rho] = T_s[\rho] + V_{ne}[\rho] + J[\rho] + E_{xc}[\rho] \tag{1.17}$$

where T_s is the Kohn–Sham (KS) kinetic energy, V_{ne} the nuclear–electron energy, J the Coulomb energy, and E_{xc} the exchange-correlation functional. The first three terms have analogous expressions in the Hartree–Fock energy (cf. Eq. 1.10). KS-DFT connects the functional expression with the wavefunction by the electronic density and the KS kinetic energy T_s defined from the molecular orbitals as

$$T_s[\rho] = \sum_{i}^{N} n_i\langle\psi_i| - \frac{1}{2}\widehat{\Delta}|\psi_i\rangle\,. \tag{1.18}$$

The nuclear–electron potential energy V_{ne} can be expressed by

$$V_{ne}[\rho] = \int \rho(\mathbf{r})\nu(\mathbf{r})\,d\mathbf{r} \tag{1.19}$$

where ν is the external potential by the nuclei, $\nu(\mathbf{r}) = -\sum_A (Z_A/|\mathbf{r} - \mathbf{R}_A|)$. The Coulomb energy J is given by the classical electron–electron repulsion energy

$$J[\rho] = \frac{1}{2}\iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}' \,. \tag{1.20}$$

The crucial difference in KS-DFT however is the exchange-correlation energy E_{xc}, which captures the antisymmetry of the wavefunction as well as the correlation effects. For the exchange-correlation energy no exact expression is known, however several approximation have been proposed since the advent of KS-DFT. The approximate exchange-correlation functional is commonly partitioned into the exchange and correlation functional, here a further partition into semi-local (SL) and non-local (NL) components will be introduced. An important NL contribution is the long-range London-dispersion interaction. However, it should be noted that this partitioning is arbitrary.

$$E_{xc}[\rho] = E_x^{SL}[\rho] + E_x^{NL}[\rho] + E_c^{SL}[\rho] + E_c^{NL}[\rho] \,. \tag{1.21}$$

While interrelated, for each of the contributions individual approximations are available. Using a complete description including semi-local and non-local components offers the best theoretical model, however especially the non-local components introduce a significant computational demand. Therefore, most exchange-correlation functional approximations at most include expressions for the semi-local contributions. A commonly used categorization of density functional approximations is to group exchange-correlation functional approximations in rungs as proposed by Perdew.[55] The respective rungs are grouped by the information they use, such as the local density, its gradient or higher derivatives or non-local components in exchange or correlation as shown schematically in Fig. 1.2. With additional components the computational cost as well as the expected accuracy of the approximation is increated.

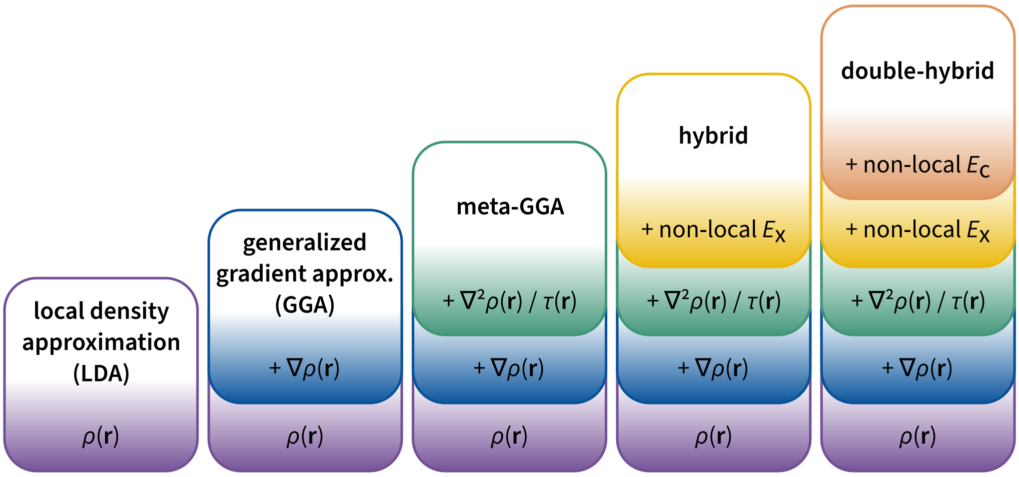

local density approximation (LDA)
$\rho(\mathbf{r})$
generalized gradient approx. (GGA)
$+\nabla\rho(\mathbf{r})$
$\rho(\mathbf{r})$
meta-GGA
$+\nabla^2\rho(\mathbf{r})$ / $\tau(\mathbf{r})$
$+\nabla\rho(\mathbf{r})$
$\rho(\mathbf{r})$
hybrid
+ non-local E_X
$+\nabla^2\rho(\mathbf{r})$ / $\tau(\mathbf{r})$
$+\nabla\rho(\mathbf{r})$
$\rho(\mathbf{r})$
double-hybrid
+ non-local E_C
+ non-local E_X
$+\nabla^2\rho(\mathbf{r})$ / $\tau(\mathbf{r})$
$+\nabla\rho(\mathbf{r})$
$\rho(\mathbf{r})$

Figure 1.2: Schematic representation of the density functional rungs and the used information.

Starting from the lowest rung of exchange-correlation functionals which is the local density approximation (LDA), the exchange functional for the uniform electron gas[56] is known by the expression

$$E_x^{\mathrm{LDA}}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int \rho^{4/3}\mathrm{d}\mathbf{r}\,. \tag{1.22}$$

LDA-type functionals are a suitable approximation for metallic systems, however not suitable for chemical relevant systems, which show strongly varying densities especially for the exponential decay of the electron density away from the atoms. The next rung of functionals addresses this by including the density gradient $\nabla\rho$ additionally to the density. Functionals of this rung are termed generalized gradient approximation (GGA). For example the PBE exchange functional[57,58] is given by

$$E_x^{\mathrm{PBE}}[\rho] = -\int \rho^{4/3}\left[\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} + \frac{\mu s^2}{1+\mu s^2/\kappa}\right]\mathrm{d}\mathbf{r} \tag{1.23}$$

where μ and κ are functional specific parameters and $s = |\nabla\rho|/(2k_F\rho)$ is the reduced density gradient. GGA functionals can provide a suitable description for many chemical systems in the ground state, but usually insufficient at reaction barriers or thermochemistry, which can be partly remedied by advancing to the third rung including second derivatives of the density and is represented by meta-GGAs. An example for an exchange functional of the meta-GGA family is the r^{2}SCAN functional[21] given as

$$E_x^{\mathrm{r^2SCAN}}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int \rho^{4/3}\left\{g_x(p)\left(h_x^1(p) + f_x(\bar{\alpha})\left[h_x^0 - h_x^1(p)\right]\right)\right\}\mathrm{d}\mathbf{r} \tag{1.24}$$

where the dimensionless kinetic energy variable $\bar{\alpha} = \frac{\tau-\tau_w}{\tau_u+\eta\tau_w}$ introduces the dependency on the Laplacian of the density with the kinetic energy density τ, here $\tau_w = |\nabla\rho|^2/(8\rho)$ and $\tau_u = 3(3\pi^2)^{2/3}$ are the von Weizsäcker and uniform electron gas kinetic energy, respectively, and η is a regularization parameter. The functions $g_x(p)$, h_x^0, $h_x^1(p)$ and $f_x(\bar{\alpha})$ provide the meta-GGA enhancement factors from the square of the reduced density gradient $p = s^2$ or the iso-orbital indicator $\bar{\alpha}$. Functionals of the meta-GGA category use the most information available for a semi-local approximation which allows for a computational efficient functional while missing most effects which are not available from the semi-local description of the electron density. To further improve upon the available functional approximations the usage of non-local components provides the necessary descriptors to account for information not available in a purely semi-local picture. However, the introduction of non-local contributions in many cases leads to a significant increased computational cost of the functional.

The inclusion of non-local exchange E_x^{NL} by using HF-like exchange contributions (cf. Eq. 1.10) advances a functional to the fourth rung, called hybrids. Including non-local exchange at a constant fraction results in global hybrids, like PBE0.[59,60] Other strategies include but are not limited to range-separated hybrids using non-local exchange at large inter-electron distances, like ωB97M,[61–63] screened exchange hybrids removing non-local exchange at large distances, like MN12-SX,[64] or local hybrids with spacial dependent non-local exchange fraction, like Lh20t.[32] The advantage of hybrid functionals is the vastly improved description of thermochemistry and barrier heights compared to semi-local functionals.

Similar to the exchange functional, non-local correlation E_c^{NL} can be introduced in the correlation functional. Non-local correlation, especially long-range London-dispersion, is crucial for the description of any medium-sized or larger chemically relevant system.[25] Most semi-local functionals up to the forth rung do not include London-dispersion or can only partially account for it by parametrization. One mean for this is the inclusion of virtual orbitals using second-order Görling–Levy perturbation theory (GL2), which is analogous to MP2 in the context of HF (cf. Eq. 1.15), leading to fifth rung double hybrid functionals. Functionals of the double hybrid category provide especially accurate computational methods but are computationally very demanding due to the evaluation of the wavefunction based correlation energy expression. An alternative approach for inclusion of long-range correlation is the usage of van-der-Waals functionals like the VV10 functional[28] which introduce a non-local correlation kernel at the GGA level.

$$E_c^{NL}[\rho] = \iint \rho(\mathbf{r})\Phi(\rho, \rho', |\nabla\rho|, |\nabla\rho'|, |\mathbf{r}-\mathbf{r}'|)\rho(\mathbf{r}')\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{1.25}$$

here Φ is the correlation kernel for describing the long-range correlation of the densities and their gradients. In the limit of long electron distances most van-der-Waals functionals[65,66] can be described by

$$E_c^{NL}[\rho] = \frac{3}{32\pi^2}\iint \frac{1}{|\mathbf{r}-\mathbf{r}'|^6}\frac{\omega_p\omega_p'}{\omega_p+\omega_p'}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{1.26}$$

where $\omega_p = \sqrt{4\pi\rho}$ is the plasmon frequency, yielding the typical distance dependence of the leading term in London-dispersion. The distance dependence of the long-range correlation energy can be described in a simplified way using the distance R_{AB} of interacting systems

$$E_c^{NL}(R_{AB}) \propto \frac{C_6^{AB}}{R_{AB}^6} \tag{1.27}$$

where the C_6^{AB} coefficient describes the strength of the dispersion interactions between the two charge densities A/B.

Since the asymptotic behavior of the dispersion energy is known, it can be included by adding a semi-classical energy expression to the DFT total energy. This scheme is known as dispersion correction and is common for modern computational chemistry. The most widely used scheme to obtain C_6 coefficients are the D3 model[26,27] and its successor D4,[7,9] where the atomic C_6 coefficients are interpolated from precalculated dynamic polarizabilities by a geometric descriptor of the chemical environment of the atom, like the coordination number (CN) or the atomic partial charge. Central to the determination of the C_6^{AB} coefficients is the information of the local environment of the atomic side, which is captured via the coordination number in D3 and additionally the partial charge in D4. The computation of the coordination numbers employs a short-range counting function f_{CN} to determine the number of neighbors as differentiable quantity given by

$$CN_A = \sum_{\substack{B \\ B\neq A}}^{N_{at}} f_{CN}(R_{AB}) \tag{1.28}$$

D3 and D4 differ in the choice of the counting function. While D4 uses an error function, a smoother exponential function is used in D3. The coordination number is used to interpolate among a set of dynamic polarizabilities which are precalculated by time-dependent DFT for compounds with the respective atom species in different chemical environments. Partial charges in D4 are calculated by a charge equilibration procedure by minimizing an auxiliary energy expression for the electrostatic energy of fluctuation charges.

$$\begin{aligned} J_{\mathrm{EEQ}} = \frac{1}{2}\sum_{\mathrm{AB}}^{N_{\mathrm{at}}} q_{\mathrm{A}} q_{\mathrm{B}} \frac{\mathrm{erf}\left[R_{\mathrm{AB}}/\sqrt{R_{\mathrm{A},0}^{2}+R_{\mathrm{B},0}^{2}}\right]}{R_{\mathrm{AB}}} \\ + \sum_{\mathrm{A}}^{N_{\mathrm{at}}} \left(q_{\mathrm{A}}^{2} U_{\mathrm{A}}^{\mathrm{EEQ}} + q_{\mathrm{A}}\left(\chi_{\mathrm{A}} - k_{\mathrm{A}}^{\mathrm{CN}}\sqrt{\mathrm{CN}_{\mathrm{A}}}\right)\right) \end{aligned} \tag{1.29}$$

where $q_{\mathrm{A/B}}$ are the partial charges, $U_{\mathrm{A}}^{\mathrm{EEQ}}$ the Hubbard parameters for the charge model which are a measure for the chemical hardness of an atom, χ_{A} the electronegativities, $k_{\mathrm{A}}^{\mathrm{CN}}$ the scaling factor for the environment dependency of the electronegativities, and $R_{\mathrm{A/B},0}$ the widths of the Gaussian charge distributions on each atom. By minimizing this energy expression under the constraint of the total charge of the system a set of partial charges is obtained, which are employed in the D4 dispersion model to scale the atomic polarizabilities. The resulting dynamic polarizabilities $\alpha(iu)$ are integrated using the Casimir–Polder scheme for set of predefined frequencies

$$C_{6}^{\mathrm{AB}} = \frac{\pi}{3}\sum_{\mathrm{A,ref}}^{N_{\mathrm{ref}}}\sum_{\mathrm{B,ref}}^{N_{\mathrm{ref}}}\sum_{u}^{23} w_{u} w_{\mathrm{A}}^{\mathrm{CN}} w_{\mathrm{B}}^{\mathrm{CN}} w_{\mathrm{A}}^{q} w_{\mathrm{B}}^{q} \alpha_{\mathrm{A,ref}}(iu)\alpha_{\mathrm{B,ref}}(iu) \tag{1.30}$$

with w_{u} being the weight of the frequency integration grid, $w_{\mathrm{A/B}}^{\mathrm{CN}}$ the weight of the interpolation over the coordination number of atom A/B, and $w_{\mathrm{A/B}}^{q}$ the extrapolation based on the partial charge of atom A/B. The charge scaling function is given as

$$w_{\mathrm{A}}^{q} = \exp\left[\beta_{\max}\left(1-\exp\left[2U_{\mathrm{A}}^{\mathrm{D4}}\left(1-\frac{Z_{\mathrm{A}}+q_{\mathrm{A}}^{\mathrm{ref}}}{Z_{\mathrm{A}}+q_{\mathrm{A}}}\right)\right]\right)\right] \tag{1.31}$$

where $\beta_{\max}$ defines the maximum charge scaling by large negative charges, $U_{\mathrm{A}}^{\mathrm{D4}}$ are Hubbard parameters for each atom in D4, Z_{A} are the nuclear charges of the atoms accounting for effective core potentials used in reference calculation, and $q_{\mathrm{A}}^{\mathrm{ref}}$ are the partial charges obtained for the reference calculation. The semi-classical correlation energy is calculated from the pairwise sum of the dispersion coefficients using

$$E_{c}^{\mathrm{NL}} = -\frac{1}{2}\sum_{\mathrm{AB}}^{N_{\mathrm{at}}}\sum_{n}^{6,8,\ldots} s_{n}\cdot f_{n}(R_{\mathrm{AB}})\cdot\frac{C_{n}^{\mathrm{AB}}}{R_{\mathrm{AB}}^{n}} \tag{1.32}$$

where s_{n} are scaling parameters usually fixed to unity and f_{n} is a damping function to remove the dispersion energy in the short-range regime already covered by the semi-local correlation functional. One possible choice is the rational damping function proposed by Becke and Johnson[67,68]

$$f_n(R_{AB}) = \frac{R^n_{AB}}{R^n_{AB} + R^n_{AB,\,\text{crit}}} \tag{1.33}$$

with the critical radius $R_{AB\,\text{crit}}$ calculated by $a_1 \cdot (C_8^{AB}/C_6^{AB})^{1/2} + a_2$ with the functional specific parameters a_1 and a_2. For D3 and D4 the series is truncated at the C_8^{AB} coefficient and the s_8 parameter is made functional specific to implicitly account for higher order terms. Furthermore the leading contribution for the non-additive many-body dispersion energy is evaluated by the Axilrod–Teller–Muto[69,70] formula

$$E^{NL}_{c,ATM} = -\frac{1}{6} \sum_{ABC}^{N_{at}} s_9 \cdot f_9(R_{AB}, R_{BC}, R_{AC}) \cdot \frac{C_9^{ABC}\,(1 + 3\cos[\vartheta_A]\cos[\vartheta_B]\cos[\vartheta_C])}{(R_{AB}R_{BC}R_{AC})^3} \tag{1.34}$$

where $\vartheta_{A/B/C}$ are the angles for the respective triple A, B, C. For an extensive discussion on the D4 dispersion model, it is referred to Ref. [9].

1.2.3 Basis set expansion

Until now the functional form of the one-particle wavefunctions, also known as molecular orbitals (MOs), has not been discussed in detail. Since the exact mathematical shape is only known for limiting cases like the free hydrogen atom, the molecular orbitals are usually expanded in a basis set. While different choices of basis sets are possible, this discussion will focus on the linear combination of atomic orbital (LCAO) approach[48,49] here. Atom-centered orbitals have the advantage of being well-defined for both molecular and periodic systems. For the latter basis set expansion the Bloch theorem is satisfied as

$$\psi_{j\sigma}(\mathbf{r} + \mathbf{L}, \mathbf{k}) = \psi_{j\sigma}(\mathbf{r}, \mathbf{k}) \exp[i\mathbf{k} \cdot \mathbf{L}] \tag{1.35}$$

where $\mathbf{k}$ is a point in momentum space, $\mathbf{L}$ a multiple of the lattice vectors, j is the orbital/band index, and σ the spin channel, *i. e.*, α or β. A suitable choice for the basis set expansion is the Bloch function

$$\psi_{j\sigma}(\mathbf{r}, \mathbf{k}) = \sum_\mu C_{\mu j\sigma}(\mathbf{k}) \frac{1}{\sqrt{N_L}} \sum_{\mathbf{L}}^{N_L} \varphi^{\mathbf{L}}_\mu(\mathbf{r}) \exp[i\mathbf{k} \cdot \mathbf{L}] \tag{1.36}$$

where $\varphi^{\mathbf{L}}_\mu$ is the atomic orbital in cell $\mathbf{L}$ normalized over the number of included image cells N_L. A common choice for the atomic orbital are contracted Gaussian type basis functions

$$\varphi^{\mathbf{L}}_\mu = \frac{1}{\mathcal{N}} \sum_m^{N_{prim}} c_m \exp\left[-\alpha_m \left(\mathbf{r} - (\mathbf{R}_\mu + \mathbf{L})\right)^2\right] \tag{1.37}$$

where $\mathcal{N}$ is the normalization constant such that $\langle \varphi_\mu | \varphi_\mu \rangle = 1$, c_m are the contraction coefficients and α_m are the exponents for the N_{prim} primitive Gaussian type basis functions. Inserting the expanded molecular orbitals in the Schrödinger equation results in a set of general eigenvalue equations

$$\sum_{\mathbf{L}} \exp[i\mathbf{k} \cdot \mathbf{L}] \sum_\mu^{N_L} \left(\mathbf{H}^{\mathbf{0L}}_{\mu\nu} - \varepsilon_{j\sigma}(\mathbf{k}) \cdot \mathbf{S}^{\mathbf{0L}}_{\mu\nu}\right) \mathbf{C}_{\mu j\sigma}(\mathbf{k}) = 0 \tag{1.38}$$

using the orthogonality between the different points in $\mathbf{k}$-space, the resulting full matrices are block-diagonal. For each of the $\mathbf{k}$-points the resulting Roothaan–Hall equation can be solved

$$\mathbf{H}(\mathbf{k})\mathbf{C}(\mathbf{k}) = \varepsilon(\mathbf{k})\mathbf{S}(\mathbf{k})\mathbf{C}(\mathbf{k}) \tag{1.39}$$

due to the dependence of the Hamiltonian matrix **H** on the density and orbitals the solution of the eigenvalue problem is iterated until self-consistency is reached. The density matrix **P** can be evaluated from the orbital coefficients and the occupation numbers **n** as

$$\mathbf{P}(\mathbf{k}) = \mathbf{C}(\mathbf{k})\mathbf{n}(\mathbf{k})\mathbf{C}^{\mathrm{T}}(\mathbf{k}) \tag{1.40}$$

The occupation numbers are usually chosen by the aufbau principle, however for metallic systems or systems with small band gap a finite electronic-temperature distribution is preferred. A possible choice is the Fermi-distribution[71] given as

$$n_{j\sigma}(\mathbf{k}) = \Big(\exp[(\varepsilon_{j\sigma}(\mathbf{k}) - \varepsilon_{\text{Fermi},\,\sigma})/(k_B T_{\text{el}})] - 1\Big)^{-1} \tag{1.41}$$

where $\varepsilon_{\text{Fermi},\,\sigma}$ is the Fermi level and $k_B T_{\text{el}}$ is the Boltzmann constant times the temperature. Since the Fermi-distribution represents an ensemble of electronic structures, the electronic entropy of this ensemble must be accounted for in the total energy using the Fermi free energy defined as

$$G_{\text{Fermi}} = k_B T_{\text{el}} \sum_{\mathbf{k}} \sum_{\sigma}^{\alpha,\beta} \sum_{j}^{N_{\text{AO}}} \Big(n_{j\sigma}(\mathbf{k}) \ln[n_{j\sigma}(\mathbf{k})] + (1 - n_{j\sigma}(\mathbf{k})) \ln[1 - n_{j\sigma}(\mathbf{k})]\Big)\,. \tag{1.42}$$

The finite-temperature treatment provides an efficient way to approximately handle static correlation effects without sacrificing computational efficiency for a method.[72]

1.3 Semiempirical molecular orbital methods

While density functional theory, especially in the framework of composite electronic structure methods[34], has become a crucial tool for computational chemistry, it is still too computationally expensive for explorative work or screening application. Semiempirical methods however provide a pragmatic approach to reduce the computational complexity by directly parametrizing the Hamiltonian for a given basis set expansion.

The tight-binding approach is based on a semiempirical approximation to KS-DFT, where the energy function from Eq. 1.17 is expanded around a known reference density ρ_0 as

$$E[\rho] = E^{(0)}[\rho_0] + E^{(1)}[\rho_0, \delta\rho] + E^{(2)}[\rho_0, (\delta\rho)^2] + E^{(3)}[\rho_0, (\delta\rho)^3] + O((\delta\rho)^4)\,. \tag{1.43}$$

The reference density ρ_0 is set here to the superposition of atomic densities and the energy expressed in terms of charge fluctuations $\delta\rho$. The energy contributions are grouped by the order of the charge fluctuation with the series expansion being truncated at third order.

The zeroth order contributions to the energy are independent of the charge density but include geometry dependent contributions like the Coulomb repulsion energy from screened nuclei given by

$$E_{\mathrm{rep}}^{(0)} = \frac{1}{2}\sum_{AB}^{N_{\mathrm{at}}} \frac{Z_A^{\mathrm{eff}} Z_B^{\mathrm{eff}}}{R_{AB}} \exp\left[-\sqrt{a_A a_B}\cdot R_{AB}^{k_{\mathrm{rep}}}\right] \tag{1.44}$$

where $Z_{A/B}^{\mathrm{eff}}$ are the effective nuclear charges, $a_{A/B}$ are the atomic radii, and k_{rep} is an element-pair specific parameter mostly kept constant. Similarly, the D3 dispersion energy is part of the zeroth order contribution to the energy being independent of the charge-fluctuation (cf. Eq. 1.32).

The contribution from the core Hamiltonian is the main contribution of the first order energy

$$E_{\mathrm{EHT}}^{(1)} = \frac{1}{2}\sum_{\kappa\lambda} P_{\kappa\lambda} H_{\lambda\kappa} \tag{1.45}$$

where $H_{\lambda\kappa}$ is the effective one-electron Hamiltonian describing the interaction of neutral atoms. The density matrix is obtained by solving the Roothaan–Hall equation (cf. Eq. 1.39) and constructing the density matrix (cf. Eq. 1.40). The Hamiltonian matrix elements $H_{\kappa\lambda}$ are obtained by scaling the average of the on-site level energies $H_{\kappa\kappa/\lambda\lambda}$ with the overlap matrix $S_{\kappa\lambda}$ and a shell-pair and distant dependent polynomial $\Pi_{\kappa\lambda}$

$$H_{\kappa\lambda} = \frac{H_{\kappa\kappa} + H_{\lambda\lambda}}{2} \cdot S_{\kappa\lambda} \cdot \Pi_{\kappa\lambda} \tag{1.46}$$

where the polynomial term takes the form of

$$\Pi_{\kappa\lambda} = k_{\kappa\lambda} \cdot \left(1 + p_\kappa \sqrt{\frac{R_{\kappa\lambda}}{R_{\kappa\lambda}^{\mathrm{vdw}}}}\right) \cdot \left(1 + p_\lambda \sqrt{\frac{R_{\kappa\lambda}}{R_{\kappa\lambda}^{\mathrm{vdw}}}}\right) \cdot \left(\frac{2\sqrt{\zeta_\kappa \zeta_\lambda}}{\zeta_\kappa + \zeta_\lambda}\right)^{w_{\mathrm{exp}}} \tag{1.47}$$

with $k_{\kappa\lambda}$ being a scaling for the respective atoms κ/λ are centered on, $\zeta_{\kappa/\lambda}$ the Slater exponents of the basis functions, w_{exp} is the weight of the Slater exponent dependent term, $p_{\kappa/\lambda}$ are the scaling factors for the distant dependent scaling, and $R_{\kappa\lambda}^{\mathrm{vdw}}$ are the van-der-Waals radii between the respective atoms. Overall the construction of the Hamiltonian hopping elements is the most sophisticated term in the xTB method with respect to the number of parameters used. The atomic level energies $H_{\kappa\kappa/\lambda\lambda}$ are further dependent on their local environment by

$$H_{\kappa\kappa} = h_\kappa + k_\kappa^{\mathrm{CN}} \mathrm{CN}_\kappa \tag{1.48}$$

where CN_κ is the coordination number defined as in Eq. 1.28, k_κ^{CN} the local level shift factor, and h_κ the self energy the orbital in the isolated atom.

Contributions from second and higher order require the self-consistent solution of the Roothaan–Hall equations. Rather than working with the full density matrix, the density fluctuations are expanded in multipole moments

$$\delta\rho = -\sum_\kappa q_\kappa - \sum_\kappa \mu_\kappa - \sum_\kappa \Theta_\kappa + O(\xi^{(3)}) \tag{1.49}$$

with q_κ being the partial charges for each atomic orbital, μ_κ the orbital dipole moment and Θ_κ the traceless orbital quadrupole moment. The partial charges are obtained by Mulliken population analysis using the overlap matrix **S**

$$q_\kappa = n_\kappa^{\mathrm{ref}} - \sum_\lambda P_{\kappa\lambda} S_{\lambda\kappa} \tag{1.50}$$

where n_κ^{ref} is the population of the neutral atom reference for the orbital κ. The orbital dipole moments are obtained in a similar way using

$$\mu_\kappa = -\sum_\lambda P_{\kappa\lambda} \underbrace{\langle\varphi_\lambda|\mathbf{r} - \mathbf{R}_\kappa|\varphi_\kappa\rangle}_{D_{\lambda\kappa,\kappa}} \tag{1.51}$$

using the dipole moment integral element $D_{\lambda\kappa,\kappa}$ evaluated with the dipole operator on the center of basis function $\mathbf{R}_\kappa$. The (traceless) quadrupole moments are computed from

$$\Theta_\kappa = -\sum_\lambda P_{\kappa\lambda} Q_{\lambda\kappa,\kappa} \tag{1.52}$$

with the quadrupole moment integral elements $Q_{\lambda\kappa,\kappa}$ being defined as for the dipole moment integrals. The resulting multipole moments are used to define the electrostatic energy. Most notably, the isotropic contribution to the Coulomb electrostatic is given by

$$E_{\text{ies}}^{(2)} = \frac{1}{2}\sum_{\kappa\lambda} q_\kappa T_{\kappa\lambda} q_\lambda \tag{1.53}$$

and the anisotropic contributions to the electrostatic energy collected up to the distance dependency of R^{-3} is provided by

$$E_{\text{aes}}^{(2)} = \sum_{\kappa\lambda} \mu_\kappa^i T_{\kappa\lambda}^i q_\lambda + \frac{1}{2}\sum_{\kappa\lambda} \mu_\kappa^i T_{\kappa\lambda}^{ij} \mu_\lambda^j + \sum_{\kappa\lambda} \Theta_\kappa^{ij} T_{\kappa\lambda}^{ij} q_\lambda + O(R_{\kappa\lambda}^{-4}) \tag{1.54}$$

where $T_{\kappa\lambda}, T_{\kappa\lambda}^i, T_{\kappa\lambda}^{ij}$ are the interaction tensors for the respective multipole moments $q_{\kappa/\lambda}, \mu_{\kappa/\lambda}^i, \Theta_{\kappa/\lambda}^{ij}$, Einstein summation convention for the Cartesian indices $i, j, \ldots$ is assumed. The interaction tensor for the charge–charge interaction is defined by

$$T_{\kappa\lambda} = \left(R_{\kappa\lambda}^2 + f_{\text{av}}(U_\kappa, U_\lambda)^{-2}\right)^{-\frac{1}{2}} \tag{1.55}$$

where $U_{\kappa/\lambda}$ are the Hubbard parameters of the respective orbitals, also known as chemical hardnesses, and f_{av} is an averaging function like the arithmetic or harmonic average. Due to the spherical atomic reference the chemical hardness is unique for each angular momentum ℓ and chemical species, allowing to the use of the same value for orbitals from the same shell. Shell-resolved partial charge $q_{A,\ell}$ are defined by accumulating the orbital partial charges with

$$q_{A,\ell} = \sum_{\kappa\in A,\ell} q_\kappa\,. \tag{1.56}$$

The shell-resolved partial charges are used in the third-order on-site contributions to the electrostatic energy

$$E_{\text{ies},\ell}^{(3)} = \frac{1}{3}\sum_A \sum_{\ell\in A} \Gamma_{A,\ell} q_{A,\ell}^3 \tag{1.57}$$

where $\Gamma_{A,\ell}$ is the shell-resolved derivative of the chemical hardness, known as Hubbard derivative. The multipole interactions are given by the charge–dipole interaction tensor

$$T^{i}_{\kappa\lambda} = -f(R_{\kappa\lambda})\frac{R_{\kappa\lambda,i}}{R^{3}_{\kappa\lambda}} \tag{1.58}$$

and the dipole–dipole, charge–quadrupole interaction tensor

$$T^{ij}_{\kappa\lambda} = f(R_{\kappa\lambda})\frac{3R_{\kappa\lambda,i}R_{\kappa\lambda,j} - \delta_{ij}R^{2}_{\kappa\lambda}}{R^{5}_{\kappa\lambda}} \tag{1.59}$$

where $f(R_{\kappa\lambda})$ is a function to remove the interaction at short distances. The damping function is given as

$$f(R_{\kappa\lambda}) = \left(1 + 6\left(\frac{R_{\kappa,0} + R_{\lambda,0}}{2R_{\kappa\lambda}}\right)^{k}\right)^{-1} \tag{1.60}$$

with $R_{\kappa/\lambda,0}$ as the radii of the respective atoms and k the exponent for the damping. Since the interaction kernel for the higher multipole moments does not contain parameters dependent on the angular momentum, the expressions are unique for each atom and can define atomic partial charges and multipole moments. The atomic partial charges q_A, atomic dipole moments μ_A, and atomic quadrupole moments Θ_A are obtained by summing the orbital resolved quantities from the respective atoms.

$$q_A = \sum_{\kappa\in A} q_\kappa\,, \qquad \mu_A = \sum_{\kappa\in A} \mu_\kappa\,, \qquad \Theta_A = \sum_{\kappa\in A} \Theta_\kappa\,. \tag{1.61}$$

A special contribution is the self-consistent dispersion, which is due to the non-linear dependency on the partial charges in Eq. 1.31 not clearly associated with a single order in the density fluctuation. Instead, the series expansion of Eq. 1.31 includes all orders of density fluctuations in the self-consistent dispersion energy, if the C_6^{AB} coefficients depend on the Mulliken partial charges. In summary, the final energy expression for GFN2-xTB[8] is given as

$$E_{\text{GFN2}} = E^{(0)}_{\text{rep}} + E^{(1)}_{\text{EHT}} + E^{(2)}_{\text{ies}} + E^{(2)}_{\text{aes}} + E^{(3)}_{\text{ies},\ell} + E^{(\infty)}_{\text{D4}} \tag{1.62}$$

In comparison, the GFN1-xTB[41] energy is given by

$$E_{\text{GFN1}} = E^{(0)}_{\text{rep}} + E^{(0)}_{\text{D3}} + E^{(0)}_{\text{XB}} + E^{(1)}_{\text{EHT}} + E^{(2)}_{\text{ies}} + E^{(3)}_{\text{ies}} \tag{1.63}$$

with the most notable differences are in the dispersion correction, using D3, requiring an additional force-field-like correction $E^{(0)}_{\text{XB}}$ for describing halogen bonding and using only atom-resolved partial charges in the third-order onsite electrostatic $E^{(3)}_{\text{ies}}$. Notable is that GFN2-xTB in comparison to GFN1-xTB does not require additional corrections for describing hydrogen and halogen bonding due to the anisotropic electrostatic capturing those interactions naturally. For a more in-depth discussion of the xTB methods Ref. [1] is recommended.

1.4 Free energy

To compare with experimental data, which is usually a free energy difference ΔG the computed energies by electronic structure methods are incomplete. The theoretical obtained energies are computed without considering finite temperature effects so far, neglecting contributions arising from the kinetics of the nuclei which were kept fix by the Born–Oppenheimer approximation. Within the Born–Oppenheimer approximation the kinetics of the nuclei can be added by evaluating the partition function of the system. The Gibbs free energy G for a given temperature T is obtained by

$$G = \underbrace{E + E_{ZPV} + H(0\,K \rightarrow T) + PV}_{H} - TS \tag{1.64}$$

where E is the energy recovered in the electronic Hamiltonian, E_{ZPV} is the zero point vibrational energy arising from the quantum mechanical motions at $0\,K$, $H(0\,K \rightarrow T)$ is the temperature dependent contribution from the motion of the system in translation, rotation and vibration, and PV is the volume work contribution. These contributions together form the enthalpy H, which combined with the ensemble entropy TS yields the Gibbs free energy.

The enthalpy and ensemble entropy can be obtained from the system's partition function Z by evaluating

$$G = \frac{V}{\beta}\left(\frac{\partial \ln Z}{\partial V}\right)_T - \frac{1}{\beta}\ln Z \tag{1.65}$$

for convenience β^{-1} will be used to express the dependency on the thermal energy k_BT, with k_B being the Boltzmann constant. Furthermore, the partition function will be considered in its logarithmic form $\ln Z$ in the following equations. The total partition function can be calculated from the individual contributions by

$$\ln Z = \ln Z_{el} + \ln Z_{tr} + \ln Z_{rot} + \ln Z_{vib} \tag{1.66}$$

where Z_{el} is the electronic partition function, Z_{tr} the translational one, Z_{rot} the rotational one, and Z_{vib} the vibrational one. The electronic partition function is calculated from the electronic states $\{I\}$ as

$$\ln Z_{el} = \ln\left[\sum_I g_I \exp[-\beta E_I]\right] \tag{1.67}$$

with g_I being the degeneracy of the electronic state degeneracy, and E_I its energy. For most applications only the ground state is relevant as the excitation energy to the first excited state is usually larger than β^{-1}. The translational partition function of an ideal gas is given by

$$\ln Z_{tr} = \ln\left[\left(\frac{1}{h}\sqrt{\frac{2\pi M}{\beta}}\right)^3 V\right] \tag{1.68}$$

where M is the mass of the particle. For the rotational partition function a rigid rotor approximation is chosen

$$\ln Z_{rot} = \ln\left[\frac{1}{\sigma}\sqrt{\frac{\pi T^3}{I_A I_B I_C}}\right] \tag{1.69}$$

with σ being the rotational symmetry number and $I_{A/B/C}$ as the principal moments of inertia. Both the translational and rotational partition function are most important if the number of particles changes and translational and rotational degrees of freedom of the individual compounds become vibrational modes of the complex. Finally, the vibrational partition function is commonly approximated by an ensemble of m uncoupled harmonic oscillators (HO) as in

$$\ln Z_{\mathrm{vib}}^{\mathrm{HO}} = \sum_{m}^{N_{\mathrm{vib}}} \ln \left[\frac{\exp[-\hbar\omega_m \beta]}{1 - \exp[-\hbar\omega_m \beta]} \right] \tag{1.70}$$

where ω_m are the reduced frequencies of the respective harmonic oscillator. However, this approximation is insufficient for large systems with soft modes and internal rotations. A modification scheme to account for the entropic contribution of the low-lying modes as hindered rotor (HR) has been proposed.[73] The modified partition function[6] is using a hindered rotor for low frequencies and a harmonic oscillator for high ones given by

$$\ln Z_{\mathrm{vib}} = \sum_{m}^{N_{\mathrm{vib}}} f(\omega_m) \ln \left[Z_{\mathrm{vib}}^{\mathrm{HO}}(\omega_m) \right] + (1 - f(\omega_m)) \ln \left[Z_{\mathrm{vib}}^{\mathrm{HR}}(\omega_m) \right] \tag{1.71}$$

where the hindered rotor partition function is calculated as

$$\ln Z_{\mathrm{HR}}^{\mathrm{vib}}(\omega_m) = \frac{1}{2} \ln \left[\frac{2\mu'(\omega_m)}{\pi\hbar^2\beta} \right] \qquad \text{with} \quad \mu'(\omega_m) = \frac{\mu(\omega_m)\bar{I}}{\mu(\omega_m) + \bar{I}} \quad \text{and} \quad \mu(\omega_m) = \frac{\hbar}{4\pi\omega_m}, \tag{1.72}$$

with $\bar{I}$ being the average moment of inertia of the molecule and μ as the moment of inertia corresponding to the normal mode. For the switching function $f(\omega_m)$ the Chai–Head-Gordon damping function[74] is chosen, which is defined by

$$f(\omega_m) = \left(1 + (\omega_0/\omega_m)^4\right)^{-1} \tag{1.73}$$

where ω_0 is an predefined rotor cutoff value typically between 20 to 100 cm^{-1}.

1.4.1 Solvation free energy

Since chemistry mostly takes place in solution, accounting for solvent effects in computational simulations is crucial. The contribution to the solvation free energy arising from the transfer of the chemical system from the gas phase to the condensed phase can be accounted for by assuming an additive contribution to the total free energy:

$$G = G_{\mathrm{elec,\,vac}} + \delta G_{\mathrm{solv}} \tag{1.74}$$

where δG_{solv} is the free energy difference required to bring the solute from the gas phase into solvation. The contribution can be calculated in a straight-forward way

$$\delta G_{\mathrm{solv}} = G_{\mathrm{elec,\,sol}} - G_{\mathrm{elec,\,vac}} + \delta G_{\mathrm{state}} \tag{1.75}$$

where $G_{\text{elec, solv/vac}}$ is the system energy in solution or gas phase, respectively, the δG_{state} arises from the volume work of bringing the solute from the gas phase into the condensed phase. This approach is shown schematically in Fig. 1.3.

A computational efficient strategy for computing solvation free energies is the usage of implicit solvation models. In an implicit solvation model the ensemble averaged solvent molecules are replaced by a polarizable dielectric continuum surrounding the solute. The interaction of the solute with the continuum can be described by the Poisson–Boltzmann (PB) equation

$$\varepsilon \Delta \psi = -4\pi\rho \tag{1.76}$$

where ε is the dielectric constant of the continuum, ψ the electrochemical potential and the ρ the electron density. A common approach to approximately solve the (linearized) PB equation is based on a set of spheres with different internal and external dielectric constants, this approach and the resulting solvation models are described in Ch. 5.

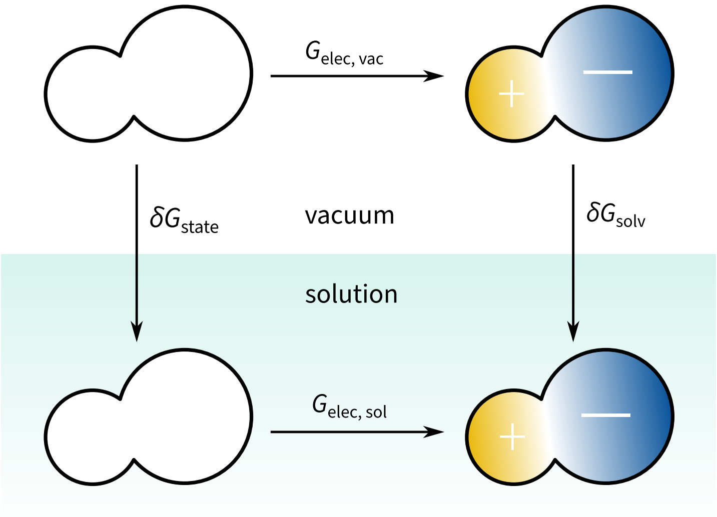


Figure 1.3: Schematic representation of the calculation of the solvation free energy δG_{solv} from a gas phase and an implicitly solvated calculation.

CHAPTER 2

r^2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications

Sebastian Ehlert,[*] Uwe Huniar,[†] Jinliang Ning,[‡] James W. Furness,[‡] Jianwei Sun,[‡] Aaron D. Kaplan,[§] John P. Perdew,[§ ¶] and Jan Gerit Brandenburg[‖]

Received 21 December 2020, Published online 09 February 2021.

Reprinted in Appendix A (adapted) with permission[**] from
S. Ehlert, U. Huniar, J. Ning, J. W. Furness, J. Sun, A. D. Kaplan, J. P. Perdew, and J. G. Brandenburg, *r^2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications*, J. Chem. Phys. **154** (2021) 061101, DOI: 10.1063/5.0041008.

Own manuscript contributions

- performing all calculations except for the periodic DFT optimizations
- interpretation of the results
- writing of the manuscript

[*] Mulliken Center for Theoretical Chemistry, University of Bonn, Beringstr. 4, 53115 Bonn, Germany
[†] Biovia, Dassault Systèmes Deutschland GmbH, Imbacher Weg 46, 51379 Leverkusen, Germany
[‡] Department of Physics and Engineering Physics, Tulane University, New Orleans, Louisiana 70118, United States
[§] Department of Physics, Temple University, Philadelphia, Pennsylvania 19122, United States
[¶] Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States
[‖] Enterprise Data Office, Merck KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany

[**] Permission requests to reuse material from this chapter should be directed to AIP Publishing.

From the recently developed new density functional approximations the strongly constrained and appropriately normed (SCAN) functional[75] is a promising candidate for broad chemical applications due to its rigorous physical construction. As a meta-general gradient approximation (meta-GGA) the SCAN functional is in a favorable spot following Perdew's classification in Jacob's ladder[55] as it retains the favorable computational efficiency of formal cubic scaling with the system size similar to other pure functionals. While many meta-GGAs have been proposed so far, few functionals of this category can actually leverage their potential. The SCAN functional especially exhibits a strong sensitivity to the numerical integration grid,[76] which hampers its applicability for computational simulations. With the regularization of the SCAN functional and subsequent restoration of the exact constraints, the r^{2}SCAN[21] functional promises the same physically rigorous foundation while reducing the numerical instabilities.

The state-of-the-art semi-classical London dispersion correction[7,9] is combined with the semi-local functional to capture crucial long-range correlation effects. For the resulting r^{2}SCAN-D4 functional we find an overall accuracy approaching those of hybrid functionals for a wide range of chemical systems. Notably, the^{2}SCAN-D4 remains numerical robust for molecular geometries, general main group, and organo-metallic thermochemistry as well as for non-covalent interactions in molecular crystals and supramolecular complexes. Additionally to r^{2}SCAN-D4, we also introduce r^{2}SCAN-D3(BJ) and r^{2}SCAN-V based on the established DFT-D3[26,27] and VV10[28,29,62] dispersion corrections, respectively.

For the large GMTKN55 benchmark collection[23] of about 1500 data points, the weighted mean absolute deviation (WTMAD2) is exceptionally small with 7.5 kcal/mol compared to other meta-GGA functionals, which are in a range of 8–9 kcal/mol.[23] This accuracy can of r^{2}SCAN-D4 is moreover transferable to chemically distinct systems like organo-metallic reactions. The performance is especially remarkable for the mindless benchmark (MB16-43) which tests the robustness of a method for dealing with unusual chemistry in artificial molecules. While generally lower errors are found for r^{2}SCAN-D4 and other non-empirical functionals in this test set, their empirical counterparts, like the dispersion corrected B97M functional,[22,63] show significantly larger errors. Bond lengths of main group compounds and transition metal ones show errors of only 0.8% making the functional competitive with hybrid functionals for main group molecules and even outperforming them for transition metal complexes. Also, for condensed systems the r^{2}SCAN-D4 method provides accurate lattice energies of molecular crystals with errors below 1 kcal/mol.

While non-covalent interactions are a weak spot for SCAN-D4, leading to overbinding in molecular crystals and difficulties with hydrogen and halogen bonded systems, r^{2}SCAN-D4 can significantly improve for those categories. We find a systematic improvement over the original SCAN-D4 functional keeping the already good performance and improving the weak spots of the method while improving the numerical stability. Furthermore, we can provide insights into the improvements, as we can attribute the improved descriptions of non-covalent interactions to the regularization going from SCAN-D4 to rSCAN-D4, while the improved thermochemistry and barrier heights result from the exact constraint restoration when going from rSCAN-D4 to r^{2}SCAN-D4. The overall performance of r^{2}SCAN-D4 makes it a consistently accurate density functional approximation for a large variety of chemical problems. Remarkably, since neither SCAN nor r^{2}SCAN were fitted to molecules, their accuracy and predictive power results from the rigorous construction following the exact constraints and appropriate norms. This makes r^{2}SCAN-D4 the first meta-GGA truly accessing its full potential while retaining numerical stability and favorable computational efficiency of a pure functional.

CHAPTER 3

Extended Tight-Binding (xTB) Quantum Chemistry Methods

Christoph Bannwarth,[*] Eike Caldeweyher,[†] Sebastian Ehlert,[†] Andreas Hansen,[†] Philipp Pracht,[†] Jacob Seibert,[†] Sebastian Spicher,[†] and Stefan Grimme[†]

Received 22 April 2020, Published online 09 August 2020.

Reprinted in Appendix B (adapted) with permission[‡] from
C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, *Extended tight-binding quantum chemistry methods*, WIREs Comput. Mol. Sci. **11** (2021) e1493, DOI: 10.1002/wcms.1493.

Own manuscript contributions

- writing the manuscript parts on solvation, implementation and parts of the xTB theory
- development of significant parts of the DFT-D4, GBSA and GFN2-xTB methods
- development and maintenance of the `xtb` and `dftd4` projects

[*] Department of Chemistry and The PULSE Institute, Stanford University, Stanford, CA 94305, United States of America
[†] Mulliken Center of Theoretical Chemistry, Bonn, Germany

[‡] Permission requests to reuse material from this chapter should be directed to Wiley Periodicals LLC.

Semiempirical methods provide computational efficient and reasonably accurate atomistic models for the description of large chemical systems in the gas or condensed phase. From the family of semiempirical quantum mechanics (SQM) the density functional derived density functional tight binding (DFTB) and extended tight binding (xTB) methods offer a framework for consistent parametrizations over a broad range of elements. Designed from the very beginning as special-purpose tools with a focus on structural properties, the geometry, frequency, and non-covalent interaction (GFN) parametrizations of the xTB Hamiltonian provide a physically sound description at the SQM level of theory. The general applicability together and the high robustness make the GFN family of methods, including the xTB parametrizations and the GFN-FF method, very attractive for many fields of chemical research employing computational simulations. All GFN methods share the availability for a very large part of the periodic table (up to radon, $Z \leq 86$) and are obtained by a consistent optimization against accurate gas-phase theoretical references. While the GFN methods are currently mostly applied for molecular systems, the extension to periodic boundary conditions, at least for molecular crystals is well within scope for the available parametrizations.

Furthermore, the GFN2-xTB method introduces improvements in the underlying tight binding theory by using anisotropic electrostatics and charge-dependent dispersion contributions.[8] GFN2-xTB provides the first model of multipole electrostatic for a chemically diverse range of elements in the context of tight binding theory. It resolves the long-standing issues of describing noncovalent interactions (NCIs), like hydrogen bonding or halogen bonding, which are especially difficult to capture in the minimal valence atomic orbital basis set usually employed with tight binding Hamiltonians. The inclusion of charge-dependent dispersion in the Hamiltonian provides an additional edge for accurately capturing non-additive effects relevant to NCIs in large systems. For this purpose, the D4 dispersion model is introduced selfconsistently in the tight binding formalism allowing to vary the dispersion contributions dependent on the local charge of the atomic centers. Together with the xTB Hamiltonian also a quantum mechanical charge model for the D4 model is devised which can provide a more accurate description of the local electronic structure than any classical charge model as used with D4 by default. Overall, the GFN2-xTB provides a tight binding method free from any pairwise parameters and force-field-like contributions.

Important for the wide adoption of the GFN methods is the freely available `xtb` program package, which provides an efficient and user-friendly implementation of the GFN methods. Central point in the development of `xtb` was the possibility to integrate with workflow drivers, like CREST,[19] or existing community-developed frameworks, like QCEngine[11] or ASE,[77] and was quickly adapted there. This enables full geometry optimizations, frequency calculations, or conformational searches on standard workstations or even laptop computers for systems composed of hundreds to about a thousand atoms in minutes to a few hours. The xTB methods also found interest in massively parallel applications like the recent exascale computation of a several million atom system based on the GFN1-xTB method.[40] Another relevant application is the large-scale screening of medium-sized compounds with several thousand and more candidates,[20] where the computational efficiency of the SQM method is of most importance.

In summary, the xTB Hamiltonian provides a good framework for developing new SQM methods due to its flexibility and strict global and element-specific parametrization. Especially, avoiding the tedious pairwise parametrization which has seriously hampered the creation of SQM methods widely applicable for the chemical space, provides an intrinsic advantage for xTB-based methods over other SQM approaches like PM*x* or DFTB. Finally, the GFN methods provide access to free energy computations as further discussed in Chapter 5.

CHAPTER 4

Conformational Energy Benchmark for Longer *n*-Alkane Chains

Sebastian Ehlert,[*] Stefan Grimme,[*] and Andreas Hansen[*]

Received 09 April 2022, Published online 26 May 2022.

Reprinted in Appendix C (adapted) with permission[†] from
S. Ehlert, S. Grimme, and A. Hansen, *Conformational Energy Benchmark for Longer n-Alkane Chains*,
J. Phys. Chem. A **126** (2022) 3521, DOI: `10.1021/acs.jpca.2c02439`.

Own manuscript contributions

- performing all calculations except for the wavefunction reference calculations
- interpretation of the results
- writing of the manuscript

[*] Mulliken Center of Theoretical Chemistry, Bonn, Germany

[†] Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

Evaluating conformational energies is a unique challenge for every ensemble, especially for flexible systems with lots of conformers and nearly contiguous ensembles. Long, unbranched alkane chains provide a prototypical example of a highly flexible system and form the basis for the construction of the ACONFL benchmark set. Due to the small energy differences in the nearly contiguous conformer ensemble, very accurate reference values by DLPNO-CCSD(T1)/CBS are required to allow a statistically meaningful comparison of even the best available density functional approximations. ACONFL comprises three conformational ensembles with 53 conformers and 50 relative energies up to about 8 kcal/mol. The three subsets are build from the conformers of *n*-dodecane $C_{12}H_{24}$, *n*-hexadecane $C_{16}H_{34}$, and *n*-icosane $C_{20}H_{42}$ covering the transition from linear to hairpin structures as energetically lowest conformers. Compared to the ACONF test set[78] the conformers used in the ACONFL benchmark make the use of dispersion corrected methods indispensable, while on the ACONF benchmark even dispersion uncorrected functionals could yield small errors.

The spectrum of assessed methods includes established and modern density functional approximations as well as wavefunction theory methods like second-order Møller–Plesset (MP2) or Hartree–Fock (HF). For the majority of assessed density functionals we find that the best dispersion corrected methods are using the D4 dispersion correction,[9] providing on average smaller errors compared to D3[26,27] and VV10.[28,29,62] Among the overall best performing methods are the recently introduced functionals r^2SCAN-V[4] and its hybrid variant r^2SCAN0-V.[5] The overall impact of exact exchange for alkane chain conformers is for the tested methods minor, only double hybrids like the DSD-BLYP-D3(BJ) can further improve on the already excellent performance of many (meta-)GGAs. However, we find that already composite methods like B97-3c[79] and r^2SCAN-3c[14] can provide an accurate description of the conformational energy. The "3c" composite methods can maintain an excellent cost–accuracy ratio compared to computationally demanding hybrid or double hybrid functionals in large atomic orbital basis sets for comparable results with 2–3 orders of magnitude reduced cost. Surprisingly good performance was observed with HF-D4 while the perturbation-based wavefunction methods, like MP2, perform rather badly. The failure of MP2 can be attributed to the uncorrelated HF-based dispersion coefficients and missing higher-order contributions which can be efficiently included with the D4 dispersion correction. Attempts to remedy this issue by introducing correlated dispersion coefficients or higher-order correlation effects are only partially successful compared to the excellent HF-D4 performance.

Furthermore, we investigate commonly used semiempirical and force field methods, which are used in the generation of conformation ensembles or large-scale molecular dynamics simulations. Although it should be a seemingly straightforward problem due to the simple electronic structure of alkanes, only a few of the tested methods performed convincingly. Among the tested methods providing reliable results, GFN2-xTB[8] and PM6-D3H4[80] show good performance at the semiempirical quantum mechanical level of theory. An outstanding cost–accuracy ratio is available at the force field level with the recently introduced general force field, GFN-FF.[81] However, for several commonly used force field methods like the universal force field[82] or the MMFF94[83,84] too shallow potential energy surfaces and larger errors were observed. Therefore, the ACONFL provides a meaningful validation set for testing existing and new semiempirical and force field methods for the description of conformational energies.

CHAPTER 5

A robust and efficient implicit solvation model for fast semiempirical methods

Sebastian Ehlert,[*] Marcel Stahn,[*] Sebastian Spicher,[*] and Stefan Grimme[*]

Received 12 May 2021, Published online 29 June 2021.

Reprinted in Appendix D (adapted) with permission[†] from
S. Ehlert, U. Huniar, J. Ning, J. W. Furness, J. Sun, A. D. Kaplan, J. P. Perdew, and J. G. Brandenburg, *r^2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications*, J. Chem. Phys. **154** (2021) 061101, DOI: 10.1063/5.0041008.

Own manuscript contributions

- performing all calculations except for the parameter optimization
- interpretation of the results
- writing of the manuscript

[*] Mulliken Center of Theoretical Chemistry, Bonn, Germany

[†] Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

Describing solvation effects at a semiempirical or force field level of theory is essential for generating conformational ensembles correctly accounting for environment effects. The solvation can impact the conformational ordering as well as the most stable structures. To account for solvation effects in the conformer generation process robust and computationally efficient solvation models are needed.

We propose an implicit solvation model based on the analytical linearized Poisson–Boltzmann (ALPB) model[85] and parametrize it for extended tight binding (xTB),[1] density functional tight binding (DFTB), and the general force field, GFN-FF.[81] By using an extension to generalized Born (GB) theory[86,87] to account for finite dielectric constants,[85] a newer interaction kernel for the Born matrix,[88] and a smooth surface area (SA) integration for the non-polar contribution,[89] the developed solvation model combines recent advances for reaction field implicit solvation models.

We parametrize the models for a broad range of different nonpolar and polar as well as protic and aprotic solvents covering a wide range of dielectric constants. For the parametrization, we optimize against reference solvation free energies either obtained by experiment as collected in the MNSOL database[90–92] or calculated at the COSMO-RS[93] level of theory. The parametrization also includes references from charged compounds to obtain a consistent parametrization for neutral and ionic solutes. For the xTB and GFN-FF methods, a complete parametrization of all elements of the periodic table up to Radon ($Z \leq 86$) is available. While for Slater–Koster based DFTB the inherently available number of elements is much smaller, we can exploit the implicit nature of the solvation model to enable the description of solvents that would be unavailable with the respective parametrization in an explicitly solvated approach.

The combination of the new ALPB and established GBSA methods with xTB, DFTB and GFN-FF are tested on a broad range of systems and applications, from conformational energies over transition-metal complexes to large supramolecular association reactions of charged species. In our tests GFN1-xTB(ALPB) is reaching the accuracy of sophisticated explicitly solvated approaches for calculating hydration free energies of small molecules on the FreeSolv database,[94,95] yielding a mean absolute deviation of only 1.4 kcal/mol compared to experiment. For logarithmic octanol–water partition coefficients ($\log K_{\text{ow}}$) we find good agreement between GFN2-xTB(ALPB) and experiment with a mean absolute deviation of 0.65, which emphasizes the consistent description of different solvents. Furthermore, we propose a set of reaction solvation free energies for the supramolecular S30L benchmark set,[96] which can be well reproduced by the best available solvation model, like COSMO-RS.[97–99] The ALPB-based solvation models provide a reasonably well accuracy even for such large complexes.

Due to the ready availability of analytical gradients, the proposed ALPB and GBSA solvation models routinely allow for energy calculations, geometry optimizations, molecular dynamics, and vibrational frequency computations. We further check the influence of the solvation models on the geometry relaxation in a qualitative and semi-quantitative benchmark set by comparing the difference in geometry parameters like bond lengths and angles between gas phase and implicitly solvated structures using DCOSMO-RS.[93] Especially for medium-sized charged solutes, we find noticeable changes which can be reproduced by the ALPB solvation model. For the investigated systems we also find that the solvation contributions have only a minor impact on the magnitude of the thermostatistical contributions and vibrational frequencies. Overall, the combination of tailored implicit solvation models with semiempirical methods opens a wide range of chemical applications.

CHAPTER 6

Summary and Outlook

The development and discovery of chemical compounds provide the drive for creating new experimental techniques to synthesize or measure these compounds as well as theoretical models to verify or predict their properties. With the experimental possibilities of investigating larger or chemically more diverse systems, the demand for theory to provide methods able to efficiently handle more atoms and more elements steadily increases. On the other hand, the predictive power of theoretical models has been steadily improving and can, for certain system sizes, compete with or even outperform experimental accuracy. This trend leads to computational simulations targeting more challenging cases, following or even setting the research directions for larger and chemically more diverse systems in the field of chemistry. Here the development of the extended tight binding (xTB) methods provides new possibilities from exploratory computational chemistry and early-stage research up to large-scale screening and long-running simulations. Of importance is the readily availability for the majority of chemical elements providing the freedom to explore many branches of chemical applications. While the xTB methods introduced new possibilities, they are by no means the solution to all computational chemistry problems. For many computational simulations, a reranking or verification at a higher level of theory is required, which is usually found with dispersion corrected density functional theory (DFT). Also for the development of DFT the ongoing challenge is to devise new and better density functionals or mend known deficiencies like the absence of long-range correlation effects or the infamous self-interaction in semi-local functionals. Having access and knowledge to the best possible density functional is crucial for the success of a computational study. Benchmarking of density functional methods is a crucial step toward providing reliable measures for selecting the best method for the task at hand.

This work shows the importance of employing dispersion corrections together with newly developed density functionals to accurately assess their performance over a wide range of chemical applications. State-of-the-art dispersion corrections like the charge-dependent D4 model provide a consistent and robust performance for non-covalent interactions or medium-to-large-sized systems where long-range correlation effects are crucial. With extensive benchmarking on the now established GMTKN55 benchmark collection, a good measure for the performance of a newly proposed method for general main group thermochemistry and reaction barriers as well as non-covalent interactions can be provided. Furthermore, testing for transferability to organometallic systems or structural properties allows for assessing the overall robustness of a method. It could be shown that the proposed r^{2}SCAN-D4 provides one of the best non-empirical meta-GGA functionals for a wide range of chemical applications as captured by the GMTKN55 database. The additionally investigated D3 and VV10 corrected functional variants provide mostly comparable if somewhat worse results. For the mindless molecule benchmark in particular the performance of non-empirical functionals adhering to exact constraints was found to be superior to empirically optimized ones. The r^{2}SCAN-D4 functional and its derivatives like the composite r^{2}SCAN-3c or its hybrid variants will provide a solid foundation for computational simulation. Together with its variants r^{2}SCAN-based methods cover many relevant applications and provide a reliable default for theoretical modeling. It is expected that many applications currently relying on established non-empirical functionals will adopt r^{2}SCAN based methods in the future. Also, the comprehensive benchmarking of r^{2}SCAN-D4 is hopefully setting a standard for the testing and verification of future new functional developments to provide reliable statistical measures over a broad spectrum of chemical tasks.

Another central topic of this thesis is the development of benchmark sets and their application to computational methods. Benchmark studies try to answer a spectrum of questions, ranging from which methods are sufficient for solving a specific problem to which category of methods provides better or worse results on average. Furthermore, the investigation of exceptional performance, both better and worse, provides valuable insights for the general choice of computational methods. An example provided in this thesis is the failure of MP2-based methods for saturated hydrocarbons due to the inaccurate description of long-range correlation effects as a result of uncoupled HF-based dispersion coefficients and missing higher-order contributions. Such a benchmark for prototypical flexible molecules like unbranched alkane chains is presented in this work together with accurate local coupled-cluster references allowing the assessment of even small energy differences as found in large conformer ensembles. A special topic of interest is the choice of dispersion corrections with a particular functional for such electronically simple systems, as the influence of density changes and charges is small. The D4 dispersion correction for this class of non-covalent interaction performs surprisingly better than the D3 or VV10 models. However, the better performance of D4 corrected functionals compared to D3 and VV10 can be found in the updated reference polarizabilities compared to D3 and the inclusion of many-body dispersion effects compared to VV10, additionally the overall more consistent parametrizations give the D4 dispersion correciton an edge over the other models mentioned here. While usual benchmark studies limit themselves to a single method category, like density functional theory, a wider range of tested methods is investigated to provide insights for semiempirical quantum mechanics (SQM) and force field (FF) methods. This is of special importance since SQM and FF methods become more widely used in the generation of conformer ensembles, which are further processed in a multilevel workflow. The selection of a suitable method for the conformer generation step is important as it determines the safe energy threshold for including candidate structures by the accuracy of the used method. Furthermore, the possible sampling length is determined by the intrinsic computational cost of the SQM and FF methods and determines whether it is feasible to use them for ensemble generation. With the criterium of ensemble completeness the cost–accuracy ratio of a method is central to the successful and efficient development of conformational sampling tools. This work for one aims to provide a new and helpful conformer benchmark, which can be used for testing SQM and FF methods. Also, it tries to provide an example for future benchmark studies for the comprehensive coverage of a wide range of methods going beyond a single method category while still providing detailed insights into the exceptional cases for all categories.

While the development of new methods for the computation of the electronic energy is important, it provides only a partial solution to real-life chemical problems. For many if not most chemical problems the knowledge about the free energy is crucial where the electronic contribution is only one part of the full picture. The thermostatistical contributions to the free energy are well understood and accessible through evaluating higher derivatives of the electronic energy, where especially the GFN parametrizations excel.

More fundamental is the influence of the environment in the condensed phase, like solvation, which impacts the electronic structure as well as the molecular geometry and is challenging to sample in an explicit approach using free energy perturbation theory or thermodynamic integration techniques, or enhanced sampling techniques like metadynamics or replica exchange. A simple yet cost-efficient solution is proposed to handle solvation effects in the framework of SQM methods without sacrificing the computational efficiency of the methods using implicit solvation models of the generalized Born (GB) type. The proposed analytical linearized Poisson–Boltzmann (ALPB) solvation model is based on the best available theories to approximate the Poisson–Boltzmann equation in the context of GB theory. This includes additional terms arising from finite dielectric constants, like the electrostatic shape-dependent contribution for ionic compounds, the P16 interaction kernel for the evaluation of the dielectric screening of the Coulomb interactions, or a smooth surface integration technique for the surface area contribution. While primarily designed for FF methods, the implicit solvation model was introduced self-consistently for the tight binding methods to allow for the efficient evaluation of gradients. The resulting model was parameterized for the extended tight binding and density functional tight binding Hamiltonians as well as for general FF methods, like the GFN-FF. Polarizable continuum models (PCMs) have been proposed for tight binding in the past, however their adoption has been hampered by the relative high computational cost in the self-consistent evaluation of the tight binding Hamiltonian, which is even more impactful in combination with general FFs. The ALPB solvation model will provide the go-to solution for handling environment effects for a wide range of applications with tight binding methods as it provides a computationally efficient and quite natural integration within the framework of tight binding theory. Partition coefficients to calculate environmental relevant distribution properties become accessible at the SQM level of theory and open the possibility for large-scale screening in combination with conformational ensemble sampling techniques. Additionally, for linear-response and time-dependent tight binding calculation, the ALPB opens a new avenue for including solvation effects.

Most if not all work in this thesis is tied to the development and implementation of appropriate software packages enabling the usage of the newly devised methods in computational chemistry workflows. While the success of most of the methods here is tied to their theoretical soundness backed by extensive testing, the availability of implementations in program packages and distributions has been crucial for accelerating their adoption in the community. Here, the spectrum is as diverse and heterogeneous as the computational methods themselves, ranging from standalone programs to libraries in large distribution channels like conda-forge. Special focus is set on the long-term maintainability as well as easy accessibility in form of open-source software. For several of the major program packages like `xtb` or `dftd4` an active community developed along with this thesis.

To summarize, the work presented in this thesis provides a solid foundation for computational chemistry, starting from the conception and exploratory work over the screening and sampling to the actual research and production stage of a computational simulation workflow. The methods developed here are well tested, providing a guide and standard for new developments in the field of computational chemistry to benchmark against. Most importantly, the connection from accurate theory to approximate models has been highlighted in several aspects of this work providing an integral view on the toolbox for theoretical modeling.

Appendix

APPENDIX A

r^2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications

Sebastian Ehlert,[*] Uwe Huniar,[†] Jinliang Ning,[‡] James W. Furness,[‡] Jianwei Sun,[‡] Aaron D. Kaplan,[§] John P. Perdew,[§ ¶] and Jan Gerit Brandenburg[‖]

Received 21 December 2020, Published online 09 February 2021.

Reprinted in Appendix A (adapted) with permission[**] from
S. Ehlert, U. Huniar, J. Ning, J. W. Furness, J. Sun, A. D. Kaplan, J. P. Perdew, and J. G. Brandenburg, *r^2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications*, J. Chem. Phys. **154** (2021) 061101, DOI: 10.1063/5.0041008.

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

[†] Biovia, Dassault Systèmes Deutschland GmbH, Imbacher Weg 46, 51379 Leverkusen, Germany

[‡] Department of Physics and Engineering Physics, Tulane University, New Orleans, Louisiana 70118, United States

[§] Department of Physics, Temple University, Philadelphia, Pennsylvania 19122, United States

[¶] Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States

[‖] Enterprise Data Office, Merck KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany

[**] Permission requests to reuse material from this chapter should be directed to AIP Publishing.

Abstract We combine a regularized variant of the strongly constrained and appropriately normed semilocal density functional [J. Sun, A. Ruzsinszky, and J. P. Perdew, *Phys. Rev. Lett.* **115**, 036402 (2015)] with the latest generation semi-classical London dispersion correction. The resulting density functional approximation r^2SCAN-D4 has the speed of generalized gradient approximations while approaching the accuracy of hybrid functionals for general chemical applications. We demonstrate its numerical robustness in real-life settings and benchmark molecular geometries, general main group and organo-metallic thermochemistry, as well as non-covalent interactions in supramolecular complexes and molecular crystals. Main group and transition metal bond lengths have errors of just 0.8%, which is competitive with hybrid functionals for main group molecules and outperforms them for transition metal complexes. The weighted mean absolute deviation (WTMAD2) on the large GMTKN55 database of chemical properties is exceptionally small at 7.5 kcal/mol. This also holds for metal organic reactions with an MAD of 3.3 kcal/mol. The versatile applicability to organic and metal-organic systems transfers to condensed systems, where lattice energies of molecular crystals are within chemical accuracy (errors $<$1 kcal/mol).

A.1 Introduction

The quantum mechanical description of physical and chemical materials at electronic resolution is an increasingly important task for *in silico* simulations. Here, density functional theory (DFT) has emerged in the past decades as one of the most versatile methodological frameworks.[53,54] This leading position in both materials and chemical applications is largely due to the excellent accuracy over computational cost ratio, as well as the broad applicability across system classes of today's density functional approximations (DFAs).[100–102]

The Jacob's ladder hierarchy[55] is commonly used to classify DFAs. In this hierarchy, DFAs are systematically improved by ascending rungs of different approximations: the local density approximation (LDA), generalized gradient approximations (GGAs), meta-GGAs, hybrid functionals (including a fraction of nonlocal exact exchange), and double-hybrid functionals (including nonlocal correlation). In terms of efficiency, meta-GGAs are in a favorable spot, as they have the same cubic scaling with system size as LDA. Yet, many of the meta-GGAs proposed so far cannot truly leverage the full potential of their rung. Some shortcomings of existing functionals are increased sensitivity to the numeric integration grid, as observed in the strongly constrained and appropriately normed (SCAN) functional[75] or several Minnesota type functionals,[103–105] purely empirical parameters, as present in the B97M functional[22], and sensitivity to the kinetic energy density.[106,107] Recent developments of semi-local DFAs combine exact constraints with various degrees of parametrization to improve descriptions of short- to medium-range electron correlation[75,108,109].

The SCAN functional[75] is constructed to rigorously satisfy all known exact constraints suitable for a meta-GGA. While the functional itself has shown excellent performance in previous studies, the severe numerical instabilities inherent to the functional impeded its adoption for many computational studies. With the recently proposed regularized SCAN (rSCAN)[76] and the subsequent restoration of exact constraints in r^2SCAN[21], the main drawback of the SCAN functional seems to be resolved. Shortcomings of the SCAN functional might still be present in its successors, rSCAN and r^2SCAN. Notably, the description of water clusters $(H_2O)_n$ ($2 \leq n \leq 8$) show the overbinding tendency of SCAN[110]. A recent work by Sharkas *et al.*[110] however demonstrates that this can be mended by the

Perdew–Zunger self-interaction correction (PZ-SIC). SCAN and r^{2}SCAN show similar behaviors for self-interaction error prone systems, which may make r^{2}SCAN amenable to a PZ-SIC correction as well. However, r^{2}SCAN is often more accurate than SCAN, as in the extensive benchmarking here, in the atomization energies of molecules[21], and in the spin-crossover energies of molecules[111].

Nevertheless, semilocal functionals cannot include long-range correlation effects like London dispersion interactions.[25] To truly judge its applicability, we extensively tested r^{2}SCAN combined with the state-of-the-art D4 dispersion correction[9], which shows unprecedented performance for a range of diverse chemical and physical properties. To investigate the development of the SCAN-type functionals we include both SCAN-D4 and rSCAN-D4 in the comparison to r^{2}SCAN-D4, and can attribute improvements in non-covalent interactions mainly to the regularization and improvements for thermochemistry and barrier heights to the restoration of the exact constraints.

We give a concise methodological overview (Section A.2) on r^{2}SCAN and D4 before testing the full method against established DFAs over a wide range of benchmarks (Section A.3), with particular focus on molecular geometries, thermochemistry, kinetics, and non-covalent interactions in small and large complexes.

A.2 Methods

The rSCAN[76] functional regularizes the severe numerical instability or inefficiency of the otherwise successful SCAN[75] functional at the expense of breaking exact constraints SCAN was constructed to obey. This problem arises in many codes that employ localized basis sets, and is less problematic in many codes that employ plane-wave basis sets. While numerical challenges are indeed resolved, a rigorous adherence to exact constraints is core to the design of the SCAN functional and likely important for transferable accuracy across domains of applicability.[111] This seems to be reflected in rSCAN's relatively poor performance for molecular atomization energies compared to other tests.[112,113] The r^{2}SCAN functional[21] combines the good accuracy of SCAN with the numerical efficiency of rSCAN by directly restoring exact constraint satisfaction to the rSCAN regularizations.

The SCAN functional is constructed as an interpolation between single orbital and slowly-varying energy densities designed to maximize exact constraint satisfaction.[75] The interpolation is controlled by an iso-orbital indicator

$$\alpha = \frac{\tau - \tau_W}{\tau_U}, \tag{A.1}$$

where $\tau_W = |\nabla\rho|^2/(8\rho)$ and $\tau_U = 3(3\pi^2)^{2/3}\rho^{5/3}/10$ are the von-Weizsäcker and uniform electron gas kinetic energy densities respectively.[114] In subsequent studies, α has been shown to contribute to numerical instability.[115,116] To remove these effects, a regularized α' was used in rSCAN that removes single orbital divergences at the expense of breaking exact coordinate scaling conditions[117–119] and the uniform density limit. These conditions are restored in r^{2}SCAN by adopting a different regularization:

$$\bar{\alpha} = \frac{\tau - \tau_W}{\tau_U + \eta\tau_W}, \tag{A.2}$$

where $\eta = 10^{-3}$ is a regularization parameter.

The second regularization made in the rSCAN functional is to substitute the twisted piece-wise exponential interpolation of the original SCAN with a smooth polynomial function. This removes problematic oscillations in the exchange-correlation potential, but introduces spurious terms in the slowly-varying density gradient expansion that deviate from the exact expansion[120,121] recovered by SCAN. A corrected gradient expansion term is used in r^{2}SCAN that cancels these spurious terms so the functional recovers the slowly-varying density gradient expansion to second order. A recent modification of SCAN for improved band gap accuracy from Aschebrock and Kümmel named “TASK”[122] is able to enforce the fourth-order gradient expansion for the exchange energy without apparent numerical problems[123], resolving the dominant source of numerical inefficiency. The importance of the fourth-order exchange terms is not established however, and we are thus satisfied using one less exact constraint compared to SCAN. TASK uses an LSDA for correlation however and consequently violates many important exact constraints for correlation, e.g. the second order gradient expansion, that are obeyed by SCAN and r^{2}SCAN.

A.2.1 Numerical stability

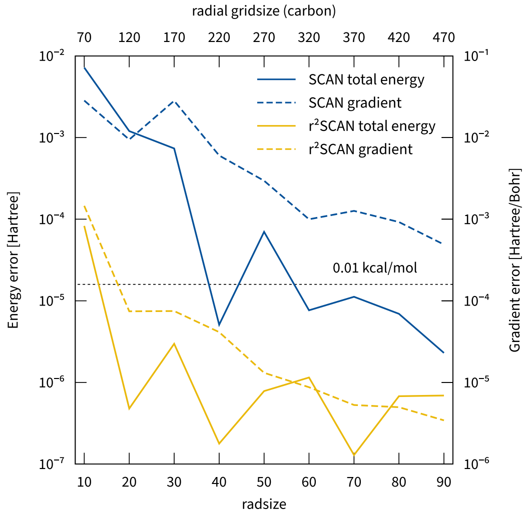


Figure A.1: Errors for FeCp$_2$ with SCAN/def2-QZVP and r^{2}SCAN/def2-QZVP using different radial gridsizes. For both methods, grid 4 and SCF convergence criteria of 10^{-7} Hartree were used and the radial gridsize was varied. The reference has a radsize of 100 or radial gridsize of 515/520/535 for hydrogen/carbon/iron, respectively. The gradient error is the sum of the absolute errors of all gradient components. For further explanation, see the end of Section A.2.1.

Numerical instabilities are revealed by SCAN’s sensitivity to the choice of numerical integration grid, often requiring dense, computationally costly grids.[76,115,124] This issue has been addressed with the rSCAN and r^{2}SCAN functionals. Fig. A.1 shows that the regularization indeed leads to two orders of magnitude error reduction when comparing r^{2}SCAN with SCAN. This holds for both total energy and nuclear gradients for all chosen numerical settings. In practice, this allows for more computationally favorable settings. To give a rough estimate of the computational cost of r^{2}SCAN compared to SCAN, we consider system 10 of the S30L[96] with 158 atoms and 8250 atomic orbitals in a def2-QZVP basis set. A SCAN calculation using Turbomole’s grid 4 and radsize 50 (8.5 million grid points) would take approximately 10 hours, while an r^{2}SCAN calculation with Turbomole grid m4 and radsize 6 (1.6 million grid points) takes only three and a half hours for the same numerical accuracy, resulting in a computational saving of a factor of three to five.[††] We recommend using r^{2}SCAN with radsize 6 and potentially increasing it to 10 for problematic geometry optimizations.[‡‡] We also compared SCAN and r^{2}SCAN with different energy cutoffs in a PAW expansion, and found that r^{2}SCAN is not as sensitive as SCAN, i.e. the total energy converges significantly faster.[§§]

A.2.2 Training of damping functions

As London dispersion interactions arise from nonlocal electron correlations, they cannot be captured by any meta-GGA. In the past years, a range of schemes have been developed to capture these interactions in the DFT framework.[25,28,125–128] Here, we combine r^{2}SCAN with the semi-classical D4 dispersion correction.[9] Its energy contribution is calculated by

$$\begin{aligned} E_{\mathrm{disp}}^{\mathrm{D4}} = &-\frac{1}{2}\sum_{n=6,8}\sum_{A,B}^{\mathrm{atoms}} s_n \frac{C_n^{AB}}{R_{AB}^n}\cdot f_n^{\mathrm{BJ}}(R_{AB}) \\ &-\frac{1}{6}\sum_{A,B,C}^{\mathrm{atoms}} s_9 \frac{C_9^{ABC}}{R_{ABC}^9}\cdot f_9^{\mathrm{BJ}}(R_{ABC},\theta_{ABC}), \end{aligned} \tag{A.3}$$

where R_{AB} is the atomic distance, C_n^{AB} is the *n*th-order dispersion coefficient, and $f_n^{\mathrm{BJ}}(R_{AB})$ is the Becke–Johnson damping function[68,129]. R_{ABC} and C_9^{ABC} denote the geometrically averaged distance and dispersion coefficient, respectively, and θ_{ABC} is the angle dependent term of the triple-dipole contribution.[69,70] The s_8 parameter for the two-body dispersion and the a_1 and a_2 parameter entering the critical radius in the damping function are adjusted to match the local description of a specific DFA. Damping parameters are fitted using a Levenberg–Marquardt least-squares minimization to reference interaction energies as described in Ref. [9]. Optimized parameters are given in Table A.1. The b parameters for rSCAN-VV10 and r^{2}SCAN-VV10 were determined to be 10.8 and 12.3, respectively, on the same set.[28,29]

Table A.1: D3(BJ) and D4 damping parameter for rSCAN and r^{2}SCAN functionals.

	model	s8	a1	a2/Bohr	RMS[*]
rSCAN	D3(BJ)-ATM	1.0886	0.4702	5.7341	0.31
	D4(EEQ)-ATM	0.8773	0.4911	5.7586	0.30
r^{2}SCAN	D3(BJ)-ATM	0.7898	0.4948	5.7308	0.28
	D4(EEQ)-ATM	0.6019	0.5156	5.7734	0.28

A.2.3 Computational details

All ground state molecular DFT calculations were performed with a development version of Turbomole.[133,134] The resolution of identity (RI) approximation[135,136] was applied in all calculations for the

[††] Wall time running on Intel(R) Xeon(R) CPU E3-1270 v5 @ 3.60GHz using four cores.

[‡‡] The 6 radial points correspond to the default settings of Turbomole's *grid m4*.

[§§] This work will be published in an upcoming paper.

electronic Coulomb energy contributions. For all functionals except SCAN, Turbomole's modified grids of type m4 were used. For all SCAN calculations, grid 4 with increased radial integration size of 50 was used instead. Self-consistent field convergence criteria of 10^{-7} Hartree were applied. Ahlrichs' type quadruple-zeta basis sets, def2-QZVP,[137] were used throughout if not stated otherwise.

The periodic electronic structure calculations were conducted with Vasp 6.1[138,139] with projector-augmented plane waves with an energy cutoff of 800 or 1000 eV (hard PAWs[140,141]). Tight self-consistent field settings and large integration (and fine FFT) grids are used. The Brillouin zone sampling has been increased to converge the interaction energy to 0.1 kcal/mol. The non-periodic directions use a vacuum spacing of 12 Å.

A.3 Results

A.3.1 Bond length and molecular geometries

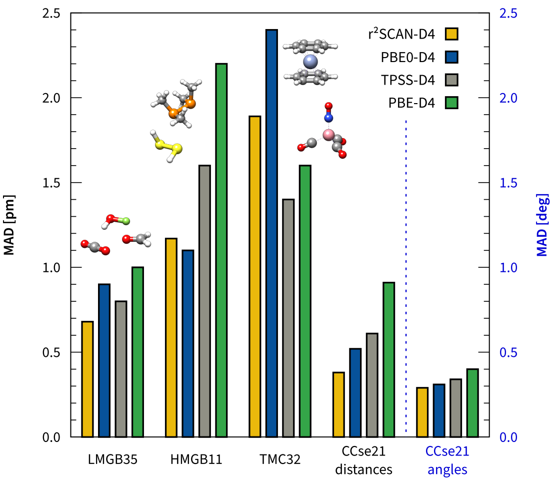


Figure A.2: Errors in bond length from r^2SCAN-D4 and other DFAs separated into light main group bonds (LGMB35[142]), heavy main group bonds (HMGB11[142]), transition metal complexes (TMC32[143]) and semi-experimental organic molecules (CCse21[144]). PBE0-D4, TPSS-D4 and PBE-D4 results for the first three sets are taken from Ref. [9].

To evaluate the description of covalent bond distances, we compare experimental and calculated ground-state equilibrium distances R_e (in pm) for 35 light main group bonds (LMGB35[142]), 11 heavy main group bonds (HMGB11[142]), and 50 bonds in 32 3d transition metal complexes (TMC32[143]). Additionally, we investigate the bond distances and angles for a set of simple organic molecules against accurate semi-experimental references.[144,145] Extended statistics and optimized geometries are made freely available.[¶¶]

We include r^2SCAN-D4, PBE0-D4[59], TPSS-D4[146], and PBE-D4[57] in the comparison shown in Fig. A.2. For organic molecules, we find exceptional performance for all functionals, with errors smaller

[¶¶] Optimized r^2SCAN-D4/def2-QZVP geometries of the LMGB35, HMGB11,and TMC32 sets, statistical performance of the LMGB35, HMGB11, TMC32, CCse21[144,145], GMTKN55, S30L, L7, C40x10 sets are provided at https://github.com/awvwgk/r2scan-d4-paper than 1 pm in the bond distances and half a degree in the bond angles. While all methods reproduce the reference values closely, we observe the best agreement from r^{2}SCAN-D4 with a mean absolute deviation (MAD) of 0.4 pm and 0.3 degree for the bond distances and angles, respectively. For light main group elements, all methods give a mean absolute deviation of less than 1 pm as well, which was also observed in previous studies[9,115]. In comparison with the other methods tested here, r^{2}SCAN-D4 also yields the lowest MAD of only 0.7 pm. Finally, for transition metal complexes r^{2}SCAN-D4 performs reasonably well with an MAD of 1.9 pm. Overall, the performance of r^{2}SCAN is similar to, and sometimes even better than, the hybrid PBE0-D4, which in turn is one of the best performing hybrid functionals for molecular geometries.[142]

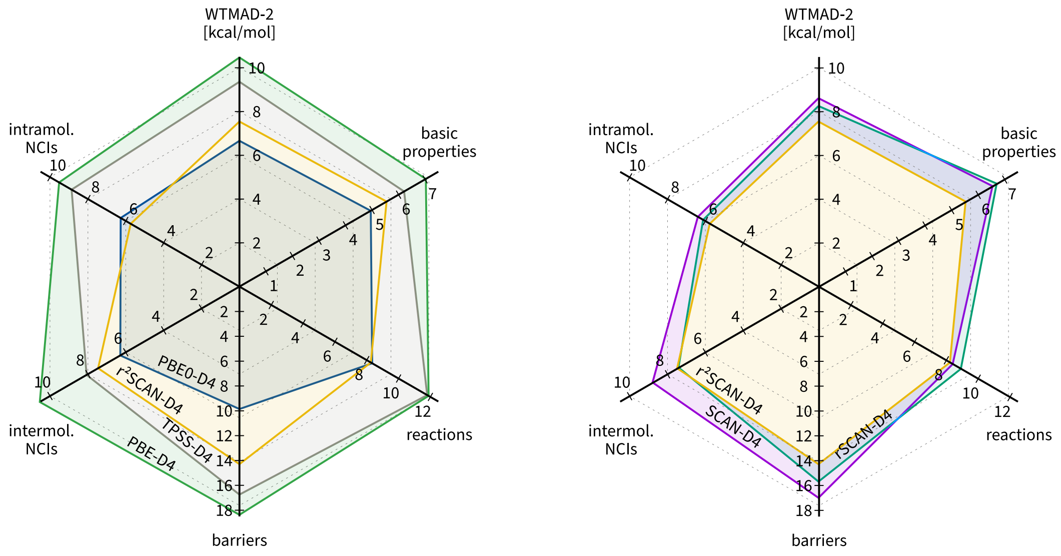


Figure A.3: Weighted mean absolute deviations of r^{2}SCAN-D4 compared to other DFAs for the large database of general main group thermochemistry, kinetics, and non-covalent interactions GMTKN55[23]. On the left-hand graphic, r^{2}SCAN-D4 is compared against functionals representative of their respective rungs. On the right-hand graphic, r^{2}SCAN-D4 is compared to other members of the SCAN family, namely rSCAN-D4 and SCAN-D4.

A.3.2 General main group thermochemistry and non-covalent interactions

To investigate the performance of r^{2}SCAN-D4 for general main group chemistry, we use the main group thermochemistry, kinetics and non-covalent interactions (GMTKN55) database.[23] The GMTKN55 database is a compilation of 55 benchmark sets to assess the performance of DFAs and allows a comprehensive comparison of DFAs. It contains five categories, namely basic properties, barrier heights, isomerisations and reactions, intermolecular, and intramolecular non-covalent interactions (NCIs). Usual weighted total MADs (WTMAD2s) range from 2–3 kcal/mol for double hybrid functionals, over 3–4 kcal/mol for hybrid functionals to 8–9 kcal/mol for (meta-)GGAs, while the lowest rung functionals like PWLDA yield WTMADs of 17 kcal/mol on the GMTKN55. With the exception of the semi-empirical B97M-V[22] (and its B97M-D4 variant[63]), r^{2}SCAN-D4 is the best non-hybrid functional on the GMTKN55 so far with a WTMAD2 of 7.5 kcal/mol, compared to other meta-GGAs like SCAN-D4 (8.61 kcal/mol) or TPSS-D4 (9.36 kcal/mol). For the isomerization and reactions category as well as for the intramolecular NCIs, r^2SCAN-D4 can even compete with the performance of the hybrid PBE0-D4 (WTMAD2 6.66 kcal/mol).

We additionally evaluated rSCAN-D4 and SCAN-D4[147] on the GMTKN55 set to monitor the development in the SCAN-family of functionals. The main difference between SCAN-D4 and rSCAN-D4 is the general improvement in the description of non-covalent interactions, while both functionals perform similarly well in all other categories. Here rSCAN-D4 improves for both NCI categories with a weighted MAD of 6.8 kcal/mol over SCAN-D4, which yields a weighted MAD of 7.6 kcal/mol. This improvement in rSCAN-D4 is mainly responsible for the smaller WTMAD2 of 8.3 kcal/mol compared to the WTMAD2 of 8.6 kcal/mol for SCAN-D4. For r^2SCAN-D4, the improved description of NCI in rSCAN-D4 is preserved (weighted MAD of 6.6 kcal/mol) but r^2SCAN-D4 bests its predecessor in all three remaining categories, resulting in its exceptional WTMAD2 of 7.5 kcal/mol. The mindless benchmark (MB16-43 subset of GMTKN55) is specifically useful for testing a methods robustness to deal with unusual chemistry in artificial molecules. Here, we see that enforcing exact constraints in non-empirical DFAs yields generally lower errors for artificial molecules than their empirical counterparts (see Table A.2).

To stress the importance of including a dispersion correction we test the plain dispersion-uncorrected r^2SCAN which yields a significantly worse WTMAD2 of 8.8 kcal/mol, a difference similar in magnitude to the improvement from SCAN-D4 to r^2SCAN-D4. In summary, r^2SCAN-D4 shows a systematic improvement over its predecessor SCAN-D4 in all categories of GMTKN55 and can preserve improvements present in rSCAN-D4. This makes r^2SCAN-D4 one of the best non-empirical meta-GGAs that have been broadly benchmarked so far.

Table A.2: Comparison of a few non-empirical and empirical dispersion corrected DFAs for the MB16-43 subset (artificial molecules) of GMTKN55. The non-empirical DFAs yield generally lower MADs (in kcal/mol) indicating better transferability across diverse systems.

Non-empirical DFA	MAD	Empirical DFA	MAD
r^2SCAN-D4	14.6	MN15L	20.5
SCAN-D4	17.3	M06L	63.9
TPSS-D4	25.8	M06L-D4	62.6
PBE-D4	25.1	B97M-D4	37.5
PBE0-D4	16.0	B3LYP-D4	28.4

A.3.3 Beyond main group chemistry

Metal organic chemistry is one of the major application areas of non-hybrid DFAs. Here, we use the MOR41 benchmark set that contains 41 closed-shell metal-organic reactions representing common chemical reactions relevant in transition-metal chemistry and catalysis[148]. We compare the statistical deviations from high-level references of r^2SCAN-D4 to PBE0-D4, TPSS-D4, and PBE-D4 in Tab. A.3. The r^2SCAN-D4 functional is one of the best meta-GGAs tested so far on the MOR41 benchmark set, with an MAD of 3.3 kcal/mol. While the SCAN-D4 method provides a slightly lower MAD, the analysis of other statistical quantities, like the standard deviation (SD) and the maximum absolute error (AMAX), suggest less systematic results compared to r^2SCAN-D4. This is confirmed by the Gini coefficient[149], which is 0.44 for r^2SCAN-D4 and 0.50 for SCAN-D4. Compare this to B97M-D4, one of the best meta-GGAs tested on the GMKTN55 set, which yields a larger MAD of 3.8 kcal/mol[63].

Table A.3: Reaction energies of 41 metal-organic reactions compared to high-level references.[148] The MD, MAD, SD and AMAX are given in kcal/mol, while the GINI coefficient is dimensionless.[149]

	MD	MAD	SD	AMAX	GINI
r^2SCAN	2.1	4.4	5.6	17.3	0.46
r^2SCAN-D4	−0.2	3.3	4.3	14.0	0.44
SCAN-D4	−0.8	3.2	4.5	14.1	0.50
TPSS-D4	−1.5	3.5	4.4	22.6	0.39
PBE0-D4	−0.3	2.3	3.1	14.2	0.46
PBE-D4	−0.1	3.5	4.8	22.7	0.45

A.3.4 Non-covalent interactions in large complexes and molecular crystals

With the improved description of non-covalent interactions (NCIs), while retaining the computational efficiency of a meta-GGA, r^2SCAN-D4 is a promising choice for interaction and association energies of large complexes. The results for the S30L[96], L7[150] and X40×10[151] benchmark set are shown in Fig. A.4.

We choose the recently revised L7 benchmark[150] set to assess the performance of r^2SCAN-D4 against converged LNO-CCSD(T)/CBS interaction energies[152]. Close agreement with an MAD of 0.9 kcal/mol is reached for r^2SCAN-D4. This is a significant improvement over other meta-GGAs like SCAN-D4 and TPSS-D4 with MADs of 1.3 and 1.4 kcal/mol, respectively.

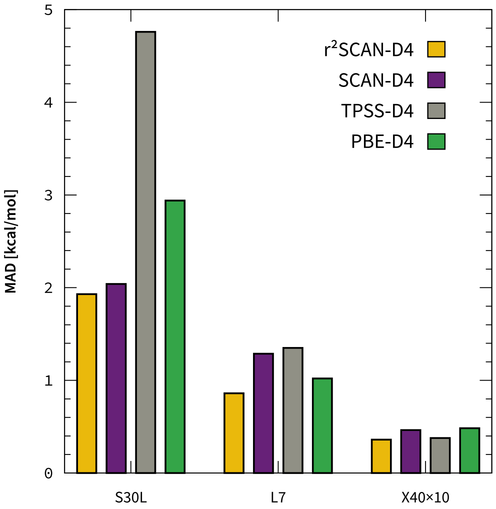


Figure A.4: Non-covalent interaction energies of host–guest systems, large systems and halogen-bonded systems from r^2SCAN-D4 compared to high-level references as well as other DFAs.

We also investigated the description of association energies for large supramolecular complexes using the S30L benchmark set[96]. SCAN-D4 proved to be one of most accurate meta-GGAs in the previous benchmarks[9], giving a remarkable MAD of 2.0 kcal/mol, close to the uncertainty of the provided reference interactions; r^2SCAN-D4 further improves upon this.

In particular, the association energies of the halogen-bonded complexes (15 and 16) are improved with r^2SCAN-D4. The same trend can be observed in the HAL59 benchmark set of the GMTKN55, which shows an MAD of 1.0 kcal/mol with SCAN-D4 and improves with r^2SCAN-D4 to an MAD of 0.8 kcal/mol. To confirm this trend we additionally evaluated the X40×10 benchmark[151] containing 40 halogen bond dissociation curves with SCAN-D4 and r^2SCAN-D4. Again, r^2SCAN-D4 gives the lowest MAD of 0.36 kcal/mol, showcasing on overall improved description of this kind of NCIs.

To evaluate if the good performance for non-covalent interactions transfers from the gas phase to solids, molecular crystals and their polymorphic forms provide useful test cases.[153–155] Here, we investigate the lattice energy benchmark DMC8[156] shown in Table A.4. The DMC8 benchmark contains a subset of the X23[154,157–159] and ICE10[160] benchmark sets with accurate structures and corresponding highly-accurate fixed node diffusion Monte Carlo (FN-DMC) results. Due to SCAN's tendency to overbind hydrogen bonded systems, like ice polymorphs or hydrogen bonded molecular crystals, dispersion corrected SCAN was problematic for these systems. With the improved description of non-covalent interactions in r^2SCAN-D4, this issue is mitigated and we find an overall improved MAD of 0.7 kcal/mol. This MAD is only half of the SCAN-D4 error of 1.5 kcal/mol for these systems and close to the very good performance of the hybrid PBE0-D4 of 0.5 kcal/mol.[7] Only the ice polymorphs are systematically overbound by r^2SCAN-D4, which is, however, a problem of many functionals,[161,162] and may be a self-interaction error.[110] Both r^2SCAN-D4 and SCAN-D4 yield similar results for the self-interaction subset (SIE4x4) of the GMTKN55, therefore a recent work investigating self-interaction corrections for SCAN might also be transferable to r^2SCAN as well.[163] In contrast, the relative stability of the ice polymorph is reproduced correctly. The energy difference of ice II and ice VIII with respect to ice Ih is 0.03 kcal/mol and 0.70 kcal/mol, respectively, agreeing well with the reference of 0.05 kcal/mol and 0.41 kcal/mol.

Table A.4: Lattice energies (kcal/mol) of eight diverse molecular crystals compared to high-level references.[156] Note the significant improvement from r^2SCAN to r^2SCAN-D4 for the dispersion-bound solids.

	ref.	TPSS-D4	r^2SCAN	r^2SCAN-D4
Ice Ih	-14.2	-15.6	-14.6	-15.4
Ice II	-14.1	-14.6	-14.3	-15.4
Ice VIII	-13.7	-12.5	-13.4	-14.7
CO_2	-6.7	-5.5	-4.7	-6.9
Ammonia	-8.9	-8.6	-8.1	-9.5
Benzene	-12.7	-12.0	-5.6	-12.3
Naphthalene	-18.8	-18.5	-7.5	-18.6
Anthracene	-25.2	-24.8	-9.9	-24.7
MD		0.3	4.6	-0.4
MAD		0.8	5.7	0.7
SD		0.9	6.0	0.8
AMAX		1.4	15.4	1.3

The benzene crystal has been frequently used for electronic benchmark purposes.[164–166] Here, we evaluated the equation of state (EOS) to compare with experimental measurements and the Murnaghan EOS fit to the FN-DMC from Ref. [156]. The resulting EOS is shown in Fig. A.5 and agrees excellently with the high-level method as well as the experimental estimate. A slight underestimation of the unit cell volume by 2.6% and overestimation of the bulk modulus by 6.4% can be seen. To highlight once again the importance of London dispersion on properties beyond the mere energy, we report plain r^2SCAN results as well. The r^2SCAN EOS has a significant offset equilibrium volume that is overestimated by 5.4% and a bulk modulus underestimated by 34.0%.

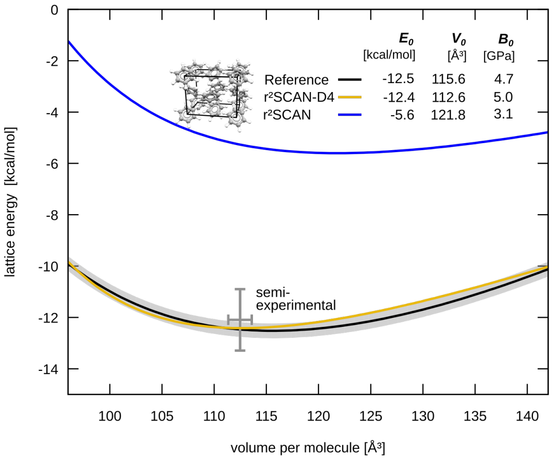


Figure A.5: Equation of state for the benzene crystal from r^2SCAN-D4 compared to experimental measurements and high-level references taken from Ref. [156], the gray area highlights the 1σ confidence interval.

A.4 Conclusions

We have presented an accurate and robust combination of the non-empirical r^2SCAN DFA with the state-of-the-art D4 dispersion correction. The resulting r^2SCAN-D4 electronic structure method shows exceptional performance across several diverse categories of chemical problems assessed by thousands of high-level data-points in a number of comprehensive benchmark sets. Included in the assessment were molecular thermochemistry for both main group and transition metal compounds, barrier heights, structure optimizations, lattice energies of molecular crystals, as well as both inter- and intramolecular non-covalent interactions of small to large systems, creating an extensive coverage of chemically relevant problems.

For the large GMTKN55 benchmark collection of about 1500 data points, r^2SCAN-D4 is one of the most accurate meta-GGAs tested so far. Unlike the best meta-GGA on this set, the dispersion corrected B97M functional, r^2SCAN-D4 can transfer this accuracy to chemically distinct systems like metalorganic reactions. We find significant improvements in NCIs, which were one of the weak-spots of SCAN based methods. More detailed analysis showed that improvements can mainly be found in the description of hydrogen and halogen bonded systems. The same trend is found for molecular crystals, where SCAN-D4's tendency to overbind is mostly resolved in r^2SCAN-D4, giving close to hybrid DFT results for lattice energies.

We found r^2SCAN-D4 to be an accurate and (more importantly) consistent DFA for a large variety of problems and chemical systems. The already good performance of the original SCAN functional is kept and systematically improved in r^2SCAN, while the numeric stability is almost on par with established GGA functionals. We were able to gain some insight in the improvement from SCAN over rSCAN to r^2SCAN, where we can attribute the improved description of non-covalent interactions to the regularization in the step from SCAN to rSCAN, and the improved thermochemistry and barrier heights to the constraint restoration in the step from rSCAN to r^2SCAN. Like SCAN, r^2SCAN is not fitted to molecules, so its accuracy in extensive molecular tests demonstrates the predictive power of its exact constraints and appropriate norms.

With r^2SCAN-D4, a meta-GGA method is finally available that truly leverages the advantages of its rung in Jacob's ladder, while retaining favorable numerical properties and fulfilling important exact constraints. We anticipate r^2SCAN-D4 to be a valuable electronic structure method with broad applications in computational chemistry and material science.

A.5 Data Availability

The data that supports the findings of this study are available within this article or are openly available in https://github.com/awvwgk/r2scan-d4-paper. Further details are available upon request.

Acknowledgement

We thank Stefan Grimme for valuable discussions. SE is supported by the DFG in the framework of the priority program 1807 ”Control of London dispersion interactions in molecular chemistry”. JN, JWF, and JS acknowledge the support of the U.S. DOE, Office of Science, Basic Energy Sciences Grant No. DE- SC0019350 (core research). ADK acknowledges the support of the U.S. DOE, Office of Science, Basic Energy Sciences, through Grant No. DE- SC0012575 to the Energy Frontier Research Center: Center for Complex Materials from First Principles. JPP was supported by the U.S. National Science Foundation under Grant No. DMR-1939528 (CMMT with a contribution from CTMC).

APPENDIX B

Extended Tight-Binding (xTB) Quantum Chemistry Methods

Christoph Bannwarth,[*] Eike Caldeweyher,[†] Sebastian Ehlert,[†] Andreas Hansen,[†] Philipp Pracht,[†] Jacob Seibert,[†] Sebastian Spicher,[†] and Stefan Grimme[†]

Received 22 April 2020, Published online 09 August 2020.

Reprinted in Appendix B (adapted) with permission[‡] from
C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, *Extended tight-binding quantum chemistry methods*, WIREs Comput. Mol. Sci. **11** (2021) e1493, DOI: 10.1002/wcms.1493.

[*] Department of Chemistry and The PULSE Institute, Stanford University, Stanford, CA 94305, United States of America
[†] Mulliken Center of Theoretical Chemistry, Bonn, Germany

[‡] Permission requests to reuse material from this chapter should be directed to Wiley Periodicals LLC.

Abstract This review covers a family of atomistic, mostly quantum chemistry (QC) based semiempirical methods for the fast and reasonably accurate description of large molecules in gas and condensed phase. The theory is derived from a density functional (DFT) perturbation expansion of the electron density in fluctuation terms to various order similar to the original DFTB model. The term 'eXtended' in their name (xTB) emphasizes the parameter availability for almost the entire periodic table of elements ($Z \leq 86$) and improvements of the underlying theory regarding, e.g., the AO basis set, the level of multipole approximation and the treatment of the important electrostatic and dispersion interactions. A common feature of most members is their consistent parameterization on accurate gas phase theoretical reference data for Geometries, vibrational Frequencies and Non-covalent interactions (GFN), which are the primary properties of interest in typical applications to systems composed of up to a few thousand atoms. Further specialized versions were developed for the description of electronic spectra and corresponding response properties. Besides a provided common theoretical background with some important implementation details in the efficient and free `xtb` program, various benchmarks for structural and thermochemical properties including (transition-)metal systems are discussed. The review is completed by recent extensions of the model to the force-field (FF) level as well as its application to solids under periodic boundary conditions. The general applicability together with the excellent cost-accuracy ratio and the high robustness make the xTB family of methods very attractive for various fields of computer-aided chemical research.

B.1 Introduction

Computational modeling at an atomistic scale is now an essential tool in natural science and has become a vital ingredient also in today's research in industry. While the field of quantum chemistry (QC) was for a long time dominated by a few experts who were using complicated software requiring super-computer resources to solve some special chemical problems, the situation has changed tremendously in the last 2-3 decades. Nowadays, routine density functional theory (DFT) calculations[101,167,168] can be conducted by experimentally working chemists on common desktop computers with user-friendly standard software for various properties and problems like structure determination, reaction mechanism exploration, or computing spectral signatures of molecules or even solids. The theories and computational tools, i.e., the "machine under the hood" should 1. provide reasonable results, 2. in short time, 3. for various systems and physical-chemical properties. This at least for larger systems difficult to achieve compromise between accuracy/robustness and computational speed can often be obtained only by applying a hierarchy of different levels of sophistication (multi-level modeling) as sketched in Figure B.1.

Here, the extraordinarily complex potential energy surface (PES) of a larger molecule is initially investigated (screened) at a relatively low theoretical level which is subsequently increased to the accurate DFT or wave function theory (WFT) level. Previously, the starting point in typical applications were classical force-fields (FFs), e.g., for initial conformation search. The drawbacks of FF approaches are manifold and in particular, missing parameterization for many elements (e.g., metals) has hampered further development of the field. At this entry point, low-level quantum chemistry methods come into play, i.e., they are proposed to replace FFs in many cases (about to a systems size of 500–1000 atoms) as the lowest theory level. Quantum chemistry approaches have many advantages but one has to keep the computational cost under control. Successful attempts to develop fast but yet reasonably efficient ab initio type procedures have been made.[169–172] We mention here our own HF-3c[173], PBEh-3c[142], and B97-3c[79] small atomic orbital (AO) basis set methods that are meanwhile applied routinely by many

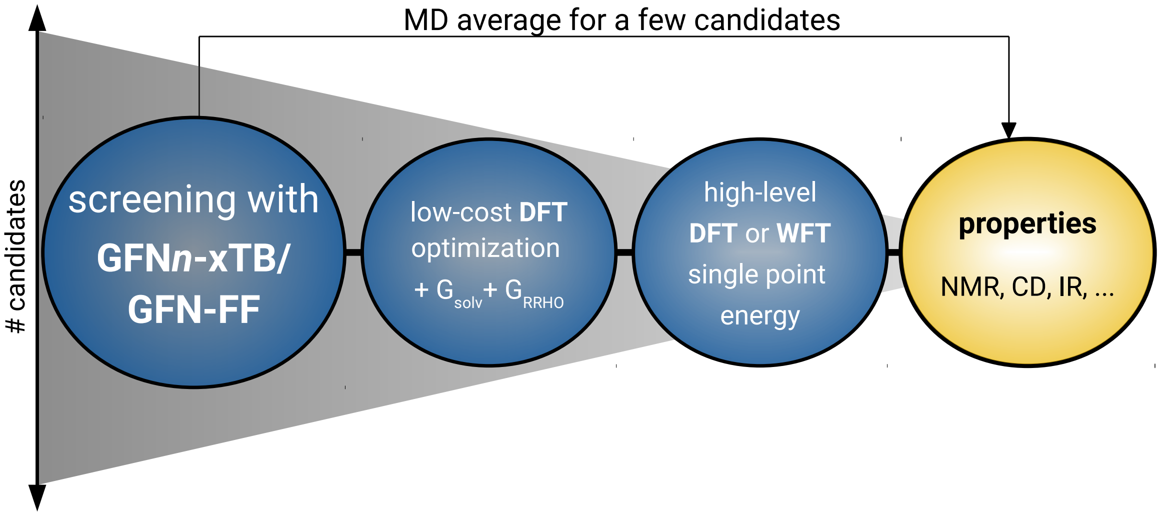


Figure B.1: This schematically depicts the proposed multi-level modeling scheme based on the GFN family of methods (1st bubble), which are described in this review. These methods are generally used at the initial stages of the multi-level workflow, where a large number of of calculations need to be carried out. This typically includes screening over numerous potential candidate molecules (visualized by the diameter of the cone) or extensive structural sampling of different molecular conformations. In subsequent steps (2nd bubble), the theory level (accuracy) is increased (electronic structure as well as including thermostatistical (RRHO) and solvation (solv) free energies (G)) while the number of considered candidate structures is reduced. At the end of this workflow, only very few structures need to be treated by high-level theory levels (3rd bubble) to accurately determine the thermodynamic state (at equilibrium) and the respective property. The latter is often a spectroscopic property and is usually obtained as a thermostatistical Boltzmann ensemble average over all remaining candidates, based on free energies computed at the high-level of theory.

groups world-wide. However, they are still too slow for larger screening purposes (e.g., thousands of energy evaluations for 50-100 atom systems) or in applications for typical biochemical structures, e.g., protein-ligand interactions[174], where often long molecular dynamics (MD) runs have to be conducted.

This situation sparked renewed interest in semiempirical QC methods which have a long history dating back to the 70s with the famous MNDO type approximations[175] or even to the Hückel Hamiltonians from the very early days of quantum mechanics. The renaissance of semiempirical methods in the last 20 years was mainly caused by the development of density functional tight binding (DFTB)[36,38,176–178] methods by Seifert, Elstner, Frauenheim and co-workers. The DFTB methods combine the efficiency of the old NDDO type minimal basis set methods with the higher (compared to a Hartree–Fock, HF) accuracy of DFT as the underlying machinery. For more details on their derivation and properties in comparison with HF and even FFs see section "Theory".

The major drawback of DFTB in its current state is (similar to FFs) the parameterization because the Hamiltonian matrix elements are formulated in an atom(element) pair-wise fashion leading to thousands of empirical values that have to be determined. This makes it already very difficult to cover consistently only a small (upper) part of the periodic table of elements and hence limits their applicability, in particular for the chemically important metals. Another issue is the design strategy of the method regarding the target properties. Although DFTB in principle approximates DFT which in turn should be able to describe any chemical system with uniformly high accuracy, the true situation is more complicated. Mainly due to the use of a small minimal AO basis set to express the simplified Kohn-Sham (KS) equations analytically, not all desired properties can be described at a similar target accuracy. The computation of accurate chemical bond or reaction energies is particularly difficult at any semiempirical level. Hence, if the parameterization procedure is forced to describe chemical (interaction) energies well, other, also important features like the computation of molecular structures necessarily deteriorate. This problem of non-generality with low-level methods is difficult to solve and has been in a certain sense just circumvented with the development of the eXtended TB (xTB) methods which are the topic of this review.

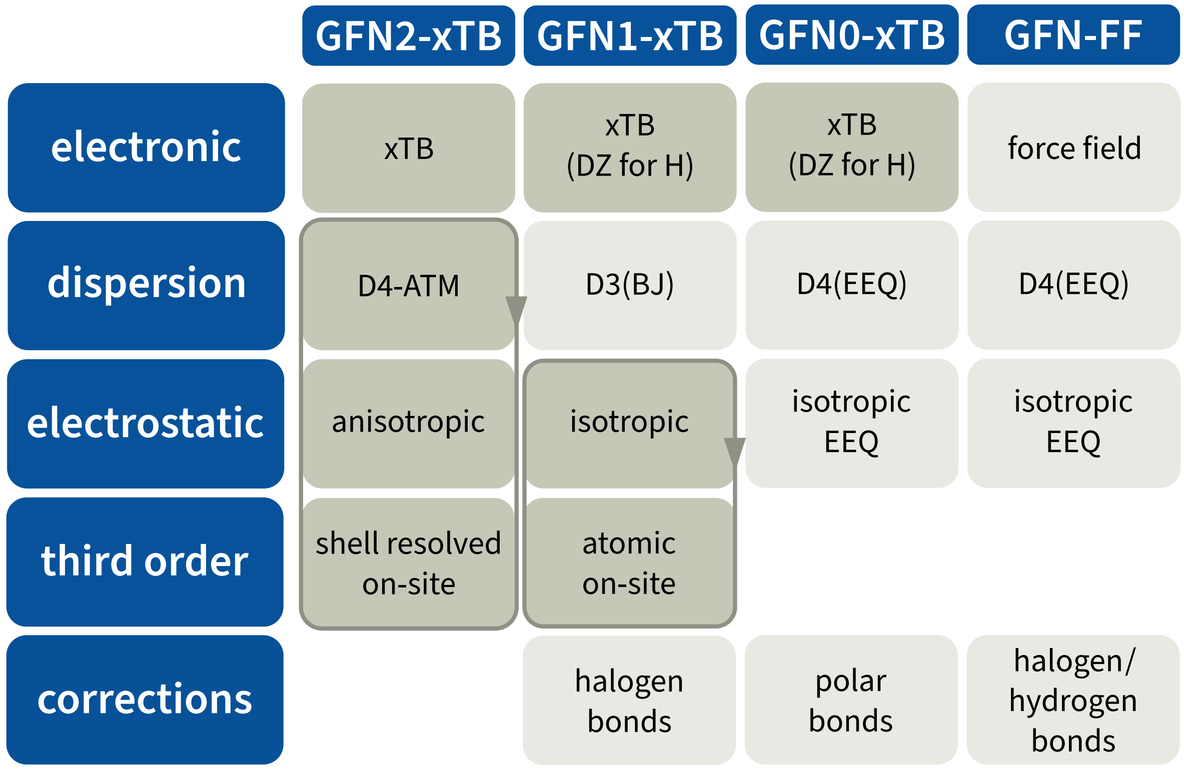
The GFNn-xTB methods ($n = 0, 1, 2$, see below) are designed from the very beginning as special purpose tools focusing on molecular properties which can relatively easily and in a physically sound manner be described at low-level, namely Geometries, (vibrational) Frequencies, and Non-covalent interactions (NCIs) leading to the acronym GFN. Chemical energies (stronger interactions) are not used as primary training data but merely define important cross-checks. Their application to solids under periodic boundary conditions (PBC) was originally not intended but is now possible with good accuracy. The first version originally termed GFN-xTB (now for better distinguishability dubbed GFN1-xTB) employs basically the same (mostly second order with some terms up to third order) approximations for the Hamiltonian and electrostatic energy as DFTB3[178] but does not rely on an atom pair-wise parameterization. Instead, as in the old NDDO type methods, mainly element specific empirical fitting is used enabling a consistent (but still tedious) parameterization covering a large part of the periodic table up to Z=86 (radon). In this respect, in 2017 GFN1-xTB filled a gap in the market of off-the-shelf atomistic models as it is fast, robust, reasonably accurate, and works for many metallic systems. It was quickly adapted by the community and implemented in various QC programs like AMS,[179] CP2K,[180] Cuby4,[181] DFTB+,[13] entos,[182] ORCA,[183,184] and TeraChem.[185,186]

For a few prototypical applications of GFN1-xTB by us and other groups see references[187–191].

Probably the most serious deficiency of GFN1-xTB/DFTB3 also regarding one of the central properties (i.e., NCI) is the monopole-type, spherically symmetric description of the atom pair-wise electrostatic interactions. This led in 2019 to the development of the successor GFN2-xTB featuring a multipole electrostatic treatment up to quadrupole terms. GFN2-xTB is built and parameterized along the same lines as the GFN1 version but incorporates better physics, the latest D4 dispersion model[9] and is completely pair-parameter-free.

Further extensions of the GFNn-xTB family of methods are also described here briefly (see Figure B.2 for a general overview).

In 2019, we tried to roughly keep the accuracy level of the successful GFN1/GFN2 versions but making them significantly faster. The key idea was to avoid the self-consistent-charge (SCC) iterations involving repeated matrix diagonalization which represents the computational bottleneck in almost all semiempirical quantum mechanical (SQM) methods. This could be achieved by formulation of a non-iterative first order variant termed GFN0-xTB[192] where the electrostatics are treated classically and only change the electronic structure to first order. In this approach only a single extended Hückel-type Hamiltonian matrix eigenvalue problem is solved, while the electrostatic terms are treated semi-classically within a so-called electronegativity equilibration (EEQ) atomic charge model.[193–196] This proposed GFN0-xTB method as all other considered GFN approaches has been implemented in the freely available `xtb` program for testing and is also briefly described here. However, because we think that the GFN0 Hamiltonian can still be substantially improved, we consider it here as a preliminary, proof-of-principle



	GFN2-xTB	GFN1-xTB	GFN0-xTB	GFN-FF
electronic	xTB	xTB (DZ for H)	xTB (DZ for H)	force field
dispersion	D4-ATM	D3(BJ)	D4(EEQ)	D4(EEQ)
electrostatic	anisotropic	isotropic	isotropic EEQ	isotropic EEQ
third order	shell resolved on-site	atomic on-site		
corrections		halogen bonds	polar bonds	halogen/ hydrogen bonds

Figure B.2: Overview of the GFN family of methods with main components and classification of the most important terms. Dark grey shaded areas denote a quantum mechanical description while light grey parts indicate a classical or semi-classical description. The parts surrounded by the arrows are treated in an iterative, self-consistent fashion. For a more detailed discussion including the definition of the acronyms see text.

version.

Another non-self-consistent xTB version was developed for the computation of excitation energies and optical spectra of huge systems. In 2013 one of us proposed a simplified Tamm-Dancoff density functional approach (sTDA) for such problems.[197] It employs a regular KS determinant as input, i.e., requires a self-consistent DFT calculation with a standard Gaussian AO basis (e.g., augmented DZ or TZ level). Although the sTDA approximations are reasonably accurate and speed-up UV- or CD-spectra computations enormously, for large systems with thousands of atoms the electronic ground state calculation of the orbitals and orbital eigenvalues still remains the bottleneck. In the so-called sTDA-xTB method,[25] this crucial DFT step is replaced by a single-shot TB calculation with an extended AO basis set including diffuse (low-exponent) functions. This approach was proposed as a very fast single-point electronic structure method already in 2016, i.e., one year before the release of GFN1-xTB and in fact was our first published xTB scheme. Meanwhile, parameters are available for almost all elements[198] and extensions to other response type properties have been reported.[199–203] The success and computational speed of sTDA-xTB motivated the development of the general GFN versions which allow full exploration of molecular and solid state PES.

Very recently, the main concepts of the GFN approach have been transferred into a non-electronic, force-field version termed GFN-FF.[81] The main point of this generic, just coordinate input-based semi-classical potential is its generality meaning that it is as the other GFN family members applicable with reasonable accuracy and robustly to almost any system with elements up to Z=86. This very fast method is also briefly mentioned here and some very large cases, for which even GFN0-xTB would be computationally too demanding, are shown.

Beside example applications and benchmarking of molecular structures and thermochemistry, transition state localization, chemical space exploration, proteins, and molecular crystals, this review covers general TB theory as well as the methodical aspects of the periodic implementation, the geometry optimization of large systems and compares some computation times. In this review, we mostly restrict details regarding implementation to the standalone `xtb` program.[204]

B.2 Theory

B.2.1 Tight-binding theory and comparison of variants

Origin and connection to other methods

Just like the closely related density functional tight-binding (DFTB) methods,[38,178,205] the extended tight-binding (xTB) methods are rooted in KS-DFT, and formally, represent a semiempirical approximation to the latter. In the following, the connection of the xTB methods to DFT, DFTB, and classical force-fields is outlined, then we will discuss the energy expressions for the individual generations of xTB methods.

It is common practice to start the derivation from a semilocal DFT energy expression,[36,38] however, we have found it more useful to use DFT that includes nonlocal correlation (dispersion) as starting point:[8]

$$\begin{aligned} E_{\text{tot}} &= E_{nn} \\ &+ \sum_i^{N_{\text{MO}}} n_i \int \psi_i^*(\mathbf{r}) \left[\hat{T}(\mathbf{r}) + V_n(\mathbf{r}) + \varepsilon_{\text{XC}}^{\text{LDA}}[\rho(\mathbf{r})] + \frac{1}{2} \int \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \Phi_{\text{C}}^{\text{NL}}(\mathbf{r}, \mathbf{r}') \right) \rho(\mathbf{r}') d\mathbf{r}' \right] \psi_i(\mathbf{r}) d\mathbf{r} \end{aligned} \tag{B.1}$$

Here, ψ_i are molecular spatial orbitals with occupation n_i, $\hat{T}(\mathbf{r})$ is the kinetic energy operator and $\hat{V}_n(\mathbf{r})$ is the Coulomb operator due to the interaction with the clamped nuclei. $\varepsilon_{\text{XC}}^{\text{LDA}}[\rho(\mathbf{r})]$ is the semilocal exchange-correlation (XC) energy per particle. The inner integral over $\mathbf{r}'$ contains the interelectronic Coulomb and nonlocal (NL) correlation via the kernel $\Phi_{\text{C}}^{\text{NL}}(\mathbf{r}, \mathbf{r}')$. By including the latter term, we find that dispersion interactions naturally occur and establish the connection between tight-binding and intermolecular force-field methods (see below). Since we are working with a KS system of formally independent particles, the density is obtained as

$$\rho(\mathbf{r}) = \sum_i^{N_{\text{MO}}} n_i \int \psi_i^*(\mathbf{r}) \psi_i(\mathbf{r}) d\mathbf{r}\,. \tag{B.2}$$

Next, we reformulate the total energy in terms of a reference density ρ_0, which ideally is close to the final converged density ρ, and a density difference $\Delta\rho$, with $\rho = \rho_0 + \Delta\rho$. Typically, a superposition of spherical, neutral atomic reference densities (SADs) $\rho_0 = \sum_A \rho_0^A$ is used.[36,38] This allows us to decompose the energy in form of a Hartree energy at the reference density, the Hartree energy difference arising from $\Delta\rho$, as well as the nonseparable exchange-correlation (local and nonlocal) energies.

$$E_{\text{tot}} = E_0^{H} + \Delta E^{H} + E_{XC}^{LDA}[\rho] + E_C^{NL}[\rho, \rho'] \tag{B.3}$$

The reason that the XC and nonlocal correlation energy are treated separately is rooted in the nonpolynomial dependency on the electron density. If we had started from a functional that includes Fock-exchange, the Hartree energy terms would correspond to Hartree–Fock-like energy expressions (cf. Ref. [206]). This is just mentioned for the sake of completeness, but not discussed further at this point.

The energy at the reference density E_0^H is given by

$$E_0^H = E_{nn} + \sum_i^{N_{MO}} n_{0,i} \int \psi_i^*(\mathbf{r}) \left[\hat{T}(\mathbf{r}) + V_n(\mathbf{r}) + \frac{1}{2} \int \frac{1}{|\mathbf{r}-\mathbf{r}'|} \rho_0(\mathbf{r}') d\mathbf{r}' \right] \psi_i(\mathbf{r}) d\mathbf{r}\,, \tag{B.4}$$

while the Hartree energy difference due to $\Delta\rho$ is given as

$$\Delta E^H = + \sum_i^{N_{MO}} \Delta n_i \int \psi_i^*(\mathbf{r}) \left[\hat{T}(\mathbf{r}) + V_0(\mathbf{r}) + \frac{1}{2} \int \frac{1}{|\mathbf{r}-\mathbf{r}'|} \Delta\rho(\mathbf{r}') d\mathbf{r}' \right] \psi_i(\mathbf{r}) d\mathbf{r}\,. \tag{B.5}$$

The reference potential $\hat{V}_0(\mathbf{r})$ is given as

$$\begin{aligned} V_0(\mathbf{r}) =& V_{e,0} + V_n(\mathbf{r}) = \int \frac{1}{|\mathbf{r}-\mathbf{r}'|} \rho_0(\mathbf{r}') d\mathbf{r}' - \sum_A^{N_{\text{nuc}}} \frac{Z_A}{|\mathbf{r}-\mathbf{R}_A|} \\ =& \sum_A^{N_{\text{nuc}}} \left(\int \frac{1}{|\mathbf{r}-\mathbf{r}'|} \rho_0^A(\mathbf{r}') d\mathbf{r}' - \frac{Z_A}{|\mathbf{r}-\mathbf{R}_A|} \right) . \end{aligned} \tag{B.6}$$

Eq. B.3 is completely equivalent to Eq. B.1, just reformulated in terms of the difference density $\Delta\rho$, and consequently, the energy can self-consistently be minimized by solving for the optimum $\Delta\rho$ (see Ref. [206]). If $\Delta\rho$ is not determined self-consistently, Eq. B.3 corresponds to a dispersion-corrected expression of the non-self-consistent Harris-Foulkes functional.[207,208]

In density functional tight-binding methods, the total energy is Taylor expanded around $\Delta\rho = 0$.

$$E[\rho] = E^{(0)}[\rho_0] + E^{(1)}[\rho_0, \delta\rho] + E^{(2)}\left[\rho_0, (\delta\rho)^2\right] + E^{(3)}\left[\rho_0, (\delta\rho)^3\right] + \cdots \tag{B.7}$$

These fluctuations are typically restricted to the valence orbitals only. The most sophisticated variants truncate this expansion after the third order term.[178,205] The same is true for the GFN1-xTB[41] and GFN2-xTB[8] approaches, while GFN0-xTB corresponds to truncation after the first order term (see below). We will shortly identify terms occurring at the different orders and then outline the various GFNn-xTB methods in a self-contained manner with the latter being presented in chronological order.

An overview over the respective tight-binding orders contained in the various GFN schemes (i.e., in the GFNn-xTB and GFN-FF methods) is given in Table B.1. We emphasize at this point that all xTB methods incorporate terms, which have a physical basis in the aforementioned Taylor expansion. Nevertheless, the semiempirical parameters are not precomputed by first principles methods as in DFTB, but optimized on a large fit set to provide the best working parameter combination for the desired GFN

Table B.1: Overview of the terms included in the different GFN-type methods. Always the highest order considered is shown.

method/order in TB expansion	$E^{(0)}$	$E^{(1)}$	$E^{(2)}$	$E^{(3)}$
GFN2-xTB	E_{rep}	E_{EHT}	$E_{\gamma} + E_{\text{AES}} + E_{\text{AXC}} + E_{\text{disp}}^{\text{D4}}$ [a]	E_{Γ}
GFN1-xTB	$E_{\text{rep}} + E_{\text{disp}}^{\text{D3}} + E_{\text{XB}}$	E_{EHT}	E_{γ}	E_{Γ}
GFN0-xTB	$E_{\text{rep}} + E_{\text{disp}}^{\text{D4}} + E_{\text{srb}} + E_{\text{EEQ}}$	$E_{\text{EHT}}(+E_{\gamma})$	–	–
GFN-FF	E_{cov} [b] $+ E_{\text{NCI}}$ [c]	–	–	–

[a] A self-consistent version of D4 based on GFN2-xTB Mulliken charges is used. [b] The classical potential energy term denoted as E_{cov} contains bonding and further short-ranged interactions usually accounted for as bending and torsion in common force-fields. [c] $E_{\text{rep}} + E_{\text{disp}}^{\text{D4}} + E_{\text{IES}} + E_{\text{XB}} + E_{\text{HB}}$. A simplified version of D4 is used for dispersion. Here, dispersion coefficients instead of the atomic polarizabilities are scaled. The isotropic electrostatics (IES) are described by two separate electronegativity equilibration (EEQ) charge models, while XB and HB refer to halogen and hydrogen bond corrections, respectively. Details are found in the main text.

target properties. So in that sense, the xTB-type methods employ physically motivated energy terms, but follow a "top-down" parameterization procedure aiming for sophisticated accuracy.

Zeroth order energy. A system of non-interacting (i.e., infinitely separated) atoms with spherical, neutral atomic reference densities is chosen as zeroth order reference. This way, the externally experienced electrostatic potentials due to the electrons and nuclei of an atom cancel exactly and the Coulomb terms in Eqs. B.4 and B.6 are reduced to on-site terms only. ΔE^{H} is then equal to zero and the total energy can be decomposed into noninteracting atomic total energies (E_A) and orbital energies, which can be precomputed for ρ_0, as well as interatomic repulsion and London dispersion interaction terms:

$$\begin{aligned} E^{(0)}\left[\rho_0\right] &= \sum_{A}^{N_{\text{nuc}}} E_A\left[\rho_0^A\right] + \frac{1}{2}\sum_{A,B}^{N_{\text{nuc}}}\left(E_{\text{rep}}\left[\rho_0^A,\rho_0^B\right] + E_{\text{disp}}\left[\rho_0^A,\rho_0^B\right]\right) \\ \Leftrightarrow E^{(0)}\left[\rho_0\right] - \sum_{A}^{N_{\text{nuc}}} E_A\left[\rho_0^A\right] &= \underbrace{\left(E_{\text{rep}}^{(0)} + E_{\text{disp}}^{(0)}\right)}_{\text{Lennard-Jones/Buckingham-type potential}} \end{aligned} \tag{B.8}$$

Here, the dispersion contribution E_{disp} results exclusively from the long-range – in a DFT sense "nonlocal" – correlation effects. The pairwise repulsion energy E_{rep} originates from overlap of the atomic reference densities ρ_0^A, which leads to changes in the Coulomb, exchange, and short-range (local) correlation energy. At zeroth order, we have hence established the connection between tight-binding methods and well-known intermolecular force-field potentials of Lennard-Jones[209] or Buckingham[210] type. In all schemes, the total energy is given relative to the energy of the free, noninteracting atoms, which is why we formally subtracted it from the zeroth order energy in the second line of Eq.B.8. If we had not included a nonlocal correlation functional to capture dispersion from the very beginning (as, e.g., in Ref. [38] and [36]), the formal equivalence of zeroth order tight-binding and intermolecular force-fields would be missing.

First order energy At first order, the following energy changes are entering:

$$\begin{aligned} E^{(1)}\left[\rho_0, \delta\rho\right] =& \Delta E^{H}\left[\Delta n_i = \delta n_i, \Delta\rho = 0\right] + \frac{1}{\partial\rho}\left(\partial E_{XC}^{LDA}[\rho_0] + \partial E_{C}^{NL}[\rho_0, \rho_0']\right)\delta\rho \\ \approx& E_{EHT}^{(1)} + E_{disp}^{(1)} \end{aligned} \tag{B.9}$$

Since first order density fluctuations are enabling changes in the energy, but not in the electrostatic potential, no interatomic Coulomb terms occur, i.e., the now charged atoms still experience the zero field from the other atoms. This is different for the dispersion energy, where the nonzero potential due to long-range correlation with ρ_0 is experienced by the fluctuation $\delta\rho$. Apart from this, only changes of the on-site energies due to reoccupation of the atomic orbital levels occur. In tight-binding methods, this is absorbed in an extended Hückel-type term, which essentially is responsible for covalent bonding. Particularly this term is different in the xTB from the DFTB methods in that it includes more empirical but chemically motivated features, which enable a sophisticated electronic structure method with only global and element-specific parameters.

Second order energy. At second order, the energy is changed by

$$\begin{aligned} E^{(2)}\left[\rho_0, \delta\rho\right] =& \Delta E^{H}\left[\Delta n_i = \delta n_i, \Delta\rho = \delta\rho\right] + \frac{1}{\partial\rho\,\partial\rho'}\left(\partial^2 E_{XC}^{LDA}[\rho_0] + \partial^2 E_{C}^{NL}[\rho_0, \rho_0']\right)\delta\rho\,\delta\rho' \\ \approx& E_{ES+XC}^{(2)} + E_{disp}^{(2)}\,. \end{aligned} \tag{B.10}$$

Due to the second-order density changes, the net electrostatic (ES) energy between two atoms becomes nonvanishing. Additionally, the (semi-)local XC energy changes in the short-range interatomic region. Both effects are typically described in form of a short-range damped Coulomb energy in tight-binding methods. The dispersion energy is corrected at second order as well. With second or higher order terms (as in GFN1- and GFN2-xTB), the tight-binding energies require a self-consistent solution.

Third order energy. At third and higher orders, no contributions from ΔE^{H} occur, because this term is no more than quadratic in the difference density. Only the nonpolynomial XC (local and nonlocal) terms lead to energy changes at third order.

$$\begin{aligned} E^{(3)}\left[\rho_0, \delta\rho\right] =& \frac{1}{\partial\rho\,\partial\rho'\,\partial\rho''}\left(\partial^3 E_{XC}^{LDA}[\rho_0] + \partial^3 E_{C}^{NL}[\rho_0, \rho_0']\right)\delta\rho\,\delta\rho'\,\delta\rho'' \\ \approx& E_{XC}^{(3)} + E_{disp}^{(3)}\,. \end{aligned} \tag{B.11}$$

While the third order dispersion changes are never explicitly considered in tight-binding methods, the (semi-)local effects are included in DFTB3, as well as in GFN1- and GFN2-xTB Hamiltonians. They typically serve the purpose to stabilize relatively highly charged atoms, e.g., electronegative elements like oxygen.

Common ingredients in the GFNn-xTB methods.

Before describing the individual xTB schemes separately, we will outline common aspects of the GFN1-, GFN2-, and GFN0-xTB methods, such as common energy terms and the employed wavefunction ansatz.

This refers solely to the mathematical form and the explicit parameters are naturally different in these schemes.

The wavefunction choice in GFNn-xTB methods. The GFN1-, GFN2-, and GFN0-xTB wavefunctions are all formulated in terms of a partially polarized, mostly minimal valence basis set, consisting of spherical Gaussian-type atomic orbital (GTO) basis functions. Each contracted GTO ϕ_μ therein approximates a Slater-type orbital (STO) ϕ_μ^{STO} as in Ref. [211]

$$\phi_\mu^{\text{STO}}(\mathbf{r}) \approx \phi_\mu(\mathbf{r}) = \sum_{z}^{N_{\text{prim}}^{\mu}} d_{z\mu}\chi_z^{\mu}(\mathbf{r}) \tag{B.12}$$

Here, the χ_z^μ are primitive GTOs that contribute to the contracted GTO ϕ_μ and $d_{z\mu}$ are the corresponding contraction coefficients.

Depending on the atomic orbital (and GFNn-xTB Hamiltonian), the number of primitives varies from three to six (see Refs.[8,41,192] for details). The AO basis set (Slater exponents) has been optimized simultaneously during the xTB parameterization and hence, is tied to a specific GFNn-xTB version. The molecular orbitals ψ_j are expanded as a linear combination in this basis of AOs (LCAO).

$$\psi_j(\mathbf{r}) = \sum_{\mu}^{N_{\text{AO}}} C_{\mu j}\phi_\mu(\mathbf{r}) \tag{B.13}$$

By derivation of the respective energy expressions (see below) with respect to the orbital coefficients, Roothaan–Hall-type[48,49] generalized eigenvalue equation is obtained.

$$\mathbf{FC} = \mathbf{SC}\varepsilon \tag{B.14}$$

Here, $\mathbf{C}$ is the matrix of orbital coefficients from Eq.B.13, ε is a diagonal matrix with orbital energies on the diagonal, $\mathbf{F}$ is the respective xTB Hamiltonian matrix, and $\mathbf{S}$ is the AO overlap matrix. In GFN1-xTB and GFN2-xTB, these equations are solved self-consistently. Just like in DFTB, the xTB methods employ a nonorthogonal basis, i.e., different from typical Hartree–Fock-based semiempirical methods, no zero differential overlap (ZDO) approximation is applied in xTB methods.

The above wavefunction ansatz can be generalized to periodic systems where ψ_j then corresponds to a crystal orbital. The Bloch function equivalent for the one-particle functions is then given by

$$\psi_j(\mathbf{r}+\mathbf{L},\mathbf{k}) = \psi_j(\mathbf{r},\mathbf{k})\exp[i\mathbf{kL}] \tag{B.15}$$

with the Born–von-Kármán cyclic boundary conditions $\psi_j(\mathbf{r}+\mathbf{L},\mathbf{k}) = \psi_j(\mathbf{r},\mathbf{k})$. Expanded in the aforementioned AOs, the crystal orbitals are then expressed as:

$$\psi_j(\mathbf{r},\mathbf{k}) = \sum_{\mu} C_{\mu j}(\mathbf{k})\frac{1}{\sqrt{N_L}}\sum_{\mathbf{L}}^{N_L}\phi_\mu^{\mathbf{L}}(\mathbf{r})\exp[i\mathbf{kL}] \tag{B.16}$$

Here the summation, in principle, goes over all N_L cells (or identical images), which are related by the translation vector $\mathbf{L}$. In the cyclic cluster model (CCM), which is employed in the `xtb` implementation, we only consider nearest-neighbors ($\mathrm{N_L} \rightarrow \mathrm{N_L^{CCM}}$) with corresponding weights[212] for the expansion of the Bloch function. The CCM assumes that all interactions beyond the Wigner–Seitz cell vanish. This condition is usually not fulfilled for electrostatic and dispersion interactions which must be treated differently, e.g., by lattice summation. Given this, we can re-write the Roothaan–Hall-type equations for the crystal orbitals as

$$\sum_{\mathbf{L}=\mathbf{0}}^{\mathrm{N_L^{CCM}}} \exp[i\mathbf{kL}] \sum_{\mu} \left(F_{\mu\nu}^{\mathbf{0L}} - \varepsilon_j(\mathbf{k}) S_{\mu\nu}^{\mathbf{0L}} \right) C_{\mu j}(\mathbf{k}) w_{\mu\nu} = 0 \,, \tag{B.17}$$

where $F_{\mu\nu}^{\mathbf{0L}}$ is the element of the Hamiltonian matrix (Kohn-Sham/Fock matrix in ab initio theories), $\varepsilon_j(\mathbf{k})$ is the energy of the jth crystal orbital and $S_{\mu\nu}^{\mathbf{0L}}$ is the overlap matrix element. The lattice translations $\mathbf{L}$ here denote the images in the Wigner–Seitz cell. The $w_{\mu\nu}$ are their pairwise weights of the AOs μ and ν to preserve the spacial symmetry of the crystal and to avoid double counting. This has been suggested by Bredow *et al.*[212] to describe crystal wavefunctions, while including only a minimum number of images compared to other approaches. In our implementation inside the `xtb` code, we only evaluate the electronic structure at the Γ-point.

To enable covalent bond dissociations, a finite electronic temperature treatment at typically low temperatures ($T_{\mathrm{el}} = 300\,\mathrm{K}$) is used.[71] This way, fractional orbital occupations are introduced and static electron correlation effects can be incorporated with formally a single-reference treatment. For a variational solution, the GFNn-xTB energy expressions given below are then augmented with an electronic entropy term:

$$G_{\mathrm{Fermi}} = k_B T_{\mathrm{el}} \sum_{\sigma=\alpha,\beta} \sum_{i} \left[n_{i\sigma} \ln(n_{i\sigma}) + (1 - n_{i\sigma}) \ln(1 - n_{i\sigma}) \right] . \tag{B.18}$$

T_{el} is the electronic temperature, k_B the Boltzmann constant and $n_{i\sigma}$ refers to the (fractional) occupation number of the spin-MO $\psi_{i\sigma}$. These are given by the Fermi-distribution.

$$n_{i\sigma} = \frac{1}{\exp[(\varepsilon_i - \varepsilon_F^{\sigma})/(k_B T_{\mathrm{el}})] + 1} \,. \tag{B.19}$$

ε_i is the orbital energy of the orbital ψ_i and ε^{σ} is the Fermi level within the respective spin orbital space (α or β). The occupation n_i for the spatial molecular orbital ψ_i is given as

$$n_i = n_{i\alpha} + n_{i\beta} \,. \tag{B.20}$$

All GFNn-xTB methods are formulated in a spin-restricted way, i.e., the spatial molecular orbitals are identical for α and β electrons. This treatment (also for open-shell systems) is rooted in the fact that no spin-dependent terms are present in the GFNn-xTB Hamiltonians. As a consequence, it is computationally efficient (only one Hamiltonian matrix diagonalization per SCF cycle) and also adds some robustness to the methods. This is beneficial in the context of extensive structural sampling, which these methods are often used for. It should, however, be clear that this treatment provides bad and even qualitatively incorrect energetic splittings for states of different multiplicity, i.e., GFNn-xTB Hamiltonians will always favor low-spin configurations.

The classical repulsion energy. The repulsion energy in all GFN-type methods is given as an atom-pairwise expression:

$$E_{\text{rep}} = \frac{1}{2}\sum_{A,B} \frac{Z_A^{\text{eff}} Z_B^{\text{eff}}}{R_{AB}} e^{-\sqrt{\alpha_A \alpha_B}(R_{AB})^{k_f}} \tag{B.21}$$

Here, Z^{eff} are element-specific constants that define the magnitude for the repulsion energy, and may coarsely correspond to effective nuclear charges (screened by the core reference density $\rho_0^{A,\text{core}}$). Though these parameters are somewhat agreeing with the latter in GFN1-xTB, they are fitted parameters in all GFN methods. Furthermore, $k_f = \frac{3}{2}$ is a global parameter, while the α exponents are element-specific parameters. There exists one exception to the value of k_f in GFN2-xTB, i.e., repulsion energies between first row (H,He) elements are using $k_f^{\text{H,He}} = 1$.

The extended Hückel energy. As mentioned above, typical tight-binding methods allow covalent bond formation through an extended Hückel (EHT)-type energy given as:

$$E_{\text{EHT}} = \sum_{\mu\nu} P_{\mu\nu} H_{\nu\mu}^{\text{EHT}} \tag{B.22}$$

Here, $P_{\mu\nu} = P_{\mu\nu}^0 + \delta P_{\mu\nu}$ is the valence electron density matrix in the nonorthogonal AO basis. In a generalized form, the EHT matrix elements read:

$$H_{\mu\nu}^{\text{EHT}} = \frac{1}{2} K_{AB}^{ll'} S_{\mu\nu} \left(H_{\mu\mu} + H_{\nu\nu}\right) X(EN_A, EN_B) \Pi(R_{AB}, l, l') Y(\zeta_l^A, \zeta_{l'}^B)\,,\ \forall\, \mu \in l(A), \nu \in l'(B) \tag{B.23}$$

Here, $K_{AB}^{ll'}$ is a shell-specific scaling constant. For a few element combinations in GFN1-xTB and in GFN0-xTB, an element pair-specific scaling parameter enters $K_{AB}^{ll'}$ as well (hence the atom labels A and B). $S_{\mu\nu}$ is the overlap matrix element of the AOs ϕ_μ and ϕ_ν. $H_{\mu\mu}/H_{\nu\nu}$ are the diagonal elements, which themselves are dependent on the chemical environment in all xTB methods (see below). The last three terms are absent in standard EHT or Wolfsberg-Helmholtz type expressions. These terms are system-specific and involve flexible scaling functions.

A common mathematical form in all GFNn-xTB schemes is found in the distance-dependent polynomial scaling function $\Pi(R_{AB}, l, l')$:

$$\Pi(R_{AB,l,l'}) = \left(1 + k_{A,l}^{\text{poly}} \left(\frac{R_{AB}}{R_{\text{cov},AB}}\right)^{\frac{1}{2}}\right)\left(1 + k_{B,l'}^{\text{poly}} \left(\frac{R_{AB}}{R_{\text{cov},AB}}\right)^{\frac{1}{2}}\right) \tag{B.24}$$

Here, the R_{AB} is the distance between the atoms on which the functions μ and ν are located. $R_{\text{cov},AB}$ are the summed covalent radii and taken from Ref. [213] without refitting. $k_{A,l}^{\text{poly}}$ are element- and shell-specific parameters.

This term allows a distance-dependent adjustment of the EHT energy in addition to the distance-dependence encoded in the overlap matrix elements $S_{\mu\nu}$. The electronegativity dependent term $X(EN_A, EN_B)$ and the basis function-dependent scaling are either not present or of non-unique mathematical form in the different GFNn-xTB Hamiltonians. The additional terms (i.e., the last three ones) rely only on global and element-specific parameters. It is clear that the EHT term in Eq. B.23 is fairly complex, but this way, the required flexibility to describe different covalent bonds is provided without resorting to an element pair-specific parameterization procedure. It is therefore not surprising that the dominant part of the GFNn-xTB methods' parameters ($> 50\%$) is incorporated here.

The isotropic electrostatic and XC energy. GFN1-xTB and GFN2-xTB share a formally equivalent isotropic electrostatic and XC energy, which originates from the second order term in the tight-binding expansion.

$$E_\gamma = \frac{1}{2} \sum_{A,B}^{N_{\text{atoms}}} \sum_{l \in A} \sum_{l' \in B} q_l q_{l'} \gamma_{AB,ll'} \tag{B.25}$$

Here, $q_l/q_{l'}$ are partial Mulliken shell charges and $\gamma_{AB,ll'}$ are short-ranged damped Coulomb interactions.[214–216]

$$\gamma_{AB,ll'} = \frac{1}{\sqrt{R_{AB}^2 + \eta_{AB,ll'}^{-2}}} \tag{B.26}$$

The form of the short-range damping term $\eta_{AB,ll'}$ is not identical in GFN1-xTB and GFN2-xTB and is given below.

The GFN1-xTB Hamiltonian

The GFN1-xTB[41] method is the first xTB version presented with a focus on the GFN properties.

Here, the energy expression is given by:

$$\begin{aligned} E_{\text{GFN1-xTB}} &= E_{\text{rep}}^{(0)} + E_{\text{disp}}^{(0)} + E_{\text{XB}}^{(0)} + E_{\text{EHT}}^{(1)} + E_{\text{IES+IXC}}^{(2)} + E_{\text{IES+IXC}}^{(3)} \\ &= E_{\text{rep}} + E_{\text{disp}}^{\text{D3}} + E_{\text{XB}}^{\text{GFN1}} + E_{\text{EHT}} + E_\gamma + E_\Gamma^{\text{GFN1}} \end{aligned} \tag{B.27}$$

In the first line of Eq. B.27, the superscript indicates the origin of the respective term in the tight-binding expansion. The repulsion energy in GFN1-xTB takes the form of Eq. B.21, while the EHT energy corresponds to Eq. B.22. In GFN1-xTB, the term $Y(\zeta_l^A, \zeta_{l'}^B) = 1$ for all cases, while the electronegativity dependent part is given by.

$$X(EN_A, EN_B) = (1 + k_{EN} \Delta EN_{AB}^2) \tag{B.28}$$

Here, ΔEN is the difference of the electronegativities (standard Pauling values) and $k_{EN} = -0.007$ is a global parameter.

The diagonal EHT matrix elements are atomic environment-dependent:

$$H_{\mu\mu} = h_A^l (1 + k_{CN,l} CN_A), \; \forall \mu \in l \in A \tag{B.29}$$

h_A^l is a shell- and element-specific parameter, while $k_{CN,l}$ are global angular momentum-specific parameters. CN_A is the geometric atomic fractional coordination number, which is taken from the D3 dispersion model.[26]

As shown in Eq. B.27, GFN1-xTB includes the isotropic energy term E_γ as given in Eq. B.25. In

GFN1-xTB, the short-range damping term is given as:

$$\eta_{AB,ll'} = 2\left(\frac{1}{\eta_A\left(1+\kappa_A^l\right)} + \frac{1}{\eta_B\left(1+\kappa_B^{l'}\right)}\right)^{-1} \tag{B.30}$$

This is the harmonic mean of the effective shell hardness values, which in turn are products of the element-specific atomic hardness and a shell-dependent scaling parameter. The last term in Eq. B.27 is an on-site third order electrostatic/XC correction:

$$E_{\Gamma}^{\mathrm{GFN1}} = \frac{1}{3}\sum_{A}^{N_{\mathrm{atoms}}} (q_A)^3 \Gamma_A \tag{B.31}$$

$q_A = \sum_{l \in A} q_l$ are atomic Mulliken partial charges and the Γ_A are element-specific atomwise parameters. This term typically enables stabilization of high partial charges, particularly for electronegative elements like O and F.

The dispersion energy is computed by means of the D3 dispersion model.[26,27]

$$E_{\mathrm{disp}}^{\mathrm{D3}} = -\frac{1}{2}\sum_{A,B}\sum_{n=6,8} s_n \frac{C_n(CN_A, CN_B)}{R_{AB}^n} f_{\mathrm{damp,BJ}}^{(n)}(R_{AB}) \tag{B.32}$$

Here, the C_n refer to the standard D3 dipole-dipole ($n = 6$) and dipole-quadrupole ($n = 8$) dispersion coefficients. These are environment-dependent by means of the geometric atomic fractional coordination numbers CN_A/CN_B. The individual scaling factors are $s_6 = 1.0$ and $s_8 = 2.4$. The Becke-Johnson (BJ) damping function is given as

$$f_{\mathrm{damp,BJ}}^{(n)}(R_{AB}) = \frac{R_{AB}^n}{R_{AB}^n + \left(a_1 R_{AB}^0 + a_2\right)^n} . \tag{B.33}$$

As in the regular BJ-damped D3 model, the damping constant is given by the ratio $R_{AB}^0 = \sqrt{C_8^{AB}/C_6^{AB}}$ and the global damping parameters, which in GFN1-XTB are $a_1 = 0.63$ and $a_2 = 5.0$.

Mainly due to the monopole approximation in GFN1-xTB, weak halogen bonds are not described well. Hence, a purely geometry dependent halogen-bond (XB) correction is added.

$$E_{\mathrm{XB}}^{\mathrm{GFN1}} = \sum_{AXB}^{N_{\mathrm{XB}}} f_{\mathrm{damp,AXB}} k_X \left[\left(\frac{k_{XR} R_{\mathrm{cov},AX}}{R_{AX}}\right)^{12} - k_{X2}\left(\frac{k_{XR} R_{\mathrm{cov},AX}}{R_{AX}}\right)^{6}\right]\left[\left(\frac{k_{XR} R_{\mathrm{cov},AX}}{R_{AX}}\right)^{12} + 1\right]^{-1} \tag{B.34}$$

Here $k_{XR} = 1.3$ and $k_{X2} = 0.44$ are global parameters, while k_X is a halogen-dependent parameter. $f_{\mathrm{damp,AXB}}$ is a damping function, which depends on the angle of the atoms involved in the halogen bond, i.e., the halogen X, its covalently bonded neighboring atom B, and the halogen-bond acceptor atom A.

$$f_{\mathrm{damp,AXB}} = \frac{1}{2}\left(1 - \frac{1}{2}\frac{R_{XA}^2 + R_{XB}^2 - R_{AB}^2}{|R_{XA}||R_{XB}|}\right)^6 \tag{B.35}$$

The typically stronger hydrogen bonds are also not well described by a minimal basis, monopole approximated tight-binding Hamiltonian, and geometry dependent hydrogen bond corrections have been used in some SQM methods.[217] In GFN1-xTB no extra term to describe hydrogen bonds is introduced. Instead, an additional s-AO function on hydrogen is used, which leads to extra stabilization of hydrogen bonds with a formally unmodified xTB Hamiltonian.

The GFN2-xTB Hamiltonian

The GFN2-xTB[8] method is the second version in the GFN framework. It is the first off-the-shelf tight-binding method with multipole electrostatics, anisotropic XC contributions and charge-dependent (D4) dispersion interactions. The energy expression is given by:

$$\begin{aligned} E_{\text{GFN2-xTB}} =& E_{\text{rep}}^{(0)} + E_{\text{disp}}^{(0),(1),(2)} + E_{\text{EHT}}^{(1)} + E_{\text{IES+IXC}}^{(2)} + E_{\text{AES+AXC}}^{(2)} + E_{\text{IES+IXC}}^{(3)} \\ =& E_{\text{rep}} + E_{\text{disp}}^{\text{D4}'} + E_{\text{EHT}} + E_{\gamma} + E_{\text{AES}} + E_{\text{AXC}} + E_{\Gamma}^{\text{GFN2}} \end{aligned} \tag{B.36}$$

The repulsion energy takes the form in Eq. B.21 in GFN2-xTB. The EHT energy term (Eq. B.22) in GFN2-xTB also includes the electronegativity dependent term as in GFN1-xTB (Eq. B.28) with $k_{\text{EN}} = 0.02$ and unmodified Pauling electronegativities are used for the difference ΔEN. The shell-exponent dependent term is

$$Y(\zeta_l^A, \zeta_{l'}^B) = \left(\frac{2\sqrt{\zeta_l^A \zeta_{l'}^B}}{\zeta_l^A + \zeta_{l'}^B} \right)^{\frac{1}{2}} . \tag{B.37}$$

Here, ζ_l^A are the STO exponents of the GFN2-xTB AO basis.[211] The exponent-dependency introduces effects similar to the kinetic energy integrals in ab initio theories. Also in GFN2-xTB, the diagonal EHT matrix elements are atomic environment-dependent:

$$H_{\kappa\kappa} = h_A^l - \delta h_{CN'_A}^l CN'_A \tag{B.38}$$

h_A^l is a shell- and element-specific parameter just like $\delta h_{CN'_A}^l$. The latter is the proportionality constant for the dependency on the modified GFN2-type coordination number CN'_A, which is more long-ranged than the D3 variant (see Ref. [8] for details).

The isotropic second order electrostatic energy E_γ is given in Eq. B.25 with the following expression for the short-range damping:

$$\eta_{AB,ll'} = \frac{1}{2} \left[\eta_A \left(1 + \kappa_A^l \right) + \eta_B \left(1 + \kappa_B^{l'} \right) \right] \tag{B.39}$$

η_A and η_B are element-specific fit parameters, while κ_A^l and $\kappa_B^{l'}$ are element-specific scaling factors for the individual shells (with $\kappa_A^l = 0$ for $l = 0$).

The dispersion energy is described by a modified D4 dispersion model

$$\begin{aligned} E_{\mathrm{disp}}^{\mathrm{D4}'} = &- \sum_{A>B} \sum_{n=6,8} s_n \frac{C_n^{AB}(q_A, CN_{\mathrm{cov}}^A, q_B, CN_{\mathrm{cov}}^B)}{R_{AB}^n} f_{\mathrm{damp,BJ}}^{(n)}(R_{AB}) \\ &- s_9 \sum_{A>B>C} \frac{(3\cos(\theta_{ABC})\cos(\theta_{BCA})\cos(\theta_{CAB})+1)C_9^{ABC}(CN_{\mathrm{cov}}^A, CN_{\mathrm{cov}}^B, CN_{\mathrm{cov}}^C)}{(R_{AB}R_{AC}R_{BC})^3} \\ &\times f_{\mathrm{damp,zero}}^{(9)}(R_{AB}, R_{AC}, R_{BC}) \end{aligned} \quad \text{(B.40)}$$

Here, $f_{\mathrm{damp,BJ}}^{(n)}(R_{AB})$ is the damping function from Eq. B.33. The term in the second line is the three-body Axilrod-Teller-Muto (ATM) term and the last line is the corresponding zero-damping function for this term. The two-body London dispersion energy is depending on the covalent coordination number CN_{cov}^A and the atomic charges q_A. Different from the regular D4 model,[9] the atomic partial charges are taken from a Mulliken population in GFN2-xTB and are solved self-consistently. The three-body term is environment-dependent through the covalent coordination numbers, but does not depend on the partial atomic charges in the D4 model. The damping and scaling parameters in the dispersion model are $a_1 = 0.52$, $a_2 = 5.0$, $s_6 = 1.0$, $s_8 = 2.7$, and $s_9 = 5.0$. The third order term in GFN2-xTB is also an on-site term, which is formulated in a shell-specific form:

$$E_{\Gamma}^{\mathrm{GFN2}} = \frac{1}{3} \sum_{A}^{N_{\mathrm{atoms}}} \sum_{l \in A} (q_l)^3 K_l^{\Gamma} \Gamma_A \quad \text{(B.41)}$$

Here, q_l is the partial shell charge and Γ_A is an element-specific parameter. K_l^{Γ} is a global shell-specific parameter (see Ref. [8]).

GFN2-xTB includes anisotropic electrostatic and XC terms. These are given as:

$$\begin{aligned} E_{\mathrm{AES}} =& E_{q\mu} + E_{q\Theta} + E_{\mu\mu} \\ =& \frac{1}{2} \sum_{A,B} \Big\{ f_3(R_{AB}) \left[q_A \left(\mu_B^T \mathbf{R}_{BA} \right) + q_B \left(\mu_A^T \mathbf{R}_{AB} \right) \right] \\ &+ f_5(R_{AB}) \Big[q_A \mathbf{R}_{AB}^T \Theta_B \mathbf{R}_{AB} + q_B \mathbf{R}_{AB}^T \Theta_A \mathbf{R}_{AB} \\ &- 3 \left(\mu_A^T \mathbf{R}_{AB} \right) \left(\mu_B^T \mathbf{R}_{AB} \right) + \left(\mu_A^T \mu_B \right) R_{AB}^2 \Big] \Big\} . \end{aligned} \quad \text{(B.42)}$$

Here, μ_A is the cumulative atomic dipole moment of atom A and Θ_A is the corresponding traceless quadrupole moment. These cumulative atomic multipole moments (CAMM)[218] describe the local atomic multipole moment contribution in a Mulliken approximation scheme. The distance dependence including damping is given by

$$f_n(R_{AB}) = \frac{f_{\mathrm{damp}}(a_n, R_{AB})}{R_{AB}^n} = \frac{1}{R_{AB}^n} \cdot \frac{1}{1 + 6 \left(\frac{R_0^{AB}}{R_{AB}} \right)^{a_n}} \quad \text{(B.43)}$$

The damping function is related to the one in the original D3 dispersion model.[26] a_n are adjusted global parameters, whereas $R_0^{AB} = 0.5\left(R_0^{A\prime} + R_0^{B\prime}\right)$ determines the damping of the AES interaction. $R_0^{A\prime}$ is made dependent on the GFN2-type coordination number (see above) for many lighter elements. GFN2-xTB includes all multipole contributions up to second order for the electrostatic energy.

The second order anisotropic XC energy in GFN2-xTB is given by

$$E_{AXC} = \sum_A \left(f_{XC}^{\mu_A} |\mu_A|^2 + f_{XC}^{\Theta_A} ||\Theta_A||^2 \right) . \tag{B.44}$$

Again, μ_A and Θ_A are the cumulative atomic multipole moments mentioned above. $f_{XC}^{\mu_A}$ and $f_{XC}^{\Theta_A}$ are fitted element-specific parameters. These terms capture changes due the anisotropic deformation of the electron density around an atom A. To some extent, shortcomings of the small AO basis set, i.e., the lack of polarization functions, may be compensated by this term.

Due to the anisotropic electrostatic and XC terms, GFN2-xTB does not require any additional hydrogen or halogen bond corrections.

The GFN0-xTB Hamiltonian

The most recent member in the family of xTB methods is GFN0-xTB. Since no terms from the tight-binding expansion beyond first order are included, no self-consistent field procedure is necessary, making the method about $5 - 20$ times faster compared to GFN2-xTB. The tradeoff is less flexibility in the electronic structure, similar as in the Harris-Foulkes functional[207,208] mentioned above in comparison to KS-DFT. Consequently, the formal choice of the reference system or its corresponding parameterization becomes more important. GFN0-xTB includes two essential types of correction procedures to recover a transferability comparable to the self-consistent xTB variants: one is a classical pairwise potential to correct for covalent bonds of heteroatoms (srb, short-range bond correction). The other, even more important one is incorporating semi-classical atomic charges determined variationally from an electronegativity equilibration model (EEQ). The latter describes the atomwise isotropic electrostatic energy (in place of the E_γ energy in GFN1- and GFN2-xTB) in form of a purely classical energy. The charges furthermore serve the purpose of modifying some of the EHT parameters. This can be regarded as a system-specific $ad\ hoc$ re-definition of the reference system parameters (i.e., of ρ_0 in Eq. B.7). In other words, GFN0-xTB works with a system-specific reference system of spherical, but in general, partially charged atoms. The GFN0-xTB energy is then given as:

$$\begin{aligned} E_{GFN0-xTB} =& E_{rep}^{(0)} + E_{disp}^{(0)} + E_{EHT}^{(1)} + \Delta E^{(0)} \\ =& E_{rep} + E_{disp}^{D4} + E_{EEQ} + E_{EHT} + E_{srb} \end{aligned} \tag{B.45}$$

The term $\Delta E^{(0)}$ in the first line of Eq. B.45 formally contains all the changes in the Hamiltonian due to the effective change of the zeroth order reference. The repulsion energy E_{rep} has the established form as in Eq. B.21. The dispersion energy is computed via the D4 dispersion model (E_{disp}^{D4})

$$E_{disp}^{D4} = - \sum_{A>B} \sum_{n=6,8} s_n \frac{C_n^{AB}(q_A, CN_{cov}^A, q_B, CN_{cov}^B)}{R_{AB}^n} f_{damp,BJ}^{(n)}(R_{AB}) . \tag{B.46}$$

Different from the default D4 dispersion model as well as its GFN2-xTB variant, the three-body term (cf. Eq. B.40) is dropped, as it would noticeably increase the computational cost of the method. As in the default D4 model, the partial charges are taken from the aforementioned EEQ model. These charges result from self-consistently solving for the E_{EEQ} energy in Eq. B.45 with the constraint that the total charge is preserved:

$$E_{EEQ} = \sum_A \left[\chi_A q_A + \frac{1}{2} \left(J_{AA} + \frac{2}{\sqrt{\pi}} \gamma_{AA} \right) q_A^2 \right] + \frac{1}{2} \sum_{A,B} q_A q_B \frac{\text{erf}(\gamma_{AB} R_{AB})}{R_{AB}} \tag{B.47}$$

Here, χ_A is the electronegativity of atom A, which is made dependent on the atomic environment via a modified coordination number (mCN_A):

$$\chi_A = EN_A - \kappa_A \sqrt{mCN_A} \tag{B.48}$$

κ_A is an element-specific fitted parameter and EN_A the Pauling electronegativity. The modified coordination number is defined as:

$$mCN_A = \frac{1}{2} \sum_{B \neq A} \left[1 + \text{erf} \left(-7.5 \left(\frac{R_{AB}}{R_{AB}^{cov}} + 1 \right) \right) \right] \tag{B.49}$$

R_{AB}^{cov} are the summed covalent radii from Ref. [219]. J_{AA} is an element-specific parameter related to the atomic hardness, and γ_{AB} is related to the inverse root mean square of the atomic radii of atoms A and B (see Ref. [192] for details). This term provides a description of the electrostatic energy at zeroth order in the tight-binding model.

The EHT energy in GFN0-xTB is given by the expressions in Eqs. B.22 and B.23. In the EHT part, the shell-exponent dependent term takes the same form as in GFN2-xTB (see Eq.B.37), while the electronegativity dependent term is shell-dependent in GFN0-xTB:

$$X(EN_A, EN_B) \rightarrow X^{ll'}(EN_A, EN_B) = 1 + k_{EN}^{ll'} \Delta EN_{AB}^2 + k_{EN}^{ll'} b_{EN} \Delta EN_{AB}^4 \tag{B.50}$$

Compared to GFN1- and GFN2-xTB, a higher power term of the electronegativity is included. Modified Pauling values for the electronegativity are used, while $k_{EN}^{ll'}$ is a shell-specific and b_{EN} is a global parameter. The diagonals of the EHT Hamiltonian are made flexible with respect to the modified coordination number as well as the atomic partial charges:

$$H_{\kappa\kappa} = h_A^l - \delta h_{mCN_A}^l mCN_A - \delta h_{q_A}^l q_A - \Gamma_{q_A}^l q_A^2 \tag{B.51}$$

While the first part of this equation is formally identical to GFN2-xTB, the last two terms describe the effective modification of the reference atom due to the charged state, which is derived from the EEQ model. The parameters $\delta h_{q_A}^l$ and $\Gamma_{q_A}^l$ are formally related to the chemical hardness and its derivative with respect to the particle number, respectively, but are simply fitted shell- and element-specific parameters in GFN0-xTB. We note at this point, that first order electrostatic effects due to neighboring atoms are currently not incorporated in the GFN0-xTB model, but might be added in a future revision.

The last term in Eq. B.45 is the short-range bond correction, similar to the “short-range basis” corrections in the HF-3c and B97-3c methods.[79,173]

$$E_{srb} = k_{srb} \sum_{A,B} \exp\left[-\eta_{srb}\left(1 + g_{scal}\Delta EN_{AB}^{2}\right)\left(R_{AB} - R_{AB}^{srb}\right)^{2}\right] \tag{B.52}$$

This correction is only applied for heteroatomic pairs with atomic number $Z_A \in [5, 9]$. Here, k_{srb}, η_{srb}, and g_{scal} are global fit parameters, The summed covalent bond radii R_{AB}^{srb} are modified by the electronegativities as:

$$R_{AB}^{srb} = \left(R_A^0 + R_B^0\right)\left(1 - c_1|\Delta EN_{AB}| - c_2\Delta EN_{AB}^{2}\right) \tag{B.53}$$

where R_A^0 are the damping radii from the D3 model[26], which are modified by the coordination number and c_1 and c_2 are global parameters. E_{srb} essentially corrects the energies and bond lengths for polar bonds of second period elements.

The GFN0-xTB variant obviously contains more empiricism in the Hamiltonian compared to GFN1- and GFN2-xTB, but still avoids pairwise parameterization. While the non-self-consistent treatment obviously makes the method less costly, it has also a practical advantage: due to the lack of Fock exchange, self-consistent tight-binding methods also suffer from self-interaction error related phenomena, which become particularly pronounced in polar systems like proteins, where the SCF calculations might not converge anymore. These problems can be remedied by including an implicit solvation model,[41] however, GFN0-xTB does not suffer from these defects and typically yields larger HOMO-LUMO gaps, as a result from the non-self-consistent treatment.

Treatment of Lanthanide Elements

In the GFN*n*-xTB Hamiltonians, the "f-in-core" approximation is used throughout for lanthanide elements. That is, they are treated as 4d transition metals with three valence electrons and no explicit consideration of the f-electrons. This treatment is motivated by ab initio calculations indicating that the f-electron shell lies below the valence shell and their implicit handling, e.g., in form of a pseudopotential or by appropriate parameterization (GFN*n*-xTB) is reasonable.[220] This is different, if spectroscopic properties are of interest, and these elements have been neglected in the simplified time-dependent xTB model for excited states for this reason (see next subsection).

B.2.2 Simplified time-dependent xTB for excited states

The computational bottleneck of calculating electronic spectra with the simplified versions of full[221] and Tamm-Dancoff approximated[197] (TDA) time-dependent (TD)-DFT (i.e., sTD-DFT or sTDA-DFT) is the calculation of the ground state orbital coefficients and energies. The very fast semiempirical ground state tight-binding method sTDA-xTB aims at eliminating this bottleneck.[25] The general work-flow of this approach is shown in Figure B.3.

The xTB ground state calculation consists of two parts: a valence tight-binding (VTB) part and an extended tight-binding (XTB) part. Note that the two parts are denoted in this chapter as XTB and VTB, respectively, to distinguish them from the other xTB methods. First, atomic charges are determined in a truncated SCC procedure by VTB. The second, single-diagonalization XTB step is using these VTB charges. The VTB part applies a minimal valence and partially polarized AO basis set as in GFN2-xTB,

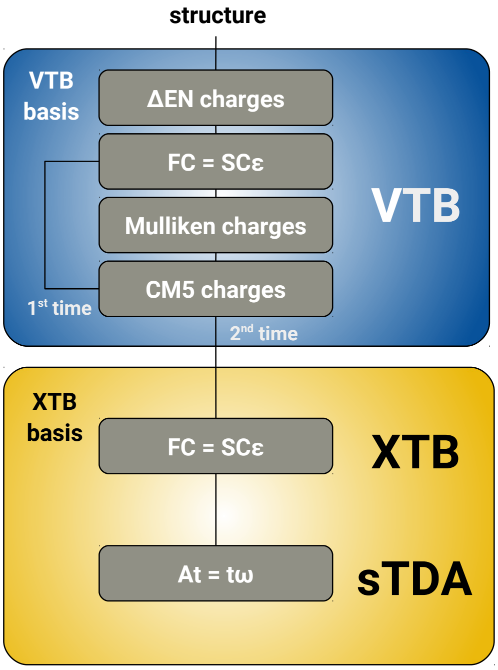


Figure B.3: Computational workflow of the sTDA-xTB method. The VTB (valence tight-binding) part uses geometry-dependent electronegativity difference-based charges as input and generates Mulliken charge-based CM5 charges in a trunctated SCC procedure. The XTB (extended tight-binding) step then uses these charges as input and generates orbitals in a minimal+diffuse basis set for a subsequent excited state calculation at the simplified Tamm-Dancoff approximated TD-DFT (sTDA) level.

whereas the second XTB part uses an augmented minimal valence expansion including diffuse functions for the treatment of states with Rydberg excitation character. The XTB Hamiltonian matrix elements have the standard GFN1 form and read

$$F_{\mu\nu} = H^0_{\mu\nu} + k_q \frac{1}{2} S_{\mu\nu} \sum_C (\gamma_{AC} + \gamma_{BC}) q_C^{VTB} \quad (\mu \in A, \nu \in B). \tag{B.54}$$

Their diagonalization yields the orbital energies and coefficients used in the sTDA or sTD formalism. γ_{AC} denotes the inter-electronic repulsion function in the Mataga-Nishimoto formulation[214] that reads as:

$$\gamma_{AC} = \frac{1}{R_{AC} + \frac{2}{\eta(A)+\eta(B)}}, \tag{B.55}$$

where R_{AC} is the interatomic distance and η the chemical hardness. The applied basis functions in the different parts are summarized in Table B.2. Virtual orbitals are shifted to resemble hybrid functional energy gaps. Apart from a local excitation correction (see Ref. [25]), the sTDA/sTD part is not modified in sTDA-xTB compared with sTDA-DFT.

The method has been parameterized to reproduce accurate theoretical reference vertical excitation

Table B.2: Description of the AO basis sets used. n denotes the principal quantum number of the valence shell of the respective element.

	part	
element	VTB	XTB
H-He	ns	ns, (n+1)sp
group I/II	nsp	nsp
B-Ne	nsp	nsp, (n+1)sp
Al,Ga,In,Zn,Cd,Hg	nsp	nsp
remaining group IV-VII non-metals	nsp, (n+1)d	nsp, (n+1)sp
d-block elements	nd, (n+1)sp	nd, (n+1)sp

energies only, and hence, it is not suited for describing ground state energetics or PES regions far away from the Franck-Condon point. Opposed to the GFN*n*-xTB methods, it is a single-point method without a gradient implementation. Therefore, its application for photochemistry and related MD studies is rather limited. The comparison of the fitted properties – atomic charges (VTB) and vertical excitation energies (sTDA-XTB) – with the theoretical reference is shown in Figure B.4. An excellent agreement between the xTB and reference data is observed.

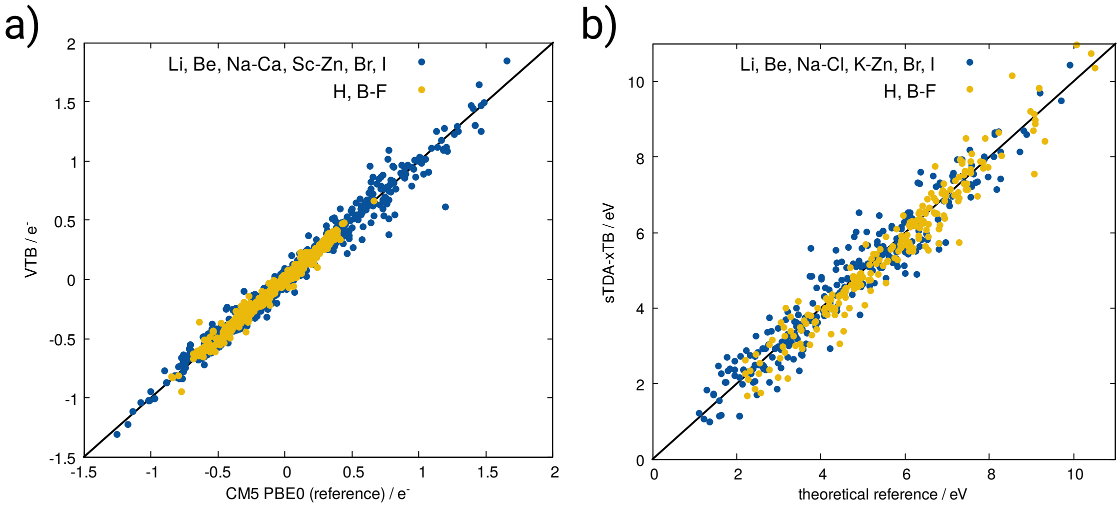


Figure B.4: Comparison of a) VTB and PBE0/TZVP[59,222] CM5[92] reference atomic charges and b) sTDA-xTB and SCS-CC2/TD-DFT[223] reference excitation energies. The black line denotes a one to one correspondence of the two data sets.

B.2.3 Non-electronic variants (GFN-FF)

The evolution of GFN1-, GFN2- and GFN0-xTB inspired the development of a generic force-field. We have already established the connection between atomistic, intermolecular Lennard-Jones-type force-fields and zeroth order tight-binding methods in Eq. B.8. Furthermore, the experience with GFN0-xTB has shown that the approach of $ad\ hoc$ adjustment of the reference system (following the picture of Eq. B.7) by adding more flexibility and more reasonably chosen parameters has some prospects for success. The development of the so-called GFN force-field (GFN-FF) approach can be regarded as in line with these findings.

From the practical point of view, the main focus of GFN-FF is directed towards the description of very large bio-macromolecular systems such as (metallo-)proteins[224], supramolecular assemblies[225] and metal-organic frameworks.[226] Screening of very many structures or treating molecules with more than a few thousand atoms routinely is impractical at any electronic GFNn-xTB level. It is intended as a versatile tool for drug design in life sciences and structure screening in various fields of chemistry.[227–229] Therefore, GFN-FF introduces an approximation to the remaining quantum mechanics in GFN0-xTB by replacing the zeroth order TB terms with classical bond, bending angle, and torsion angle potentials. To remain accurate in the description of conjugated systems, GFN-FF retains an iterative Hückel scheme for a selected set of π-atoms. The resulting bond orders affect the force constants and other energy relevant parameters of the system. To yield accurate results, the FF parameters are fitted to reproduce B97-3c[79] equilibrium geometries and frequencies. Thereby, a strictly global and element specific parameter strategy is applied and no element pair specific parameters are employed. This approach is a unique feature of all GFN methods and differs strongly from the parameterization strategies of other FFs (e.g., see Refs.[230–233]). Special attention is paid to the simple application of GFN-FF. As input only Cartesian coordinates and elemental composition are required from which fully automatically all potential energy terms are constructed.

The total GFN-FF energy expression is given by

$$E_{\text{GFN-FF}} = E_{\text{cov}} + E_{\text{NCI}}, \tag{B.56}$$

where E_{cov} refers to the bonded FF energy and E_{NCI} describes the non-covalent interactions. In the covalent part, interactions are described by asymptotically correct (dissociative) bonding, angular, and torsional terms. Repulsive terms are added for bonded and non-bonded interactions separately. Additionally, a three-body correction to the bonded part is added (abc), that extends beyond the sum of pair-wise interactions

$$E_{\text{cov}} = E_{\text{bond}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{rep}}^{\text{bond}} + E_{\text{abc}}^{\text{bond}}. \tag{B.57}$$

In the non-covalent part, electrostatic interactions are described by the EEQ model as in GFN0-xTB. Overall, two sets of EEQ charges are used. One depends as usual on the actual geometry, whereas another set of charges is exclusively covalent topology based, introducing further polarizability and leading to large simplifications in terms of gradient computations. Dispersion interactions are taken into account by a simplified version of the D4 scheme,[9] in which the dispersion coefficients are scaled by atomic charges. Without any detailed electronic information, the correct description of hydrogen and halogen bonds is challenging. Therefore, additional charge-scaled corrections (denoted HB and XB, respectively) are applied to the non-covalent energy:

$$E_{\text{NCI}} = E_{\text{IES}} + E_{\text{disp}} + E_{\text{HB}} + E_{\text{XB}} + E_{\text{rep}}^{\text{NCI}}. \tag{B.58}$$

GFN-FF reaches quadratic scaling $\mathcal{O}(\mathrm{N}^2)$ of the computation time for energy and forces, whereas all GFNn-xTB methods show cubic scaling with respect to the number of atoms. It is the computationally most efficient member of the GFN family. For illustration, a comparison between total CPU times for single point and gradient calculations for all GFN methods is given in Figure B.5 for small to medium sized proteins (from 300 up to 6000 atoms).

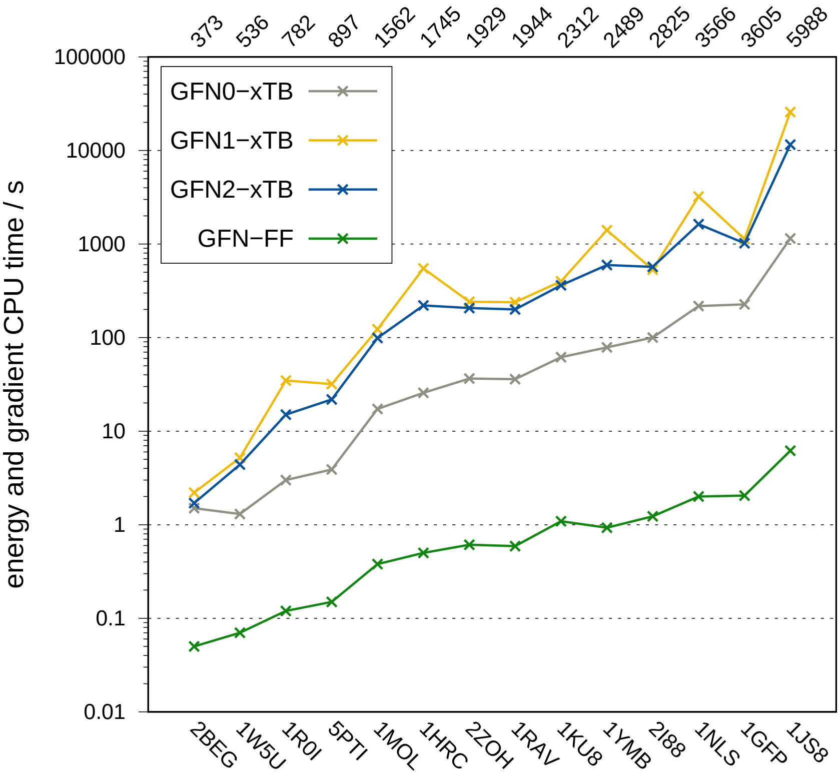


Figure B.5: CPU times (given in seconds on a logarithmic scale) for single point energy plus gradient calculations of 14 proteins. Computations were performed using a quad-core desktop machine with 4.20 GHz Intel i7-7700K CPUs. The PDB identifiers are given on the abscissa, the corresponding number of atoms is given on top.

To achieve SCC convergence, the GBSA(H_2O) solvation model (see next subsection) had to be employed for GFN1- and GFN2-xTB. With GFN0-xTB, a speed-up factor of 2-20 is achieved while GFN-FF improves on this even further. It is about two orders and three orders of magnitude faster than GFN0-xTB and GFN1-/GFN2-xTB, respectively.

B.2.4 Continuum solvation model (GBSA)

To create realistic computational models, solvent effects have to be accounted for, either by explicitly including solvent molecules (and dynamical sampling) or by a parameterized implicit solvent model. Due to its favorable computational cost the latter approach is pursued in the framework of xTB methods including GFN-FF.

Two kinds of polar implicit solvation models are suitable for xTB, either a polarizable continuum model (PCM) or a generalized Born (GB) model, where the former has the disadvantage of introducing an integration grid, which can introduce a significant overhead for large scale calculations. Therefore, we will only discuss the implementation of the GB model present in the `xtb` program.

In the GB model, a molecule is considered as a continuous region with a dielectric constant ϵ_{in} surrounded by infinite solvent with a dielectric constant ϵ_{out}.[87] The electrostatic interaction in the presence of a polarized solvent can then be expressed as the solvation energy

$$\Delta G_{\text{GB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{out}}}\right)\sum_{A=1}^{N}\sum_{B=1}^{N}\frac{q_A q_B}{\left(R_{AB}^2 + a_A a_B \exp\left[-\frac{R_{AB}^2}{4a_A a_B}\right]\right)^{\frac{1}{2}}}, \tag{B.59}$$

where $a_{A/B}$ are the effective Born radii of the atoms A/B. The GB model is introduced in the xTB Hamiltonian as a second order fluctuation in the charge density and described by the atomic potential $\mathbf{V}^{\text{GB}}$ given as

$$V_A^{\text{GB}} = -\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{out}}}\right)\sum_{B=1}^{N}\frac{q_B}{\left(R_{AB}^2 + a_A a_B \exp\left[\frac{R_{AB}^2}{4a_A a_B}\right]\right)^{\frac{1}{2}}}. \tag{B.60}$$

The Born radii are evaluated by an Onufriev–Bashford–Case (OBC) corrected pairwise approximation to the molecular volume given as

$$\frac{1}{a_A} = \frac{1}{a_{\text{scale}}}\left(\frac{1}{R_A^{\text{cov}} - R_{\text{offset}}} - \frac{1}{R_A^{\text{cov}}}\cdot \tanh\left[b\Psi_A - c\Psi_A^2 + d\Psi_A^3\right]\right), \tag{B.61}$$

where R_A^{cov} is the covalent radius of atom A, a_{scale} and R_{offset} are global parameters and $b = 1.0$, $c = 0.8$ and $d = 4.85$ are the parameters for the OBC correction, which correspond to the GB$^{\text{OBC}}$II model.[234] The OBC correction increases the Born radii for atoms buried deep inside a molecular cavity, which would usually be underestimated. Ψ_A is the pairwise approximation to the volume integral given by

$$\Psi_A = \frac{R_A^{\text{cov}} - R_{\text{offset}}}{2}\sum_{B}\Omega(R_{AB}, R_A^{\text{cov}}, s_B R_B^{\text{cov}}), \tag{B.62}$$

with Ω being the pairwise function used to approximate the volume integral, which is only dependent on the distance and the covalent radii. Note that the covalent radius of the second atom is scaled by the element-specific descreening value s_B to compensate the systematic overestimation of the volume by this approach.

In addition to this polar contribution to the solvation energy, a non-polar surface area contribution depending on the solvent accessible surface area (SASA) is given by

$$\Delta G_{\text{SASA}} = \sum_{A=1}^{N}\gamma_A\sigma_A, \tag{B.63}$$

where γ_A is the surface tension and σ_A is the SASA of atom A. To evaluate the latter we resort to the smoothly differentiable numerical approach introduced from Ref. [89] and integrate the surface area on an angular Lebedev grid.

The SASA is also used in an empirical hydrogen-bond correction to the generalized Born energy as

$$\Delta G_{\mathrm{GB+HB}} = \Delta G_{\mathrm{GB}} - \sum_{A} g_{A}^{\mathrm{HB}} q_{A}^{2} \frac{\sigma_{A}}{A_{A}}, \tag{B.64}$$

where g_A^{HB} is the strength of the hydrogen bonds between this atom and the solvent molecules and A_A is the surface area of the free atom. This simplified form has been chosen since the hydrogen-bond correction should enter the Hamiltonian as a potential due to the charge dependency.

The total solvation free energy is given by

$$\Delta G_{\mathrm{solv}} = \Delta G_{\mathrm{GB+HB}} + \Delta G_{\mathrm{SASA}} + \Delta G_{\mathrm{shift}}, \tag{B.65}$$

where an additional shift is included depending on the chosen reference state of the solution. This solvation free energy is fitted with four global parameters, the Born radius offset, the Born radius scaling, the probe radius of the solvent molecule, and the value of $\Delta G_{\mathrm{shift}}$ as well as three element specific parameters, the descreening value, the surface tension, and the hydrogen bond strength to reproduce COSMO-RS16 solvation free energies.[99]

B.3 Implementation

The GFN methods are implemented in the open-source software `xtb`, which provides a framework to use them on their own or together with other algorithms. `xtb` provides a userfriendly interface for performing geometry optimizations, vibrational frequency calculations, and molecular dynamics simulations. Usual workflows like geometry optimization, vibrational (harmonic) frequency calculation, and evaluation of thermodynamic functions are easily available as composite keywords. Thus, an additional input file, besides the input geometry, is not necessary to perform calculations with `xtb`. The usual obstacle of learning a new input format and adjusting numerical thresholds is minimized as much as possible, making it fairly straightforward to start running calculations with `xtb`. Through its unique design, the `xtb` program has become an integral part in a number of algorithms and programs, developed in our group or other groups. One of the early users of the `xtb` program is the quantum chemistry electron ionization mass spectrometry (QCEIMS) method implemented in the program of the same name.[235] `qceims` is accessing the self-consistent xTB methods to provide electronic structure information, like ionization potentials or to drive high-temperature molecular dynamics simulation to allow for first principles predictions of electron ionization mass spectra, which provides fundamental insights into the fragmentation process not available with standard machine learning-based algorithms. While not tailor-made for the prediction of transition states or reaction paths, the GFNn-xTB methods are suitable to provide a fast initial path when coupled with an appropriate algorithm. One such interface exists for the growing string method (GSM) program by the Zimmermann group implemented in the `mGSM` program.[236,237] To avoid introducing an additional interface in an established software package, the `xtb` program is wrapped in a flexible way to mimic output for an already existing and tested interface to `mGSM`. We find that this approach is most sustainable and particularly advantageous for users already familiar with the `mGSM` program, such that they can quickly integrate the GFNn-xTB methods and any future extension of the `xtb` program to their already existing workflows. In a recent publication of our group, we have highlighted this particular combination of `mGSM` and `xtb`.[238] The most prominent use of `xtb` is in the conformer-rotamer ensemble search tool `crest`,[19,239] which acts as a driver to perform and schedule calculations with `xtb`. A robust file-based input and output communication between the programs is established by using shared memory parallelization to allow for multiple `xtb` instances being driven by the scheduler provided with `crest`. Starting from an input structure, `xtb` will be used to generate the distorted structures by (biased) molecular dynamics simulations. From the generated trajectories `crest` is selecting structures for relaxations performed in parallel with multiple `xtb` instances. Each instance regularly reports back its optimization progress, such that `crest` can re-rank, filter, and prioritize the calculations most efficiently. Finally, we want to highlight the usage of both `xtb` and `crest` within the computational chemistry framework for energetic sorting of conformer-rotamer ensembles, named `enso`. The `enso` framework is designed to automate the re-ranking of conformer-rotamer ensembles at DFT level.[188] It relies implicitly on `xtb`-via-`crest`, but also directly for the calculation of thermostatistical corrections with the modified rigid-rotor-harmonic-oscillator (mRRHO) partition functions[73] or of solvation free energy contributions using the GBSA solvation models available for the GFNn-xTB methods. Those contributions to free energy are then used to refine conformer-rotamer ensembles produced by `crest` in an initial stage together with low-cost DFT methods like PBEh-3c[142] or B97-3c[79] before starting more expensive hybrid-functional calculations or property calculation like, e.g., NMR shifts (see Fig. B.1).

Besides this selection of applications and algorithms developed in the recent years by our group, the GFNn-xTB methods were quickly adopted in many existing quantum chemistry program packages, like AMS,[179] CP2K,[180] DFTB+,[13] entos,[182] Orca,[183,184] and TeraChem.[185,186] Moreover, interfaces to the `xtb` program are already available in large frameworks like ASE[77] or Cuby4.[181]

To facilitate the usability of `xtb`, an in-depth documentation is available covering all of the possible workflows.[240] This documentation is managed by established software documentation tools, namely `sphinx`[241] and `asciidoctor`,[242] which are used to automatically generate, among others, Linux manual pages, a printable PDF manual, and the static HTML code for the online-documentation. The latter is hosted by the read-the-docs project to be easily accessible to all users (see `https://xtb-docs.readthedocs.io`). Additionally, we dedicated the development of the `xtb` program to the open-source idea and hence published it under the Lesser General Public License (v3+) with the source code publicly available on GitHub (see Ref.[204]). One of the greatest burdens of quantum chemistry programs is the proper installation, which can be especially tricky in high-performance computing (HPC) environments. To actively support system operators in their efforts to make `xtb` available to their users, we make `xtb` available with the `conda` cross-platform package manager via an official conda-forge channel and also together with the easy build toolchain allowing to install `xtb` in an existing `module` system in HPC environments. For developers, we decided to offer dual support of both the established CMake build system and the relatively new but promising meson build system. A solid unit testing framework is also available and readily coupled to continuous integration services of Travis-CI and GitHub-Actions.

B.3.1 Geometry optimization of large systems

Due to the inherent computational efficiency of the underlying electronic structure method, a similarly fast and robust geometry optimization procedure had to be developed along with the GFNn-xTB methods. Otherwise, the geometry relaxation steps can amount to a significant fraction of the overall computation time, in particular at the GFN-FF level.

Special constraints on the optimization algorithm are the choice of a general and robust coordinate system, which is inexpensive in the generation, and a fast and reliable geometry displacement step with a better formal scaling and prefactor than the fastest available GFNn-xTB method. A model Hessian based on the work of Lindh *et al.*[243] is used for the generation of an Approximate Normal mode Coordinate (ANC) system. With an efficient screening, the calculation of this Hessian can be performed with a quadratically scaling algorithm. However, its diagonalization to obtain the transformation matrices scales cubically with the number of atoms. To reduce this bottleneck the coordinate system is generated from the actual Cartesian coordinates only every 20 to 40 optimization steps. For the calculation of the coordinate displacement we established a combination of the rational function algorithm[244] together with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) Hessian update step.[245]

For systems with several thousands of atoms even this setup can become too slow when combined with faster methods than GFN0-xTB, i.e., GFN-FF. Even if both displacement and update are evaluated by a L-BFGS algorithm,[246] additionally a faster diagonalization strategy had to be designed based on a Hessian fragmentation scheme (see Figure B.6).

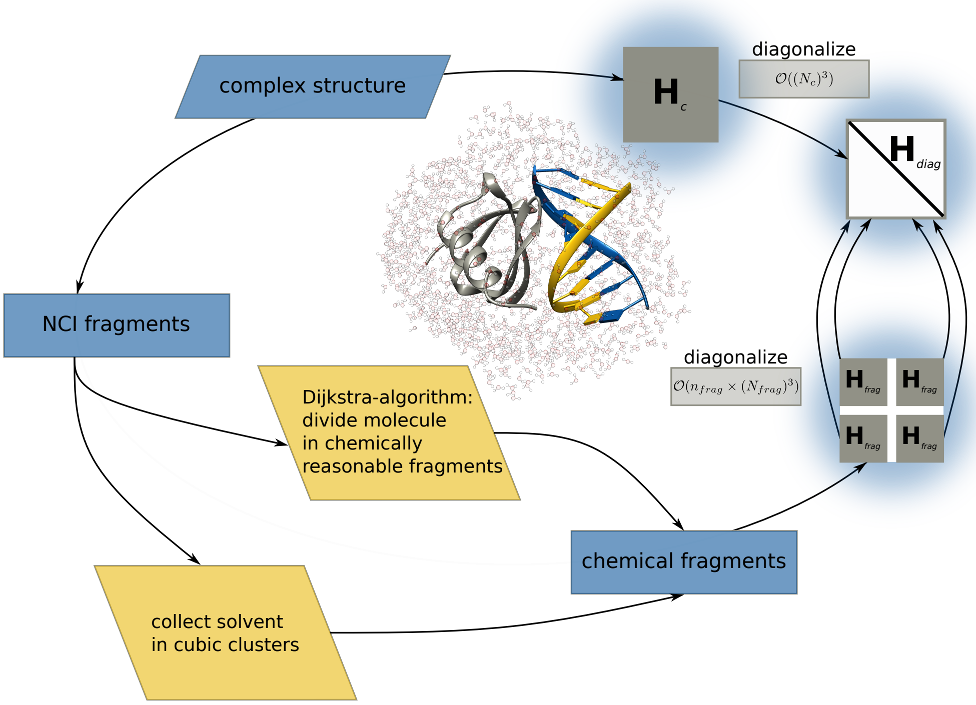


Figure B.6: Schematic representation of the fragmented Hessian algorithm. A complex molecular structure with $\mathrm{N_c}$ atoms is divided into several chemically reasonable fragments (of size $\mathrm{N}_{\mathtt{frag}}$) and the Hessian of each fragment $\mathrm{H}_{\mathtt{frag}}$ is diagonalized individually. The diagonalized Hessian $\mathrm{H}_{\mathtt{diag}}$ is constructed from the fragments, instead of diagonalizing the Hessian of the entire system $\mathrm{H_c}$, thus reducing the overall computational costs.

With the covalent topology at hand, a fragmentation scheme identifies in a first step non-covalently bound fragments. For explicit solvent molecules, this number can be rather large. Therefore, the volume occupied by the system of interest is divided in cubic boxes and all NCI fragments of small size within such a box are collected together in a cubic cluster approach. On the other hand, large NCI fragments are further divided in smaller chemically reasonable fragments upon the application of the Dijkstra algorithm.[247] This algorithm based on graph theory finds the shortest paths between two nodes in a graph. For each resulting chemical fragment, a fragmented Hessian matrix is set up and diagonalized separately. The problem of diagonalizing one large matrix is thus broken down to the diagonalization of a few much smaller matrices (divide-and-conquer type algorithm). To achieve high efficiency, this process is performed in parallel. With the fragmented Hessian scheme, the `xtb` optimizer, termed ANCopt, is capable of performing geometry optimization of large and complex macromolecular structures.

B.4 Example applications and benchmarking

B.4.1 Molecular Structures

The geometry optimization of molecular structures is a common application for SQM methods, specifically for large systems, where DFT and other first-principles methods become computationally infeasible. The GFN1- and GFN2-xTB methods already proved to be fast and robust semiempirical optimization tools for computing reasonably accurate molecular structures with elements up to radon.[8,17,41,248] To substantiate this statement, we summarize their performance for well-established and recently published benchmark sets including organic and main group molecules, non-covalent interactions, and transition-metal as well as organometallic and lanthanide complexes.

With respect to organic and main group molecules, we will first focus on the established ROT34[249,250] set of 12 small to medium-sized organic gas phase structures and only briefly mention the results for other benchmarks (for details, see Refs.[8,41]). The ROT34 set comprises 34 equilibrium rotational constants $\mathrm{B_e}$ derived from accurate spectroscopically measured rotational constants $\mathrm{B_0}$ which were back-corrected by calculated nuclear vibrational effects, such that they can be directly compared to (local) clamped-nuclei Born-Oppenheimer minimum structure values computed with the electronic structure methods, which are to be assessed. It is a sensitive test for the accuracy of calculated molecular structures since small changes in the geometry can already lead to significant deviations from the back-corrected experimental reference values. As long as conformational changes can be excluded, too large theoretical $\mathrm{B_e}$ values indicate shortened covalent bonds (or overall shrinked molecule size) w.r.t. the experimental reference.

The performance in terms of relative deviations for the GFN*n*-xTB methods is compared to the dispersion-corrected DFTB3-D3(BJ) and PM6-D3H4X[251,252] SQM methods as well as to the low-cost composite DFT method B97-3c and accurate B3LYP[253]-D3(BJ)/def2-QZVPP[254] structures (see Figure B.7a)). Well-performing DFT methods such as the two latter show mean unsigned relative deviations (MURDs) and standard relative deviation (SRDs) below 0.5%,[79,249] while SQM methods expectedly yield larger deviations. The ZDO-based PM6-D3H4X method yields SRDs and MURDs of $\leq$2.5% and, generally, less reliable geometries are predicted compared to DFTB3-D3(BJ), GFN1-xTB, and GFN2-xTB (one molecule (isoamyl-acetate) was excluded from the benchmark due to problematic conformational changes). The latter ranks second among the tested methods and is only slightly outperformed by GFN1-xTB. This could be due to the fact that the relative weight of geometries during the fitting procedure has been larger than for GFN2-xTB. The GFN1-xTB method yields an SRD of 1.1% and a mean relative deviation (MRD) of only 0.5% thus providing only slightly too small molecules on average, which is an excellent result for a SQM method. A comparably small SRD is obtained with DFTB3-D3(BJ) but at the expense of a systematic shift towards too long bonds bonds, which is common for DFTB as well as for (semi)local density functionals.[249]

GFN0-xTB and GFN-FF yield an SRD that is approximately halfway between PM6-D3H4X and DFTB3-D3(BJ). In contrast to GFN0-xTB, which shows a clear shift towards too small molecules, the MRD of GFN-FF is close to zero and, hence, the geometries calculated with this universally applicable force-field do not show a systematic error on average, though the scatter is somewhat larger compared with GFN1- and GFN2-xTB.

In addition, two benchmark sets were assessed which contain more unusual and challenging structures: the HMGB11[142] covering heavy main group molecules and the LB12[142] testing particularly long bonds. Here, PM6-D3H4X with a mean absolute deviation (MAD) of 10.1 pm for the HMGB11 set is clearly outperformed by GFN1-xTB and GFN2-xTB with MADs ≤ 3 pm. Particularly the latter predicts the long bond lengths of the LB12 fairly well with an MAD of 12.6 pm (excluding S_8^{2+} from the statistics), while PM6-D3H4X yields an MAD of 20.5 pm (excluding HAPPOD and KAMDOR from the statistics). The good performance of GFN1- and GFN2-xTB for these "unusual" cases may be attributed to their element-specific parameterization.

Furthermore, molecular structures with dominating non-covalent interactions were assessed comprehensively employing center-of-mass (CMA) distance deviations for the fully optimized S22[131,255] complexes and relative deviations w.r.t. extrapolated CMA distances for the S22x5[256], S66x8[130], and X40x10[257] benchmark sets. Even though all tested methods delivered comparably good results for the S22 with an MAD of ≈14 pm, GFN2-xTB represents a clear improvement especially for X40x10. The MRD for the latter approaches zero and the MURD is with 2.5% only about half as large as that of PM6-D3H4X. This also indicates that the (anisotropic) electrostatic terms are well balanced with repulsive and dispersion interactions and, therefore, particularly GFN2-xTB is well suited for the optimization of non-covalent complexes and supramolecular assemblies. In contrast to organic and main group molecules, the geometry optimization of the chemically important transition-metal and organometallic complexes with SQM methods is much less common. This is due to the fact that these structures often already pose a challenge for single reference QM methods and because there are hardly any SQM methods that are parameterized for the whole ($Z \leq 86$) periodic table. With the development of the GFN*n*-xTB methods this situation has improved significantly, as a promising alternative to the ZDO-based PM methods has been made available.

To substantiate and to quantify this statement we demonstrate the performance of GFN1- and GFN2-xTB for the well-established TMC32[143] 3d transition-metal structures benchmark set comparing 50 bond lengths, as well as for two recently published comprehensive molecular structure benchmark sets, the TMG145[17] including 145 transition-metal complexes and a set of 80 challenging lanthanide complexes.[248] The complexes in both sets are treated as low-spin systems.

For the TMC32 set, GFN1-xTB is the most accurate of all investigated SQM methods (MAD = 5.0 pm), closely followed by its successor GFN2-xTB (MAD = 5.7 pm). On average, GFN1-xTB yields more systematic deviations and slightly shorter bonds compared with PM6-D3H4X. Still, all transition-metal complex structures were reproduced by all three methods without dissociation or significant chemical reorganization. The TMG145 benchmark set was compiled from 145 challenging "real-life" transition-metal complexes (25-200 atoms) up to Hg (Z = 80) with high-quality hybrid DFT (TPSSh[258]-D3(BJ)-ATM/def2-TZVPP[254]) structures as reference. In total, 941 bond lengths and 2846 bond angles are compared.

GFN1- and GFN2-xTB both predict distinctive bond angles around the metal atom as well as metal-ligand bond lengths with good accuracy (bond angles/ bond lengths : MAD = 4.0 degrees/7.5 pm and 3.9 degrees/8.3 pm for GFN1- and GFN2-xTB, respectively; cf. Figure B.7b) for the individual deviations for the subset with DFT reference structures) with high efficiency (typically below one minute computation time compared with days to weeks for the reference calculations). This is also reflected in the Cartesian heavy-atom (all elements except H) root-mean-square deviation (hRMSD), which is

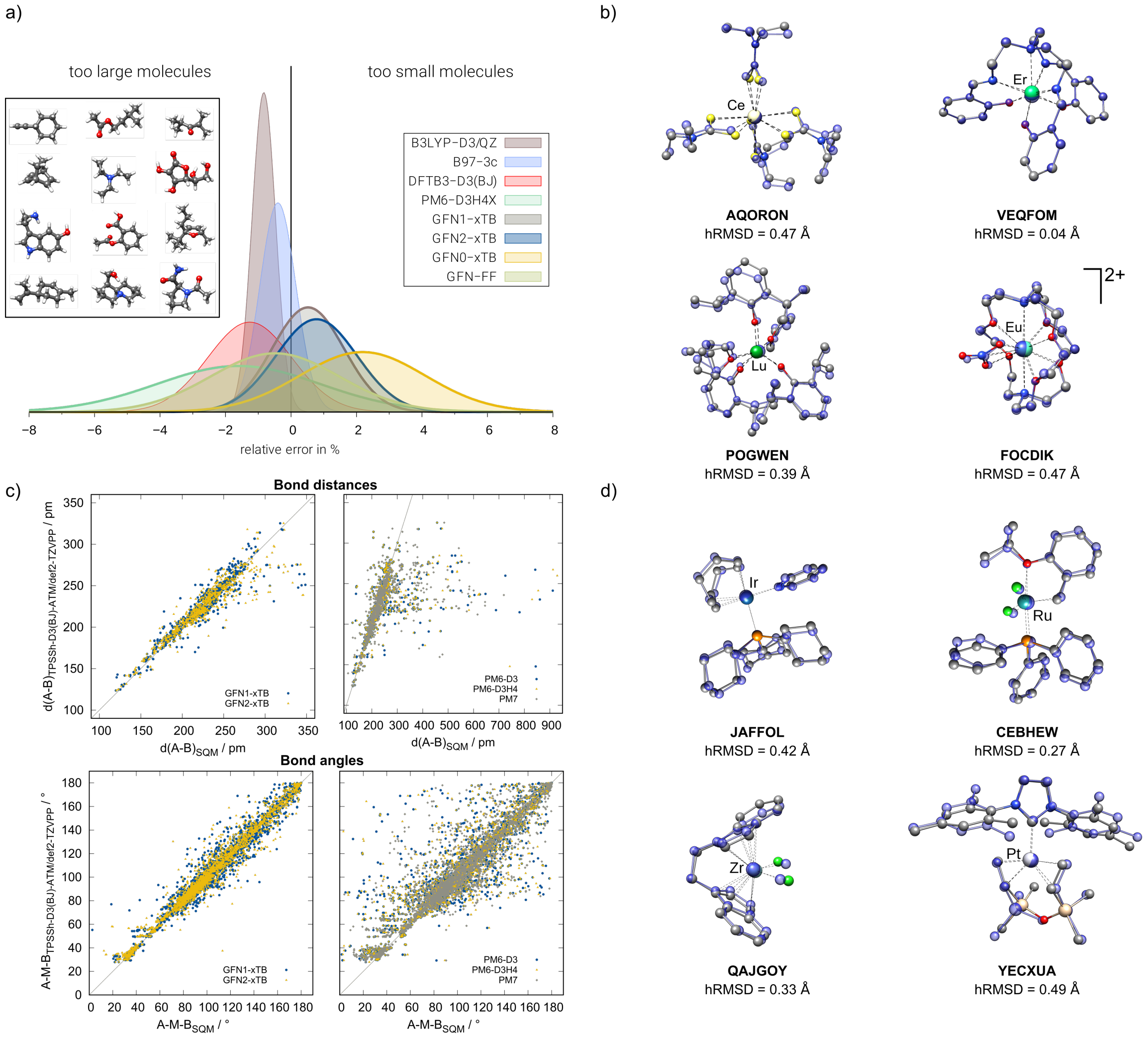


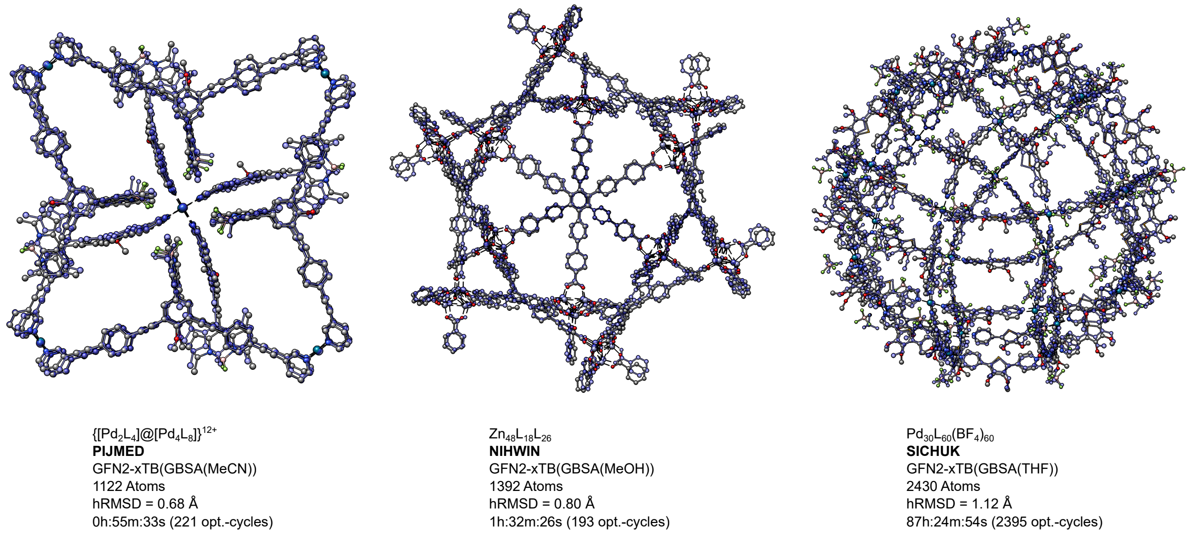
Figure B.7: In part a), normal distribution plots of deviations in calculated equilibrium rotational constants B_e for the ROT34 set[249,250] (see inset for the respective structures) are shown. In part b), structure overlays and corresponding heavy-atom RMSDs (hRMSD) in Å w.r.t. the crystal structure or DFT reference structures of four exemplary complexes from the lanthanide set[248] are depicted (GFN2-xTB-optimized structures are shown in transparent blue, carbon bound hydrogen atoms are omitted for better visibility). Part c) shows correlation plots for bond lengths and angles in pm and degree, respectively, calculated with GFN1- and GFN2-xTB for the subset of the TMG145 set.[17] In part d), structure overlays and hRMSD in Å w.r.t. DFT reference structures of four exemplary complexes from the TMG145 set[17] are shown. The respective Cambridge Structural Database (CSD) codes are placed below each structure: Crabtree catalyst (JAFFOL), Brintzinger-Kaminsky catalyst (QAJGOY), Grubbs-Hoveyda I catalyst (CEBHEW), and Karstedt's catalyst variant (YECXUA).

on average only 0.34 Å and 0.33 Å for GFN1- and GFN2-xTB, respectively. The tested ZDO-based methods PM6-D,[251] PM6-D3H4,[251,252] and PM7[259] perform clearly inferior with MADs of about 30 pm and 9 degrees for bond lengths and angles, respectively, as well as twice as large hRMSD on average, mainly due to about 40% of the complexes for which these methods predicted qualitatively incorrect molecular structures. In contrast, only ≈ 11% and ≈ 7% of the complexes could not be optimized to the chemically correct structure with GFN1- and GFN2-xTB, respectively, thus underlining their robustness even for such challenging systems. Four representative structure overlays together with the respective hRMSD are shown in Figure B.7d).

The applicability of the GFN*n*-xTB methods for all elements up to Z=86 also allows the optimization of molecular structures with rather unusual elements important for special applications such as luminescent lanthanide-based metal-organic frameworks.[260] Thus, the correct computation of such lanthanide-containing structures is of practical relevance. For GFN1-xTB, this was assessed on a challenging and comprehensive test set of 80 lanthanide structures. Except for three promethium complexes, for which accurate PBE0-D3(BJ)/ZORA-def2-TZVP[261] (SARC2-ZORA-QZV[262] basis for Pr) reference structures were calculated, the optimized geometries are benchmarked w.r.t. high-quality X-ray structures. For more than half of the complexes, the structures optimized with GFN1-xTB yield a hRMSD < 0.6 Å. Only for a few larger multi-nuclear lanthanide clusters bridged by anionic ligands, it was more difficult to reach the default convergence criteria.

However, this would be also the expected behavior of higher-level QM methods for such challenging systems. Considering the uncertainty due to the neglect of crystal field and crystal packing effects, this means that most reference structures could be reproduced qualitatively correct and 44 out of 80 structures even show a good quantitative agreement with the reference structures. Compared to the Sparkle/PM6 method, which is the only semiempirical competitor for lanthanide chemistry, the mean hRMSD of 0.65 Å obtained with GFN1-xTB for all complexes is significantly smaller, even slightly outperforming the low-cost composite QM method HF-3c (0.68 Å). This is a very good result keeping in mind that the lanthanide atoms are treated in an "f-in-core" approximation by the GFN1-xTB method (see "Treatment of Lanthanoide Elements" part of the theory section). GFN1-xTB also clearly outperforms Sparkle/PM6 in terms of computational wall times mainly due to the fact that significantly fewer SCF and optimization cycles are required on average.

Together with the promising results for the TMG145 and TMC32 benchmark sets, GFN1-xTB and GFN2-xTB proved to describe organometallic bonding motifs significantly better compared with the ZDO-based PM methods, which are currently the only other SQM methods applicable to such systems. This is a particularly remarkable result since no specific modifications for the treatment of organometallic complexes were introduced in the development of the GFN*n*-xTB methods and they are also clearly superior in terms of computational speed. Moreover, the finite electronic temperature (Fermi smearing) along with the Hamiltonian, which is devoid of exchange/spin density terms, also enables the robust optimization of molecules with small HOMO-LUMO gaps or open shells, which are challenging for many single reference QM and most other SQM methods. The excellent accuracy/computational cost ratio of the GFN1- and GFN2-xTB methods together with their robustness enables the fast and reliable geometry optimization of typical transition-metal complexes as well as the routine application, e.g., for semi-automated conformational search procedures for larger organometallic complexes and the optimization of very large systems with several thousands of atoms such as extended metal-organic polyhedra at a SQM level of theory (see Figure B.8 for three representative examples taken from Ref. [17]). GFN2-xTB fully optimizes the structures of medium-sized polyhedra (≈ 1100–1400 atoms) in about one hour (≈ 200 optimization cycles). The relatively small hRMSD (0.68 Å and 0.80 Å respectively) for such large systems show impressively how robust, efficient, and accurate such optimizations with GFN2-xTB in combination with a GBSA solvation model are. Even the $Pd_{30}L_{60}(BF_4)_{60}$ polyhedron with almost 2500 atoms could be optimized in a few days yielding a qualitatively correct molecular structure.



$\{[Pd_2L_4]@[Pd_4L_8]\}^{12+}$
PIJMED
GFN2-xTB(GBSA(MeCN))
1122 Atoms
hRMSD = 0.68 Å
0h:55m:33s (221 opt.-cycles)

$Zn_{48}L_{18}L_{26}$
NIHWIN
GFN2-xTB(GBSA(MeOH))
1392 Atoms
hRMSD = 0.80 Å
1h:32m:26s (193 opt.-cycles)

$Pd_{30}L_{60}(BF_4)_{60}$
SICHUK
GFN2-xTB(GBSA(THF))
2430 Atoms
hRMSD = 1.12 Å
87h:24m:54s (2395 opt.-cycles)

Figure B.8: Structure overlays of the GFN2-xTB optimized (transparent blue; the GBSA solvation model was applied) and X-ray reference structures (color code; the respective CSD code is given) for three metal-organic polyhedra (carbon bound hydrogen atoms are omitted for better visibility). The heavy-atom RMSDs (hRMSD) are given and the timings were obtained with *normalopt* settings on 14 CPU Intel® Xeon® E5-2660 v4 2.00 GHz CPU cores.

In summary, it can be stated that the general applicability for almost all elements of the periodic table, the efficient and robust treatment even of very large and complicated electronic structures, a reliable description of the major part of the PES, especially for the non-covalent interactions, as well as the coupled implicit GBSA solvent model are strong points of the methods. Hence, GFN1- and GFN2-xTB methods are perfectly suited for optimizing molecular structures for various chemical applications. Since only comparatively low computational resources are required for this purpose, these methods provide an unprecedented quantum-mechanical tool which can be of great benefit for chemical research, e.g., for the calculation of large organometallic structures. GFN0-xTB performs slightly worse, but may be of great interest in a revised form for optimizations under periodic boundary conditions. First assessments of the new universal force field GFN-FF yielded promising results[81] for geometry optimizations of proteins. This is remarkable, since GFN-FF was not specifically adjusted to proteins. In Figure B.9, the performance of GFN-FF along with the universal (UFF[82]) and highly specialized FFs (OPLS2005[263], AMBER*[264,265]), respectively, is assessed for geometry optimizations of 70 protein structures.[266] For the dihedral angles ϕ, ψ, and χ, which act as soft descriptors for local displacements in the respective protein backbone, GFN-FF provides similar or even better accuracy than the specialized FFs OPLS2005 and AMBER*, respectively. Only for ω, significant deviations in the larger protein structures were observed with GFN-FF, indicating that the barrier for rotation around the respective C-N bonds was underestimated. For both the hRMSD and the C_α RMSD, GFN-FF provides comparably accurate

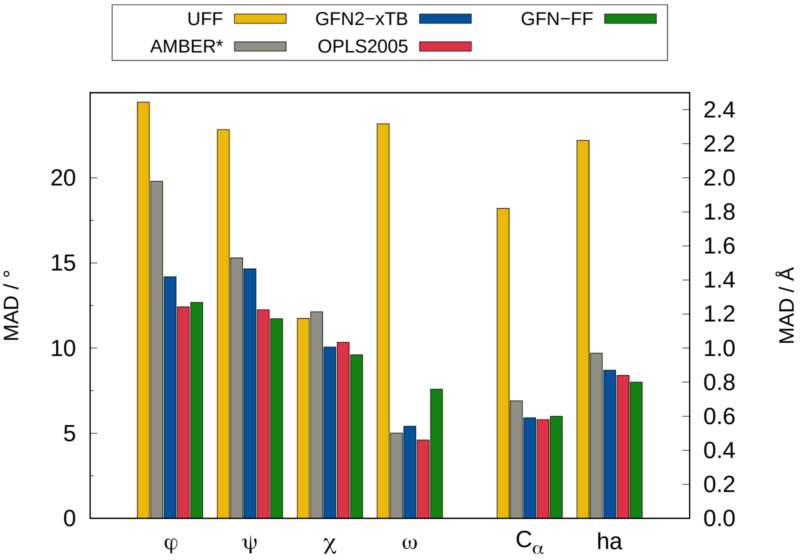


Figure B.9: Performance of GFN-FF in comparison to the universal (UFF) and highly specialized FFs (OPLS2005 and AMBER*) for a set of 70 protein structures[266]. Average hRMSD, C_α RSMD, and average deviations of four distinctive dihedral angles w.r.t. the corresponding crystal structure given in Å and degree, respectively.

results to the computationally much more demanding SQM method GFN2-xTB. UFF, on the other hand, performes clearly worse than GFN-FF for all statistical measures and is therefore, in contrast to the latter, not recommended for protein structure optimizations. Furthermore, the parameterization up to Z = 86 allows the optimization of metalloprotein structures with GFN-FF.

B.4.2 Thermochemistry and Kinetics

The GFNn-xTB family of methods does not reach the aforementioned accuracy level for the calculation of (covalent) thermochemical and kinetic properties. However, this is typical for SQM methods in general,[267,268] which are therefore not generally recommended for accurate routine calculations of, e.g., covalent reaction energies and barrier heights. However, particularly for the latter, GFN2-xTB performs surprisingly well for a SQM method (*vide infra*, cf. Figure B.10d)). Nonetheless, the semiempirical approximations on which the GFNn-xTB methods are based have the clearest implications for covalent bonds and hence, they (and other SQM methods) should only be used for qualitative estimates of covalent thermochemistry. However, there are many applications where the GFNn-xTB methods, due to their high efficiency combined with the general parameterization and high robustness, prove to be very useful, e.g., as a starting point for multi-level, large-scale screening applications (see subsection "Chemical Space Exploration") or the semi-automatic localization of transition states presented in the next subsection. Moreover, the performance for non-covalent interactions is very good, as will be summarized in the following for eight intermolecular non-covalent interaction energies benchmark sets from the comprehensive GMTKN55[23] database. They are composed of various small molecules including also heavier elements and more unusual interaction motifs. Specifically, the interaction energies in rare gas complexes (RG18), n-alkane (ADIM6) and various other non-covalently bound dimers (S22 and S66) as well as between heavy element hydrides (HEAVY28) were assessed. Furthermore, hydrogen-bonded complexes between H_2O, NH_3, or HCl and carbene analogues (CARBHB12), pnicogen-containing dimers (PNICO23), and halogenated dimers including also halogen bonds (HAL59) were tested (for original references to the used subsets of the GMTKN55 database, see Ref. [23]). The respective MADs obtained for the GFNn-xTB methods, GFN-FF, and PM7 with respect to the very accurate coupled cluster reference values are shown in Figure B.10a) together with average MADs over all tested sets. Values for PM6-D3H4X as well as low cost composite and large basis set DFT methods are also shown.

Except for the ADIM set, for which GFN2-xTB is the worst of all tested methods and the CARBHB12, for which GFN1-xTB slightly outperforms GFN2-xTB, the latter is the most accurate method in all other considered benchmark sets. This is also reflected in its remarkably small average MAD of only 0.9 kcal/mol, which is slightly better than that of the composite methods HF-3c and B97-3c and even comparable to some large basis dispersion corrected GGAs. This is especially true for the rather special HEAVY28 and RG18 test sets, where GFN2-xTB achieves excellent results (MAD = 0.6 kcal/mol and 0.2 kcal/mol, respectively), most likely due to a more advanced treatment of dispersion interactions within the D4 scheme compared to D3(BJ). In contrast, PM7 shows much larger MADs for both sets (HEAVY28: 2.9 kcal/mol, RG18: 0.6 kcal/mol).

The benefit of the AES (see subsection “The GFN2-xTB Hamiltonian”) is already visible for the commonly applied S22 and S66 test sets, which are also often used for fitting purposes. For example, the MAD for S66 obtained with GFN2-xTB (0.7 kcal/mol)) is 0.6 kcal/mol lower than that of the monopole based GFN1-xTB method, which illustrates the importance of AES for such systems. Moreover, GFN2-xTB does not show any larger deviations for the hydrogen bonded systems, although in contrast to GFN1-xTB, no special terms for hydrogen bonds are included. PM7 performs comparably well as GFN2-xTB for the S22 and S66 sets and is with an MAD of only 0.2 kcal/mol clearly the most accurate SQM method for the ADIM6 test set. However, for CARBHB12 and especially for the PNICO23 and HAL59 sets, the errors for PM7 are much larger (up to five times higher MAD than GFN2-xTB) and significant outliers occurred. The good performance of the GFN2-xTB method for the PNICO23 and HAL59 sets can be attributed to the inclusion of higher multipole electrostatic terms, since in these molecules a good description of the anisotropic electron density of the bound pnicogen and halogen atoms is crucial for the accuracy of the respective interaction energies. The GFN1-xTB method, despite its monopole approximation, gives comparably good results for HAL59 and still reasonably accurate interaction energies for the PNICO23 systems which can be explained by its well balanced parameterization and special halogen bond correction.

GFN0-xTB and GFN-FF perform similar to GFN1-xTB (average MAD = 1.1 kcal/mol) with slightly larger average MADs of 1.4 kcal/mol and 1.1 kcal/mol, respectively. Only for the CARBHB12 test set, significantly larger deviations were observed for both methods. Considering its very low computational cost, the performance of GFN-FF for intermolecular NCIs is outstanding.

The excellent performance of GFN2-xTB, in comparison to other SQM methods, for non-covalent interactions of smaller molecules is also observed for larger and more difficult supramolecular complexes, as could be shown with the S30L[96] benchmark set. It tests 30 association energies of non-covalently bound neutral and charged complexes with up to 200 atoms for which accurate DLPNO-CCSD(T)[269,270]/CBS*[271] reference values are available.[79] With an MAD of only 4.0 kcal/mol, GFN2-xTB can compete with some large basis set dispersion-corrected DFT methods, which is an outstanding result for an SQM method for such a challenging benchmark. Especially the very small errors for the charged systems are striking, whereas PM6-D3H4X, among others, shows up to 20% deviation from the corresponding reference association energies. This again confirms that electrostatic and polarization interactions are described much more accurately in GFN2-xTB than with other SQM methods. Larger deviations for GFN2-xTB were only found for complexes **7-12** with conjugated π systems and dominant van der Waals interactions, for which the respective association energies were overestimated, probably due to non-additive dispersion interactions that are not yet described accurately enough with the ATM term (see subsection "The GFN2-xTB Hamiltonian"). The universal force-field GFN-FF yield similarly accurate results (MAD = 4.1 kcal/mol) as GFN2-xTB and even deviates less from the reference values for van der Waals interaction-dominated conjugated π systems, although only a simplified version of the D4 scheme is applied[9], thus, indicating a good and well balanced parameter fit. For the charged systems, however, the errors are, as expected, somewhat larger compared to GFN2-xTB. GFN1-xTB yields a MAD (6.1 kcal/mol) between that of PM6-D3H4X (5.1 kcal/mol) and DFTB3-D3(BJ) (6.9 kcal/mol) and hence does not reach the accuracy of GFN2-xTB and GFN-FF.

Overall it can be stated that the GFN methods, especially GFN2-xTB and GFN-FF, are well-suited to investigate non-covalently bound systems and, due to their very low computational effort compared to dispersion-corrected DFT methods with large basis sets, reliable calculation of such systems with several thousand of atoms becomes feasible.

Next, we discuss the performance of the GFNn-xTB and other SQM methods for eight conformational energy benchmark sets taken from the GMTKN55 database, specifically relative energies of alkane (ACONF), amino acid (Amino20x4), butane-1,4-diol (BUT14DIOL), inorganic (ICONF), melatonin (MCONF), tri- and tetrapeptide (PCONF21) and sugar conformers (SCONF) as well as energy differences between RNA-backbone conformers (UPU23[271]) and one additional test set (MALT205[272]) comprising 205 conformers of α-maltose (for original references to the used subsets of the GMTKN55 database, see Ref. [23]). This is a challenging test for SQM methods since the accurate description of these small relative energies requires a well balanced accuracy for both intramolecular non-covalent and covalent interactions. Figure B.10b) shows the MADs for the tested methods and the average MADs over all nine sets in comparison with low-cost composite QM methods and larger basis set DFT results. Among the SQM methods tested, GFN2-xTB is on average the most accurate, closely followed by GFN1-xTB and GFN-FF. Especially the significant improvement compared to GFN1-xTB for more polar and hydrogen bonded systems, with the exception of the UPU23 set, led to an excellent performance of GFN2-xTB for the Amino20x4 (MAD = 0.9 kcal/mol), PCONF21 (MAD = 1.8 kcal/mol), and SCONF (MAD = 1.6 kcal/mol) test sets. For the BUT14DIOL and MCONF test sets, similarly good results for all assessed SQM methods were observed, with GFN-FF performing best for the latter (MAD = 0.5 kcal/mol). For the MALT205 set, which is challenging due to the many involved intramolecular hydrogen bonded, GFN0-xTB and PM6-D3H4X (MAD = 5.2 kcal/mol and 5.1 kcal/mol, respectively) showed significantly larger deviations than GFN1-xTB, GFN2-xTB, and GFN-FF with the latter yielding the lowest MAD of 2.8 kcal/mol. Except for the PCONF and UPU23 benchmark sets, the mean deviations (MD) obtained with GFN1- and GFN2-xTB are negative for all tested conformer energy sets. In general, all tested SQM methods and also GFN-FF tend to underestimate the high-level coupled-cluster conformational energies that serve as a reference values, especially for the higher-energy conformers. Furthermore, the MADs of the SQM methods are on average 3–4 times larger than those of DFT methods.

Nevertheless, the investigated conformational energy benchmarks clearly show that GFN2-xTB provides more reliable conformational energies than PM6-D3H4X, PM7, and even the computationally more expensive HF-3c method. For polar and hydrogen bonded systems, GFN2-xTB also outperforms GFN0-xTB, GFN1-xTB, and GFN-FF, potentially due to the inclusion of the AES term (see subsection "The GFN2-xTB Hamiltonian"). The universal force-field GFN-FF performs on average similar to GFN0-xTB and GFN1-xTB, which is particularly remarkable and offers a valuable alternative to

GFN2-xTB for the conformer search of large molecules with several hundreds of atoms. The proper energy ranking of conformers is an important application field for SQM methods (see subsection "Chemical Space Exploration") and the results for the conformational energy benchmarks suggest that particularly GFN2-xTB and GFN-FF should be well-suited for this purpose.

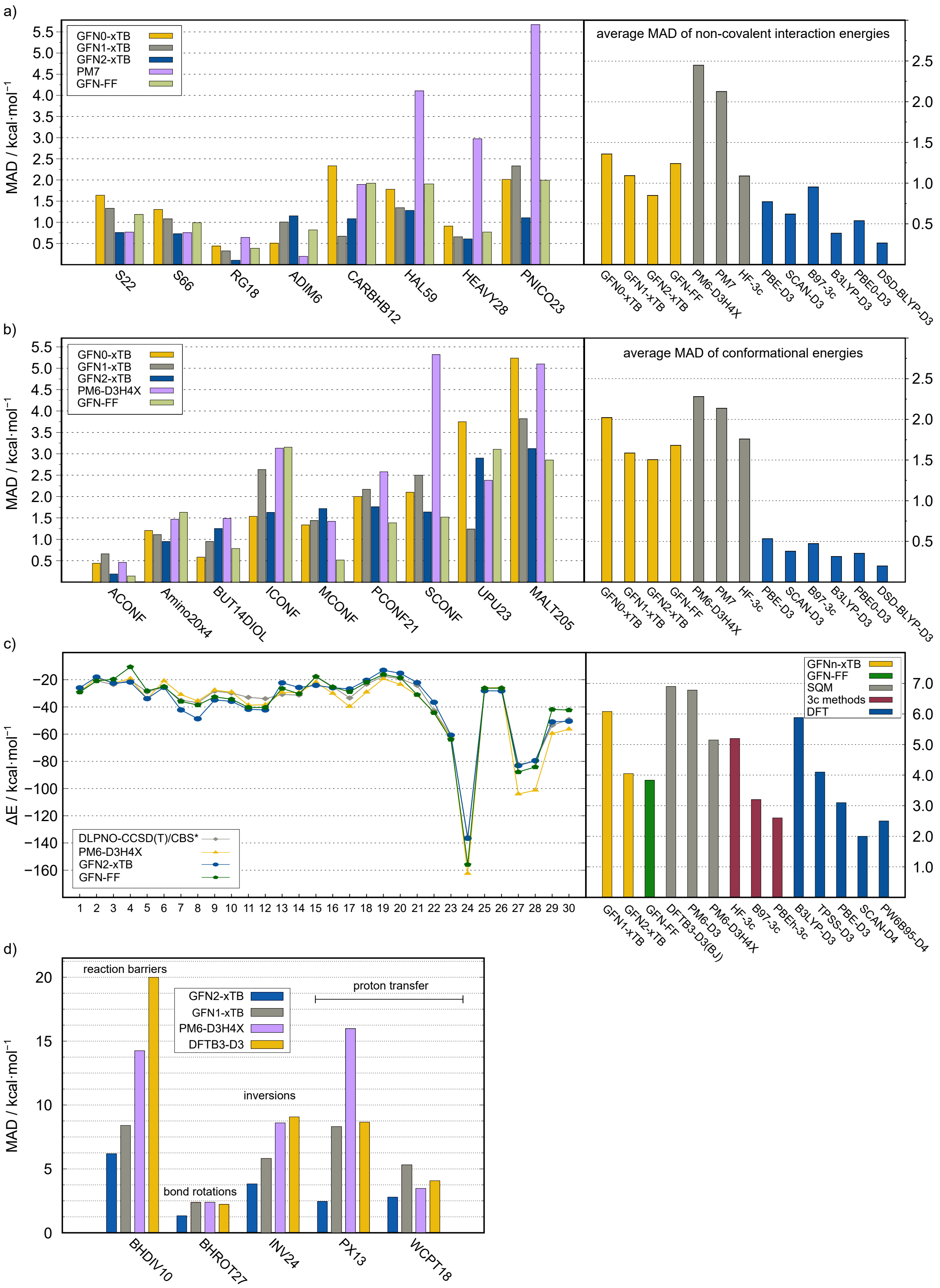


Figure B.10: (Caption next page.)

Figure B.10: (Previous page.) Mean absolute deviations (MADs) in kcal/mol of GFNn-xTB methods, GFN-FF, PM6-D3H4, and PM7 for various benchmarks sets comprising non-covalent interaction energies (part a) and conformational energies (part b). The respective inset shows the average MAD over all tested sets compared to the low-cost composite QM methods HF-3c and B97-3c as well as a few large basis DFT results taken from Ref. [23]. Part c) shows the association energies for 30 large supramolecular complexes (S30L[96] test set) computed with GFN2-xTB, GFN-FF, and PM6-D3H4 together with the respective DLPNO-CCSD(T)[269,270]/CBS*[271] reference values. The inset depicts the MADs for the GFNn-xTB methods and GFN-FF as well as further SQM, low-cost composite, and large basis set DFT methods. Part d): MADs in kcal/mol of GFN1-xTB, GFN2-xTB, PM6-D3H4, and DFTB3-D3(BJ) for five reaction barrier height test sets. Except for the S30L and MALT205[272] sets, the geometries and reference values are taken from the GMTKN55[23] database.

Finally, we turn to five of the barrier height oriented benchmark sets of the GMTKN55 database (for original references to the used subsets, see Ref. [23]). The MADs of the assessed methods (GFN1- and GFN2-xTB, PM6-D3H4, and DFTB3-D3(BJ) for the five test sets are shown in Figure B.10d)). Among all methods considered, GFN2-xTB clearly performs best with the lowest MAD for the diverse reaction barriers set (BHDIV10: MAD = 8.1 kcal/mol), the barrier heights for rotations around single bonds (BHROT27: MAD = 1.2 kcal/mol), and inversions (INV24: MAD = 3.5 kcal/mol) as well as for barriers in proton transfer reactions. (PX13: MAD = 2.7 kcal/mol). Only for barriers of tautomerization reactions (WCPT18), GFN2-xTB (MAD = 3.8 kcal/mol) is slightly outperformed by PM6-D3H4 (MAD = 3.5 kcal/mol). This performance of GFN2-xTB is remarkable for an SQM method, since no barrier heights were included in the fitting procedure.

Especially for PX13, for which PM6-D3H4 gives quite large deviations (MAD = 16.0 kcal/mol), and for BHDIV10, for which DFTB3-D3(BJ) performs relatively poorly (MAD = 13.3 kcal/mol), GFN2-xTB even yields slightly smaller MADs than PBE0 with a large basis set, which is an outstanding result for SQM methods. GFN1-xTB, although slightly worse than GFN2-xTB in all considered benchmark sets, still predicts reasonably accurate barrier heights.

Overall, the GFNn-xTB methods, particularly GFN2-xTB, yield acceptable to good (for non-covalent interactions and barrier heights) accuracy for thermochemistry applications, even though this was not the main focus (except for non-covalent systems) in the development of the GFNn-xTB methods.

B.4.3 Reaction Mechanism Exploration

Reliable SQM methods also offer new perspectives for the fast screening of reaction mechanisms and transition states (TS) for transition-metal and organometallic systems, if they are interfaced with state-of-the-art transition state localization algorithm such as, e.g., the double-ended growing string algorithm[236,237] (see subsection “Implementation”). However, this necessitates that the corresponding SQM method yields not only reliable transition state geometries but also a sufficiently accurate thermochemistry for such systems. In a recent study,[238] this was assessed employing two organometallic reaction energy benchmark sets, the MOR41[148] and the WCCR10.[273] Although the development of the GFNn-xTB methods was not specifically targeted to predict accurate reaction energies, GFN1-xTB (MOR41: MAD = 13.2 kcal/mol, WCCR10: MAD = 10.9 kcal/mol) and GFN2-xTB (MOR41: MAD = 11.8 kcal, WCCR10: MAD = 10.7 kcal/mol) achieve reasonably accurate relative energies for these challenging reactions (with GFN2-xTB performing slightly better), comparable to dispersion uncorrected DFT methods and not so much worse than, e.g., M06-2X[64]/def2-QZVPP with an MAD = 7.3 kcal/mol for the MOR41 reactions. The PM*x* methods are the only other SQM approach that can currently be employed for the calculation of such systems, but they yield substantially larger deviations (PM6-D3H4: MAD = 21.7 kcal/mol, PM7: MAD = 22.7 kcal/mol).

The performance of the GFNn-xTB methods for barrier heights of organometallic reactions was assessed with a similarly encouraging result for a slightly modified version[238] of the MOBH35[274,275] benchmark set comprising 29 backward and forward barriers. The GFN1- and GFN2-xTB methods in combination with `mGSM` localized the correct transition states in 89.7% and 86.2% of all investigated reactions, respectively and predicted the barrier heights with reasonable accuracy (GFN1-xTB: MAD = 8.8 kcal/mol, GFN2-xTB: MAD = 8.2 kcal/mol). The PM*x* methods are significantly less robust (only 72.4% and 69.0% of the transition states were correctly localized with PM6-D3H4 and PM7, respectively) and they are also clearly outperformed by the GFNn-xTB methods in terms of accuracy for the barrier heights (PM6-D3H4: MAD = 17.1 kcal/mol, PM7: MAD = 19.6 kcal/mol), while their computational costs are more than twice as high.

In addition, the GFNn-xTB methods provide reasonably accurate TS geometries[17] (cf. also subsection "Molecular Structure") thus allowing for an efficient reoptimization on a higher level of theory. The computationally involved initial reaction path generation could be significantly accelerated with reliable SQM methods until a certain residual gradient is reached. Further computational savings can be achieved by avoiding superfluous reaction path searches at the significantly more expensive DFT level, because the GFNn-xTB methods allow a reliable and efficient chemical plausibility check of the possible paths.

Averaged over all 29 reactions, the tested GFNn-xTB methods need about five minutes to obtain a converged reaction path compared to several hours at low-cost DFT level (TPSS[146]-D3(BJ)/SVP[276]), which clearly demonstrates the great benefit of this workflow for investigating organometallic reactions. An example reaction is shown in Figure B.11. Although the GFN2-xTB reaction barrier of the rate determining TS is significantly underestimated compared to TPSS-D3(BJ) and the coupled-cluster reference values, the predicted reaction energy is clearly more accurate than that calculated with the significantly more computationally expensive DFT method. Furthermore, if a rough estimate of an upper limit for the respective barrier height is sufficient, the RMSD-push-pull path optimizer, which is described in Ref.[277] and implemented in the `xtb` code, provides an even faster alternative for this purpose. In summary, this workflow opens up new perspectives for the fast and sufficiently accurate theoretical investigation of challenging organometallic reaction mechanisms. Only minimal user input is required which offers a wide range of possible applications, e.g., in hybrid multi-level schemes aiming at the fully automated exploration of the reaction space.

B.4.4 Chemical Space Exploration

The exploration of the chemical space is an important task in computational chemistry. Macroscopic properties of physical observables (e.g., pKa values, NMR, CD, or IR spectra) can be well approximated as a thermostatistical average of the respective properties of low-energy chemical species under thermal equilibrium conditions. This implies that the underlying compound space (i.e., the three-dimensional structures of the molecules) is known, leading to a sampling problem in a space of huge dimensionality. A good balance between computational cost and accuracy is thus required.

The prerequisite for sampling is a continuous, well-behaved PES which has to be explored by the underlying computational method involving typically several thousand to millions of energy and gradient evaluations. Therefore, carrying out this initial exploration step with QM methods is only possible for very small molecules at a DFT or WFT level and other methods have to be chosen for larger

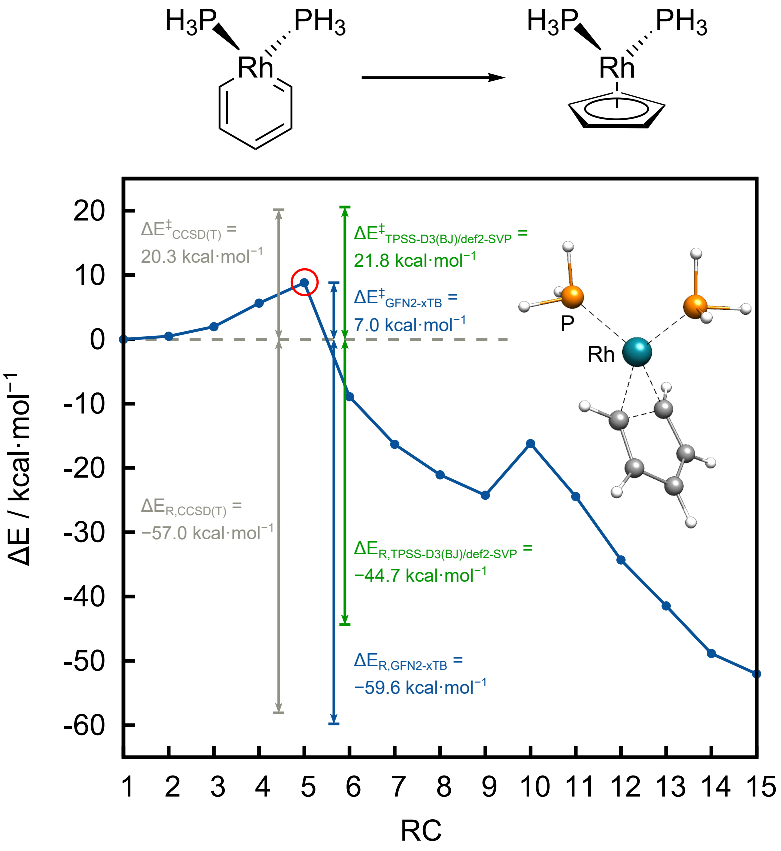


Figure B.11: Exemplary transition state (TS) localization with GFN2-xTB driven `mGSM` for reaction **15** (the Lewis structures are given) of the MOBH35[274,275] test set. The `mGSM` reaction path along the reaction coordinate (RC) is depicted in blue whereas the relevant TS, whose structure is shown as inset, is marked with a red circle. The respective reaction energies ΔE_R and reaction barriers $\Delta E^{\ddagger}$ predicted by GFN2-xTB (blue) and TPSS-D3(BJ)/SVP (green) as well as the corresponding coupled-cluster reference values (grey) are also shown.

molecules. As shown in the previous sections, the methods of the GFN*n*-xTB family provide an excellent cost-to-accuracy ratio and should enable extensive sampling procedures. In fact, the exploration of the low-energy chemical space is one of the main application fields for the GFN*n*-xTB level of theory.

In the following, we discuss two practically important examples: finding molecular conformations and preferred protonation sites. These and several other screening algorithms have been implemented in a standalone program called `crest`,[19] that acts as a driver for the `xtb` program.

B.4.5 Conformations

Molecular conformations are the primary example for the low-energy chemical space. Conformations are defined by minima on the PES with the same covalent bonding topology, obtained by rotation around bonds or inversion-type processes. Because of the huge number of possible conformations already for medium-sized molecules, finding those minima is a challenge for any exploration algorithm.

Furthermore, the description of the PES, i.e., basically the conformational energies, has to be qualitatively correct. From recent benchmark studies it is known[8,41] that this is difficult to achieve and that one has to apply a relatively large error margin (energy window) for the SQM pre-selected structures such that important conformations are not sorted out by mistake. This further increases the space of the considered structures. In recent publications,[19,277] GFN*n*-xTB was successfully combined with Cartesian RMSD-based meta-dynamics (MTD) simulations into an efficient, automated workflow for the task of finding molecular conformations. Note that the generation of conformers by rotation around all dihedral angles becomes impractical already for small systems and furthermore requires the *a priori* definition of conformational coordinates, which is avoided entirely by the general MTD sampling procedure.

As an example the well known pain-relieving medication drug 2-(4-isobutylphenyl)propanoic acid, also known as ibuprofen, is shown in Figure B.12.

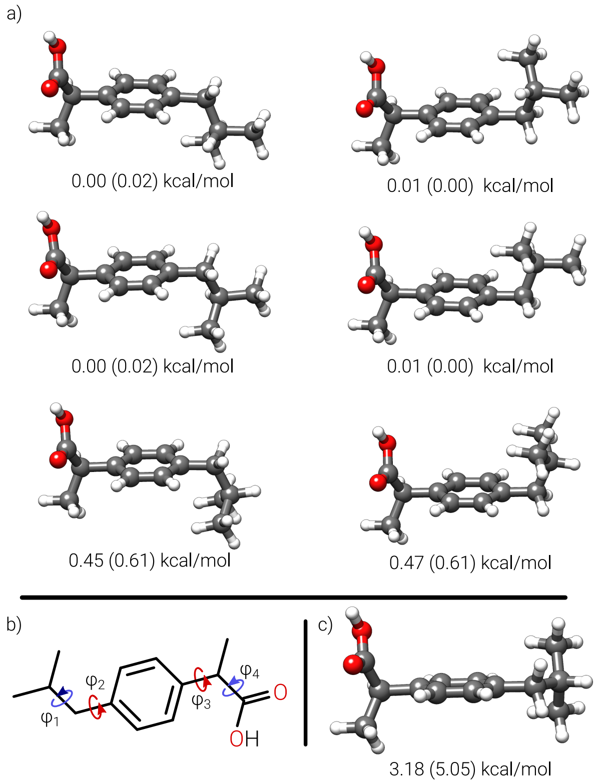


Figure B.12: The ibuprofen molecule: a) most populated conformers of ibuprofen in the gas phase obtained at the GFN2-xTB level, b) Lewis structure of the molecule including highlighted dihedral angles, c) highest lying conformer of ibuprofen in the gas phase obtained at the GFN2-xTB level. Relative energies below the structures correspond to the GFN2-xTB level, energies in parenthesis refer to PBE0-D3/def2-TZVPP results obtained for the fully optimized DFT structures.

Here, all relevant conformations are generated with the above mentioned MTD approach, but without the need to define and rotate any of the four dihedral angles shown in Figure B.12b) explicitly. From chirped-pulse Fourier transform microwave (CP-FTMW) spectroscopy,[278] four ibuprofen conformations are observable in the gas-phase. These four conformers coincide with the four lowest conformers found at the GFN2-xTB, as well as at the DFT (PBE0-D3(BJ)/def2-TZVPP) level (cf. Figure B.12a)). Furthermore, for the low-lying (i.e., populated) conformers of ibuprofen, the GFN2-xTB and DFT PES seem to be almost perfectly parallel as there is only a minor change ($\ll$1 kcal/mol) between the relative energies of the structures at the two levels. Only for higher lying conformations (Figure B.12c)), the difference in relative energy becomes more pronounced. It is a typical observation that conformational energies for organic molecules at the GFN*n*-xTB level are underestimated (see subsection "Thermochemistry and Kinetics"), i.e., the PES appears to be too flat.

As a larger much more complicated example, the protonated polypeptide Ac-Ala$_{19}$-Lys-H$^+$ is taken from Ref. [19]. The gas-phase conformations of this peptide were previously studied in a combined theoretical and experimental effort.[279] In Ref. [19] it was shown that the conformational screening at GFN2-xTB level is able to correctly predict important conformational features, such as the change from an α-helical structure into a folded 3_{10}-helical structure depicted in Figure B.13a) by protonation.

Furthermore, by re-ranking the GFN2-xTB ensemble at the PBE0-D4/def2-TZVPD//PBEh-3c level of theory even slightly better conformations than the previously known ones could be found. By visualization of the conformational energies at the GFN2-xTB and DFT level (see Figure B.13b)), the aforementioned issue of too flat SQM method PES becomes clear. Noticeably, the narrowly spaced conformational energy levels of GFN2-xTB are spread out over a much larger energy window at the DFT level. However, low-energy conformers on the GFN2-xTB PES are often also among the low-energy conformers at higher theoretical levels and vice versa. This relation is essential for a practically useful muli-level approach.

A great advantage of the GFN*n*-xTB methods in combination with automated screening procedures is their computational robustness and parameterization for almost all elements of the periodic table. This makes it possible to investigate also systems that would normally pose problems for either the applied theoretical method or the screening algorithm. As an example, Figure B.14 shows a macrocyclic molecule containing a Pd^{2+}-ion taken from Ref. [280].

Finding conformations for this metallomacrocyclic molecule is highly challenging for three reasons: 1) the chemical composition (i.e., the metal ion), 2) interdependent dihedral angles in the macrocyclic part of the molecule, and 3) the rigid NNN pincer-backbone of the molecule. While point 1) will most likely prevent the usage of, e.g., standard FFs for the conformational screening, the combination of 2) and 3) will make it nearly impossible to obtain conformations from simple conformer generators based on dihedral angles rotation. In contrast, the automated MTD based screening at GFN2-xTB[GBSA(acetonitrile)] level is able to provide a global minimum conformer closely resembling the crystal structure (cf. Figure B.14b)), as well as a rather diverse conformational ensemble (cf. Figure B.14c)) for the complex.

In summary, automatized conformational screening at the GFN*n*-xTB level can be applied over a broad range of chemical systems. The main advantage lies in the robustness and overall well performance of the methods, but also in the computational speed. The computation of several thousand of energies is impractical at DFT or WFT with the current technological limitations, but becomes feasible at the SQM level. The implementation in the `xtb` program also allows for some special applications, such as constrained conformational sampling, which extends the capabilities of this screening procedures even further, e.g., to adsorption problems or transition states. For a more detailed discussion and more examples see Ref. [19].

B.4.6 Protonation Sites

From recently published studies it is known that TB methods reasonably accurately describe relative proton affinities (PA),[36,41,187] e.g., for the PArel subset of the GMTKN55 data base. The description

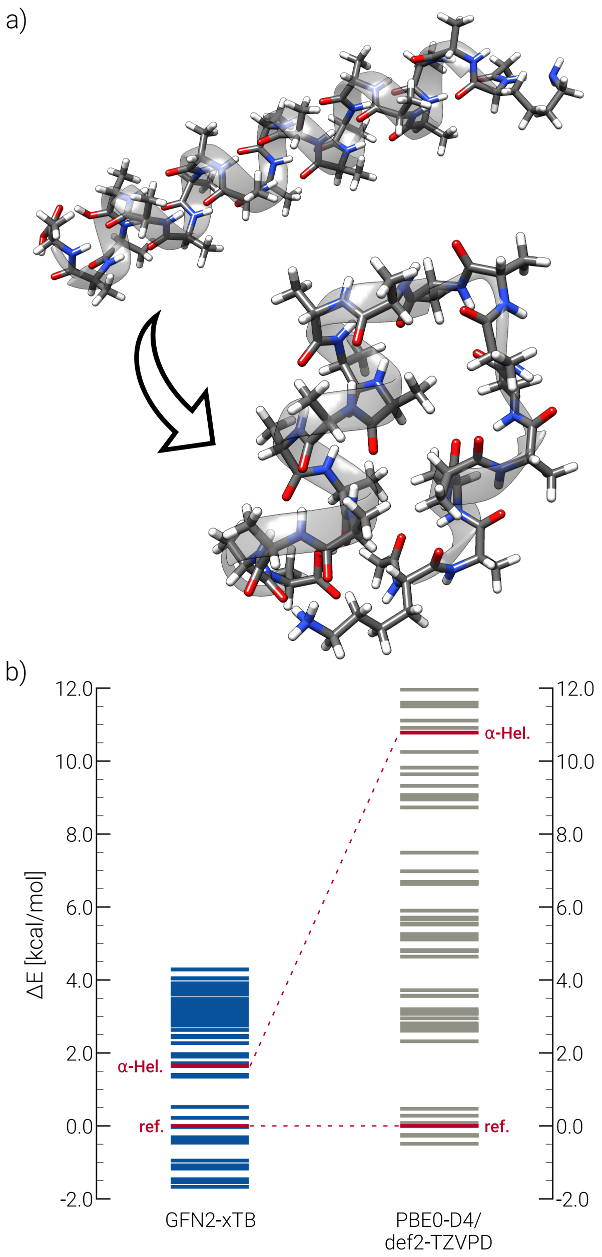


Figure B.13: The Ac-Ala$_{19}$-Lys-H^{+} molecule. a) Most stable α-helical and folded conformations of Ac-Ala$_{19}$-Lys-H^{+} in the gas phase at GFN2-xTB level. As indicated by the arrow, both forms can interchange depending on the protonation site (Ac-H^{+} vs. Lys-H^{+}). b) Comparison of corresponding conformational energies at the GFN2-xTB and PBE0-D4/def2-TZVPD//PBEh-3c levels in the gas phase.

of relative proton affinities is important, i.e., to rank different protonation sites of a molecule (also referred to as protomers). The first automated protomer screening procedure at the GFN1-xTB level was discussed in Ref. [187]. Here, the idea was to generate all protomers of a molecule automatically and rank them based on their energy, i.e., the relative PA, to obtain estimated preferred protonation

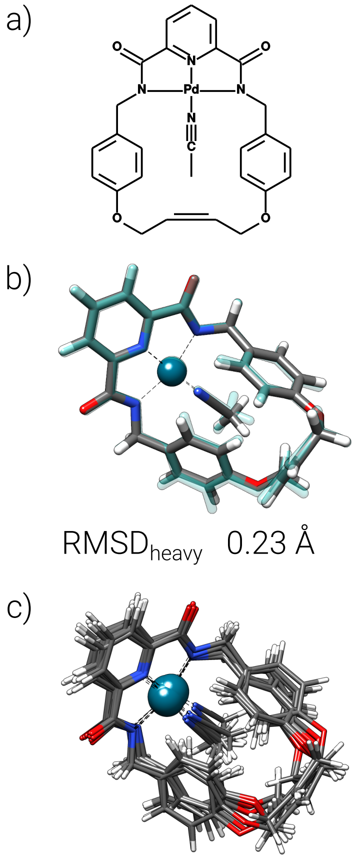


Figure B.14: Pd-pincer metallomacrocycle: a) Lewis structure of the metallomacrocycle, b) overlay between the crystal structure (transparent blue) and the lowest conformer generated at the GFN2-xTB[GBSA(acetonitrile)] level, c) overlay of the 10 lowest energy conformers.

sites. For an automation this requires the generation of initial protonated candidate structures. Common protonation sites are typically π- and lone pair (LP) centers, which can be obtained easily from an SQM calculation. In Ref. [187], it was demonstrated that this procedure works reasonably well for almost arbitrary chemical systems. As an electronically rather complicated example, the (PCP)Ir(N_2) complex and its protomers generated at the GFN2-xTB[GBSA(THF)] level are shown in Figure B.15.

This complex is known to be able to activate N_2 upon protonation, which was studied by combined NMR and cyclic voltammetry experiments.[281] It was concluded that the protonation selectively occurs at the metal center of the complex and no protonation at the dinitrogen group is observed. In a completely automatized fashion this can be validated at the GFN2-xTB level, where the protonation occurs either in a three-center-two-electron (3c-2e) bond between the metal and the pincer, or directly at the metal (cf. Figure B.15b) and B.15c)). All other possible protomers of the complex, including protonation at the dinitrogen ligand are significantly higher in energy (see Figure B.15c) and B.15e)).

The automatized protonation site identification at the GFN*n*-xTB level was extended in Ref. [19] to other (mononuclear) ions. The resulting capability, e.g., to screen for alkalization sites is a further proof of the robustness of GFN*n*-xTB methods. Note, that the automated generation of protomers (or similar ion adducts) is only possible due to the QM nature of the approach: starting structures are generated

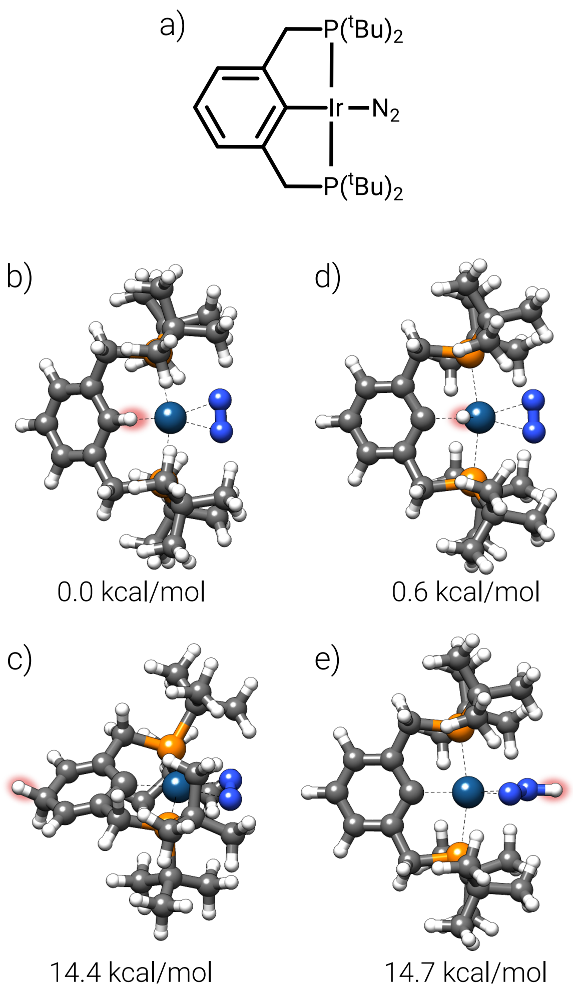


Figure B.15: The (PCP)Ir(N_2) pincer complex and its protomers: a) Lewis structure of (PCP)Ir(N_2) in the unprotonated state, b-e) most stable protomers of $[(\mathrm{PCP})\mathrm{Ir}(N_2)\mathrm{H}]^+$. The proton position is highlighted for better visibility. Relative energies below the structures correspond to the GFN2-xTB[GBSA(THF)] level.

from localized molecular orbital (LMO) centers without other prior information and new bonds can be formed freely during a structure optimization, which is both impossible at FF or chemoinformatical level. From the examples above and the previous publications[19,187,277,282] it is concluded that the automation of GFN*n*-xTB-based screening procedures provide reliable workflows for the generation and ranking of various low-energy structures. Typically, they require further treatment at higher DFT or WFT levels of theory as discussed in the introduction (cf. Figure B.1). Furthermore, the protonation feature described here has been extended in Ref. [19] to an automatized procedure for generating low-energy tautomers as well.

B.4.7 Electron Impact Mass Spectra Simulations

One of the important special features of TB methods at finite electronic temperature is that covalent bonds can be dissociated properly into atoms (for some example potential energy curves see Ref. [8]).

This was exploited already in 2013 (originally with DFTB3-D3(BJ)) for the first principles computation of electron impact mass spectra (EI-MS) of molecules[283] by combination of QC with stochastically initiated MD simulations. Since in this approach, a large chemical space resulting from fragmentations is automatically explored, we consider it as an example here. Actually, the long term idea to be able to simulate EI-MS was one of the reasons that led to the development of the GFN*n*-xTB methods and very recently, GFN1- and GFN2-xTB was employed on a large scale for this purpose.[235,284] Note, that the whole procedure termed QCEIMS would hardly be technically and computationally feasible without the herein described robust SQM methods. The principle and main features of QCEIMS are briefly described in the following paragraph and depicted in Figure B.16.

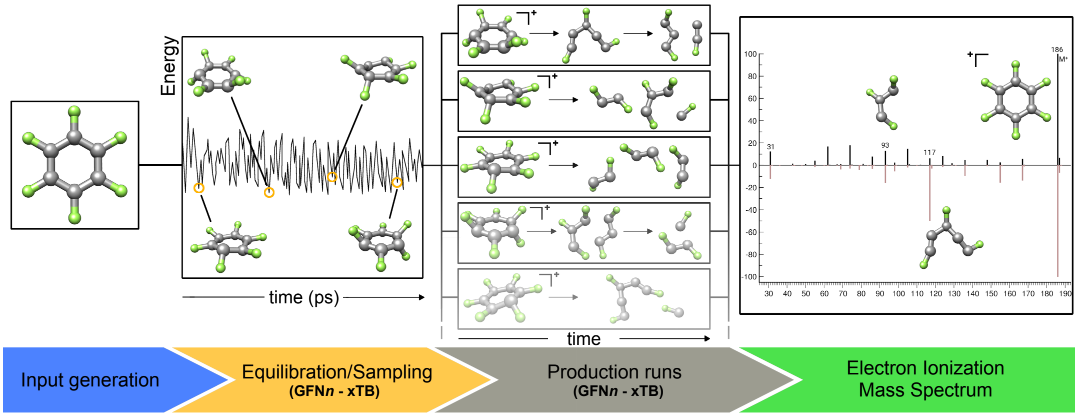


Figure B.16: Automatic workflow for GFN*n*-xTB based EI-MS calculations (modified from Ref. [284]). The hexafluorobenzene molecule is taken as an example.

In the first step, the (neutral) input molecular structure is equilibrated in an MD run from which a predefined number (typically a few hundred) snapshots are randomly selected and saved as starting coordinates. For complicated cases, a preceding detailed conformational analysis could be conducted and the entire procedure is started separately for the found conformers.

For each snapshot, the molecular orbital spectrum is calculated and a Mulliken population analysis is performed. With this information, the internal excess energy (IEE) and internal conversion (IC) time are estimated for proper initial (randomized) conditions. Then, the snap-shots are instantaneously (valence) ionized and independently propagated in time on the cationic GFN*n*-xTB PES until a reaction occurs in the simulation. The ionization potential (IP) of the (neutral) fragments is calculated and used to determine their statistical charge. The fragment with the highest charge is selected for propagation in a cascade fashion. It can undergo further fragmentation until either no internal energy is left, or the fragment is getting too small. All charged fragments are counted, stored, and, by gathering data from all production runs, the mass spectrum is obtained as shown on the right part of the Figure B.16 in comparison with the respective experimental data (inverted for better visibility). At a typical electron impact energy of 70 eV, the IEE of the target molecule is several eV so that very high reaction rates in the few ps regime are initiated. Hence, although one has to sample over hundreds to thousand of trajectories, each of them can be computed rather fast such that the overall computational effort is manageable at an SQM (but not at DFT) level.

The calculations carried out in this way are basically following first principles and fully theoretical, i.e., not based upon any experimental results and represent a viable, “black-box” type alternative to rule-based, chemoinformatical approaches. Typically, a very reasonable (semi-quantitative) agreement between theory and experiment is observed.[285] This is somewhat surprising since barriers and reaction thermodynamics (for which the GFN methods are *not* parameterized) strongly influence the sampled relative reaction rates. Overall, GFN1- and GFN2-xTB (and also DFTB3-D3(BJ) for organic molecules) perform rather similar for a broad set of molecules. In addition to the spectrum, the reactive MD trajectories yield detailed chemical information on the reaction mechanisms leading to specific fragments (spectral peaks). Because overall millions of energy/force evaluations are required in this approach, this is only possible with fast, robustly converging, and dissociative QC methods like GFN*n*-xTB.

B.4.8 Further Applications

Thermostatistical Corrections

The computation of harmonic vibrational frequencies by quantum chemistry methods is a common application mostly conducted to obtain thermostatistical corrections from energy to enthalpy (H) or free energy (G). This requires knowledge of the equilibrium molecular structure (and atomic masses) as well as (at least) the second derivatives of the energy with respect to all nuclear displacements (Hessian matrix), both of which are provided rather accurately by the GFN methods.

Here, GFN2-xTB-computed zero-point vibrational energies (ZPVE, which contribute dominantly to H) and total molecular free energies at 298 K (G_{298}) are compared to corresponding values at the low-cost B97-3c DFT theoretical level. Because such standard DFT calculations are roughly two orders of magnitude slower, a very substantial reduction of the needed computational resources may be achieved if they could routinely be replaced by SQM. The B97-3c vibrational frequencies are on average close to experimental fundamental ones and normally require no scaling to be comparable to experimental data.[286] They are here taken as a reasonable reference to benchmark GFN2-xTB (the GFN1-xTB variant performs similar but slightly worse and is therefore not discussed here).

Figure B.17 shows a comparison of ZPVE values at B97-3c and GFN2-xTB values for a set of 39 medium sized organic molecules taken from a benchmark study of Li et al.,[287] which was originally employed to establish computational procedures for computing molecular entropies. The corresponding free energy values are also depicted in Figure B.17 which (compared to the ZPVE) emphasize more structural aspects as well as the low frequency part of the vibrational spectrum. The G_{298} values refer to the modified rigid-rotor-harmonic-oscillator (mRRHO) treatment from Ref. [73] with a rotor-cutoff of 20 cm^{-1}. As can be clearly seen from the graph, there is a very good reproduction of the DFT reference thermostatistical properties by GFN2-xTB. The MAD for the ZPVE data is only 0.34 kcal/mol (MD = 0.29 kcal/mol) with a maximum error of 1.6 kcal/mol. In actual applications where normally differences of the values for reactants and products are taken, the effective error may be even smaller because of cancellation. The performance for the free energies is similar with an MAD of 0.5 kcal/mol and a maximum error of 2.0 kcal/mol. These deviations are small compared to other sources of error in typical thermochemical studies for larger systems like the intrinsic error of the DFT/WFT electronic energies[23,288] or an inaccurate treatment of solvation effects.[289,290] Tab. B.3 shows that this also applies to larger or electronically more complex cases. Here, the absolute values (and hence also the absolute deviations) are larger than for the previous benchmark set but nevertheless, still small and rather systematic deviations of less than 5 % (< 2 % for the two anti-viral drugs) are obtained with GFN2-xTB. Typically, the predicted values are too small which could be remedied by a simple scaling

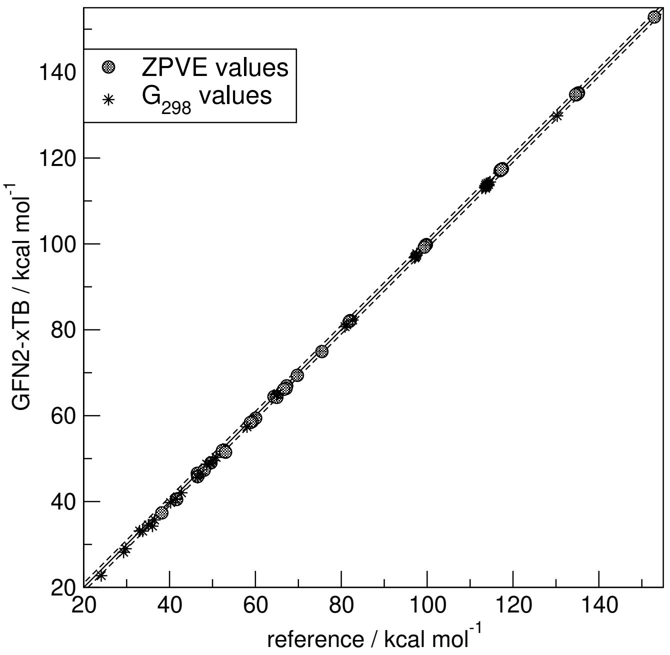


Figure B.17: Comparison of GFN2-xTB thermostatistical data with corresponding B97-3c DFT reference values for 39 organic molecules ranging from ethane (smallest) to n-octane (largest). The solid line shows the one-to-one correspondence and the dashed ones indicate a common error range for chemical accuracy, i.e., ±1 kcal/mol.

Table B.3: Comparison of GFN2-xTB thermostatistical data (in kcal/mol) with corresponding B97-3c DFT reference values for a set of six molecules (two antiviral drugs, one large conjugated π-system with a coordinated metal atom, and two transition-metal complexes).

	ZPVE		G_{298}	
molecule	GFN2-xTB	B97-3c	GFN2-xTB	B97-3c
lopinavir	491.3	500.4	441.1	451.3
remdesivir	377.8	385.1	330.0	338.3
Mg-porphine	165.9	169.2	138.5	143.1
ferrocene	100.5	104.8	79.1	85.1
$Cr(CO)_6$	29.7	31.7	1.9	7.0

procedure. Note that for the largest system with 94 atoms (lopinavir) the frequency calculation took about 7.5 and 0.03 hours, respectively, at the DFT and SQM levels showing the tremendous speed-up at very little loss of accuracy. Furthermore, the severe sensitivity of DFT-computed thermostatistical data (vibrational partition function) on the numerical integration grid for low-frequency modes has been pointed out recently.[291] Eliminating this issue would require the use of very large, computationally costly integration grids in the calculation of the nuclear Hessians with DFT methods. The GFNn-xTB energy expressions are fully analytic, and hence, represent a robust electronic structure scheme for the calculation of harmonic frequencies to be used in free energy calculations. For these reasons, the application of GFN2-xTB is highly recommended instead of costly and numerically sensitive DFT methods in the computation of vibrational frequencies in large scale computational projects or early stages of extensive mechanistic studies.

Protein Examples

The accurate simulation of large biomolecular systems like proteins remains one of the "holy grails" in computational chemistry.[167] Efficient theoretical models can be applied in tandem with the experiment, e.g., to elucidate molecular mechanisms *in vivo*[292], to compute target-drug interactions[174,293] or to simulate secondary structure rearrangements.[294–296] The GFNn-xTB family of methods is a promising candidate for such tasks and has already been successfully applied multiple times in the context of biomolecular systems.[189,266,297] One example is the comprehensive study of Schmitz et al.[266] that evaluates the performance of GFN1/2-xTB variants on structure optimizations of 90 protein structures (of which 20 contain metal atoms). The experimentally derived X-ray structures serve as reference in this study. GFN2-xTB performs for various standard geometrical descriptors very similar to special-purpose force-fields (OPLS2005 and AMBER*) and outperforms even the wave function-based HF-3c method. For the subset of metalloproteins – for which standard FFs and HF-3c could not be applied – GFN2-xTB shows remarkable performance in reproducing the secondary structure motifs and the coordination sphere around the metal centers.

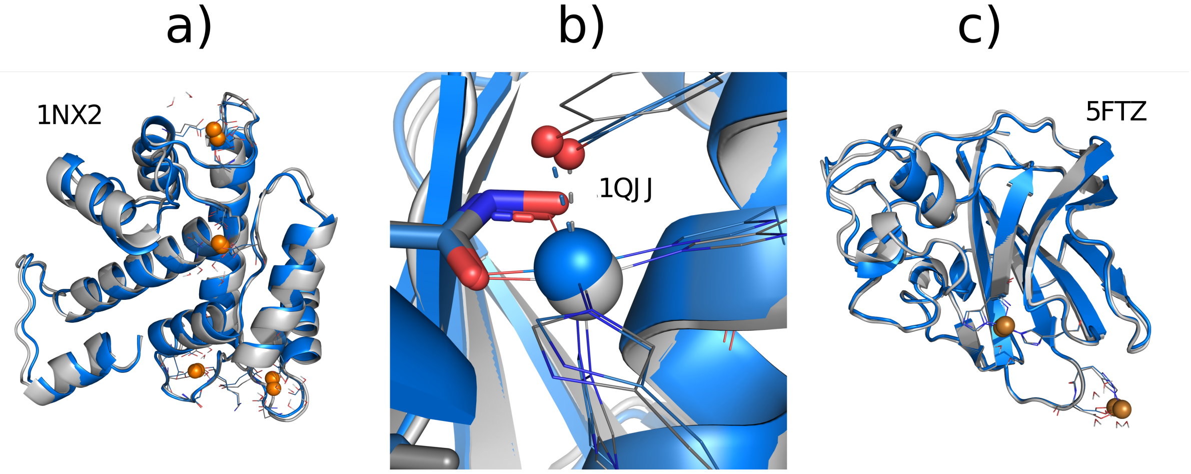


Figure B.18: Structural overlay of experimental crystal (grey) and GFN2-xTB/GBSA(H_2O)-optimized metalloprotein structures (blue). 1NX2 and 1QJJ were computed as closed-shell systems with charges equal to +1 and −11, respectively. 5FTZ was treated as a doublet with a net charge of +6 (see Ref.[266] for details).

Figure B.18a) depicts the structural overlay of the calcium-containing hydrolase 1NX2 shown as a first example. The secondary structure and metal sites are preserved during optimization. The C_α RMSD of 0.89 Å is reasonably small and the structures differ mainly in the unstructured loop regions.

Figure B.18c) depicts the Cu-containing lyase protein (5FTZ). The overlay of the X-ray and GFN2-xTB optimized structure shows a remarkable agreement in the secondary structure. The RMSD of only 0.48 Å is one of the smallest within the whole metalloprotein subset. GFN2-xTB reproduces – also for this protein – all hydrogen bond stabilized structural motifs like helices and sheets very well. The reason for the structural differences in the unstructured regions remains elusive. Figure B.18b) depicts the coordination sphere of the zinc metal center of the hydrolase 1QJJ. The mean absolute error of all metal-ligand bond lengths is 0.07 Å. GFN2-xTB perfectly reproduces the coordination sphere with respect to the experimental X-ray structure.

The combination of GFN*n*-xTB variants for structural problems with the sTDA-xTB method for excited states opens interesting possibilities for investigation of large biomolecular systems. Some of us have presented the applicability and performance of the sTDA-xTB method for electronic excitation spectra like ECD and UV-vis for proteins and DNA fragments.[189] Figure B.19 shows the ECD spectrum of cytochrome c computed with sTDA-xTB on a GFN2-xTB optimized structure as an example.

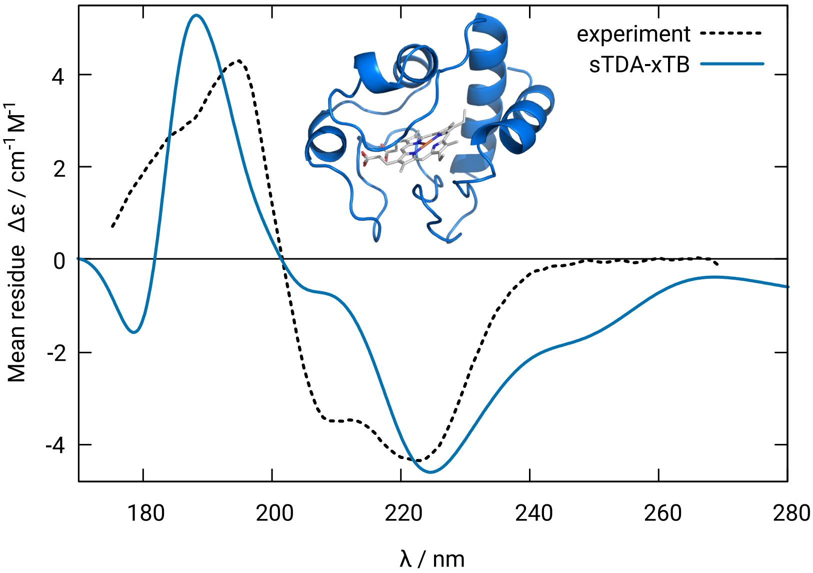


Figure B.19: Calculated ECD spectrum of cytochrome c (blue solid line). The individual transition strengths are broadened by Gaussian functions with a full width at 1/e maximum of 0.5 eV and the spectrum is red-shifted by 0.5 eV. The experimental spectrum in water (gray solid line) is taken from Ref. [298].

The characteristic ECD bands of the α-helical secondary protein structure are reproduced very well. This is remarkable, considering that no molecular fragmentation procedure is applied, i.e., the entire protein is computed fully quantum mechanically. The computation takes 5 h for the optimization and 33 h for the spectrum calculation (involving 30405 excited electronic states) on a standard compute node with four Intel® Xeon® CPU E3-1270 v5 @ 3.6GHz cores.

Previous studies have shown that for some optical properties, MD sampling is essential.[189,198,202,299–301] Here, an MD simulation is performed, and snapshots are taken equidistantly from the resulting trajectory. The desired property is then computed for each snapshot and averaged. The MD simulation at the tight-binding level is not feasible for systems with more than 1000 atoms and for the necessary simulation lengths of several nanoseconds. Figure B.5 shows the timings for single-point plus gradient calculations for the different GFN methods for 14 proteins. The GFN-FF method, with a three orders of magnitude lower computational cost compared to GFN2-xTB, enables MD simulations of biomolecular systems with up to 5000-10000 atoms. Even the inclusion of explicit solvent molecules for structure optimization and MD simulation is now possible. Figure B.20 shows the structure of hemocyanin (PDB: 1JS8) with an explicit water shell of 6 Å distance to the protein surface (10074 atoms). The structure of this system is optimized on loose thresholds within 1058 cycles and in 9 hours and 33 minutes at the GFN-FF level of theory (31 seconds per optimization cycle). The heavy atom and C_α-RMSD are 1.0 and 0.9 Å, respectively, compared to the experimental crystal structure.

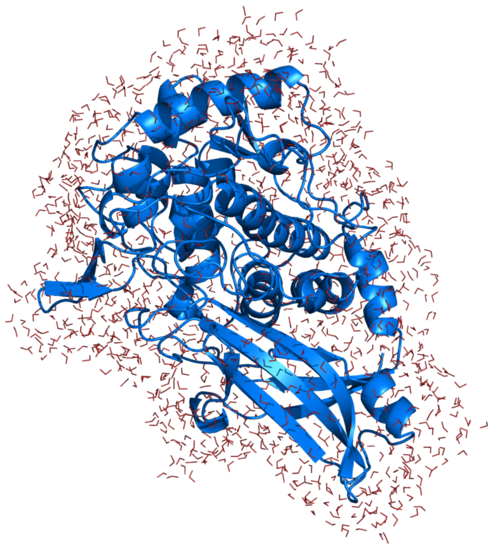


Figure B.20: Cartoon representation of the GFN-FF optimized (blue) structure of hemocyanin (PDB: 1JS8), including an explicit water solvent shell of 6 Å (red).

Periodic systems

Recently, the GFN0- and GFN1-xTB methods were extended to periodic boundary conditions (PBC) in two program packages[302,303] enabling the routine computation of solids and surfaces. As examples for application to periodic systems, we describe the computation of lattice energies for various molecular crystals in the common X23 benchmark set,[154,157] as well as structural data of 222 zeolite frameworks taken from a standard database.[304,305]

Molecular crystals are an important research area for material science as well as in pharmaceutical chemistry.[306,307] Their perhaps most important property is the lattice energy, E_{lat}, which reflects how much energy is released per molecule upon sublimation. It is defined as

$$E_{\text{lat}} = \frac{1}{n} E_{\text{crystal}} - E_{\text{gas}}, \tag{B.66}$$

where E_{crystal} is the total energy of the crystal including overall n molecules within the primitive cell and E_{gas} is the energy of the isolated molecule in its lowest energy conformation. A commonly used benchmark set for lattice energies, which comprises a diverse set of experimentally well-determined organic molecular crystals, is the X23 dataset.[154] Reference values are experimental sublimation enthalpies, which have been back-corrected for vibrational contributions in the harmonic approximation. It includes various intermolecular binding motifs as, e.g., hydrogen bonding, electrostatic interactions, as well as London-dispersion dominated unsaturated hydrocarbons and is therefore ideally suited to assess how well the underlying theoretical method is able to describe non-covalent interactions in a dense environment.

The error of the applied Γ-point only approximation applied in the periodic GFN1-xTB method can become quite large for small cells. Therefore, we constructed different supercells (2x2x2 and 3x3x3) for every molecular crystal to minimize this error and noted that the best results are obtained within the largest supercells which are reported here. Furthermore, we employ the low-cost DFT composite HSE-3c method[308] for comparison. The average experimental value of E_{lat} over the whole test set is 20.3 kcal/mol.

Table B.4: Errors in lattice energies for the X23 benchmark set. All statistical measures are given in kcal/mol. GFN1-xTB values are obtained using the `ADF` program,[302] DFTB3 values are obtained using the `DFTB+` program[309] (excluding the ammonia and cytosine systems which did not converge). HSE-3c values were taken from Ref. [308].

Measure	HSE-3c	GFN1-xTB	DFTB3-(3ob)-D3(BJ)
MD	0.3	−1.0	2.4
MAD	1.3	2.7	3.3
RMSD	1.6	3.5	3.4
AMAX	3.8	8.6	8.8

As expected, the overall best performance is obtained by HSE-3c with a mean absolute deviation of 1.3 kcal/mol. GFN1-xTB doubles the MAD of HSE-3c but outperforms its direct competitor DFTB3-(3ob)-D3(BJ) by more than 0.5 kcal/mol. Since the GFN1-xTB method is orders of magnitudes faster than HSE-3c, this result is especially promising for future screening studies, e.g., in crystal structure prediction. Further enhancements have to be developed, especially the proper inclusion of a k-point sampling is necessary for smaller cell sizes. Nevertheless, the applicability of the GFN1-xTB method is superior to the DFTB methods enabling the calculation of systems with almost every element combination.

With this in mind, we discuss how cell volumes of small- to large-sized zeolite frameworks are reproduced in comparison to higher-level theoretical reference data. Overall, 222 zeolite frameworks were considered,[304,305] for which the number of atoms inside the primitive cells range from 15 atoms (IZA code: EDI) to the biggest with 4320 atoms (IZA code: MWF).

All structures have been optimized by Baerlocher and co-workers using the distance least squares (DLS-76) level of theory,[310] which we use for comparison. Due to the large size of some systems, we apply the cost-effective periodic GFN0-xTB method as implemented under PBC in the `xtb` program. Note that the DLS-76 method has been fitted to reproduce experimental data thereby incorporating thermal effects implicitly.

Figure B.21 depicts a correlation plot between GFN0-xTB-computed and DLS-76 reference volumes, which shows that GFN0-xTB is able to accurately reproduce most of the DLS-76 structures rather accurately. Overall, the GFN0-xTB method slightly underestimates the cell volumes, which can be explained by missing zero-point vibrational and thermal contributions. For molecular crystals this contribution amounts to approximately 2%.[34] The reference cell angles are mostly conserved in the optimizations. The MAD values of cell lengths, angles, and volumes for GFN0-xTB with respect to the DLS-76 structures are given in Figure B.21, where one significant outlier (PAR) is shown as inset (cell volume decreased by approximately 14 %).

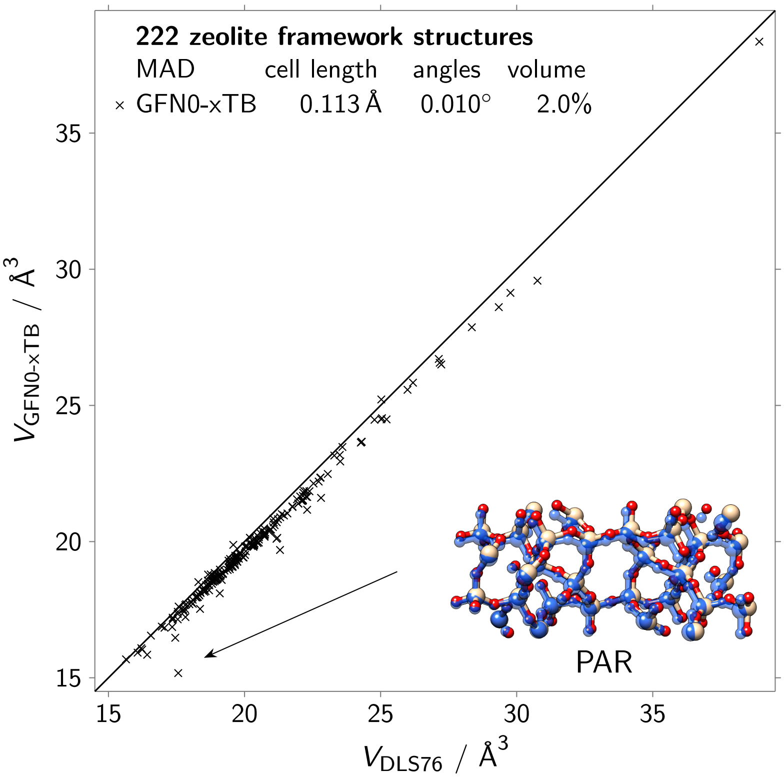


Figure B.21: Comparison between GFN0-xTB and DLS-76 volumes for 222 experimentally known zeolite framework structures. The inset shows a structure overlay of the only significant outlier (see text).

B.5 Conclusions

In this review, the theory, development, and implementation as well as prototypical applications of the GFN family of atomistic, mostly quantum chemistry-based semiempirical methods is described. Their main purpose is the fast, robust and reasonably accurate calculation of large molecules in gas and condensed phase with an emphasis on a good description of ground state structures, non-covalent interactions, low-energy chemical transformations like molecular conformations and vibrational frequencies. A common feature and strong point consistent among the methods (including the recently developed GFN-FF universal force field) is that practically any chemically interesting species can be treated, since parameters exist for a very large part of the periodic table (up to radon). Opposed to the only existing competing family of semiempirical methods with broad parameterization (PMx), the theory is derived from a density functional theory background, which makes the methods applicable also in electronically more complicated cases like transition metal compounds or other systems involving stronger electron correlation effects. Although we currently have less experience with the methods applied under periodic boundary conditions, it is indicated that due to their consistent design strategy and physically very reasonable energy terms, they should be also applicable to solids, surfaces or dense materials in general. The theoretically weakest point of the xTB methods in their current formulation (but also an essential reason for their high efficiency) is the use of an (almost) minimal atomic orbital basis set for the expansion of the Kohn-Sham-type electronic eigenvalue equations.

Overall, the xTB methods, especially the most sophisticated GFN2-xTB version, provide an exceptionally good accuracy-performance/computational cost ratio for the target properties. On standard workstations or even laptop computers, full geometry optimizations, frequency calculations or conformational searches can be conducted in minutes to a few hours for systems composed of hundreds to about a 1000 atoms. Although the xTB methods can be applied in principle also to much larger systems composed of about 5000-10000 atoms, in this regime force-fields like GFN-FF are more appropriate. Alternatively, they allow fast and robust screening of compound libraries in about the same computation time, e.g., thousands of candidate species with 50-100 atoms. This applies even to other interesting chemical (off-target) properties not discussed here like, e.g., polarity (dipole moments), gaps (electrochemistry), vibrational anharmonicity or bonding (electron density) information. Calculations can be carried out in a user-friendly black-box style as implemented in the efficient `xtb` program, which is freely available[204] and accompanied by a continuously updated online documentation.[240]

Not part of the GFN family is another xTB-based scheme, which is discussed in this review: in the sTDA-xTB method, a non-self-consistent xTB approach is combined with a subsequent simplified Tamm-Dancoff-approximated time-dependent linear response treatment. This is a single-point method, which is suited for the computation of excitation energies and optical spectra of huge systems, such as full proteins without resorting to a QM/MM or fragment approach.

We are confident that the described methods will form the basis for many successful studies in various fields of computational science. Application in automated workflows may open up new perspectives for molecular property design or chemical reaction space exploration. Combination with modern machine learning techniques, e.g., as input or data generator, also seems to be a promising research direction and first attempts based on the xTB methods have already been made.[311]

B.6 Funding information

This work was supported by the DFG in the framework of the “Gottfried Wilhelm Leibniz Prize” awarded to S. G., by the DFG priority Program No. SPP 1807 “Control of Dispersion Interactions in molecular Chemistry”, by the Bayer AG, and the Novartis AG foundation. C.B. acknowledges funding from the German National Academy of Sciences Leopoldina through the Leopoldina Fellowship Program (Project No. LPDS 2018-09).

B.7 Acknowledgments

The authors would like to thank Markus Bursch and Jeroen Koopman for help in creating high-quality graphics. Furthermore, the authors thank the numerous beta users and all working students who have tested the `xtb` software extensively. Last but not least the authors would like to thank all staff members of the Mulliken Center for Theoretical Chemistry for their great cooperation.

B.7.1 Further reading

The `xtb` and `crest` programs can be obtained free of charge from GitHub (see Ref. [204] and [239], respectively). Instructions for both programs can be found in Ref. [240].

APPENDIX C

Conformational Energy Benchmark for Longer *n*-Alkane Chains

Sebastian Ehlert,[*] Stefan Grimme,[*] and Andreas Hansen[*]

Received 09 April 2022, Published online 26 May 2022.

Reprinted in Appendix C (adapted) with permission[†] from
S. Ehlert, S. Grimme, and A. Hansen, *Conformational Energy Benchmark for Longer n-Alkane Chains*,
J. Phys. Chem. A **126** (2022) 3521, doi: 10.1021/acs.jpca.2c02439.

[*] Mulliken Center of Theoretical Chemistry, Bonn, Germany

[†] Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

Abstract We present the first benchmark set focusing on the conformational energies of highly flexible, long *n*-alkane chains, termed ACONFL. Unbranched alkanes are ubiquitous building blocks in nature, so the goal is to be able to calculate their properties most accurately to improve the modeling of, e.g, complex (biological) systems. Very accurate DLPNO-CCSD(T1)/CBS reference values are provided, which allow for a statistical meaningful evaluation of even the best available density functional methods. The performance of established and modern (dispersion corrected) density functionals is comprehensively assessed. The recently introduced r^{2}SCAN-V functional shows excellent performance, similar to efficient composite DFT methods like B97-3c and r^{2}SCAN-3c, which provide an even better cost-accuracy ratio, while almost reaching the accuracy of much more computationally demanding hybrid or double hybrid functionals with large QZ AO basis sets. In addition, we investigated the performance of common wavefunction methods, where MP2/CBS surprisingly performs worse compared to simple D4 dispersion corrected Hartree–Fock. Furthermore, we investigate the performance of several semiempirical and force field methods, which are commonly used for the generation of conformational ensembles in multilevel workflows or in large scale molecular dynamics studies. Outstanding performance is obtained by the recently introduced general force field, GFN-FF, while other commonly applied methods like the universal force field yield large errors. We recommend the ACONFL as a helpful benchmark set for parameterization of new semiempirical or force field methods and machine learning potentials as well as a meaningful validation set for newly developed DFT or dispersion methods.

C.1 Introduction

Conformers are defined as distinct minima on a molecular potential energy surface with fixed covalent topology and can be converted into each other by rotations about formally single bonds. Their structure and relative (conformational) energies are of great importance in organic molecules[312] and for biological activity.[313] This is especially true for open-chain compounds, as they usually feature many possible internal rotation axes. Such flexible molecules are key for targeting compounds with specific spatial properties and understanding how folding of biological polymers and peptides is controlled.[314] Often, many conformers cover a rather narrow energy range, so that these systems exist as a thermally populated mixture of conformers at room or physiological temperature. Since measured equilibrium properties correspond to the Boltzmann average of the microstates, representing all relevant molecular structures is essential for a reliable prediction of molecular properties.[20,315]

Unbranched *n*-alkanes, which are ubiquitous in nature, are the simplest examples of this class of molecules, so the goal is to be able to calculate their properties most accurately to improve the modeling of complex (biological) systems such as membrane proteins[316] and to allow for more reliable protein ligand docking libraries.[317] *n*-alkanes and aliphatic chains in general are of particular importance, not only as basic building blocks of organic chemistry and part of fossil fuels, but also as components of lipids and polymers.

The existence of multiple conformers for *n*-alkanes for $n > 3$ has been known since the seminal work of Pitzer published 85 years ago.[318] Since then, a number of experimental and theoretical studies have been carried out with the aim of determining conformation energies, *i. e.* the energy difference between two different conformers, of *n*-alkanes as accurately as possible. Shorter *n*-alkanes are typical subjects of experimental studies investigating conformational enthalpies[319] or low-energy conformers.[320,321] Also theoretical investigations of torsional and conformational energies feature mainly short alkane chains.[322–324] For longer alkane chains ($n > 10$), previous theoretical studies focus mainly on the lowest lying conformer,[325–328] investigating the balance between repulsive hydrogen contacts and attractive London-dispersion to predict the change from the linear zig-zag conformation to a hairpin like, closed conformer.

The most comprehensive theoretical studies on conformational energies of *n*-alkanes up to *n*-hexane were carried out by Gruzman *et al.*,[78] augmented in a later by Martin with a detailed potential energy surface investigation of *n*-pentane beyond equilibrium structures.[329] However, the conformational ensemble of short *n*-alkanes chain is not fully representative for longer *n*-alkanes, which have particular relevance for modeling biological systems, *e. g.* in lipids, due to the much stronger attractive London-dispersion.[326–328]

While *n*-alkanes are seemingly simple systems for wavefunction theory (WFT) due to the large HOMO-LUMO gap and absence of static correlation, the need for large and diffuse basis sets to accurately capture long-range dynamic correlation effects, *i. e.* London dispersion, and to reduce the residual intramolecular basis set superposition error (BSSE), which cannot be removed by standard counterpoise correction schemes, posed a computational challenge. Moreover, since the differences in conformational energies should be described with an accuracy of at least 0.1 kcal/mol to allow a correct assessment of the conformational order or to calculate the Boltzmann populations reasonably well at room temperature,[20] sophisticated WFT methods such as CCSD(T) are essential for a reliable theoretical reference.

To provide this level of accuracy, the cumulative medium and long range intramolecular nonconvalent interactions (NCI) in long saturated chains need to be described accurately at the same footings.[78,330–332] This balance is problematic for density functional theory (DFT) as has been, e.g., shown for 1,3 interactions in branched alkanes by one the present authors[333] even if London dispersion is captured by a suitable dispersion correction. To assess the description of intramolecular NCIs in semilocal density functionals and semiempirical methods, various related conformational benchmarks sets were devised, e.g., for melatonin,[334] butane-1,4-diol,[335] RNA backbone models,[271] amino acids,[336] and many more, for example composed in the GMTKN55 database.[23] The latter also includes the ACONF set, which comprises 15 relative energies of n-butane, n-pentane and n-hexane conformers taken from the work of Gruzman *at al.*[78]

Nowadays, the long-range NCI problem for DFT, semiempirical quantum mechanics (SQM) and also force fields methods is largely solved,[9,25,28] but this has not yet been extensively assessed for conformer ensembles of longer *n*-alkane chains. Moreover, especially in folded *n*-alkane chains, many NCI contacts of H atoms due to Pauli repulsion become important and existing theoretical studies focusing on shorter *n*-alkanes or just equilibrium conformer structures could not evaluate these interactions comprehensively. The recently introduced combined tools GFN-FF,[337] GFN*n*-xTB,[1,8,41] CREST,[19] and CENSO[20] have filled a gap in the field of quantum chemical modeling, specifically for generating conformer ensembles of larger molecules. Here, we make use of these new capabilities to generate suitable conformer ensembles of longer *n*-alkanes, which we have combined into a new benchmark set termed ACONFL (see C.3. Recent developments of accurate low-order scaling local coupled cluster methods[184] enabled us to generate high level theoretical reference values close to the basis set limit suitable for a statistical meaningful evaluation of much more approximate methods.

After summarizing the computational details in the next section, the generation of our new benchmark set and its reference values will be described followed by an extensive evaluation of various FF, SQM, DFT and WFT methods. Finally, general conclusions and method recommendations will be given.

C.2 Computational details

Conformer ensembles were obtained with the advanced conformer rotamer ensemble sampling tool[19,20,239] (crest, version 4) employing the GFN-FF[81] method and default settings. Subsequently, selected conformers (see C.3.1) were re-optimized at the B97-3c[79] level of theory utilizing the Turbomole program package version 7.5.1.[133,338]

The ORCA program package version 5.0.1[184,339] was used to perform the calculations with double hybrid functionals, MP3, DLPNO-CCSD and the meta-GGA B97M, while the Hartree-Fock (HF), second order Møller-Plesset perturbation theory (MP2) and local coupled cluster and calculations were carried out employing ORCA 4.2.1[340]. All other DFT calculations were executed with Turbomole 7.5.1. MP2D[341] calculations were conducted with Psi4.[342] κOO-MP2 ($\kappa = 1.1$)[343] and MP2.5[344] were evaluated with QChem 5.4.2.[12] The resolution of identity (RI) method was employed to accelerate the evaluation of Coulomb (RIJ) and exchange integrals (RIJK).[345,346] Except for the ”3c” methods, which use the respective stripped and optimized basis sets, Ahlrichs' type quadruple-ζ def2-QZVPP[347] basis sets with matching auxiliary basis sets for RIJ and RIJK[136,348] were applied in the DFT calculations. The numerical quadrature grid option *DefGrid3* and *TightSCF* convergence criteria were applied as implemented in ORCA 5.0.1, while the *m4* grid was used in the Turbomole calculations.

The RI and frozen core approximations for the correlation part as well as *TightSCF* convergence criteria for the HF part as implemented in ORCA 4.2.1 were employed. The domain based, local pair natural orbital coupled cluster method[269] in its ORCA 4.2.1 closed-shell, sparse maps[270] iterative triples[349] implementation (DLPNO-CCSD(T1)) together with *VeryTightPNO* threshold settings (i.e. ORCA 4.2.1 *TightPNO* settings with *TCutMKN*, *TCutPNO*, and *TCutPairs* tightened to 10^{-4}, 10^{-8}, and 10^{-6}, respectively) was applied. An aug-cc-pVTZ/aug-cc-pVQZ[350] complete basis set (CBS) extrapolation according to the scheme introduced by Helgaker and Klopper[351] was carried out separately for the HF and correlation energy for all MP2 and parts of DLPNO-CCSD(T1) energies. The same extrapolation scheme was used for the MP2/CBS and PWPB95-D4/CBS energies. For all other DLPNO-CCSD(T) energies, a special CBS extrapolation scheme (see Section C.3.2) was employed.

DFTB calculations were conducted with the DFTB+ program package (version 21.1),[13,352] the *3ob* parameterization[353–356] was used for the third-order DFTB Hamiltonian, while the *mio* parameterization[38,178,357] was used with the second-order DFTB Hamiltonian. The LC-DFTB Hamiltonian was applied with the *ob2-base* parameterization.[358] The GFN1-xTB,[41] GFN2-xTB,[8] and GFN-FF[81] calculations were carried out using the xtb program (version 6.4.1)[1,204]. The MOPAC program (version 2016)[359] was used to perform PM6-D3H4[80] and PM7 calculations.[360] The SMIRKS Native Open Force Field (SMIRNOFF) based methods[361] were used to evaluate the SMIRNOFF99Frosst-1.1.0[362], OpenFF-1.0.0[363] and OpenFF-2.0.0[364] as implemented in the OpenFF toolkit.[365] The UFF[82] and the MMFF94[83,84] were evaluated with RDKit.[366] Calculations with the OpenFF toolkit and RDKit were driven via QCEngine.[11]

All tested DFT methods were evaluated in combination with one out of the following three London dispersion corrections, D3,[26,27] D4,[7,9] or VV10[28,29,62] (also called NL or V). The two former were applied together with the rational (Becke–Johnson) damping function,[67,68,129] except for the M06-L functional, where zero (Chai–Head-Gordon) damping[367] was employed. Furthermore, revised damping parameters proposed by Smith *et al.*[368] and the optimized power damping function[369] were tested if available for the respective functionals. D3 and D4 corrections were calculated with the s-dftd3 (version 0.5.1)[370] and `dftd4` (version 3.3.0)[371] standalone programs and consistently include three-body Axilrod–Teller–Muto (ATM)[69,70] dispersion contributions. The non-local density-dependent VV10 dispersion correction was calculated non-selfconsistently as implemented in `Turbomole 7.5.1` or in case of B97M-V, `ORCA` 5.0.1.

Table C.1: Tested semi-empirical and force field methods

method	dispersion	reference
FF		
smirnoff99Frosst-1.1.0	LJ	[361, 362]
OpenFF-1.0.0	LJ	[361, 363]
OpenFF-2.0.0	LJ	[361, 364]
UFF	LJ	[82]
MMFF94	Buf-14-7	[83, 84]
GFN-FF	D4	[81]
SQM		
GFN1-xTB	D3(BJ)	[41]
GFN2-xTB	D4	[8]
PM7	D2	[360]
PM6	D3(BJ)	[80]
DFTB3	D3(BJ), D4	[353–356]
DFTB2	D4	[38, 178, 357]
LC-DFTB2	D3(BJ), D4	[358]

C.3 The ACONFL test set

We present a new benchmark set termed ACONFL (Alkane CONFormers Large) to evaluate QC, SQM, and FF methods concerning their performance in predicting alkane conformer energies. It extends the well-established and commonly used ACONF benchmark set by employing longer n-alkanes and more diverse conformer ensembles. While the largest alkane used in the ACONF benchmark is n-hexane, the ACONFL set is composed of n-dodecane (ACONF12 subset), n-hexadecane (ACONF16 subset), and n-icosane conformers (ACONF20 subset), including 50 conformational energies in total.

C.3.1 Construction of the test set

The conformer potential energy surface of an unbranched alkane is characterized by torsional twists that lead from linear chains to highly deformed structures dominated by intramolecular dispersion forces. At temperatures less than 300 K, short alkanes ($n = 4 - 8$) in the gas phase are well-known to prefer the linear all-trans conformation. However, as the length of the alkane grows, there is a point where the attractive NCIs will cause the chain to "self-solvate" into a folded conformer.[325] A cross-gauche-cross rotation combination introduces an energetically unfavorable syn-pentane-like conformation. In addition, the chain ends are not parallel in this conformation, thus reducing the possible stabilization due to van der Waals attraction. A hairpin conformation with four gauche rotations minimizes the number of

Table C.2: Tested methods and dispersion corrections

method	D3	D4	VV10	reference
Composite ("3c")				
HF-3c	✓	—	—	[173]
PBEh-3c	✓	—	—	[142]
B97-3c	✓	—	—	[79]
r^{2}SCAN-3c	—	✓	—	[14]
(meta-)GGA				
PBE	✓	✓	✓	[57, 58]
TPSS	✓	✓	✓	[146, 372]
B97M	✓	✓	✓	[22, 62, 63]
r^{2}SCAN	✓	✓	✓	[4, 21, 373]
M06L	✓	✓	—	[103]
(rs-)(meta-)Hybrid				
B3LYP	✓	✓	✓	[253, 374]
PBE0	✓	✓	✓	[59, 60]
PW6B95	✓	✓	✓	[375]
M06-2X	✓	—	—	[64]
MN12-SX	—	✓	—	[376]
ωB97M	✓	✓	✓	[61–63]
ωB97X	✓	✓	✓	[74, 108, 367]
Lh20t	—	✓	—	[32]
Double-hybrid				
B2PLYP	✓	✓	✓	[333]
PWPB95	✓	✓	✓	[377]
DSD-BLYP	✓	✓	—	[377]
revDSD-BLYP	—	✓	—	[378]
WFT				
HF	✓	✓	✓	
MP2	—	—	—	
MP3	—	—	—	
MP2D	✓	—	—	[341]
κOO-MP2	—	—	—	[343, 379]
MP2.5	—	—	—	[344]
DLPNO-CCSD	—	—	—	

strained bonds. This allows an energetically favorable parallel arrangement of the chain ends, yielding the suggested global minimum for longer alkanes.[326–328] The zig-zag-hairpin stability turning point appears to be around hexadecane.

Since the conformational ensembles become nearly continuous in energy for longer alkanes, compiling a benchmark set out of the lowest conformers up to a certain energy threshold is not practicable since even with the CCSD(T) at the estimated basis set limit to reliably predict conformational energies smaller than about 0.1 kcal/mol. Therefore, we created a conformational ensemble at the GFN-FF level of theory using the version four conformer search as implemented in CREST and selected conformers in a 5–6 kcal/mol energy window with a decent (i.e., clearly distinguishable at the CCSD(T) level of theory (*vide infra*)) equidistant spacing of the conformational energies. This approach keeps the total number of conformers in the ACONFL set reasonably small while still including as much of the diversity of the complete conformational ensemble as possible. Those conformers were re-optimized at the B97-3c level of theory and used as the starting point for performing the reference calculations. B97-3c provides sufficiently accurate geometries for our purpose, although the overall accuracy of the geometries is secondary due to the *de facto* continuum of structures in the conformers for these highly flexible systems and the use of the same structures also for the reference CC calculations. In total, 53 single point calculations are required to evaluate the complete ACONFL, and 50 conformational energies with respect to the respective energetically lowest conformers are overall assessed for the three subsets, ACONF12, ACONF16, and ACONF20. Compared to the ACONF set with an mean absolute conformational energy of 1.83 kcal/mol the complete ACONFL set has a higher mean of 4.62 kcal/mol.

The ACONF12 subset shown in Fig. C.1 contains twelve relative conformational energies, with the lowest conformer being the linear n-dodecane molecule and the mean absolute conformational energy being 4.28 kcal/mol. This set was already successfully used in several studies to test the performance of new DFT methods[5,14] as well as in a recent perspective on the description of conformational ensembles.[20] The numbering of the conformers results from the initial conformational search rather than the final energetic ordering.

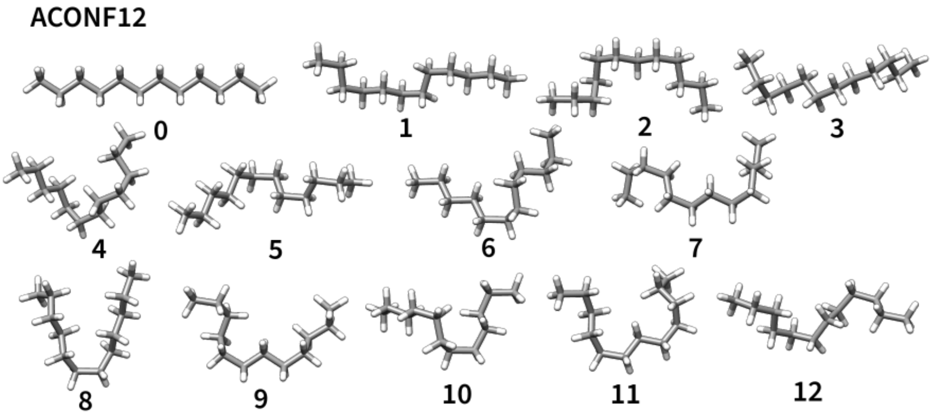


Figure C.1: The 13 n-dodecane conformers of the ACONF12 subset. The conformer **0** is the lowest conformer, the numbering of the conformers does not necessarily correspond to their energetic order.

For the ACONF16 subset in Fig. C.2 17 relative conformational energies are included with the folded n-hexadecane as the energetically lowest conformer, which is in line with previous studies,[327,328,380] but the linear conformation being the second lowest is only slightly higher energy by 0.09 kcal/mol. The mean absolute conformational energy of ACONF16 is 3.98 kcal/mol. Finally, the ACONF20 subset contains 21 relative conformational energies (cf. Fig. C.3). For this set, the hairpin-like conformer of n-icosane is with 1.31 kcal/mol already significantly more stable the linear conformation. Here, the mean absolute conformational energy is 5.32 kcal/mol. These subsets allow for assessing a balanced description of dispersion and repulsion by the tested methods, since the shorter n-dodecane features

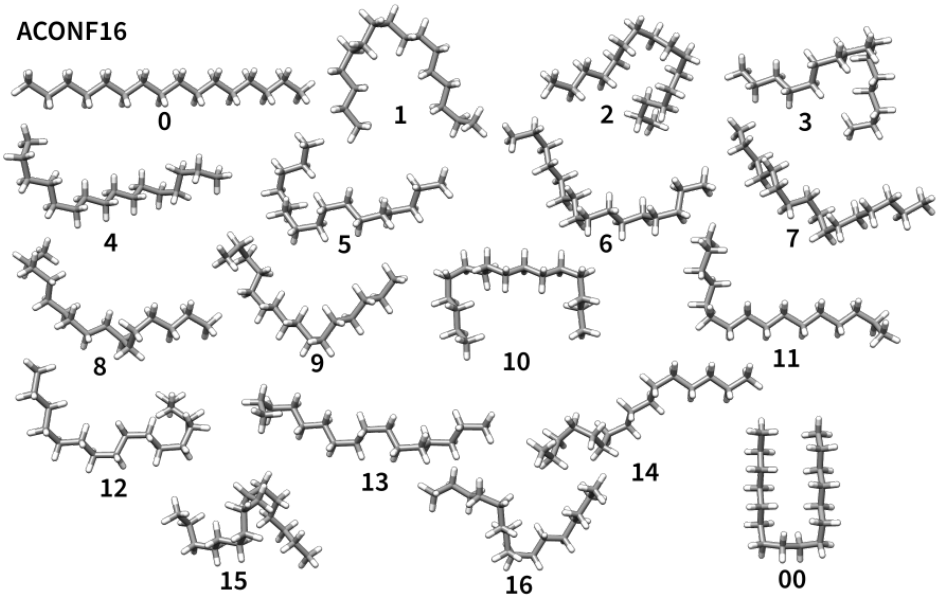


Figure C.2: The 18 n-hexadecane conformers of the ACONF16 subset. The conformer **0** is the lowest conformer, the numbering of the conformers does not necessarily correspond to their energetic order.

incomplete attractive intramolecular NCIs to favor a closed form, while the longer n-icosane chains with favor a closed hairpin-like conformation due to cummulated effect of intramolecular NCIs, and the intermediate n-hexadecane set refers to the point, where both forms are very close in energy. It should be noted that the mean signed error (MSE) statistical descriptor depends on whether the lowest conformer is the linear or folded structure. Therefore, it is more meaningful to analyse the MSE, e.g. looking for a systematic of the linear form, for each subset separately rather than for the complete ACONFL as the MSEs of the subsets could inevitably cancel each other due to different signs.

C.3.2 Generation of the reference values

Previous studies of alkane conformers proved that effects beyond CCSD(T) are not important for the accurate description of conformational energies.[78] Since canonical CCSD(T) is computationally prohibitive for the target systems, the DLPNO-CCSD(T1) method with very tight PNO thresholds is used to approximate the canonical result as close as possible. For the extrapolation to the complete basis set limit[351,381] the aug-cc-pVTZ and aug-cc-pVQZ basis sets were used, dubbed aT and aQ respectively in the following paragraphs. This level of theory was shown to serve as reliable reference in various NCI benchmarks.[131,337,382] To also verify the accuracy the suggested reference protocol for the ACONFL set, we evaluate ACONF and compare to the highly accurate W1hval[78,383,384] reference values of the latter. The deviation of DLPNO-CCSD(T1)/CBS(aTaQ) from the W1hval conformational energies is shown in Fig. C.4 on the left. With an negligible MSE of 0.04 kcal/mol and an error range of only 0.10 kcal/mol, sufficient accuracy for ACONFL can be expected with our reference protocol.

However, the full basis set extrapolation is still too expensive to be applicable for the ACONF16 and ACONF20 subsets. Therefore, we resort to a CBS extrapolation scheme based on focal-point analysis[385]

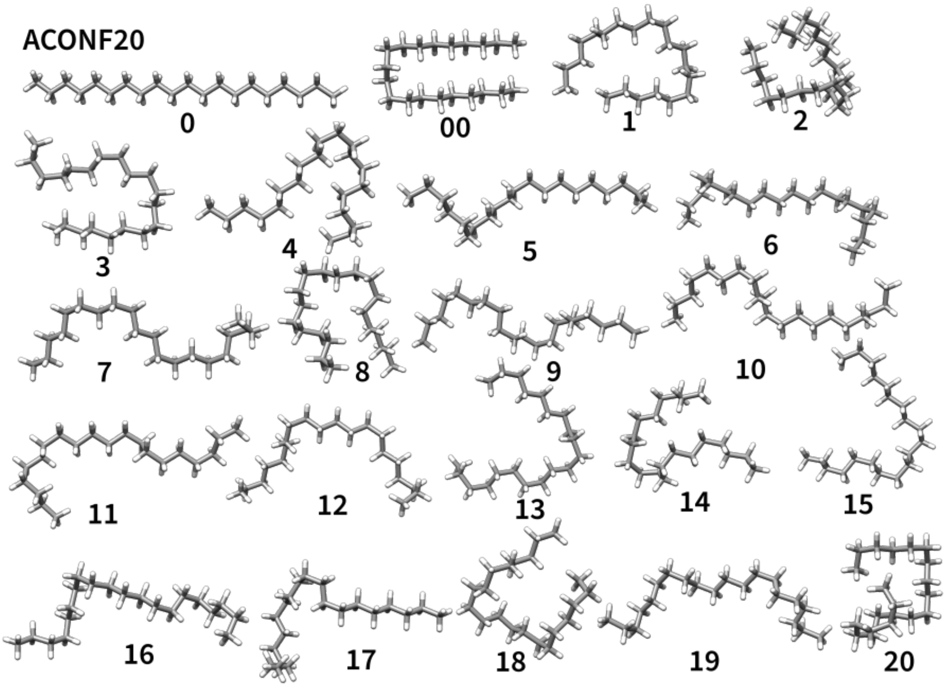


Figure C.3: The 22 n-icosane conformers of the ACONF20 subset. The conformer **00** is the lowest conformer, the numbering of the conformers does not necessarily correspond to their energetic order.

for the latter. Following Marshall *et al.*,[131] the respective "δCBS" basis set extrapolation scheme is given by

$$\delta\text{CBS} = \mathrm{E}(\text{MP2/CBS(aTaQ)}) + \mathrm{E}_\mathrm{c}(\text{DLPNO-CCSD(T1)/aT}) - \mathrm{E}_\mathrm{c}(\text{MP2/aT})\,, \tag{C.1}$$

where E_c is the correlation energy part of the total energy E. Additionally, we introduce a similar but multiplicative scheme dubbed xCBS, which represents a refined variant of the multiplicative CBS* scheme,[79,271] and is defined in the following way:

$$\text{xCBS} = \mathrm{E}(\text{HF/CBS(aTaQ)}) + \mathrm{E}_\mathrm{c}(\text{MP2/CBS(aTaQ)}) \cdot \left(\frac{\mathrm{E}_\mathrm{c}(\text{DLPNO-CCSD(T1)/aT})}{\mathrm{E}_\mathrm{c}(\text{MP2/aT})} \right)\,. \tag{C.2}$$

The xCBS protocol is typically less sensitive to more severe MP2 errors compared to the δCBS protocol. To estimate the additional error introduced by the more approximate basis set extrapolation, we compared the two schemes with the full CBS(aTaQ) conformational energies for the ACONF12 subset (see Fig. C.4, on the right). While the xCBS(aTaQ) yields a slightly positive MSE of 0.05 kcal/mol, the δCBS(aTaQ) a slightly negative MSE of −0.08 kcal/mol. Hence, we find that the arithmetic mean of both schemes agrees exceptionally well with the full CBS(aTaQ) scheme. Therefore, this average was chosen as reference for the ACONF16 and ACONF20 subsets, for which the full CBS(aTaQ) extrapolation was computationally unfeasible.

The maximum residual error of the ACONFL reference conformational energies, resulting from the local DLPNO approximations, the basis set incompleteness and intramolecular superposition errors, and the additonal error from the focal point analysis for the larger subsets is conservatively estimated

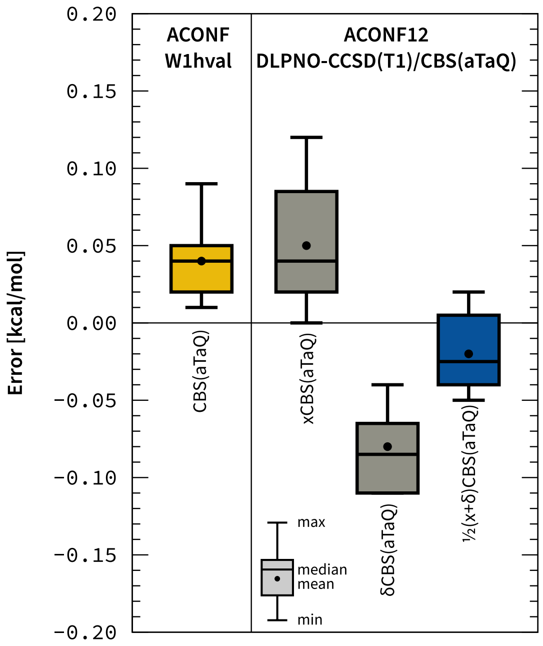


Figure C.4: Deviation of the reference method DLPNO-CCSD(T1)/CBS(aTaQ) against W1hval for the ACONF benchmark set as well as deviation of CBS extrapolations on the ACONF12 benchmark set compared to a CBS(aTaQ) extrapolation scheme. The average conformational energy for the ACONF benchmark is given as 1.83 kcal/mol.

to be 0.35 kcal/mol. This uncertainty of the reference values is largely averaged in the analysis of the statistical descriptors for the entire ACONFL set. The square root of the sum of the squares of the estimated maximum error divided by the number of conformational energies, yielding 0.05 kcal/mol for the ACONFL set, can be used as an estimate for statistically distinguishable values of the analyzed descriptors (see Sec. C.4). With the given accuracy of the reference values, we are thus able to distinguish statistically significant errors of any method above 0.05 kcal/mol.

C.4 Results and discussion

In this section, the performance of all tested methods for the ACONFL set is presented and discussed. Specifically, DFT and the respective dispersion corrections, as well as WFT methods, are assessed in subsection C.4.1, while SQM and FF methods are evaluated in subsection C.4.2. Finally, a performance analysis in terms of computation times vs. accuracy is given in subsection C.4.3. To assess the methods we will mainly discuss the mean absolute error (MAE), the analysis of other statistical quantities, like the mean signed error (MSE), standard deviation (SD), and the error range were investigated as well, but will only be discussed if they show deviating trends from the MAE. The consistency of the conformational ordering is measured by the Pearson r_p and Spearman correlation coefficients r_s, besides the MAE for the conformational energies and correctly identifying the lowest-lying conformer. For the precise definition of the employed statistical measures see the supporting information.

C.4.1 Assessment of DFT and WFT methods

For the following discussion, we selected five (meta-)GGAs, eight hybrids, and four DHDFs, which represent either commonly used or best performing members[23] of the respective functional rungs.[55] Established functionals like PBE,[57,58] TPSS,[146,372] and B3LYP[253,374] are included as well as modern functionals like B97M,[22,62,63] r^{2}SCAN,[4,21,373] and revDSD-BLYP.[378] In the hybrid class of functionals we have included global hybrids like PBE0,[59,60] range-separated hybrids like ωB97M,[61–63] screened exchange hybrids like MN12-SX,[64] as well as local hybrids like Lh20t[32] to access a broad range of different construction strategies in this functional class. We also evaluated wavefunction methods like HF and MP2 in the overall comparison. Finally, we include several composite electronic structure methods of the "3c" scheme, namely B97-3c[79] (GGA), r^{2}SCAN-3c[14] (meta-GGA), PBEh-3c[142] (hybrid), and HF-3c[173] (HF), which use a tailored basis set, in combination with D3 or D4 dispersion correction and the geometrical counter-poise correction (gCP)[386,387] or a short-range basis correction (SRB)[79] to allow efficient yet accurate calculations (for a detailed overview on the "3c" type of methods we refer to Refs. [34] and [14]). All methods tested here were combined with correction schemes to capture long-range dispersion interactions, which are absent in semi-local DFT.[25] We apply the D3 dispersion correction[26,27] with the rational Becke–Johnson (BJ) and zero Chai–Head-Gordon (0) damping schemes, the recently developed D4 dispersion correction,[7,9] and the nonlocal dispersion correction via the VV10 functional in its non-selfconsistent variant.[28,29,62] In case D4 damping parameters were not available, we determined them following the procedure described in Ref. [9].

Before assessing the general performance on the ACONFL, we will investigate the influence of different London dispersion corrections for a selected number of methods, including those which are available with D4, D3, and VV10. First, we want to stress that the MAE for all dispersion corrected methods is well below 1 kcal/mol, while MAEs of non-dispersion corrected functionals are significantly higher yielding an average MAE larger than 2 kcal/mol. Therefore, we will generally only consider dispersion corrected functionals in the following discussion. Dispersion interactions are crucial for the correct description of the investigated alkane conformers and due to their electronically simple structure semi-classical geometry dependent models should be sufficient. To check the influence of the dispersion correction we choose twelve methods for which D3, D4 and VV10 are available. For the tested methods shown in Fig. C.5, we find that in seven cases the D4 corrected variant performs best, while for three methods D3 results in the best performing method, and only in two cases the VV10 corrected functional yields the lowest MAE. Similarly, the average MAE for D4 corrected methods is with 0.29 kcal/mol lowest compared to an average MAE of 0.36 and 0.42 kcal/mol for D3 and VV10 corrected methods, respectively. We find a generally better performance for D4 compared to D3, which is most likely related to the improved parameterization strategy introduced together with D4.[9] Investigating different damping functions for D3, we find a worse performance with the zero-damping and usually an on-par performance with re-parameterized damping functions, for a full comparison see the ESI. This can be seen for a method like r^{2}SCAN where the same parameterization strategy was employed for all three dispersion corrections. Notably, the performance of VV10 with r^{2}SCAN is remarkably good and with an MAE of only 0.18 kcal/mol the best performing method in the (meta-)GGA class while outperforming all tested hybrid functionals. Further, we find for HF-D4 especially good performance with an MAE of 0.14 kcal/mol compared to its D3 and VV10 variant with larger errors, indicating that the consistent parameterization of the dispersion correction is crucial. Overall the D4 dispersion correction shows to be a reliable choice over a wide range of functionals in agreement with previous studies.[9,337,388]

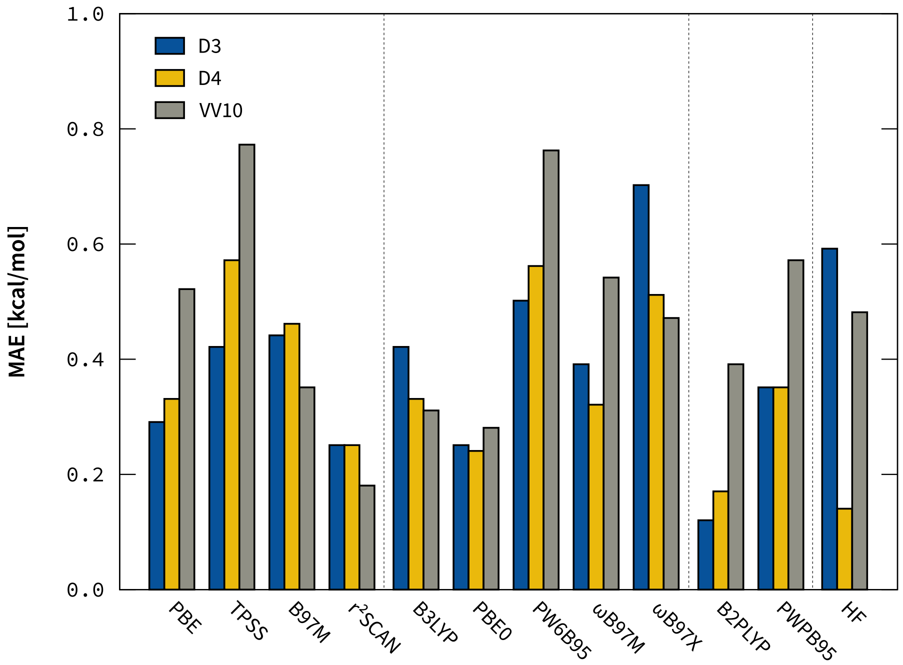


Figure C.5: MAE of all twelve methods available with D4, D3, and VV10 dispersion corrections. The methods are grouped in respective rungs.

To reduce the complexity of the further discussion and focus on the difference in the methods rather than dispersion corrections, we will select the best dispersion correction in each case for the DFT methods discussed in the next paragraphs. The complete statistics for all corrected methods are given in the supporting information. The error spread of all tested DFT, composite DFT and wavefunction methods is shown in Fig. C.6. Notably, many methods are below an error range of ±1 kcal/mol, with the best method DSD-BLYP-D3(BJ) even reaching an error range of only ±0.22 kcal/mol approaching the accuracy of the coupled cluster reference values.

Overall, we find the best performing (meta-)GGA to be the newly developed r^{2}SCAN-V with an MAE of only 0.18 kcal/mol, the second best is B97M-V with 0.35 kcal/mol MAE. In the hybrid class the best performing method is r^{2}SCAN0-V with an MAE of 0.17 kcal/mol, which performs as good as the best (meta-)GGAs. Since the investigated systems are electronically simple, the quality of the base functional is more important than the admixture of Fock exchange here, and for reasons of computational efficiency, a good (meta-)GGA like r^{2}SCAN-V is therefore preferable. We can recover the hierarchy of Jacob's ladder[55] at the highest rung with the DHDFs, where the best-performing method on the entire ACONFL is DSD-BLYP-D3(BJ) with an MAE of only 0.06 kcal/mol.

Notably, HF-D4 performs very well with an MAE of only 0.14 kcal/mol and thus, as good as the best tested hybrid functional. However, while dispersion corrected HF performs well, we find that MP2 at the estimated basis set limit (CBS(aug-TZ/aug-QZ)) results in a large MAE value of 0.59 kcal/mol, *i. e.* worse than most of the assessed dispersion corrected DFT methods, except for some Minnesota-type functionals. The significance of post-MP2 contributions was already observed for n-hexane in the ACONF set.[78] While for MP2 and correlated WFT methods in general, large and diffuse basis sets are

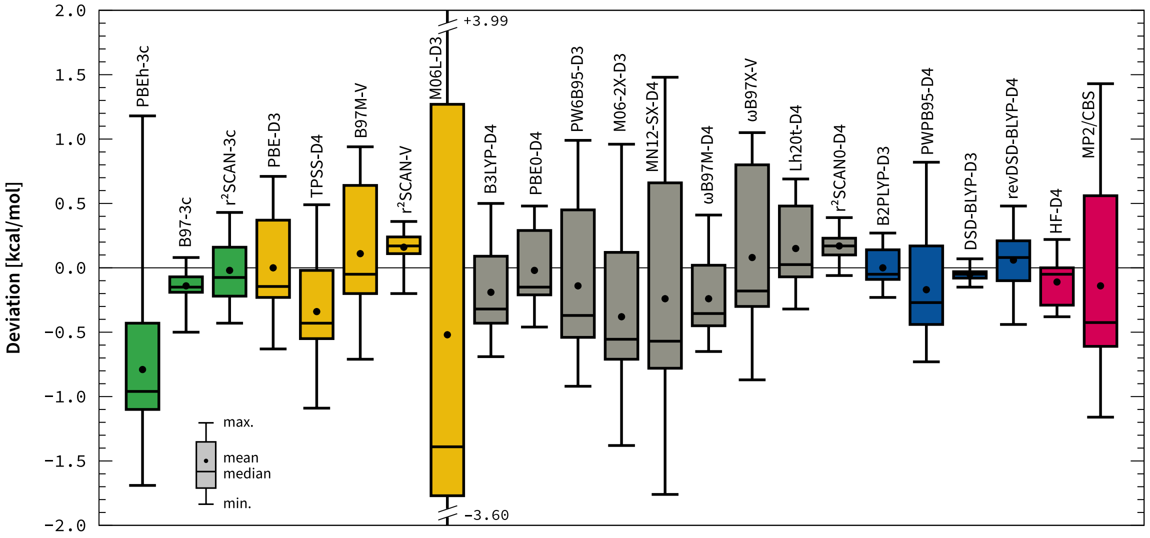


Figure C.6: Deviations for the ACONFL set for selected DFT and WFT methods.

necessary to fully recover long-range dispersion, the physically correct behaviour at the mean field HF level is included more conveniently by a suitable dispersion correction. Moreover, in D4 we can approximately recover three-body dispersion contributions, which would require a higher order treatment than MP2. To further analyse the rather poor performance of MP2, we compare MP2/def2-QZVPP with the recently introduced regularized κOO-MP2 (κ=1.1)/def2-QZVPP, MP2.5[344]/def2-TZVPP, and MP2D[341]/def2-QZVPP for the ACONF12 subset. With the κ regularization and orbital-optimization we only find a small improvement 0.12 kcal/mol in the MAE, as expected for closed-shell systems with large HOMO–LUMO gaps.[343,379] The MAE is reduced to 0.24 kcal/mol by mixing in third-order terms via the MP2.5 scheme containing 50% MP3 correlation energy. The full MP3/def2-TZVPP method yields an outstandingly small MAE of 0.02 kcal/mol due to a fortunate compensation of the MP3 overshooting and the residual BSSE of the triple-ζ basis set. Finally, employing the MP2D approach to correct the uncoupled HF dispersion treatment by DFT-D3 ones reduces the MAE by 0.33 kcal/mol compared to the original MP2. The remaining residual BSSE can be estimated by comparing the QZ results with the CBS result reducing the MAE by 0.17 kcal/mol. Notably, the combination of MP2 and a dispersion correction contribution to recover the proper long-range dispersion also significantly reduces the error of DHDFs (*vide supra*). However, the MAE for ACONF12 with MP2D is still 0.21 kcal/mol higher than for HF-D4 due to residual basis set incompleteness and superposition errors. Comparable to the general performance of the series MP2/MP3/MP4 for NCIs,[344] DLPNO-CCSD overstabilizes the linear structure (MSE of 0.60 kcal/mol), verifying that the connected triples correction including contributions from MP4 and MP5 is essential for accurate coupled cluster results. The comparison of all tested WFT methods for the ACONF12 subset is shown in Fig. C.7.

To assess the potentially larger residual basis set incompleteness and superposition error in DHDF functionals we evaluated PWPB95-D4 with CBS(aTaQ) basis extrapolation and compared the results on the ACONF12 set with the values obtained in the def2-QZVPP basis set. The MAE reduces for this subset reduces from 0.46 to 0.32 kcal/mol, which is a statistically significant improvement. However,

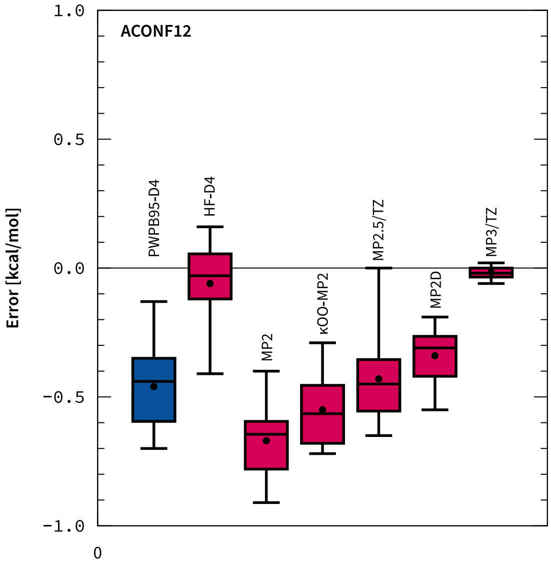


Figure C.7: Comparison of wavefunction methods on the ACONF12 using a def2-QZVPP basis set if not noted otherwise.

DHDFs are usually not extrapolated to the approximated basis set limit (i.e., CBS(aTaQ) due to the increased (about eight times in this case) computational effort compared to the def2-QZVPP calculation. Therefore, we primarily investigated the performance with the commonly applied def2-QZVPP basis set.

Compared to the generally good performance of DFT across all functional classes, we note that only the Minnesota-type functionals tested here show significantly increased deviations, which is rather unusual. While they incorporate short and medium-range dispersion implicitly via their parameterization, the semi-local functionals still cannot fully account for long-range dispersion.[25] However, we find for most functionals of this type the combination with long-range London dispersion corrections is not beneficial for large *n*-alkanes. For example, for MN12-SX the uncorrected functional yields an MAE of 0.48 kcal/mol while that of the D4 corrected functional is with 0.76 kcal/mol significantly larger.

With the best functionals identified close to their basis set limit, we now want to investigate how much of their performance can be recovered with more cost-efficient composite electronic structure methods, namely of the "3c" construction scheme. The "3c" methods are well-suited in a multilevel model scheme to re-rank or re-optimize an ensemble created at a lower level of theory, like SQM or force fields. Both B97-3c and r^{2}SCAN-3c provide a very good description of ACONFL with an MAE of 0.15 and 0.20 kcal/mol, respectively, approaching the accuracy of the best performing methods in a quadruple-ζ basis set. B97-3c is even the best among the tested GGA functionals for this benchmark set. PBEh-3c performs somewhat worse with 0.87 kcal/mol, which can mainly be attributed to the small modified double-ζ basis set and the respective gCP error, while B97-3c and r^{2}SCAN-3c employ a larger modified triple-ζ basis set. Note that the base functional PBE already performs worse compared to modern functionals like r^{2}SCAN. Overall, the composite methods r^{2}SCAN-3c and B97-3c prove to be sufficiently accurate in a very cost-effective way. Therefore we clearly recommend their usage in multilevel workflows, *e. g.*, for conformer ranking of flexible molecules with *n*-alkanes as building blocks.

C.4.2 Assessment of SQM and FF methods

SQM and FF methods are often employed for large scale screening purposes. Due to their much lower computational cost the calculation of large conformational ensembles becomes possible, including challenging tasks like determining the absolute conformational entropy.[389] However, due to approximations inherent to SQM and FF methods the accuracy is often significantly lower and hence a re-ranking of generated ensembles at a higher level of theory becomes necessary.[20] The margin of the energetic threshold to include structures and therefore the amount of structures from the lower level of theory within such refinement workflows is crucial for the overall computational efficiency. Especially, the prediction of the correct energetically most-favorable conformer is important to avoid sorting out structures with major contributions to the final ensemble. The ACONFL provides the chemically most simple yet most flexible molecules for assessing the quality of SQM and FF methods in this context. The deviations of all semiempirical methods and force fields are shown in Fig. C.8 For the three best and the three worst performing methods of this category we also show their conformational energies in Fig. C.9.

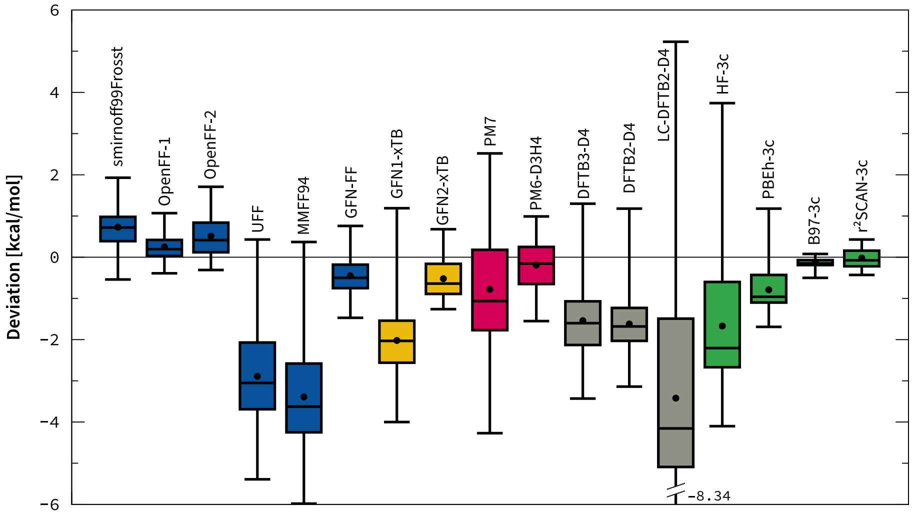


Figure C.8: Comparison of all tested semiempirical quantum mechanical methods and force field methods. The “3c” composite method are included as point of reference to Fig. C.6.

A widely used method is the universal force field (UFF).[82] UFF yields a good correlation ($r_p = 0.98$ and $r_s = 0.95$) but the overall MAE of 2.91 kcal/mol is large given the mean energy of 4.62 kcal/mol. Furthermore, the UFF conformational energies are systematically too small (MSE of −2.89 kcal/mol), indicating an overall too shallow potential energy surface. A strong systematic error and too small conformational energies even with a good correlation results in larger conformational ensembles, which negatively impact the computational cost of later refinement steps at a higher level of accuracy. A similar behavior is observed for the MMFF94 force field with an MAE of 3.41 kcal/mol and also a large negative MSE.

GFN-FF is another general force field which we have tested. It yields very good agreement with the reference, at least for an FF, with an MAE of 0.55 kcal/mol and a good correlation of the conformer ordering ($r_p = 0.97$ and $r_s = 0.96$). Most importantly, it correctly identifies all the lowest-lying

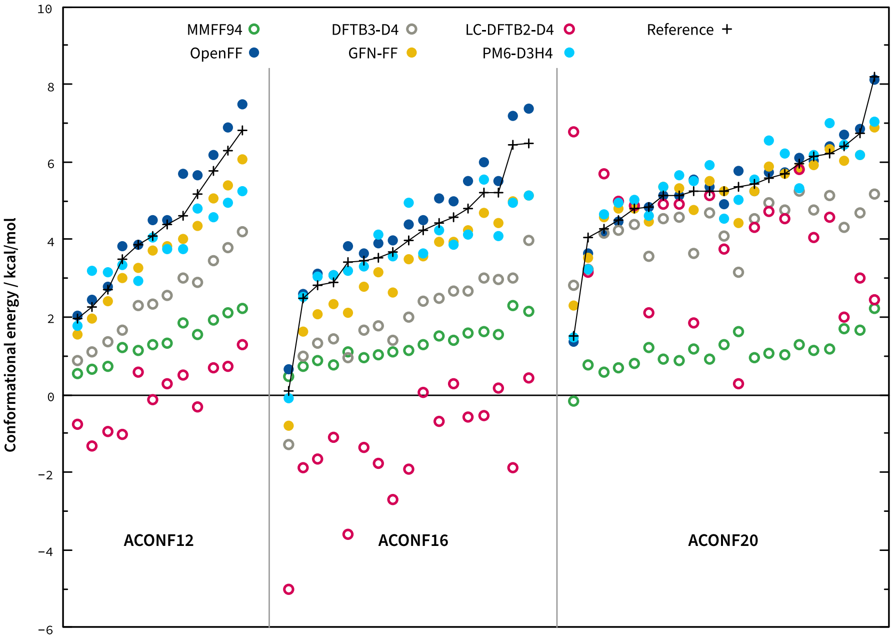


Figure C.9: Conformational energies for the subsets ACONF12, ACONF16, and ACONF20 for three of the best performing and three (filled dots) of the worst performing (open dots) semiempirical and force field methods tested. The reference energies are given as black crosses, the connecting line serves only for better visibility.

conformers in the respective subsets while getting close to the performance of some DFT methods.

From all tested force field methods, the OpenFF-1.0.0 performs best with an MAE of only 0.31 kcal/mol. Also, the SMIRNOFF99Frost and OpenFF-2.0.0 force fields yield small MAEs of 0.76 and 0.56 kcal/mol, respectively. Moreover, all SMIRNOFF methods yield an excellent Pearson correlation coefficient of 0.99. Most force fields can also correctly identify the linear form of *n*-hexadecane to be lower in energy than the lowest-lying folded structure. The individual conformational energies for the OpenFF-1.0.0 and GFN-FF as well as MMFF94 are shown in Fig. C.9, emphasizing the correct conformational ordering produced by the former methods and the too shallow potential energy surface produced by the latter method.

After investigating force fields we will focus on SQM methods as the next more sophisticated level of theory explicitly including electronic structure effects, like the HF based NDDO methods of the PM*x* family and the DFT based tight-binding methods of either the DFTB or xTB flavor. The PM6-D3H4 method provides with an MAE of 0.55 kcal/mol a reasonably accurate description. This good performance seems to be in line with the very good results obtained by dispersion-corrected HF (MAE of 0.14 kcal/mol), on which PM6 is formally based on. However, in contrast its successor PM7 yields with 1.48 kcal/mol a much larger MAE.

From the tested tight binding methods, we find that GFN2-xTB performs best with an MAE of 0.58 kcal/mol. Compared to GFN-FF the GFN2-xTB MAE shows similar performance, however the error range is with 0.29 kcal/mol smaller than for GFN-FF (1.94 compared to 2.23 kcal/mol). Similarly, the error range obtained by GFN2-xTB is by 0.59 kcal/mol smaller than with PM6-D3H4. While the performance of DFTB2-D4 and DFTB3-D4 is quite similar to each other (MAEs of 1.66 and 1.60 kcal/mol, respectively) the better performance in GFN2-xTB may originate from the improved description of the anisotropic electrostatics. In contrast, the GFN1-xTB method performs with an MAE of 2.06 kcal/mol worse, which is could be related to the basis set on hydrogen and the resulting worse description of repulsive NCI contacts. A remarkably weak performer is the LC-DFTB2-D4 method, which introduces spurious large errors in the conformational energies and almost no correlation of the energetic order with the reference ($r_p = 0.37$ and $r_s = 0.36$). Visual inspection of the conformational energies in Fig. C.9 for LC-DFTB2-D4 shows severe errors for each of the subset, where the conformational energies are systematically too low (ACONF12, ACONF16) or spread over a wide range (ACONF20). The large MAE 3.71 kcal/mol results from the distorted potential energy surface. Whether this originates from the parameterization or is a more fundamental problem remains an open question due to lack of alternative long-range corrected DFTB methods to compare with.

Although alkanes are thought to be very prototypical, especially SQM and FFs result in a rather unusual performance order at least for the longer alkane conformers (ACONF16 and ACONF20). Overall, among the force fields and semiempirical methods tested here, GFN-FF provides the best compromise between speed and accuracy.

C.4.3 Performance comparison

Besides the accuracy of the method, an important factor for conformational sampling is its computational cost. For a representative number of methods, we show the computation time to evaluate the whole ACONFL benchmark set, together with their MAE. The wall times were obtained by parallel calculations using four CPU cores and are shown in Fig. C.10. The evaluation of single point energies is representative for a reranking of an ensemble generated by a lower level method in a multilevel workflow, however less suitable for semiempirical methods as those are usually used in the generation of the ensemble in geometry optimizations as additional overhead from the restart or program invocation can be already substantial compared to the total runtime.

Still, the relative time required for the single point evaluation is representative for comparing the computational efficiency of different semiempirical methods with each other. We find that the evaluation of the single point energies for SMIRNOFF methods takes 1.1 min on the entire ACONFL, while the GFN2-xTB method requires less than a second runtime. This difference results from the AM1-BCC charge calculation performed as part of the setup of the SMIRNOFF parameterization. In practice, this calculation has to be done only once per structure, subsequent energy and gradient evaluations will be significantly faster but require proper caching via the compute engine to remain feasible.

Note, the calculation of the reference values at DLPNO-CCSD(T1)/CBS level of theory took about four months cumulative wall time for the whole set.

C.5 Conclusions

We introduced the first benchmark set focusing on the conformational ensembles of long alkane chains, which are a prominent structural motif in many technically and biologically relevant molecules. This new set is termed ACONFL, indicating its relation to the ACONF benchmark introduced by Gruzman *et al.* in 2009,[78] which only includes alkane conformers up to n-hexane. ACONFL comprises conformational ensembles (53 conformers and 50 relative energies up to about 8 kcal/mol) of the n-alkanes $C_{12}H_{24}$,

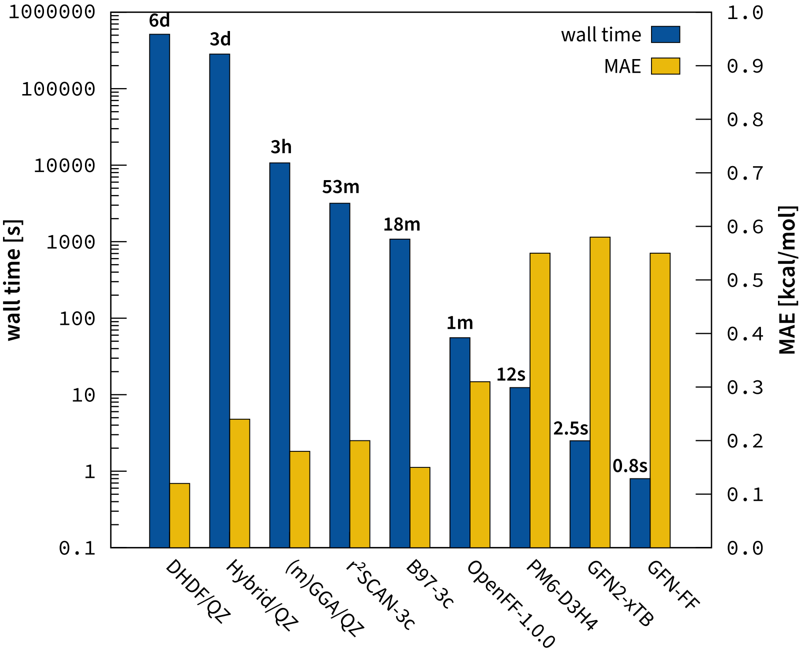


Figure C.10: Wall time for evaluation of the complete ACONFL benchmark set on four Intel Core i7-7700K CPU cores. Due to the vastly different scales present over the wide range of methods assessed here we show the timings in seconds on a logarithmic scale.

$C_{16}H_{34}$, and $C_{20}H_{42}$ that cover the transition from linear to hairpin structures as energetically lowest conformers thus providing a more realistic picture than the ACONF set. We generated reliable reference conformational energies employing high level coupled cluster theory close to the basis set limit (DLPNO-CCSD(T1)/VeryTightPNO/CBS(aug-cc-pVTZ/aug-cc-pVQZ)) allowing for a statistically meaningful evaluation for lower level methods with MAE differences larger than 0.05 kcal/mol.

Using this highly accurate reference data, we explored the performance of a hierarchy of density functionals, the “3c” family of density functional theory (DFT) composite methods, the wavefunction-based approaches HF and MP2, semiempirical approaches (SQM), as well as standard and recent force field (FF) methods. It bears pointing out that of those methods, only the latter (SQM and FF-based) are sufficiently efficient to comprehensively explore the conformational space of these flexible molecules, and are thus indispensable to accurately calculate properties like their absolute entropy.[389]

Concerning the DFT-based methods, we found that (meta-)GGA and hybrid functionals are similarly accurate. In other words, the inclusion Fock exchange does not lead to significant improvements which would justify the increased computational demands. Only DHDFs significantly reduce the error in the conformational energies further. However, even in this case it is questionable whether the small gain in accuracy (0.05 kcal/mol on average) satisfies the massively increased computational cost. The best tested method is DSD-BLYP-D3(BJ) with an MAE of 0.06 kcal/mol while the worst tested functional is M06L-D4 with an MAE of 1.84 kcal/mol. In the ACONFL benchmark we are able to quantify the impact of dispersion, while in the smaller ACONF benchmark set, many dispersion uncorrected functionals perform only slightly worse than their dispersion corrected counterparts. In this respect, the composite DFT methods B97-3c and r^{2}SCAN-3c provide an outstanding cost/accuracy ratio as they perform on part with DFT/QZ methods for a small fraction (about 2–3 orders of magnitude faster compared to DHDF/QZ) of the computational cost.

Regarding correction-schemes in general, we want to point out that for DFT (and also HF) the application of a dispersion correction is crucial for conformational energies of longer alkane chains, which is consistent with previous studies.[327,328,380] Especially the influence of dispersion corrections on the conformational energies cannot be assessed with the smaller ACONF benchmark set[78] alone. Comparing commonly applied dispersion correction schemes, we find that methods with D4 perform on average slightly better (MAE about 0.1 kcal/mol lower) than D3 or VV10 corrected methods. Further, we notice a very good performance for HF-D4 (MAE of 0.14 kcal/mol), which can be attributed to the accurate parameterization of D4 as well as the approximate inclusion of many-body dispersion effects in the latter. While saturated systems with large gaps are usually well described by MP2, we find surprisingly poor performance for MP2/CBS, which we largely attribute to the uncoupled HF dispersion coefficients.[390] This shortcoming of the MP2 can be partially overcome by using the MP2D method, however the performance was found to be still worse compared to computationally less demanding HF-D4 method. Finally, the combination of MP3 in a triple-ζ basis set profits from fortunate error compensation, which makes DHDFs using KS-MP3 correlation[391] worth exploring in the future.

Moving to SQM and FF methods, we find that the inherent additional approximations of those methods also increase the overall error (average MAE 1.55 kcal/mol) compared to DFT significantly. However, GFN2-xTB and PM6-D3H4, the best-performing among the tested SQM methods (0.58 and 0.55 kcal/mol MAE, respectively) are sufficiently accurate to retain the energetic ordering of the conformer ensemble reasonably well. Older standard FFs like UFF and MMFF94 yield generally too shallow potential energy surfaces and, in turn, much too large conformer ensembles in a given energy window. These methods thus require re-ranking and re-optimization at a higher level of theory, which makes them unsuitable in practice, especially if the global energy minimum conformer is incorrectly predicted (*i. e.* preference for the linear over the folded conformer for hexadecane and larger). This also raises the question of whether these common force fields are able to distinguish lipid side-chain conformations crucial for modeling biological systems in solution. Significantly higher accuracy is obtained with the recently introduced GFN-FF and the OpenFF-1.0.0 from the SMIRNOFF FF method, both outperforming all tested SQM methods, even approaching the accuracy of some hybrid DFT/QZ methods (with MAEs of 0.55 and 0.31 kcal/mol, respectively). In our experience, however, the freely available implementation of the SMIRNOFF FFs via QCEngine is not optimal, requiring an overhead of computer time by two orders of magnitude compared to GFN-FF that render their use impractical. Hence, GFN-FF provides both, fast and accurately conformational ensembles and outperformed several SQM methods on this benchmark set, which is quite surprising.

After all, due to its most favorable cost-accuracy ratio, we recommend GFN-FF for conformational searches of alkane conformers for large scale screening applications or to model extended systems with long alkyl chains. However, depending on the other details of the system in question, it may be required to move to a more robust and accurate DFT based method. Here, the efficient composite methods r^{2}SCAN-3c and B97-3c performed particularly well. Although it should be a seemingly straightforward problem for SQM and FF methods due to the simple electronic structure of alkanes, only few of the tested methods performed convincingly and thus we recommend the ACONFL as a helpful fit set for parameterization of new SQM and FF as well as machine learning potentials. Further, the ACONFL provides a meaningful validation set for newly developed DFT and MP2-type WFT methods, especially since the accurate description of conformational energy poses a unique challenge for every investigated ensemble.

Acknowledgement

We thank J.-M. Mewes for carefully proofreading the manuscript and performing the QChem calculations. We thank C. Bannwarth for helpful discussions. This work was supported by the DFG in the framework of the “Gottfried Wilhelm Leibniz Prize” awarded to S. G. and the SPP 2363 on “Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning.”

APPENDIX D

A robust and efficient implicit solvation model for fast semiempirical methods

Sebastian Ehlert,[*] Marcel Stahn,[*] Sebastian Spicher,[*] and Stefan Grimme[*]

Received 12 May 2021, Published online 29 June 2021.

Reprinted in Appendix D (adapted) with permission[†] from
S. Ehlert, U. Huniar, J. Ning, J. W. Furness, J. Sun, A. D. Kaplan, J. P. Perdew, and J. G. Brandenburg, *r^2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications*, J. Chem. Phys. **154** (2021) 061101, DOI: 10.1063/5.0041008.

Abstract We present a robust and efficient method to implicitly account for solvation effects in modern semiempirical quantum mechanics and force-fields. A computationally efficient yet accurate solvation model based on the analytical linearized Poisson–Boltzmann (ALPB) model is parameterized for the extended tight binding (xTB) and density functional tight binding (DFTB) methods as well as for the recently proposed GFN-FF general force-field. The proposed methods perform well over a broad range of systems and applications, from conformational energies over transition-metal complexes to large supramolecular association reactions of charged species. For hydration free energies of small molecules GFN1-xTB(ALPB) is reaching the accuracy of sophisticated explicitly solvated approaches, with a mean absolute deviation of only 1.4 kcal/mol compared to experiment. Logarithmic octanol–water partition coefficients ($\log \mathrm{K}_{ow}$) are computed with a mean absolute deviation of about 0.65 using GFN2-xTB(ALPB) compared to experimental values indicating a consistent description of differential solvent effects. Overall, more than twenty solvents for each of the six semiempirical methods are parameterized and tested. They are readily available in the `xtb` and `dftb+` programs for diverse computational applications.

[*] Mulliken Center of Theoretical Chemistry, Bonn, Germany

[†] Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

D.1 Introduction

Solvation is ubiquitous in biological systems and plays an important role in many aspects of chemistry. Therefore, any computational method targeting to describe structures or interactions under realistic conditions must account for solvation effects. Solute–solvent interactions for example in hydration processes[392,393] or binding free energy computation[394] can be evaluated explicitly by free energy methods like thermodynamic integration[395–397] or free energy perturbation.[398–401] Techniques like metadynamics,[402] umbrella sampling[403] or replica exchange[404,405] allow to enhance the efficiency of the configuration space sampling determining the precision in the free energy computation.[406] Beside molecular dynamic based sampling techniques, Monte-Carlo methods can be used to effectively sample the conformational landscape.[392,407,408] While those methods are a suitable for accurate free energy computation, they require significant computational effort, which can be prohibitive for detailed investigations of chemical reaction mechanisms or high throughput computational workflows.

A practical compromise is usually found in the application of implicit solvation models.[99,409–414] In this approach, the contributions to the solvation free energy are often partitioned into polar (electrostatic) and non-polar (dispersion and cavity) solvation free energies, which allows devising tailored models for each contribution separately. The polar contribution is mostly approximated by conductor-like screening (COSMO)[415] or polarizable continuum models (PCM).[416–419] Implicit solvation models like the conductor-like screening model for real solvents (COSMO-RS)[97–99,420], the solvent models based on solute electron density (SMD)[421] or the three-dimensional reference interaction side model (3D-RISM)[422–424] enable accurate computation of solvation free energies based on standard (mostly DFT) electronic structure input data. However, while low-scaling implementations of COSMO/PCM have been proposed,[425] they still yield a noticeable computational overhead for force-field (FF) or semiempirical quantum mechanical (SQM) methods. For example the PM6 and PM7 methods have been successfully combined with COSMO[426,427] in previous studies using linear scaling algorithms.

A promising alternative are generalized Born (GB) models,[86,428,429] which are used in the reaction field based solvation models (SM), like SM6[90], SM8[430,431] or SM12[432] and for FFs as generalized Born and surface area model (GBSA).[85,88,234] These models allow devising computationally efficient but yet accurate schemes to include solvation effects in large scale simulations.[87]

For the many SQM methods available, robust solvation models are needed, as this is tightly bound to the applicability of the respective methods in computational chemistry or biology.[433,434] One of the most promising members of the family of SQM methods are tight binding (TB) approaches, like density functional tight binding (DFTB)[13,36] or extended tight binding (xTB).[1] While the xTB methods were from the very beginning coupled with an implicit solvation model,[8,41] only a few DFTB implementations account for solvation effects.[268,435–437]

In this work we revisit the implicit solvation models tied with the xTB methods and try to systematically improve the underlying theory as well as the parametrization. Both experimental data from the MNSOL database[90–92] and theoretical values based on COSMO-RS are used as reference data in the fit. To allow for a fair comparison with DFTB, we also implemented, parameterized, and tested the implicit solvation models developed for xTB with three DFTB Hamiltonians. Furthermore, we include the recently devised general force field GFN-FF[45] which is not equipped with a tailored solvation model yet, but employed the GFN2-xTB solvation model in Ref. [45].

We investigate the accuracy of solvation free energies using the implicit solvation models based on the curated FreeSolv database[94,95] for experimental hydration free energies. Since accurate experimental data are usually only available for small compounds, we devise a benchmark set of back-corrected experimental association solvation free energies for large supramolecular complexes as well, building upon the established S30L set.[96] Furthermore, we employ a set of experimental octanol–water partition coefficients for 26 organic compounds and investigate the differential description of two solvents by the new models.

This paper is organized as follows. In section D.2 the theory and algorithms used to implement the models with the TB methods are presented. Technical details of the parametrization and generation of reference data are described in section D.3. In section D.4 performance comparisons for a wide range of systems and benchmarks of the investigated methods are shown. Finally, in section D.5 conclusions and perspectives are presented.

D.2 Theory

To describe a solute in a given solvent or dielectric medium the solvation free energy ΔG_{solv} is partitioned into a polar contribution ΔG_{polar}, which depends on the electrostatic potential, a non-polar contribution ΔG_{npol}, which depends on the shape of the solute cavity and a constant shift ΔG_{shift} depending on the reference state for the solvation process. The solvation free energy is therefore given as

$$\Delta G_{\text{solv}} = \Delta G_{\text{polar}} + \Delta G_{\text{npol}} + \Delta G_{\text{shift}}. \tag{D.1}$$

This partitioning is commonly used in many popular solvation models like GBSA[86,87] or SMD[421] resulting in energy expressions of different complexity for the individual contributions. Changes in the internal free energy of the molecule (rotation, vibration and conformational partition function) upon solvation is mostly not accounted for explicitly but absorbed into the empirical parameters of the model (see section D.4.1).

A suitable form for an efficient evaluation of the polar contributions is the analytical linearized Poisson–Boltzmann (ALPB) model.[85], where the polar contribution is given by

$$\Delta G_{\text{polar}}^{\text{ALPB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{out}}}\right)\frac{1}{1+\alpha\beta}\sum_{A=1}^{N}\sum_{B=1}^{N} q_A q_B \left(\frac{1}{f(R_{AB}, a_A, a_B)} + \frac{\alpha\beta}{\mathcal{A}_{\text{det}}}\right), \tag{D.2}$$

and ϵ_{in} is the dielectric constant of the solute, ϵ_{out} the dielectric constant of the solvent, $q_{A/B}$ are atomic partial charges, f is the interaction kernel, $a_{A/B}$ are the atomic Born radii for atoms A and B and $\mathcal{A}_{\text{det}}$ is the electrostatic size of the solute. The value of α is fixed at 0.571214 following the work of Sigalov *et al.*[85] and β is $\epsilon_{\text{in}}/\epsilon_{\text{out}}$. The electrostatic size $\mathcal{A}_{\text{det}}$ and its derivative with respect to atomic displacement is calculated from the inertia tensor as proposed in Ref. [85], while for ϵ_{in} the dielectric constant of the vacuum, *i. e.*, one, is chosen. Since GB models are derived in the limit of $\epsilon_{\text{out}} \to \infty$ or $\beta \to 0$ we can effectively cast them into the same energy expression as the ALPB model by setting $\alpha\beta$ to zero.

We note that in GFN2-xTB the charge density is expanded up to quadrupoles,[8] while Eq. D.2 only captures the leading term of the polar contribution to the solvation free energy. The solvation model could be improved for GFN2-xTB by generalizing Eq. D.2 to include higher multipole moments for the polar solvation contribution.

In this work we will test two different interaction kernels, first the canonical interaction kernel proposed by Still[86]

$$f_{AB}^{\text{Still}} = \left(R_{AB}^2 + a_A a_B \exp\left[-\frac{R_{AB}^2}{4 a_A a_B} \right] \right)^{\frac{1}{2}} \tag{D.3}$$

and the more recently proposed P16 kernel[88]

$$f_{AB}^{\text{P16}} = R_{AB} + \sqrt{a_A a_B} \left(1 + \frac{1.028 \cdot R_{AB}}{16\sqrt{a_A a_B}} \right)^{16} . \tag{D.4}$$

The evaluation of the interaction kernel requires the atomic Born radii $a_{A/B}$, which are obtained by carrying out an integral over the molecular volume of the solute. For computational efficiency, we employ a pairwise approximate scheme, namely the Onufriev–Bashford–Case (OBC) corrected integrator termed $\text{GB}^{\text{OBC}}\text{II}$[234]. The Born radius in this integrator is calculated by

$$\frac{1}{a_A} = \frac{1}{a_{\text{scale}}} \left(\frac{1}{R_A^{\text{vdw}} - R_{\text{offset}}} - \frac{\tanh[b\Psi_A - c\Psi_A^2 + d\Psi_A^3]}{R_A^{\text{vdw}}} \right), \tag{D.5}$$

where R_A^{vdw} are the D3 van-der-Waals radii[26], a_{scale} is a global scaling parameter for the Born radii, R_{offset} is a global shift parameter to introduce more flexibility for the van-der-Waals radii and Ψ_A is the pairwise approximation to the integral over the molecular volume given by

$$\Psi_A = \frac{R_A^{\text{vdw}} - R_{\text{offset}}}{2} \sum_{B \neq A}^{N} \Omega(R_{AB}, R_A^{\text{vdw}}, s_B \cdot R_B^{\text{vdw}}), \tag{D.6}$$

where Ω is the pairwise contribution to the approximate volume. To compensate for the overestimation of the molecular volume in Ω, an element-specific de-screening parameter s_B is introduced for the van-der-Waals radius of the other atoms. Analytical first derivatives with respect to atomic displacements for all geometry dependent quantities have been derived and implemented.

D.2.1 Non-polar Surface Area Contribution

To account for non-polar contributions to the solvation free energy we include a surface area (SA) model using the solvent-accessible surface area (SASA) to compute the free energy needed to form the solute–solvent cavity and to account for solute–solvent dispersion interactions by

$$\Delta G_{\text{npol}}^{\text{SA}} = \sum_{A}^{N} \gamma_A \sigma_A, \tag{D.7}$$

where γ_A is the surface tension of each atom and σ_A is its SASA. The surface tension is fitted as an element-specific parameter. To evaluate the complete SASA of the molecule we use

$$\sigma_{\text{total}} = \int_V \left| \nabla \left(\prod_A H_A(|\mathbf{R}_A - \mathbf{r}|) \right) \right| d^3\mathbf{r}, \tag{D.8}$$

where $\mathrm{H_A}(|\mathbf{R}_\mathrm{A} - \mathbf{r}|)$ is the atomic volume exclusion function,[89] given by

$$\mathrm{H_A}(r) = \begin{cases} 0, & \text{if} \quad r \le R_A^{\mathrm{surf}} - w \\ \frac{1}{2} + \frac{3}{4w}(r - R_A^{\mathrm{surf}}) - \frac{1}{4w^3}(r - R_A^{\mathrm{surf}})^3, & \text{if} \quad R_A^{\mathrm{surf}} - w < r < R_A^{\mathrm{surf}} + w \\ 1, & \text{if} \quad R_A^{\mathrm{surf}} + w \le r \end{cases} \tag{D.9}$$

The extent of the boundary region is defined by the smoothing parameter w and equals 0.3 Å. The R_A^{surf} are the combined van-der-Waals radii R_A^{vdw} with the solvent probe radius R_{probe}, the latter being a global parameter.

Since the volume exclusion function and its gradient are constant everywhere except for a narrow region around the surface area, we discretize the integral in Eq. D.8 on an angular Lebedev–Laikov grid[438] for an efficient numerical implementation on a sparse surface grid, as

$$\sigma_A = 4\pi \sum_{g}^{N_{\mathrm{ang}}} w_g w_A \prod_{B \ne A}^{N} \mathrm{H_B}(R_{Bg}), \tag{D.10}$$

where w_g and w_A are the angular and radial integration weights, respectively. Analytical derivatives of the surface area with respect to atomic displacements have been implemented following Ref. [89].

D.2.2 Hydrogen Bonding Correction

Specific interactions due to the molecular structure of the solvent, like hydrogen bonding (HB) between the solute and solvent molecules are not accounted for due to the implicit description of the solvent as a dielectric medium. We partition this important interaction component in atom-resolved contributions by

$$\Delta G_{\mathrm{polar}}^{\mathrm{HB}} = \sum_{A}^{N} \Delta G_A^{\mathrm{HB}}. \tag{D.11}$$

To approximate the HB interaction we use the Keesom interaction[439] of hydrogen-acceptor and donor dipoles, μ_{AH}, and μ_D, respectively, with an average distance $\bar{R}$, which is similar to the distance of the first solvation shell. The final interaction term includes the probability $\delta\rho_{\mathrm{solv}}^{\mathrm{HB}}$ that a solvent molecule interaction site is close to the respective atom. This contribution is proportional to the SASA of the interacting atom and therefore, we parameterize the HB contribution as

$$\Delta G_A^{\mathrm{HB}} = -\frac{2}{3} \frac{\mu_{AH}^2 \mu_D^2}{\epsilon_{\mathrm{out}}^2 k_B T \bar{R}^6} \delta\rho_{\mathrm{solv}}^{\mathrm{HB}} \approx -g_A^{\mathrm{HB}} q_A^2 \frac{\sigma_A}{4\pi (R_A^{\mathrm{surf}})^2}, \tag{D.12}$$

where g_A^{HB} is the HB strength, absorbing most of the constants, which is used as an element-specific parameter. Since not all methods have the dipole moment readily available we resort to approximate the quadratic dipole moment by the quadratic site atomic charge instead. Using this formula allows us to include the HB contribution with the polar electrostatic energy as a potential in the Hamiltonian or Lagrangian when minimizing the electrostatic energy with no additional cost as we can reuse the already calculated SASA from the non-polar solvation free energy and its derivative.

D.3 Computational Details

All DFT calculations were conducted with the Turbomole 7.5.1[134] software package. COSMO-RS calculations were done with COSMOtherm 19[97–99] using the BP86[440]/def-TZVPD[347] default level of theory. We employ the 2015, 2016, 2017, 2018 and 2019 versions of the BP86/def-TZVP and BP86/def-TZVPD fine parametrization in this work. The presented solvation models were implemented in the open source software packages `xtb`[1,204] and DFTB+[13,352] and subsequently released with `xtb` version 6.3.2 and DFTB+ version 20.1, which were used to conduct all xTB and DFTB calculations, respectively. For the third-order variant of DFTB, the 3ob parametrization is employed[353–356] and the mio parametrization is used for the second-order DFTB variant.[38,178,357] The ob2-base parametrization is used for the LC-DFTB Hamiltonian.[358] All DFTB Hamiltonians are combined with the D4(EEQ) London dispersion model without three-body contributions[9] using the parameters published in Ref. [13]. Conformational searches and analysis were conducted with the conformer–rotamer ensemble search tool (CREST) version 2.10 using the iMTD-GC algorithm.[19,441] For high-level reference calculations the CRENSO workflow version 1.0.0 was used.[20,442,443]

D.3.1 ALPB Training Set

The training set for the ALPB Solvation model consists of a mixture of experimental and theoretical solvation free energies. For the experimental part, we chose the Minnesota Solvation Database (MNSOL) version 2012[90–92] to serve as reference data. This database includes two types of solvation free energies: absolute solvation free energies and relative solvation free energies between various organic solvents and water. For the parametrization of ALPB, only the absolute solvation free energies were taken. In total the MNSOL Database consists of experimental solvation free energies and optimized geometries for 520 neutral and ionic solutes, including the elements H, C, N, O, F, Si, P, S, Cl, Br, and I. The contained compounds range from small to medium sized organic solutes with a maximum atom count of 46, as well as 31 clustered ions containing a single water molecule. These experimental values were used for the parametrization of the solvents hexadecane, octanol, and water for the mentioned elements, as well as the global empirical parameters in the ALPB model.

Finding sufficient experimental solvation free energies for other elements of the periodic table and other solvents proved to be difficult. For this reason, additionally to the compounds in the MNSOL Database, we used the fit set created for the parametrization of the xTB Hamiltonians as it covers a wide range of elements.[8,41] In total the set used here contains about 2500 compounds for 67 elements. The geometries were optimized with the functional PBEh-3c[142]. The δG_{solv} reference values were obtained using COSMO-RS with the default version 2016 parametrization.[97–99]

D.3.2 Parametrization

Other than the physical constants ε, m, and ϕ, which describe the dielectric constant, the mass and the mass density of the solvent. The solvent mass and mass density are only used when calculating the free energy shift for infinite dilution. ALPB uses two types of parameters: global parameters and element specific parameters. The global parameters are G_{shift} in equation D.1, α_{scale} and R_{offset} in equation D.5 and the probe radius R_{Probe}, which is included in R_A^{surf} in equation D.10. The element specific parameters are s_X in equation D.6, γ_A in equation D.7 and g_A^{HB} in equation D.12. The parameterized solvents are shown in Tab. D.1. The ALPB solvation model employs the P16 interaction kernel, while we use the Still interaction kernel for the GBSA solvation model in this work.

Table D.1: Parameterized solvents for ALPB in combination with GFN1-xTB, GFN2-xTB, GFN-FF, DFTB3-D4, DFTB2-D4, and LC-DFTB2-D4 or GBSA in combination with GFN1-xTB or GFN2-xTB.

	solvents	
acetone	acetonnitrile	aniline[a]
benzaldehyde[a]	benzene	dichlormethane
chloroform	carbondisulfid	dioxane[a]
dimethylformamide[a]	dimethylsulfoxide	ether
ethylacetate[a]	furan[a]	hexadecane[a,c]
hexane[b]	methanol	nitromethane[a]
octanol[a,c]	octanol (wet)[a,c]	phenol[a]
toluene	tetrahydrofuran	water

a: Not parameterized for GFN1-xTB(GBSA) and GFN2-xTB(GBSA),
b: Not parameterized for GFN1-xTB(GBSA),
c: Not parameterized for DFTB3-D4(ALPB), DFTB2-D4(ALPB), LC-DFTB2-D4(ALPB)

For the parameter fit, a fully automated workflow was implemented. Therefore, the compounds of the ALPB test set were split into several subgroups, depending on the contained elements. The reference solvation free energies (MNSOL database or COSMO-RS) refer to the process of transferring the molecule from the gas phase to the liquid phase, and hence, the total molecular energy was first calculated for the gas phase, using the respective method. Subsequently, another full self-consistent calculation for the same gas phase geometry was conducted with added implicit solvation yielding ΔE in equation D.1. In the fitting process, first the global parameters for the elements of the organic subgroup, containing H, C, N, O, F, P, S and Cl, were determined by calculating the difference between the respective free energies in the two states

$$\delta G_{\text{solv}} = G_{\text{gas}} - G_{\text{sol}}. \tag{D.13}$$

This value was then fitted to the reference data utilizing the Levenberg–Marquardt non-linear least squares minimization algorithm[444,445]. The other element-specific parameters were then fitted on a subset for each specific element with all previously determined parameters kept fixed. When developing the workflow, care had to be taken to ensure that the element specific subsets contained only the corresponding element and those that had already been considered. While the structures for the gas phase and solvation calculation are not necessarily identical, we chose to perform both calculations on the gas phase structure without relaxing the geometry. This is mainly done to reduce the computational effort in the parametrization and to avoid potential instability of the fit due to artificial intermediate parameters in the geometry relaxation with implicit solvation. The influence of the geometry relaxation will be discussed in detail in Sec. D.4.1.

D.4 Results

In the following sections, we discuss the performance of the new ALPB solvation models on a broad range of test sets. We also compare the parametrizations of this work with the originally published GFN1-xTB(GBSA)[41] and GFN2-xTB(GBSA)[8] models, which have been shipped with the `xtb` program distribution[1] but were not thoroughly benchmarked so far. As statistical measures we use the mean signed deviation (MSD), mean absolute deviation (MAD), and the standard deviation (SD) and the error range as maximum deviation minus minimum deviation for free (solvation)energies. While the discussion mainly focuses on the MAD we also investigated the MSD, SD and error range for all sets and will discuss those measures if they show deviating behavior from the trends in the MAD. We note that we include the volume work in the solvation free energy when comparing against experimental results rather than the thermostatistical contributions.

D.4.1 Influence of geometry relaxation

To assess the possible error of neglecting geometry relaxations we evaluate the water subset of the MNSOL database. The hydration free energy for all systems was evaluated in three variants of the models. First, by using the gas phase optimized geometry and ignoring geometry relaxations from the implicit solvation model. Second, by relaxing the geometry with the implicit solvation model. And third, by explicitly computing the free energy of the molecule thermostatistically in the modified rigid rotor, harmonic oscillator (mRRHO) approximation[73] with a rotor-cutoff of 50 cm^{-1} to account for the changes in rotational (structure) and vibrational (frequency) contributions. This requires two full geometry optimizations and Hessian calculations.

The results with GFN2-xTB using the parametrization for GBSA and ALPB are shown in Tab. D.2. The overall error range is at 30 kcal/mol for all compounds, while only around 18 kcal/mol for neutral solutes. To better interpret different error sources the set is split in neutral solutes as well as positive and negative ions.

First, we find an overall mean MAD for the hydration free energies of 1.95 and 1.88 kcal/mol for the GBSA and ALPB solvation model, respectively. The account for geometry relaxation only leads to small changes in the MAD for neutral solutes, slightly deteriorating the MAD by less than 0.1 kcal/mol for both models.

For the charged solutes we find a larger MAD in the hydration free energies of 10.05 and 7.06 kcal/mol for GBSA and ALPB solvation models, respectively. While GBSA and ALPB show similar performance for neutral solutes, ALPB represents a significant improvement for the charged solutes due to additional charge dependent terms which are absent in GBSA. A notable observation is that GFN2-xTB(ALPB) reduces the MAD of the hydration free energies by half for cationic solutes compared to GFN2-xTB(GBSA). Furthermore, we find an overall improvement of the hydration free energies when geometry relaxations are included for charged solutes by approximately 0.2 kcal/mol for both models.

Including rotational and vibrational contributions deteriorates the performance for hydration free energies slightly but consistently for both neutral and charged solutes. Tentatively, this slightly diminished accuracy can be attributed to the translational and rotational partition functions of the ideal gas and the rigid rotor which are more approximate for the solvated system. Approaches like a harmonic solvation model[446] or heuristic corrections to the partition function[447] could improve the description but are beyond the scope of this work.

To further investigate the influence of geometry relaxation we select 25 neutral solutes which show

Table D.2: Mean absolute deviation in kcal/mol for the hydration free energies of GFN2-xTB(ALPB) and GFN2-xTB(GBSA) for different evaluation strategies.

		GFN2-xTB(GBSA)			GFN2-xTB(ALPB)		
subset	entries	SP only	opt.	freq.	SP only	opt.	freq.
neutral	390	1.95	2.01	2.08	1.88	1.96	2.07
positive	60	9.49	9.38	9.37	4.65	4.78	4.95
negative	83	11.32	10.92	11.05	9.39	8.98	9.13
all charged	143	10.05	9.78	9.84	7.06	6.85	7.02
all	533	4.26	4.23	4.30	3.36	3.36	3.49

a significant change in the hydration free energy upon optimization in solution. Notable motifs with larger geometry changes are hydroxy, amid and nitro groups as well as sulfur or phosphorous containing groups, *i. e.* especially polar groups. To establish a benchmark set we optimized all 25 compounds using r^{2}SCAN-3c[14] in the gas phase and with DCOSMO-RS[93] in water. To minimize the influence of the underlying electronic structure methods, we compare changes in bond lengths and angles between the gas phase and solvated structure rather than absolute bond lengths and angles. Since changes in the hydration free energy due to geometry relaxations are smaller than 0.5 kcal/mol and the root mean square deviation between the gas phase and the solvated structure is on average only 0.15 Å, small overall geometry changes are expected. In order to put the SQM results into some perspective and to establish a lower error bound we compute the geometry changes in addition with r^{2}SCAN-3c(COSMO).

The MAD for the solvent induced geometry changes with GFN1-xTB, GFN2-xTB, DFTB3-D4 and GFN-FF and the ALPB solvation model as well as PM6-D3H4(COSMO) are shown in Tab. D.3. First, because of the overall small magnitude of geometry changes upon solvation and correspondingly, the small errors involved in this test, the assessment of the methods here should be considered of qualitative nature at most. For the TB methods in combination with ALPB, we find a similar error compared to r^{2}SCAN-3c(COSMO), which we consider as the lower bound for possible errors on this structures. PM6-D3H4(COSMO) performs only slightly worse. A notable exception is GFN-FF(ALPB) which shows doubled errors compared to the SQM methods. From this test we can conclude that the ALPB solvation model semi-quantitatively computes small geometry changes compared to more sophisticated models like DFT(DCOSMO-RS).

Table D.3: MAD in geometry differences for 25 neutral solutes.

	distances [10^{-3} Å]	angles [°]
r^{2}SCAN-3c(COSMO)	1.6	0.14
GFN1-xTB(ALPB)	2.6	0.29
GFN2-xTB(ALPB)	1.7	0.28
GFN-FF(ALPB)	4.5	0.39
DFTB3-D4(ALPB)	2.0	0.20
PM6-D3H4(COSMO)	2.2	0.33

Solvation effects on bond lengths and bond angles are found to be small but can be much larger and chemically relevant when the entire three-dimensional structure of a molecule is considered. For conformational changes, the solvent accessibility of the polar groups in the solute as well as the SASA may change drastically and hence solvation effects can be crucial for the relative energetic ordering in conformational ensembles. As an example, the PES of the antibiotic drug erythromycin was investigated using the recently developed CRENSO[20] workflow with a final conformational energy threshold of 3.0 kcal/mol. The resulting optimized structure ensemble consists of six conformers within this energy window. The solvation free energies were calculated as the difference in energy between the structure in the gas phase and in the liquid phase with full geometry relaxations.

Table D.4 shows the deviation between r^{2}SCAN-3c(COSMO-RS) and the tested methods for calculated free solvation energies for these six conformers. It is apparent, that the solvation free energies, as well as the deviations, significantly differ depending on the investigated conformer. Note, that changes of conformational energies on the order of 1–2 kcal/mol strongly affect thermal populations and average thermal molecular properties. With an MAD of 1.5 kcal/mol and 1.5 kcal/mol as well as an SD of 1.2 kcal/mol, GFN2-xTB(ALPB) and GFN1-xTB(ALPB) perform reasonably well. GFN-FF(ALPB) and GFN2-xTB(GBSA) produce significantly too positive solvation free energies with an MAD of 7.0 kcal/mol and 5.2 kcal/mol. However, the small SD of 0.4 kcal/mol and 1.2 kcal/mol, respectively, indicates rather systematic errors. While GFN1-xTB(GBSA) yields slightly smaller deviations than GFN2-xTB(ALPB) and GFN1-xTB(ALPB) with an MAD of 0.9 kcal/mol, the SD is a bit larger for the latter (1.3 kcal/mol) indicating a slightly lower robustness.

Table D.4: Free solvation energies δG_{solv} in kcal/mol for the six erythromycin conformers calculated with r^{2}SCAN-3c(COSMO-RS) and the deviation δG_{solv}(model)-δG_{solv}(COSMO-RS) for the GFN methods with ALPB and GBSA.

	δG_{solv}			deviation		
	r^{2}SCAN-3c (COSMO-RS)	GFN2-xTB (ALPB)	GFN1-xTB (ALPB)	GFN-FF (ALPB)	GFN2-xTB (GBSA)	GFN1-xTB (GBSA)
1	−27.7	1.3	−1.5	7.4	5.3	−0.1
2	−27.7	1.3	−1.5	7.3	5.3	−0.1
3	−26.5	2.1	−0.1	7.1	6.1	0.9
4	−26.4	2.2	−0.2	7.0	6.1	0.9
5	−23.9	−1.2	−2.0	6.2	2.9	−0.4
6	−28.3	0.8	−3.5	6.8	5.2	−2.7
	MAD	1.5	1.5	7.0	5.2	0.9
	SD	1.2	1.2	0.4	1.2	1.3

We have quantified the impact of geometry relaxations on the hydration free energies and observed only a minor influence for small to medium sized solutes. While we can verify that excluding geometry relaxations in the ALPB parametrization to reduce the computational effort and enhance the stability of the fit is reasonable, we also note that already for medium-sized charged solutes, neglecting geometry relaxations can increase the error in the calculated solvation free energies substantially. This also holds even more for solvation effects on conformational ensembles for flexible molecules, where we refer the reader to Ref. [20] for a more detailed discussion. Thus, for general consistency we recommend to always include geometry relaxation when calculating solvation free energies. Unless noted otherwise all solvation free energies discussed from here on include full geometry relaxations in the respective solvents.

D.4.2 Hydration Free Energies for the FreeSolv database

Water is one of the most commonly used solvents in (bio)chemistry. Yet the description of water is difficult for implicit solvation models due to its high polarity and the importance of HB as well as many-body (polarization) effects. To model chemistry in aqueous solution, an accurate description of the hydration free energies is an important test. We assessed the performance of the solvation models for hydration free energies by comparing our methods on the curated FreeSolv database, which contains currently 642 experimental values for neutral molecules.[94,95] About 250 of these molecules are also contained in the Minnesota Solvation Database, which was used in the fitting process. Starting from the provided geometries the structures were optimized with the respective methods, once with the implicit solvation model, and once in gas phase. We also evaluated the contributions of rotational and vibrational thermostatistical functions to the solvation energies by Hessian calculations for both, the optimized gas phase structure, and the optimized structure in implicit solvation. Yet, only minor effects on the overall statistics were obtained.

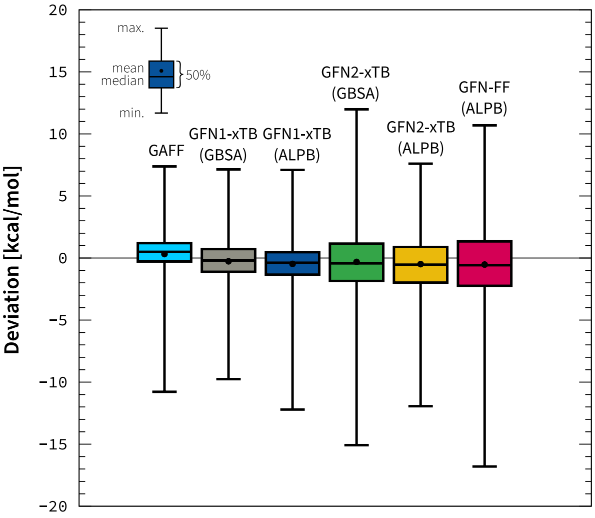


Figure D.1: Statistical analysis of hydration free energies for the neutral species of the FreeSolv database. The values of the GAFF method in explicit water are taken from Ref. [94].

Fig. D.1 shows the deviation of the calculated hydration free energies from the experimental values. Besides the models presented here, we include the explicit solvation approach (GAFF) from Ref. [94], which yields an MAD of 1.1 kcal/mol. In comparison, GFN1-xTB(ALPB) provides only a slightly larger MAD of 1.4 kcal/mol, which is encouraging considering the implicit nature of the solvation model and the parametric treatment of hydrogen-bonding. GFN1-xTB(GBSA) performs best from tested TB methods with an MAD of 1.3 kcal/mol, while GFN2-xTB(ALPB) performs slightly worse with an MAD of 1.8 kcal/mol, which is still reasonably accurate. Overall, the hydration free energies are slightly underestimated with an MSD of −0.5 kcal/mol for GFN2-xTB(ALPB). The GFN2-xTB(GBSA) method performs slightly worse than the ALPB variant with an MAD of 1.9 kcal/mol. The smaller standard deviation of ALPB compared to GBSA for GFN2-xTB with 2.3 kcal/mol and 2.5 kcal/mol, respectively, indicates higher robustness and less outliers. DFTB(3ob)-D4(ALPB) yields an MAD of 1.7 kcal/mol and an MSD of −0.3 kcal/mol, similar to the xTB variants.

With GFN-FF(ALPB) a respectable MAD of 2.2 kcal/mol is obtained, which is larger than for any of the tested SQM methods but still acceptable considering the about hundred-fold speed-up in typical applications. Evaluating the complete database with GFN-FF, including full geometry optimizations for solution and gas phase, takes about 36 s on one core of an Intel Xeon E5-4620 CPU, while the same calculation with GFN2-xTB takes 11 min.

D.4.3 Partition coefficients

An important property to characterize the distribution and accumulation of organic compounds and contaminants in the environment are n-octanol/water (K_{OW}) partition coefficients, which correlate with observed biochemical and toxic effects[448] and are related to internal partitioning between biological tissues and body fluids.[449] While it may be used as a single descriptor in a linear free energy relationship (LFER), different log K relationships are also of interest to form poly-parameter LFERs. Experimental partition coefficients for octanol–water are mostly determined for the transition of a compound from a wet octanol (30% water) phase to a water phase. Partition coefficients can be calculated thermodynamically from the difference in molecular free energy between these two phases. Including thermostatistical contributions (see Section D.4.1), the value is obtained by

$$\log K_{OW} = \frac{1}{k_B T \log e}\left(G_{\text{water}} + \Delta G^{T}_{\text{mRRHO, water}} - \left(G_{\text{octanol}} + \Delta G^{T}_{\text{mRRHO, octanol}}\right)\right), \tag{D.14}$$

where G is the total energy including solvation effects, G^{T}_{mRRHO} is the thermostatistical contribution, k_B is Boltzmann's constant, T is the temperature, and e Euler's number.

Fig. D.2 shows calculated octanol–water partition coefficients with reference to experimental values for 26 typical organic compounds. The calculations for the GFN methods were performed for dry octanol and wet octanol (30% water). For such a complicated property, GFN2-xTB(GBSA) shows reasonably small MAD values of 0.61 and 0.66, for dry and wet octanol, respectively. With an MAD of 0.88 and 0.67, GFN2-xTB(ALPB) performs slightly worse than GFN1-xTB(ALPB) which is in line with the results for the hydration free energy benchmark in section D.4.2. Both methods show the expected slight overestimation of the log K_{ow} values for dry compared to wet octanol.

To classify the results, we also included calculated log K_{OW} values with r^{2}SCAN-3c(COSMO-RS) with the '19 parametrization. Although r^{2}SCAN-3c(COSMO-RS) shows an outlier for cytosin, it yields overall more consistent results with an SD of 0.60 units and a very good MAD of 0.54 units, compared to the GFN methods coupled with the ALPB and GBSA solvation models. While the GFN methods are not able to reach the accuracy of the more sophisticated DFT based solvation model, given the semi-empirical nature of these methods, the results are reasonable and useful for screening or high-throughput studies.

D.4.4 Supramolecular host–guest binding reactions

To assess the accuracy of the TB methods in combination with our parameterized solvation models for larger systems, which are the main target application, we propose a benchmark set of experimental, back-corrected solvation free energies for realistic host–guest binding reactions based on the existing

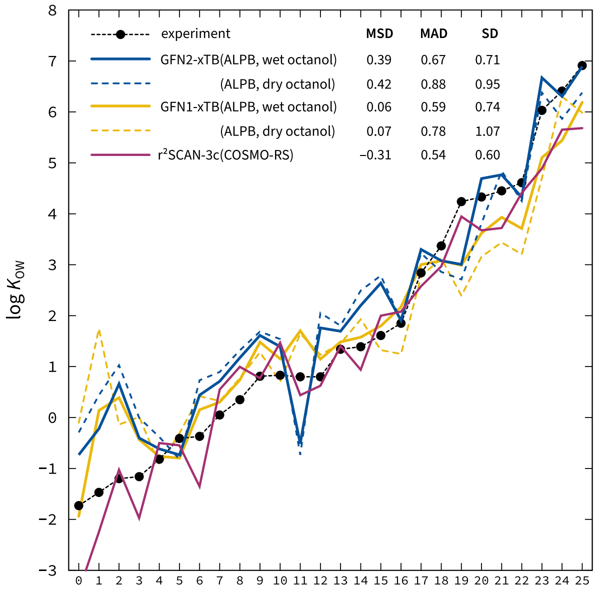


Figure D.2: log K_{OW} partition coefficients of 26 organic compounds ordered according to increasing values. The values are once calculated using wet octanol parameters and once using dry octanol. Statistical measures are given in kcal/mol.

S30L benchmark set.[96] For each of the 30 association reactions host + guest ⟶ guest@host an experimental binding free energy $\Delta G_{\text{a, exp}}$ was taken from the original work. The reference solvation free energies $\Delta\delta G_{\text{solv}}$ are obtained according to

$$\Delta\delta G_{\text{solv}} = \Delta G_{\text{a, exp}} - \Delta E_{\text{a}} - \Delta G^{\text{T}}_{\text{mRRHO}}, \tag{D.15}$$

where ΔE_{a} are accurate DLPNO-CCSD(T) reference values from Ref. [79] and ΔG_{RRHO} are thermostatistical corrections taken from Ref.[450]. The volume work is included in the solvation free energy. The resulting backcorrected association solvation free energies range from +90 kcal/mol for $[\text{Ad}_2(\text{NMe}_3)_2\text{@CB7}]^{2+}$ (24) to −2.7 kcal/mol for AdOH@CB7 (21). The estimated accuracy of the reference values is about 3–4 kcal/mol. For the association reactions studied here, too strong solvation of the individual compounds will result in a too large association solvation energy and therefore a positive MSD. Similarly, too weak solvation of the individual compounds will result in an overall negative MSD. For comparison, results are presented for the COSMO-RS parametrization based on BP86/def2-TZVP densities from 2016 and the more recent 2019 parametrization for BP86/def2-TZVP densities (for both the BP86/def2-TZVPD fine parametrization was employed for water), and the SMD solvation model. The SMD values are based on BP86/def2-SVP calculations and were extracted from the original S30L publication.[96] The solvation free energy contribution to the association free energy for each complex of the S30L is plotted in Fig. D.3.

For the S30L set COSMO-RS(2016) yields an MAD of 2.83 kcal/mol and a small MSD of 0.54 kcal/mol indicating somewhat too strong solvation of the reactants (guest and empty host). Nevertheless, this relatively small MAD (considering the huge range of values) indicates the reliability of the reference

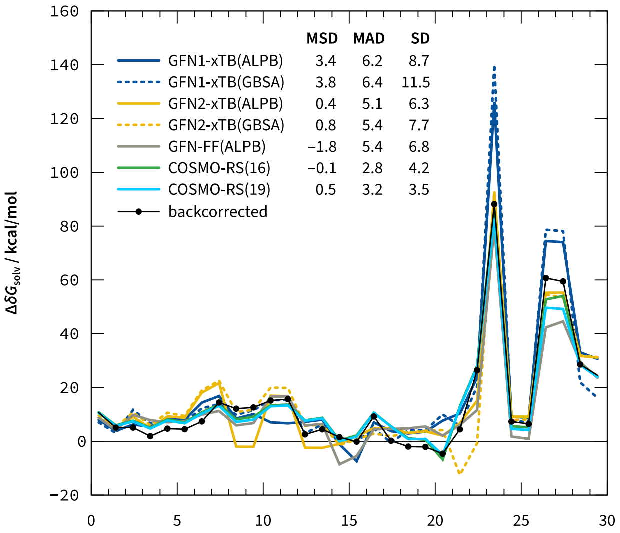


Figure D.3: Deviation of COSMO-RS with 2016 and 2019 parametrization, GFN2-xTB with GBSA and ALPB, GFN1-xTB with GBSA and ALPB and GFN-FF(ALPB) to the back-corrected, experimental association solvation free energies. The statistical measures are given in kcal/mol for each of the selected methods.

values. The results over the whole set differ only slightly compared to the 2015, 2017 and 2018 parametrizations while the 2019 version in contrast shows an increased MAD of 3.2 kcal/mol. The worse performance of the latest COSMO-RS parametrization is mainly caused by the underestimation of the reaction solvation free energies for water with the fine parametrization. For this variant, the MSD is shifted to −0.1 kcal/mol. For SMD an MAD of 3.4 kcal/mol and an MSD close to zero (0.1 kcal/mol) is obtained.

With the GFN2-xTB(ALPB) solvation model a very good MAD of 5.1 kcal/mol is achieved for an SQM method. GFN2-xTB(ALPB) slightly overestimates (MSD of 0.4 kcal/mol) the association solvation free energy. The DFTB(3ob)-D4(ALPB) model shows a rather poor performance with an overall MAD of 7.2 kcal/mol, which can be attributed to the poor description of the mainly dispersion bound complex 1–14, while the remaining systems are described sufficiently well. We also evaluate the DFTB(mio)-D4(ALPB) model for the 27 systems, which can be described with the base mio-parametrization, excluding systems 4, 15 and 16 because of missing parameters for halogens. With an MAD of 7.4 kcal/mol it performs similar to the 3ob-parametrization. Both versions significantly underestimate the values with an MSD of −4.5 kcal/mol and −5.1 kcal/mol, respectively. The LC-DFTB(ob2)-D4(ALPB) model was excluded from this test set due to missing parameters.

The GFN2-xTB(GBSA) model yields an MAD of 5.4 kcal/mol and is slightly worse compared to the new GFN2-xTB(ALPB) model. The trend of overestimating the association solvation energies is also present in the GFN2-xTB(GBSA) model as seen from the MSD of 0.8 kcal/mol. We note that GFN2-xTB(ALPB) reduces the error range significantly compared to GFN2-xTB(GBSA) from 38.8 kcal/mol to only 25.3 kcal/mol.

For GFN1-xTB(GBSA) we exclude systems 15 and 16 due to missing parametrization data for the solvent cyclohexane, for the remaining 28 systems the MAD is 6.4 kcal/mol. GFN1-xTB(ALPB) preforms slightly better with an MAD of 6.2 kcal/mol on the 28 systems and somewhat worse with an

MAD of 6.2 kcal/mol on the complete set compared to GFN2-xTB. Both ALPB and GBSA employed together with GFN1-xTB yield a systematic overestimation of the association solvation free energies with an MSD of 3.4 and 3.8 kcal/mol, respectively. Again, we find that the error range with GFN1-xTB(ALPB) is significantly reduced compared to its GBSA variant from 60.3 to 47.1 kcal/mol. GFN-FF(ALPB) performs with an MAD of 5.4 kcal/mol almost as good as GFN2-xTB(ALPB) on this set. Even the error range for GFN-FF(ALPB) is similarly small with a value of 26.0 kcal/mol compared to the SQM method.

Overall, the the ALPB solvation model together with the GFN methods yields a good description for the solvation contributions for this challenging supramolecular reactions.

D.4.5 Transition Metal Chemistry

The here presented solvation models have been thoroughly investigated for neutral and charged solutes comprised of main group elements. Here we extend the investigation to neutral and charged solutes containing transition metal elements as well.

We evaluate all reaction solvation free energies for the reactions in the MOR41 benchmark set[148] using COSMO-RS for the three representative solvents, water, acetonitrile (ACN) and tetrahydrofuran (THF). The MOR41 benchmark set consists of metal-organic reactions featuring 3d and late transition metals and covers a wide range of possible d-block elements. Since we are comparing with COSMO-RS, we neglect geometry relaxations consistently in the reference method and in the tested solvation models. Due to missing parametrization data the DFTB methods cannot be considered here.

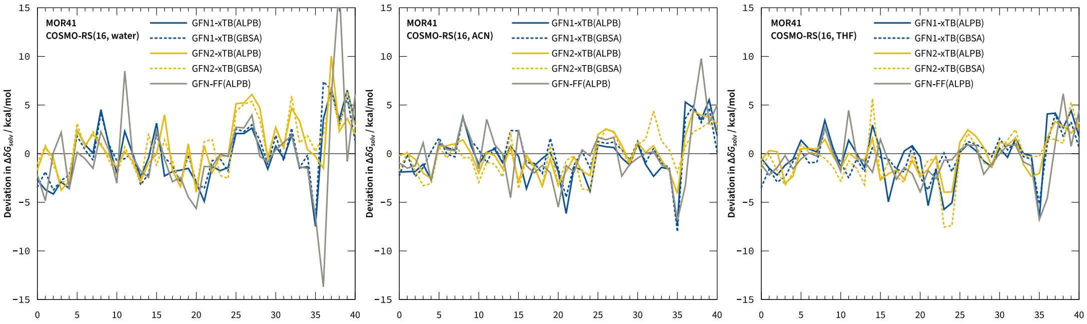


Figure D.4: Left: deviation in the MOR41 reaction solvation free energies from the COSMO-RS reference for water. Center: deviation in the MOR41 reaction solvation free energies from the COSMO-RS reference for acetonitrile (ACN). Right: deviation in the MOR41 reaction solvation free energies from the DCOSMO-RS reference for tetrahydrofuran (THF). PM6-D3H4(COSMO) and PM7(COSMO) were not included in the graphic due the large error range of 25 and 48 kcal/mol, respectively.

The deviation of the tested methods for each of the 41 reaction solvation free energies is shown in Fig. D.4 (left panel). Overall, we find MAD values in the range of 2.3 to 3.0 kcal/mol for the tested solvation models. The best performing methods are GFN2-xTB, with both ALPB and GBSA, and GFN1-xTB(GBSA) all with an MAD of 2.3 kcal/mol. Only the ALPB solvation model for GFN1-xTB gives a slightly larger MAD of 2.5 kcal/mol. The GFN-FF(ALPB) method yields a slightly worse MAD of 3.0 kcal/mol. For comparison we included PM6-D3H4(COSMO) and PM7(COSMO), which perform badly for this kind of systems with an MAD of 3.7 and 5.3 kcal/mol, respectively.

Additionally, we have investigated the same systems for ACN and THF with COSMO-RS and the statistical data are shown in Fig. D.4 (center and right panel). The overall trend of the deviation in the reaction solvation free energies is similar compared to the reaction hydration free energies, while the magnitude of the overall deviation is reduced with the polarity of the solvent. For ACN we find an MAD ranging from 1.6 to 2.1 kcal/mol. GFN2-xTB(ALPB) is performing best for this solvent with an MAD of 1.6 kcal/mol, while the GBSA variant yields a larger MAD of 1.8 kcal/mol. For GFN1-xTB we find a similar good agreement using the GBSA solvation model with an MAD of 1.7 kcal/mol and a slightly deteriorated performance with GFN1-xTB(ALPB) (MAD 1.9 kcal/mol. PM6-D3H4(COSMO) and PM7(COSMO) yield a rather large MAD compared to this of 3.4 and 5.1 kcal/mol. In case of THF as solvent deviations are further reduced with MADs ranging from 1.5 to 1.8 kcal/mol for the presented methods.

The overall performance of the semiempirical methods is reasonable, considering that they inherently yield a much larger error for the reaction energies, as seen in the MAD for the MOR41 set with is 13.2 kcal/mol and 11.8 kcal/mol for GFN1-xTB and GFN2-xTB, respectively, compared to <5 kcal/mol for well performing DFT methods.

Furthermore, we investigated the tetrakis(isonitrile)rhodium(I) cation,[451] which has been previously analyzed under different aspects in theoretical studies.[9,17,452] Due to its relatively high charge it is an interesting and challenging example for the computation of solvation free energies. Here, we focus on the formation of the dication complex from two (mono)cations as shown in Fig. D.5.

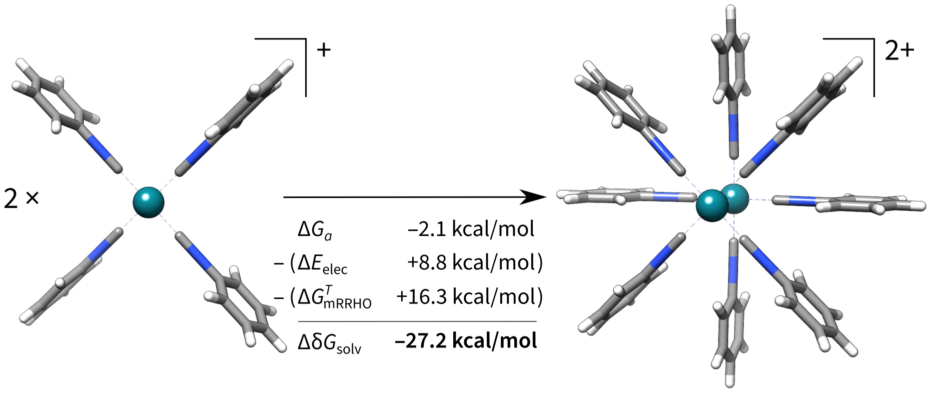

+
2 ×
2+
ΔG_a –2.1 kcal/mol
– (ΔE_{elec} +8.8 kcal/mol)
– ($\Delta G^T_{\text{mRRHO}}$ +16.3 kcal/mol)
$\Delta\delta G_{\text{solv}}$ –27.2 kcal/mol

Figure D.5: Formation of the Rhodium dication complex. The experimental association energy is backcorrected using the electronic association energy ΔE_{elec} at the DLPNO-CCSD(T)/CBS* level of theory and thermostatistical correction to the reaction free energy $\Delta G^{T}_{\text{mRRHO}}$ at the r^{2}SCAN-3c level of theory.

To obtain a backcorrected reaction solvation free energy $\Delta\delta G_{\text{solv}}$ we use the DLPNO-CCSD(T)/CBS* electronic reaction energy ΔE_{elec} of 8.8 kcal/mol taken from Ref. [9] and calculate the reaction at the r^{2}SCAN-3c level to obtain a thermostatistical correction to the reaction free energy $\Delta G^{T}_{\text{mRRHO}}$ of 16.3 kcal/mol. With the experimental association free energy ΔG_{α} of −2.1 kcal/mol[451,452] we obtain a backcorrected reaction solvation free energy $\Delta\delta G_{\text{solv}}$ of −27.2 kcal/mol as our benchmark value. The results are shown in Tab. D.5.

For this example, clear differences are observed between the GBSA and ALPB solvation models. In general, we find that ALPB provides generally less negative reaction solvation free energies compared to GBSA. GFN2-xTB(ALPB) yields a reaction solvation free energy of −26.2 kcal/mol very close to the reference value, while GBSA is slightly over-shooting with −29.8 kcal/mol. A similar trend between GBSA and ALPB is observed for GFN1-xTB, which results in more positive values compared to GFN2-

Table D.5: Reaction solvation free energies in kcal/mol for the formation of the rhodium dication complex.

method	$\Delta\delta G_{solv}$
reference	−27.2
GFN1-xTB(GBSA)	−28.3
GFN2-xTB(GBSA)	−29.8
GFN1-xTB(ALPB)	−23.8
GFN2-xTB(ALPB)	−26.2
GFN-FF(ALPB)	−24.5
PM6-D3H4(COSMO)	−33.3
PM7(COSMO)	−34.0

xTB. GFN1-xTB with GBSA solvation model also reaches a quite good agreement of −28.3 kcal/mol, while the ALPB model is gives an overall to positive reaction solvation free energy of −23.8 kcal/mol. We mainly attribute this difference of 4–5 kcal/mol to the additional charge dependent contributions in the ALPB solvation model, which are absent in most other implicit solvation models. Furthermore, GFN-FF(ALPB) yields a very reasonable value of −24.5 kcal/mol while PM6-D3H4(COSMO) as well as PM7(COSMO) perform rather badly. Overall, the ALPB based solvation models provide a decent description of the solvation effects in this challenging transition metal reaction.

D.5 Conclusion

We presented a fast and computationally efficient solvation model suitable for combination with various tight binding Hamiltonians and even general force fields. A broad range of twenty nonpolar and polar as well as protic and aprotic solvents are readily available. In combination with the GFN family of methods all elements of the periodic table up to Radon ($Z \leq 86$) are covered. For Slater–Koster based DFTB the implicit nature of the solvation model enables the description of systems which are unavailable with the respective parametrizations in an explicit approach.

The resulting methods yield consistent and reasonably accurate solvation free energies for small and large molecules with various solvents. Hydration free energies for a wide range of solutes from the FreeSolv database are in good agreement to the experimental values and close to the accuracy of explicitly solvated approaches which are clearly more elaborate and computationally expensive. Additionally, the consistent description of different solvents has been demonstrated for the accurate computation of partition coefficients, *e. g.*, K_{ow} for octanol and water. For the association energies in the supramolecular S30L benchmark set, also good results close to the backcorrected experimental values were obtained with the xTB models.

The effect of geometry relaxations with implicit solvation models was investigated qualitatively and semi-quantitatively and their importance for medium-sized charged solutes was evaluated. For properties depending on the description of a structural ensemble of flexible solutes such as conformational free energies, the inclusion of solvation effects is indispensable.

The ALPB and GBSA models parameterized here are implemented in the freely available `xtb` and `dftb+` program packages. Based on our tests we can recommend the ALPB solvation model in combination with GFN2-xTB as well as the GBSA solvation model in conjunction with GFN1-xTB as routinely and consistently applicable methods for energy calculations, geometry optimizations, molecular dynamics simulations, and vibrational frequency calculations. Furthermore, we are planning to investigate the generalization of the polar contribution in the solvation model to capture the anisotropic electrostatic of tight-binding models like GFN2-xTB. We are optimistic that the presented solvation models will, together with current and future SQM methods, be valuable in many computational chemistry studies and workflows.

Acknowledgement

The authors thank P. Shushkov for the initial implementation of the GBSA model in `xtb` and A. Hansen for proofreading the manuscript. S. E. thanks B. Hourahine for helpful discussions and code review on the DFTB+ implementations. S. S. thanks the “Fond der chemischen Industrie” (FCI) for financial support. This work was supported by the DFG in the framework of the “Gottfried Wilhelm Leibniz Prize” awarded to S. G.

Bibliography

[1] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, and S. Grimme, *Extended tight-binding quantum chemistry methods*, WIREs Comput. Mol. Sci. **11** (2021) e1493, DOI: 10.1002/wcms.1493.

[2] S. Ehlert, U. Huniar, J. Ning, J. W. Furness, J. Sun, A. D. Kaplan, J. P. Perdew, and J. G. Brandenburg, *r^2SCAN-D4: Dispersion corrected meta-generalized gradient approximation for general chemical applications*, J. Chem. Phys. **154** (2021) 061101, DOI: 10.1063/5.0041008.

[3] S. Ehlert, S. Grimme, and A. Hansen, *Conformational Energy Benchmark for Longer n-Alkane Chains*, J. Phys. Chem. A **126** (2022) 3521, DOI: 10.1021/acs.jpca.2c02439.

[4] S. Ehlert, M. Stahn, S. Spicher, and S. Grimme, *Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods*, J. Chem. Theory Comput. **17** (2021) 4250, DOI: 10.1021/acs.jctc.1c00471.

[5] M. Bursch, H. Neugebauer, S. Ehlert, and S. Grimme, *Dispersion corrected r^2SCAN based global hybrid functionals: r^2SCANh, r^2SCAN0, and r^2SCAN50*, J. Chem. Phys. **156** (2022) 134105, DOI: 10.1063/5.0086040.

[6] P. Zaby, J. Ingenmey, B. Kirchner, S. Grimme, and S. Ehlert, *Calculation of improved enthalpy and entropy of vaporization by a modified partition function in quantum cluster equilibrium theory*, J. Chem. Phys. **155** (2021) 104101, DOI: 10.1063/5.0061187.

[7] E. Caldeweyher, J.-M. Mewes, S. Ehlert, and S. Grimme, *Extension and evaluation of the D4 London-dispersion model for periodic systems*, Phys. Chem. Chem. Phys. **22** (2020) 8499, DOI: 10.1039/D0CP00502A.

[8] 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*, J. Chem. Theory Comput. **15** (2019) 1652, DOI: 10.1021/acs.jctc.8b01176.

[9] E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme, *A generally applicable atomic-charge dependent London dispersion correction*, J. Chem. Phys. **150** (2019) 154122, DOI: 10.1063/1.5090222.

[10] L. Kedward, B. Aradi, O. Certik, M. Curcic, S. Ehlert, P. Engel, R. Goswami, M. Hirsch, A. Lozada-Blanco, V. Magnin, A. Markus, E. Pagone, I. Pribec, B. Richardson, H. Snyder, J. Urban, and J. Vandenplas, *The State of Fortran*, Comput. Sci. Eng. **24** (2022) 63, DOI: 10.1109/MCSE.2022.3159862.

[11] D. G. A. Smith, A. T. Lolinco, Z. L. Glick, J. Lee, A. Alenaizan, T. A. Barnes, C. H. Borca, R. Di Remigio, D. L. Dotson, S. Ehlert, A. G. Heide, M. F. Herbst, J. Hermann, C. B. Hicks, J. T. Horton, A. G. Hurtado, P. Kraus, H. Kruse, S. J. R. Lee, J. P. Misiewicz, L. N. Naden, F. Ramezanghorbani, M. Scheurer, J. B. Schriber, A. C. Simmonett, J. Steinmetzer, J. R. Wagner, L. Ward, M. Welborn, D. Altarawy, J. Anwar, J. D. Chodera, A. Dreuw, H. J. Kulik, F. Liu, T. J. Martínez, D. A. Matthews, H. F. Schaefer, J. Šponer, J. M. Turney, L.-P. Wang, N. De Silva, R. A. King, J. F. Stanton, M. S. Gordon, T. L. Windus, C. D. Sherrill, and L. A. Burns, *Quantum Chemistry Common Driver and Databases (QCDB) and Quantum Chemistry Engine (QCEngine): Automation and interoperability among computational chemistry programs*, J. Chem. Phys. **155** (2021) 204801, DOI: 10.1063/5.0059356.

[12] E. Epifanovsky et al., *Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package*, J. Chem. Phys. **155** (2021) 084801, DOI: 10.1063/5.0055522.

[13] B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu, and T. Frauenheim, *DFTB+, a software package for efficient approximate density functional theory based atomistic simulations*, J. Chem. Phys. **152** (2020) 124101, DOI: 10.1063/1.5143190.

[14] S. Grimme, A. Hansen, S. Ehlert, and J.-M. Mewes, *r^2SCAN-3c: A "Swiss army knife" composite electronic-structure method*, J. Chem. Phys. **154** (2021) 064103, DOI: 10.1063/5.0040021.

[15] 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*, Chem. Eur. J. **27** (2021) 12104, DOI: 10.1002/chem.202100835.

[16] X. Jie, C. G. Daniliuc, R. Knitsch, M. R. Hansen, H. Eckert, S. Ehlert, S. Grimme, G. Kehr, and G. Erker, *Aggregation Behavior of a Six-Membered Cyclic Frustrated Phosphane/Borane Lewis Pair: Formation of a Supramolecular Cyclooctameric Macrocyclic Ring System*, Angew. Chem. Int. Ed. **58** (2019) 882, DOI: 10.1002/anie.201811873.

[17] M. Bursch, E. Caldeweyher, A. Hansen, H. Neugebauer, S. Ehlert, and S. Grimme, *Understanding and Quantifying London Dispersion Effects in Organometallic Complexes*, Acc. Chem. Res. **52** (2019) 258, DOI: 10.1021/acs.accounts.8b00505.

[18] L. Trombach, S. Ehlert, S. Grimme, P. Schwerdtfeger, and J.-M. Mewes, *Exploring the chemical nature of super-heavy main-group elements by means of efficient plane-wave density-functional theory*, Phys. Chem. Chem. Phys. **21** (2019) 18048, DOI: 10.1039/c9cp02455g.

[19] P. Pracht, F. Bohle, and S. Grimme, *Automated exploration of the low-energy chemical space with fast quantum chemical methods*, Phys. Chem. Chem. Phys. **22** (14 2020) 7169, DOI: 10.1039/C9CP06869D.

[20] 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*, J. Phys. Chem. A **125** (2021) 4039, DOI: 10.1021/acs.jpca.1c00971.

[21] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, *Accurate and numerically efficient r2SCAN meta-generalized gradient approximation*, J. Phys. Chem. Lett. **11** (2020) 8208, DOI: 10.1021/acs.jpclett.0c02405.

[22] N. Mardirossian and M. Head-Gordon, *Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V*, J. Chem. Phys. **142** (2015) 074111, DOI: 10.1063/1.4907719.

[23] 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*, Phys. Chem. Chem. Phys. **19** (2017) 32184, DOI: 10.1039/C7CP04913G.

[24] J. Rezac, *Non-covalent interactions atlas benchmark data sets: Hydrogen bonding*, J. Chem. Theory Comput. **16** (2020) 2355, DOI: 10.1021/acs.jctc.9b01265.

[25] S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth, *Dispersion-corrected mean-field electronic structure methods*, Chemical reviews **116** (2016) 5105.

[26] 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*, J. Chem. Phys. **132** (2010) 154104.

[27] S. Grimme, S. Ehrlich, and L. Goerigk, *Effect of the Damping Function in Dispersion Corrected Density Functional Theory*, J. Comput. Chem. **32** (2011) 1456, DOI: 10.1002/jcc.21759.

[28] O. A. Vydrov and T. Van Voorhis, *Nonlocal van der Waals density functional: The simpler the better*, J. Chem. Phys. **133** (2010) 244103, DOI: 10.1063/1.3521275.

[29] W. Hujo and S. Grimme, *Performance of Non-Local and Atom-Pairwise Dispersion Corrections to DFT for Structural Parameters of Molecules with Noncovalent Interactions*, J. Chem. Theory Comput. **9** (2013) 308, DOI: 10.1021/ct300813c.

[30] S. Song, S. Vuckovic, E. Sim, and K. Burke, *Density-corrected DFT explained: Questions and answers*, Journal of chemical theory and computation **18** (2022) 817, DOI: 10.1021/acs.jctc.1c01045.

[31] J. L. Bao, L. Gagliardi, and D. G. Truhlar, *Self-interaction error in density functional theory: An appraisal*, J. Phys. Chem. Lett. **9** (2018) 2353, DOI: 10.1021/acs.jpclett.8b00242.

[32] M. Haasler, T. M. Maier, R. Grotjahn, S. Gückel, A. V. Arbuznikov, and M. Kaupp, *A Local Hybrid Functional with Wide Applicability and Good Balance between (De) Localization and Left–Right Correlation*, J. Chem. Theory Comput. **16** (2020) 5645, DOI: 10.1021/acs.jctc.0c00498.

[33] J. Kirkpatrick, B. McMorrow, D. H. Turban, A. L. Gaunt, J. S. Spencer, A. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, et al., *Pushing the frontiers of density functionals by solving the fractional electron problem*, Science **374** (2021) 1385, DOI: 10.1126/science.abj6511.

[34] E. Caldeweyher and J. G. Brandenburg, *Simplified DFT methods for consistent structures and energies of large systems*, J. Phys. Condens. Matter **30** (2018) 213001, DOI: 10.1088/1361-648x/aabcfb.

[35] J. Pople and D. L. Beveridge, *Approximate Molecular Orbital Theory*, 2nd ed., New York, NY: McGraw-Hill, New York, 1970.

[36] A. S. Christensen, T. Kubař, Q. Cui, and M. Elstner, *Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications*, Chem. Rev. **116** (2016) 5301, DOI: 10.1021/acs.chemrev.5b00584.

[37] M. J. Dewar and W. Thiel, *Ground states of molecules. 38. The MNDO method. Approximations and parameters*, J. Am. Chem. Soc. **99** (1977) 4899, DOI: 10.1021/ja00457a004.

[38] 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*, Phys. Rev. B **58** (11 1998) 7260, DOI: 10.1103/PhysRevB.58.7260.

[39] J. L. Gao, D. G. Truhlar, Y. J. Wang, M. J. M. Mazack, P. Loffler, M. R. Provorse, and P. Rehak, *Explicit polarization: a quantum mechanical framework for developing next generation force fields*, Acc. Chem. Res. **47** (2014) 2837, DOI: 10.1021/ar5002186.

[40] R. Schade, T. Kenter, H. Elgabarty, M. Lass, T. D. Kühne, and C. Plessl, *Breaking the Exascale Barrier for the Electronic Structure Problem in Ab-Initio Molecular Dynamics*, 2022, DOI: 10.48550/ARXIV.2205.12182.

[41] 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)*, J. Chem. Theory Comput. **13** (2017) 1989, DOI: 10.1021/acs.jctc.7b00118.

[42] T. A. Young, J. J. Silcock, A. J. Sterling, and F. Duarte, *autodE: Automated Calculation of Reaction Energy Profiles—Application to Organic and Organometallic Reactions*, Angew. Chem. Int. Ed. **133** (2021) 4312.

[43] A. Alibakhshi and B. Hartke, *Implicitly perturbed Hamiltonian as a class of versatile and general-purpose molecular representations for machine learning*, Nat. Commun. **13** (2022) 1, DOI: 10.1038/s41467-022-28912-6.

[44] V. Sinha, J. J. Laan, and E. A. Pidko, *Accurate and rapid prediction of p K a of transition metal complexes: semiempirical quantum chemistry with a data-augmented approach*, Phys. Chem. Chem. Phys. **23** (2021) 2557, DOI: 10.1039/D0CP05281G.

[45] S. Spicher and S. Grimme, *Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems*, Angew. Chem. Int. Ed. **59** (2020) 15665, DOI: 10.1002/anie.202004239.

[46] E. Schrödinger, Ann. Phys. **384** (1926) 361.

[47] M. Born and R. Oppenheimer, Ann. Phys. **389** (1927) 457.

[48] C. C. J. Roothaan, *New developments in molecular orbital theory*, Reviews of modern physics **23** (1951) 69.

[49] G. Hall, *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** (1951) 541.

[50] C. Møller and M. S. Plesset, *Note on an Approximation Treatment for Many-Electron Systems*, Phys. Rev. **46** (7 1934) 618, DOI: 10.1103/PhysRev.46.618.

[51] A. J. Cohen, P. Mori-Sánchez, and W. Yang, *Challenges for density functional theory*, Chem. Rev. **112** (2012) 289, DOI: 10.1021/cr200107z.

[52] P. Hohenberg and W. Kohn, Phys. Rev. **136** (1964) B864.

[53] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford: Oxford University Press, 1989.

[54] W. Kohn, *Electronic Structure of matter — Wave functions and density functionals*, Rev. Mod. Phys. **71** (1998) 1253.

[55] J. P. Perdew and K. Schmidt, *Jacob's ladder of density functional approximations for the exchange-correlation energy*, AIP Conference Proceedings **577** (2001) 1, DOI: 10.1063/1.1390175.

[56] P. A. Dirac, "Note on exchange phenomena in the Thomas atom", *Mathematical proceedings of the Cambridge philosophical society*, vol. 26, 3, Cambridge University Press, 1930 376, DOI: 10.1017/S0305004100016108.

[57] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized gradient approximation made simple*, Phys. Rev. Lett. **77** (1996) 3865, DOI: 10.1103/PhysRevLett.77.3865.

[58] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized gradient approximation made simple (vol 77, pg 3865, 1996)*, Phys. Rev. Lett. **78** (1997) 1396, DOI: 10.1103/PhysRevLett.78.1396.

[59] C. Adamo and V. Barone, *Toward reliable density functional methods without adjustable parameters: The PBE0 model*, J. Chem. Phys. **110** (1999) 6158, DOI: 10.1063/1.478522.

[60] M. Ernzerhof and G. E. Scuseria, *Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional*, J. Chem. Phys. **110** (1999) 5029, DOI: 10.1063/1.478401.

[61] N. Mardirossian and M. Head-Gordon, ω *B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation*, J. Chem. Phys. **144** (2016) 214110, DOI: 10.1063/1.4952647.

[62] 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*, J. Chem. Theory Comput. **14** (2018) 5725, DOI: 10.1021/acs.jctc.8b00842.

[63] A. Najibi and L. Goerigk, *DFT-D4 counterparts of leading meta-generalized-gradient approximation and hybrid density functionals for energetics and geometries*, J. Comput. Chem. **41** (2020) 2562, doi: 10.1002/jcc.2641.

[64] Y. Zhao and D. G. Truhlar, *The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals*, Theor. Chem. Acc. **120** (2008) 215, doi: 10.1007/s00214-007-0401-8.

[65] Y. Andersson, D. C. Langreth, and B. I. Lundqvist, *van der Waals Interactions in Density-Functional Theory*, Phys. Rev. Lett. **76** (1 1996) 102, doi: 10.1103/PhysRevLett.76.102.

[66] J. F. Dobson and B. P. Dinte, *Constraint satisfaction in local and gradient susceptibility approximations: application to a van der Waals density functional*, Phys. Rev. Lett. **76** (1996) 1780.

[67] E. R. Johnson and A. D. Becke, *A post-Hartree–Fock model of intermolecular interactions*, J. Chem. Phys. **123** (2005) 024101, doi: 10.1063/1.1949201.

[68] A. D. Becke and E. R. Johnson, *A density-functional model of the dispersion interaction*, J. Chem. Phys. **123** (2005) 154101, doi: 10.1063/1.2065267.

[69] B. M. Axilrod and E. Teller, *Interaction of the van der Waals Type Between Three Atoms*, J. Chem. Phys. **11** (1943) 299, doi: 10.1063/1.1723844.

[70] Y. Muto, J. Phys. Math. Soc. Jpn. **17** (1943) 629.

[71] N. D. Mermin, *Thermal properties of the inhomogeneous electron gas*, Phys. Rev. **137** (1965) A1441, doi: 10.1103/PhysRev.137.A1441.

[72] C. A. Bauer, A. Hansen, and S. Grimme, *The fractional occupation number weighted density as a versatile analysis tool for molecules with a complicated electronic structure*, Chemistry–A European Journal **23** (2017) 6150, doi: 10.1002/chem.201604682.

[73] S. Grimme, *Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory*, Chem. Eur. J. **18** (2012) 9955, doi: 10.1002/chem.201200497.

[74] J.-D. Chai and M. Head-Gordon, *Systematic optimization of long-range corrected hybrid density functionals*, J. Chem. Phys. **128** (2008) 084106, doi: 10.1063/1.2834918.

[75] J. Sun, A. Ruzsinszky, and J. P. Perdew, *Strongly constrained and appropriately normed semilocal density functional*, Phys. Rev. Lett. **115** (2015) 036402.

[76] A. P. Bartók and J. R. Yates, *Regularized SCAN functional*, J. Chem. Phys. **150** (2019) 161101, doi: 10.1063/1.5094646.

[77] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Du lak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, *The atomic simulation environment—a Python library for working with atoms*, J. Phys. Condens. Matter **29** (2017) 273002, doi: 10.1088/1361-648x/aa680e.

[78] D. Gruzman, A. Karton, and J. M. L. Martin, *Performance of Ab Initio and Density Functional Methods for Conformational Equilibria of CnH2n+2 Alkane Isomers (n = 4–8)*, J. Phys. Chem. A **113** (2009) 11974, doi: 10.1021/jp903640h.

[79] J. G. Brandenburg, C. Bannwarth, A. Hansen, and S. Grimme, *B97-3c: A revised low-cost variant of the B97-D density functional method*, J. Chem. Phys. **148** (2018) 064104, doi: 10.1063/1.5012601.

[80] P. S Brahmkshatriya, P. Dobes, J. Fanfrlik, J. Rezac, K. Paruch, A. Bronowska, M. Lepsik, and P. Hobza, *Quantum mechanical scoring: structural and energetic insights into cyclin-dependent kinase 2 inhibition by pyrazolo [1, 5-a] pyrimidines*, Current computer-aided drug design **9** (2013) 118, doi: 10.2174/1573409911309010011.

[81] S. Spicher and S. Grimme, *Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems*, Angew. Chem. Int. Ed. **59** (2020) 15665, doi: 10.1002/anie.202004239.

[82] A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III, and W. M. Skiff, *UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations*, J. Am. Chem. Soc. **114** (1992) 10024.

[83] T. A. Halgren, *Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94*, J. Comput. Chem. **17** (1996) 490.

[84] T. A. Halgren, *Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions*, J. Comput. Chem. **17** (1996) 520.

[85] G. Sigalov, A. Fenley, and A. Onufriev, *Analytical electrostatics for biomolecules: Beyond the generalized Born approximation*, J. Chem. Phys. **124** (2006) 124902, doi: 10.1063/1.2177251.

[86] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *Semianalytical treatment of solvation for molecular mechanics and dynamics*, J. Am. Chem. Soc. **112** (1990) 6127, doi: 10.1021/ja00172a038.

[87] A. V. Onufriev and D. A. Case, *Generalized Born Implicit Solvent Models for Biomolecules*, Annu. Rev. Biophys. **48** (2019) 275.

[88] A. W. Lange and J. M. Herbert, *Improving Generalized Born Models by Exploiting Connections to Polarizable Continuum Models. I. An Improved Effective Coulomb Operator*, J. Chem. Theory Comput. **8** (2012) 1999, doi: 10.1021/ct300111m.

[89] W. Im, M. S. Lee, and C. L. Brooks III, *Generalized Born model with a simple smoothing function*, J. Comput. Chem. **24** (2003) 1691, doi: 10.1002/jcc.10321.

[90] C. P. Kelly, C. J. Cramer, and D. G. Truhlar,
SM6: A density functional theory continuum solvation model for calculating aqueous solvation free energies of neutrals, ions, and solute- water clusters,
J. Chem. Theory Comput. **1** (2005) 1133, DOI: 10.1021/ct050164b.

[91] J. D. Thompson, C. J. Cramer, and D. G. Truhlar,
New universal solvation model and comparison of the accuracy of the SM5. 42R, SM5. 43R, C-PCM, D-PCM, and IEF-PCM continuum solvation models for aqueous and organic solvation free energies and for vapor pressures, J. Phys. Chem. A **108** (2004) 6532,
DOI: 10.1021/jp0496295.

[92] A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. Winget, C. J. Cramer, and D. G. Truhlar, *Minnesota Solvation Database-version 2012*, University of Minnesota, Minneapolis (2020).

[93] A. Klamt and M. Diedenhofen, *Calculation of Solvation Free Energies with DCOSMO-RS*, J. Phys. Chem. A **119** (2015) 5439, DOI: 10.1021/jp511158y.

[94] G. Duarte Ramos Matos, D. Y. Kyu, H. H. Loeffler, J. D. Chodera, M. R. Shirts, and D. L. Mobley, *Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database*, J. Chem. Eng. Data **62** (2017) 1559, DOI: 10.1021/acs.jced.7b00104.

[95] D. L. Mobley, M. Shirts, N. Lim, J. Chodera, K. Beauchamp, and Lee-Ping,
FreeSolv: Version 0.52, DOI: 10.5281/zenodo.1161245.

[96] R. Sure and S. Grimme,
Comprehensive Benchmark of Association (Free) Energies of Realistic Host–Guest Complexes,
J. Chem. Theory Comput. **11** (2015) 3785, DOI: 10.1021/acs.jctc.5b00296.

[97] A. Klamt, V. Jonas, T. Bürger, and J. C. W. Lohrenz,
Refinement and Parametrization of COSMO-RS, J. Phys. Chem. A **102** (1998) 5074,
DOI: 10.1021/jp980017s.

[98] A. Klamt,
COSMO-RS : from quantum chemistry to fluid phase thermodynamics and drug design,
Amsterdam Boston: Elsevier, 2005.

[99] A. Klamt, *The COSMO and COSMO-RS solvation models*,
WIREs Comput. Mol. Sci. **1** (2011) 699, DOI: 10.1002/wcms.56.

[100] K. Burke, *Perspective on density functional theory*, J. Chem. Phys. **136** (2012) 150901.

[101] A. D. Becke, *Perspective: Fifty years of density-functional theory in chemical physics*,
J. Chem. Phys. **140** (2014) 18A301, DOI: 10.1063/1.4869598.

[102] R. J. Maurer, C. Freysoldt, A. M. Reilly, J. G. Brandenburg, O. T. Hofmann, T. Bjöorkman, S. Lebègue, and A. Tkatchenko,
Advances in Density-Functional Calculations for Materials Modeling,
Annu. Rev. Mater. Res. **49** (2019) 3.1, DOI: 10.1146/annurev-matsci-070218-010143.

[103] Y. Zhao and D. G. Truhlar, *A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions*,
J. Chem. Phys. **125** (2006) 194101, DOI: 10.1063/1.2370993.

[104] R. Peverati and D. G. Truhlar, *M11-L: A Local Density Functional That Provides Improved Accuracy for Electronic Structure Calculations in Chemistry and Physics*, J. Phys. Chem. Lett. **3** (2012) 117.

[105] N. Mardirossian and M. Head-Gordon, *Characterizing and Understanding the Remarkably Slow Basis Set Convergence of Several Minnesota Density Functionals for Intermolecular Interaction Energies*, J. Chem. Theory Comput. **9** (2013) 4453, DOI: 10.1021/ct400660j.

[106] L. Goerigk, *Treating London-dispersion effects with the latest Minnesota density functionals: problems and possible solutions*, J. Phys. Chem. Lett. **6** (2015) 3891.

[107] E. R. Johnson, A. D. Becke, C. D. Sherrill, and G. A. DiLabio, *Oscillations in meta-generalized-gradient approximation potential energy surfaces for dispersion-bound complexes*, J. Chem. Phys. **131** (2009) 034111, DOI: 10.1063/1.3177061.

[108] 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*, Phys. Chem. Chem. Phys. **16** (2014) 9904, DOI: 10.1039/C3CP54374A.

[109] Y. Wang, X. Jin, H. S. Yu, D. G. Truhlar, and X. He, *Revised M06-L functional for improved accuracy on chemical reaction barrier heights, noncovalent interactions, and solid-state physics*, Proc. Natl. Acad. Sci. **114** (2017) 8487.

[110] K. Sharkas, K. Wagle, B. Santra, S. Akter, R. R. Zope, T. Baruah, K. A. Jackson, J. P. Perdew, and J. E. Peralta, *Self-interaction error overbinds water clusters but cancels in structural energy differences*, Proc. Nat. Acad. Sci., USA **117** (2020) 11283, DOI: 10.1073/pnas.1921258117.

[111] D. Mejia-Rodriguez and S. B. Trickey, *Spin-Crossover from a Well-Behaved, Low-Cost meta-GGA Density Functional*, J. Phys. Chem. A **124** (2020) 9889, DOI: 10.1021/acs.jpca.0c08883.

[112] D. Mejia-Rodriguez and S. Trickey, *Comment on "Regularized SCAN functional"[J. Chem. Phys. 150, 161101 (2019)]*, J. Chem. Phys. **151** (2019) 207101, DOI: 10.1063/1.5120408.

[113] A. P. Bartók and J. R. Yates, *Response to "Comment on 'Regularized SCAN functional'"[J. Chem. Phys. 151, 207101 (2019)]*, J. Chem. Phys. **151** (2019) 207102, DOI: 10.1063/1.5128484.

[114] J. Sun, B. Xiao, Y. Fang, R. Haunschild, P. Hao, A. Ruzsinszky, G. I. Csonka, G. E. Scuseria, and J. P. Perdew, *Density functionals that recognize covalent, metallic, and weak bonds*, Phys. Rev. Lett. **111** (2013) 106401, DOI: 10.1103/PhysRevLett.111.106401.

[115] J. G. Brandenburg, J. E. Bates, J. Sun, and J. P. Perdew, *Benchmark tests of a strongly constrained semilocal functional with a long-range dispersion correction*, Phys. Rev. B **94** (2016) 115144.

[116] J. W. Furness and J. Sun, *Enhancing the efficiency of density functionals with an improved iso-orbital indicator*, Phys. Rev. B **99** (2019) 041119, DOI: 10.1103/PhysRevB.99.041119.

[117] M. Levy and J. P. Perdew,
Hellmann-Feynman, virial, and scaling requisites for the exact universal density functionals. Shape of the correlation potential and diamagnetic susceptibility for atoms,
Phys. Rev. A **32** (1985) 2010, DOI: 10.1103/PhysRevA.32.2010.

[118] A. Görling and M. Levy, *Correlation-energy functional and its high-density limit obtained from a coupling-constant perturbation expansion*, Phys. Rev. B **47** (1993) 13105,
DOI: 10.1103/PhysRevB.47.13105.

[119] L. Pollack and J. P. Perdew,
Evaluating density functional performance for the quasi-two-dimensional electron gas,
J. Phys. Condens. Matter **12** (2000) 1239, DOI: 10.1088/0953-8984/12/7/308.

[120] P. Svendsen and U. von Barth,
Gradient expansion of the exchange energy from second-order density response theory,
Phys. Rev. B **54** (1996) 17402, DOI: 10.1103/PhysRevB.54.17402.

[121] J. P. Perdew and Y. Wang,
Accurate and simple analytic representation of the electron-gas correlation energy,
Phys. Rev. B **45** (1992) 13244, DOI: 10.1103/PhysRevB.45.13244.

[122] T. Aschebrock and S. Kümmel,
Ultranonlocality and accurate band gaps from a meta-generalized gradient approximation,
Phys. Rev. Research **1** (2019) 033082, DOI: 10.1103/PhysRevResearch.1.033082.

[123] F. Hofmann and S. Kümmel, *Molecular excitations from meta- generalized gradient approximations in the Kohn – Sham scheme Molecular excitations from meta-generalized gradient approximations in the Kohn – Sham scheme*, J. Chem. Phys. **153** (2020) 114106,
DOI: 10.1063/5.0023657.

[124] D. Mejia-Rodriguez and S. B. Trickey, *Meta-GGA performance in solids at almost GGA cost*,
Phys. Rev. B **102** (12 2020) 121109, DOI: 10.1103/PhysRevB.102.121109.

[125] J. Klimeš and A. Michaelides, *Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory*, J. Chem. Phys. **137** (2012) 120901.

[126] J. Hermann, R. A. DiStasio, and A. Tkatchenko, *First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications*,
Chem. Rev. **117** (2017) 4714, DOI: 10.1021/acs.chemrev.6b00446.

[127] K. Berland, V. R. Cooper, K. Lee, E. Schröder, T. Thonhauser, P. Hyldgaard, and B. I. Lundqvist,
van der Waals forces in density functional theory: a review of the vdW-DF method,
Rep. Prog. Phys. **78** (2015) 066501.

[128] R. Sabatini, T. Gorni, and S. de Gironcoli,
Nonlocal van der Waals density functional made simple and efficient,
Phys. Rev. B **87** (4 2013) 041108, DOI: 10.1103/PhysRevB.87.041108.

[129] E. R. Johnson and A. D. Becke,
A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections,
J. Chem. Phys. **124** (2006) 174104, DOI: 10.1063/1.2190220.

[130] J. Řezáč, K. E. Riley, and P. Hobza, *S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures*, J. Chem. Theory Comput. **7** (2011) 2427, DOI: 10.1021/ct2002946.

[131] M. S. Marshall, L. A. Burns, and C. D. Sherrill, *Basis set convergence of the coupled-cluster correction, δ MP2 CCSD (T): Best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases*, J. Chem. Phys. **135** (2011) 194102.

[132] D. E. Taylor, J. G. Ángyán, G. Galli, C. Zhang, F. Gygi, K. Hirao, J. W. Song, K. Rahul, O. Anatole von Lilienfeld, R. Podeszwa, et al., *Blind test of density-functional-based methods on intermolecular interaction energies*, J. Chem. Phys. **145** (2016) 124105.

[133] F. Furche, R. Ahlrichs, C. Hättig, W. Klopper, M. Sierka, and F. Weigend, *Turbomole*, WIREs Comput. Mol. Sci. **4** (2014) 91, DOI: 10.1002/wcms.1162.

[134] *TURBOMOLE V7.5 2020, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.org.*

[135] K. Eichkorn, O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs, *Auxiliary basis sets to approximate Coulomb potentials*, Chem. Phys. Lett. **240** (1995) 283.

[136] 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*, Theor. Chem. Acc. **97** (1997) 119.

[137] F. Weigend, F. Furche, and R. Ahlrichs, *Gaussian basis sets of quadruple zeta quality for atoms H to Kr*, J. Chem. Phys. **119** (2003) 12753.

[138] G. Kresse and J. Hafner, *Ab initio molecular dynamics for liquid metals*, Phys. Rev. B **47** (1993) 558.

[139] G. Kresse and J. Furthmüller, *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*, J. Comp. Mat. Sci. **6** (1996) 15.

[140] P. E. Blöchl, *Projector augmented-wave method*, Phys. Rev. B **50** (1994) 17953.

[141] G. Kresse and J. Joubert, *From ultrasoft pseudopotentials to the projector augmented-wave method*, Phys. Rev. B **59** (1999) 1758.

[142] S. Grimme, J. G. Brandenburg, C. Bannwarth, and A. Hansen, *Consistent structures and interactions by density functional theory with small atomic orbital basis sets*, J. Chem. Phys. **143** (2015) 054107, DOI: 10.1063/1.4927476.

[143] M. Bühl and H. Kabrede, *Geometries of Transition-Metal Complexes from Density-Functional Theory*, J. Chem. Theory Comput. **2** (2006) 1282.

[144] 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*, J. Phys. Chem. A **119** (2015) 2058.

[145] É. Brémond, M. Savarese, N. Q. Su, Á. J. Pérez-Jiménez, X. Xu, J. C. Sancho-García, and C. Adamo, *Benchmarking Density Functionals on Structural Parameters of Small-/Medium-Sized Organic Molecules*, J. Chem. Theory Comput. **12** (2016) 459.

[146] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, *Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids*, Phys. Rev. Lett. **91** (2003) 146401, DOI: 10.1103/PhysRevLett.91.146401.

[147] For a subset of the GMTKN55 not all of the published benchmark results with the SCAN functional were reproducible. Therefore, we reevaluated SCAN on the complete GMKTN55.

[148] S. Dohm, A. Hansen, M. Steinmetz, S. Grimme, and M. P. Checinski, *Comprehensive Thermochemical Benchmark Set of Realistic Closed-Shell Metal Organic Reactions*, J. Chem. Theory Comput. **14** (2018) 2596, DOI: 10.1021/acs.jctc.7b01183.

[149] P. Pernot and A. Savin, *Using the Gini coefficient to characterize the shape of computational chemistry error distributions*, arXiv preprint arXiv:2012.09589 (2020).

[150] R. Sedlak, T. Janowski, M. Pitoňák, J. Řezáč, P. Pulay, and P. Hobza, *Accuracy of Quantum Chemical Methods for Large Noncovalent Complexes*, J. Chem. Theory Comput. **9** (2013) 3364.

[151] M. K. Kesharwani, D. Manna, N. Sylvetsky, and J. M. L. Martin, *The X40×10 Halogen Bonding Benchmark Revisited: Surprising Importance of (n–1)d Subvalence Correlation*, J. Phys. Chem. A **122** (2018) 2184, DOI: 10.1021/acs.jpca.7b10958.

[152] Y. S. Al-Hamdani, P. R. Nagy, D. Barton, M. Kállay, J. G. Brandenburg, and A. Tkatchenko, *Interactions between Large Molecules: Puzzle for Reference Quantum-Mechanical Methods*, 2020.

[153] A. Otero-de-la-Roza, B. H. Cao, I. K. Price, J. E. Hein, and E. R. Johnson, *Predicting the Relative Solubilities of Racemic and Enantiopure Crystals by Density-Functional Theory*, Angew. Chem. Int. Ed. **53** (2014) 7879, DOI: https://doi.org/10.1002/anie.201403541.

[154] A. M. Reilly and A. Tkatchenko, *Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals*, J. Chem. Phys. **139** (2013) 024705, DOI: 10.1063/1.4812819.

[155] G. J. O. Beran, *Modeling Polymorphic Molecular Crystals with Electronic Structure Theory*, Chem. Rev. **116** (2016) 5567.

[156] A. Zen, J. G. Brandenburg, J. Klimeš, A. Tkatchenko, D. Alfè, and A. Michaelides, *Fast and accurate quantum Monte Carlo for molecular crystals*, Proc. Nat. Acad. Sci., USA **115** (2018) 1724.

[157] A. Otero-De-La-Roza and E. R. Johnson, *A benchmark for non-covalent interactions in solids*, J. Chem. Phys. **137** (2012) 054103, DOI: 10.1063/1.4738961.

[158] A. Ambrosetti, A. M. Reilly, R. A. DiStasio, and A. Tkatchenko, *Long-range correlation energy calculated from coupled atomic response functions*, J. Chem. Phys. **140** (2014) 18A508.

[159] D. J. Carter and A. L. Rohl, *Benchmarking Calculated Lattice Parameters and Energies of Molecular Crystals Using van der Waals Density Functionals*, J. Chem. Theory Comput. **10** (2014) 3423.

[160] J. G. Brandenburg, T. Maas, and S. Grimme, *Benchmarking DFT and Semiempirical Methods on Structures and Lattice Energies for Ten Ice Polymorphs*, J. Chem. Phys. **142** (2015) 124104.

[161] B. Santra, J. Klimeš, A. Tkatchenko, D. Alfè, B. Slater, A. Michaelides, R. Car, and M. Scheffler, *On the Accuracy of van der Waals Inclusive Density-Functional Theory Exchange-Correlation Functionals for Ice at Ambient and High Pressures*, J. Chem. Phys. **139** (2013) 154702.

[162] J. G. Brandenburg, A. Zen, D. Alfè, and A. Michaelides, *Interaction between water and carbon nanostructures: How good are current density functional approximations?*, J. Chem. Phys. **151** (2019) 164702, DOI: 10.1063/1.5121370.

[163] K. Wagle, B. Santra, P. Bhattarai, C. Shahi, M. R. Pederson, K. A. Jackson, and J. P. Perdew, *Self-Interaction Correction in Water-Ion Clusters*, arXiv preprint arXiv:2012.13469 (2020).

[164] S. Wen and G. J. O. Beran, *Accurate Molecular Crystal Lattice Energies from a Fragment QM/MM Approach with On-the-Fly Ab Initio Force Field Parametrization*, J. Chem. Theory Comput. **7** (2011) 3733.

[165] M. R. Kennedy, A. R. McDonald, A. E. DePrince, M. S. Marshall, R. Podeszwa, and C. D. Sherrill, *Communication: Resolving the Three-Body Contribution to the Lattice Energy of Crystalline Benzene: Benchmark Results from Coupled-Cluster Theory*, J. Chem. Phys. **140** (2014) 121104.

[166] W. J. Hehre, R. Ditchfield, and J. A. Pople, J. Chem. Phys. **56** (1972) 2257.

[167] K. N. Houk and F. Liu, *Holy grails for computational organic chemistry and biochemistry*, Acc. Chem. Res. **50** (2017) 539, ISSN: 15204898, DOI: 10.1021/acs.accounts.6b00532.

[168] S. Grimme and P. R. Schreiner, *Computational Chemistry: The Fate of Current Methods and Future Challenges*, Angew. Chem. Int. Ed. **57** (2017) 4170, DOI: 10.1002/anie.201709943.

[169] A. Otero-de-la-Roza and G. A. DiLabio, *Transferable Atom-Centered Potentials for the Correction of Basis Set Incompleteness Errors in Density-Functional Theory*, J. Chem. Theory Comput. **13** (2017) 3505, DOI: 10.1021/acs.jctc.7b00300.

[170] 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*, J. Chem. Phys. **146** (2017) 234105, DOI: 10.1063/1.4986962.

[171] J. Hostaš and J. Řezáč, *Accurate DFT-D3 Calculations in a Small Basis Set*, J. Chem. Theory Comput. **13** (2017) 3575, DOI: 10.1021/acs.jctc.7b00365.

[172] H. J. Kulik, N. Seelam, B. D. Mar, and T. J. Martinez, *Adapting DFT+U for the Chemically Motivated Correction of Minimal Basis Set Incompleteness*, J. Phys. Chem. A **120** (2016) 5939, DOI: 10.1021/acs.jpca.6b04527.

[173] R. Sure and S. Grimme, *Corrected small basis set Hartree-Fock method for large systems*, J. Comput. Chem. **34** (2013) 1672, DOI: 10.1002/jcc.23317.

[174] S. Ehrlich, A. H. Göller, and S. Grimme, *Towards full quantum-mechanics-based protein–ligand binding affinities*, ChemPhysChem **18** (2017) 898, DOI: 10.1002/cphc.201700082.

[175] W. Thiel, *Semiempirical quantum–chemical methods*, WIREs Comput. Mol. Sci. **4** (2014) 145, DOI: 10.1002/wcms.1161, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1161.

[176] G. Seifert, D. Porezag, and T. Frauenheim, *Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme*, Int. J. Quantum Chem. **58** (1996) 185, DOI: 10.1002/(SICI)1097-461X(1996)58:2<185::AID-QUA7>3.0.CO;2-U.

[177] 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*, Phys. Rev. B **51** (19 1995) 12947, DOI: 10.1103/PhysRevB.51.12947.

[178] M. Gaus, Q. Cui, and M. Elstner, *DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)*, J. Chem. Theory Comput. **7** (2011) 931, DOI: 10.1021/ct100684s.

[179] R. Rüger, M. Franchini, T. Trnka, A. Yakovlev, E. van Lenthe, P. Philipsen, T. van Vuren, B. Klumpers, and T. Soini, AMS 2019.3, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, http://www.scm.com.

[180] T. D. Kühne, M. Iannuzzi, M. D. Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borstnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, and J. Hutter, *CP2K: An Electronic Structure and Molecular Dynamics Software Package – Quickstep: Efficient and Accurate Electronic Structure Calculations*, 2020, arXiv: 2003.03868 [physics.chem-ph].

[181] J. Řezáč, *Cuby: An integrative framework for computational chemistry*, J. Comput. Chem. **37** (2016) 1230, DOI: 10.1002/jcc.24312.

[182] F. Manby, T. Miller, P. Bygrave, F. Ding, T. Dresselhaus, F. Batista-Romero, A. Buccheri, C. Bungey, S. Lee, R. Meli, and et al., *entos: A Quantum Molecular Simulation Package*, 2019, DOI: 10.26434/chemrxiv.7762646.v2.

[183] F. Neese, *The ORCA program system*, WIREs Comput. Mol. Sci. **2** (2012) 73, DOI: 10.1002/wcms.81.

[184] F. Neese, *Software update: the ORCA program system, version 4.0*, WIREs Comput. Mol. Sci. **8** (2018) e1327, DOI: 10.1002/wcms.1327.

[185] I. S. Ufimtsev and T. J. Martinez, *Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics*, J. Chem. Theory Comput. **5** (2009) 2619, DOI: 10.1021/ct9003004.

[186] N. Luehr, I. S. Ufimtsev, and T. J. Martinez, *Dynamic Precision for Electron Repulsion Integral Evaluation on Graphical Processing Units (GPUs)*, J. Chem. Theory Comput. **7** (2011) 949, DOI: 10.1021/ct100701w.

[187] P. Pracht, C. A. Bauer, and S. Grimme, *Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites*, J. Comput. Chem. **38** (2017) 2618, DOI: 10.1002/jcc.24922.

[188] S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert, and F. Neese, *Fully Automated Quantum-Chemistry-Based Computation of Spin–Spin-Coupled Nuclear Magnetic Resonance Spectra*, Angew. Chem. Int. Ed. **56** (2017) 14763, DOI: 10.1002/anie.201708266.

[189] J. Seibert, C. Bannwarth, and S. Grimme, *Biomolecular Structure Information from High-Speed Quantum Mechanical Electronic Spectra Calculation*, J. Am. Chem. Soc. **139** (2017) 11682, ISSN: 15205126, DOI: 10.1021/jacs.7b05833.

[190] L. Wilbraham, E. Berardo, L. Turcani, K. E. Jelfs, and M. A. Zwijnenburg, *High-Throughput Screening Approach for the Optoelectronic Properties of Conjugated Polymers*, J. Chem. Inf. Model. **58** (2018) 2450, DOI: 10.1021/acs.jcim.8b00256.

[191] C. A. Bauer, G. Schneider, and A. H. Göller, *Gaussian process regression models for the prediction of hydrogen bond acceptor strengths*, Mol. Inf. **38** (2019) 1800115, DOI: 10.1002/minf.201800115.

[192] P. Pracht, E. Caldeweyher, S. Ehlert, and S. Grimme, *A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules*, ChemRxiv (2019), DOI: 10.26434/chemrxiv.8326202, DOI: 10.26434/chemrxiv.8326202.

[193] A. K. Rappé and W. A. Goddard III, *Charge Equilibration for Molecular Dynamics Simulation*, J. Chem. Phys. **95** (1991) 3358.

[194] C. E. Wilmer, K. Chul Kim, and R. Q. Snurr, *An Extended Charge Equilibration Method*, J. Phys. Chem. Lett. **3** (2012) 2056.

[195] Z.-T. Cheng, T.-R. Shan, T. Liang, R. K. Behera, S. R. Phillpot, and S. B. Sinnott, *A charge optimized many-body (comb) potential for titanium and titania*, J. Phys.: Condens. Matter **26** (2014) 315007, DOI: 10.1088/0953-8984/26/31/315007.

[196] 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*, Phys. Rev. B **92** (4 2015) 045131, DOI: 10.1103/PhysRevB.92.045131.

[197] S. Grimme, *A simplified Tamm–Dancoff density functional approach for the electronic excitation spectra of very large molecules*, J. Chem. Phys. **138** (2013) 244104, DOI: 10.1063/1.4811331.

[198] J. Seibert, J. Pisarek, S. Schmitz, C. Bannwarth, and S. Grimme, *Extension of the element parameter set for ultra-fast excitation spectra calculation (sTDA-xTB)*, Mol. Phys. **117** (2019) 1104, ISSN: 13623028, DOI: 10.1080/00268976.2018.1510141.

[199] M. de Wergifosse and S. Grimme, *Nonlinear-response properties in a simplified time-dependent density functional theory (sTD-DFT) framework: Evaluation of the first hyperpolarizability*, J. Chem. Phys. **149** (2018) 024108, DOI: 10.1063/1.5037665, eprint: https://doi.org/10.1063/1.5037665.

[200] M. de Wergifosse and S. Grimme, *Nonlinear-response properties in a simplified time-dependent density functional theory (sTD-DFT) framework: Evaluation of excited-state absorption spectra*, J. Chem. Phys. **150** (2019) 094112, DOI: 10.1063/1.5080199, eprint: https://doi.org/10.1063/1.5080199.

[201] M. de Wergifosse, C. Bannwarth, and S. Grimme, *A Simplified Spin-Flip Time-Dependent Density Functional Theory Approach for the Electronic Excitation Spectra of Very Large Diradicals*, J. Phys. Chem. A **123** (2019), PMID: 31199632 5815, DOI: 10.1021/acs.jpca.9b03176, eprint: https://doi.org/10.1021/acs.jpca.9b03176.

[202] M. de Wergifosse, J. Seibert, B. Champagne, and S. Grimme, *Are Fully Conjugated Expanded Indenofluorenes Analogues and Diindeno[n]thiophene Derivatives Diradicals? A Simplified (Spin-Flip) Time-Dependent Density Functional Theory [(SF-)sTD-DFT] Study*, J. Phys. Chem. A **123** (2019) 9828, ISSN: 15205215, DOI: 10.1021/acs.jpca.9b08474.

[203] J. Seibert, B. Champagne, S. Grimme, and M. de Wergifosse, *Dynamic Structural Effects on the Second-Harmonic Generation of Tryptophane-Rich Peptides and Gramicidin A*, J. Phys. Chem. B **124** (2020), PMID: 32148035 2568, DOI: 10.1021/acs.jpcb.0c00643, eprint: https://doi.org/10.1021/acs.jpcb.0c00643.

[204] "Semiempirical Extended Tight-Binding Program Package xtb", https://github.com/grimme-lab/xtb. Accessed: 2021-05-03.

[205] 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*, J. Phys. Chem. A **111** (2007) 10861, DOI: 10.1021/jp074167r.

[206] R. M. Parrish, F. Liu, and T. J. Martinez, *Communication: A difference density picture for the self-consistent field ansatz*, J. Chem. Phys. **144** (2016) 131101, DOI: 10.1063/1.4945277.

[207] J. Harris, *Simplified method for calculating the energy of weakly interacting fragments*, Phys. Rev. B **31** (1985) 1770, DOI: 10.1103/PhysRevB.31.1770.

[208] W. M. C. Foulkes and R. Haydock, *Tight-binding models and density-functional theory*, Phys. Rev. B **39** (1989) 12520, DOI: 10.1103/PhysRevB.39.12520.

[209] J. E. Jones, *On the determination of molecular fields. –II. From the equation of state of a gas*, Proc. R. Soc. A **106** (1924) 463, DOI: 10.1098/rspa.1924.0082.

[210] R. A. Buckingham, *The classical equation of state of gaseous helium, neon and argon*, Proc. R. Soc. A **168** (1938) 264, DOI: 10.1098/rspa.1938.0173.

[211] 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*, J. Chem. Phys. **51** (1969) 2657, DOI: 10.1063/1.1672392.

[212] T. Bredow, G. Geudtner, and K. Jug, *Development of the cyclic cluster approach for ionic systems*, J. Comput. Chem. **22** (2001) 89, DOI: 10.1002/1096-987X(20010115)22:1<89::AID-JCC9>3.0.CO;2-7.

[213] M. Mantina, R. Valero, C. J. Cramer, and D. G. Truhlar, "Atomic Radii of the Elements", *CRC Handbook of Chemistry and Physics, 91nd edition*, ed. by W. M. Haynes, Boca Raton, FL: CRC Press, 2010 9-49.

[214] K. Nishimoto and N. Mataga, *Electronic Structure and Spectra of Some Nitrogen Heterocycles*, Z. Phys. Chem. **12** (1957) 335, DOI: 10.1524/zpch.1957.12.5_6.335.

[215] K. Ohno, *Some Remarks on the Pariser-Parr-Pople Method*, Theor. Chim. Act. **2** (1964) 219, DOI: 10.1007/BF00528281.

[216] G. Klopman, *A Semiempirical Treatment of Molecular Structures. II. Molecular Terms and Application to Diatomic Molecules*, J. Am. Chem. Soc. **86** (1964) 4450, DOI: 10.1021/ja01075a008.

[217] M. Korth, *Empirical Hydrogen-Bond Potential Functions—An Old Hat Reconditioned*, Chem. Phys. Chem. **12** (2011) 3131, DOI: 10.1002/cphc.201100540.

[218] A. M. Köster, M. Leboeuf, and D. R. Salahub, "Molecular Electrostatic Potentials from Density Functional Theory", *Molecular Electrostatic Potentials*, ed. by J. S. Murray and K. Sen, vol. 3, Theo. Comput. Chem. Elsevier, 1996 105, DOI: 10.1016/S1380-7323(96)80042-2.

[219] P. Pyykkö and M. Atsumi, *Molecular Single-Bond Covalent Radii for Elements 1–118*, Chem. Eur. J. **15** (2009) 186, DOI: 10.1002/chem.200800987.

[220] M. Hülsen, A. Weigand, and M. Dolg, *Quasirelativistic energy-consistent 4f-in-core pseudopotentials for tetravalent lanthanide elements*, Theor. Chem. Account **122** (2009) 23, DOI: 10.1007/s00214-008-0481-0.

[221] C. Bannwarth and S. Grimme, *A simplified time-dependent density functional theory approach for electronic ultraviolet and circular dichroism spectra of very large molecules*, Comput. Theor. Chem. **1040–1041** (2014) 45, DOI: 10.1016/j.comptc.2014.02.023.

[222] A. Schäfer, C. Huber, and R. Ahlrichs, *Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr*, J. Chem. Phys. **100** (1994) 5829, DOI: 10.1063/1.467146, eprint: https://doi.org/10.1063/1.467146.

[223] A. Hellweg, S. A. Grün, and C. Hättig, *Benchmarking the performance of spin-component scaled CC2 in ground and electronically excited states*, Phys. Chem. Chem. Phys. **10** (2008) 4119.

[224] E. Lindahl and M. S. Sansom, *Membrane proteins: molecular dynamics simulations*, Curr. Opin. Struct. Biol. **18** (2008) 425, DOI: 10.1016/j.sbi.2008.02.003.

[225] D. Fujita, Y. Ueda, S. Sato, N. Mizuno, T. Kumasaka, and M. Fujita, *Self-assembly of tetravalent Goldberg polyhedra from 144 small components*, Nature **540** (2016) 563, DOI: 10.1038/nature20771.

[226] J. R. Long and O. M. Yaghi, *The pervasive chemistry of metal–organic frameworks*, Chem. Soc. Rev. **38** (2009) 1213, DOI: 10.1039/B903811F.

[227] J. Åqvist, C. Medina, and J.-E. Samuelsson, *A new method for predicting binding affinity in computer-aided drug design*, Protein Eng. Des. Sel. **7** (1994) 385, DOI: 10.1093/protein/7.3.385.

[228] F. Ooms, *Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry*, Curr Med. Chem. **7** (2000) 141, DOI: 10.2174/0929867003375317.

[229] I. J. Enyedy, Y. Ling, K. Nacro, Y. Tomita, X. Wu, Y. Cao, R. Guo, B. Li, X. Zhu, Y. Huang, et al., *Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening*, J. Med. Chem. **44** (2001) 4313, DOI: 10.1021/jm010016f.

[230] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, *Development and testing of a general amber force field*, J. Comput. Chem. **25** (2004) 1157, DOI: 10.1002/jcc.20035.

[231] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, *Comparison of multiple Amber force fields and development of improved protein backbone parameters*, Proteins **65** (2006) 712, DOI: 10.1002/prot.21123.

[232] D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley, and W. Sherman, *Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field*, J. Chem Theory Comput. **6** (2010) 1509, DOI: 10.1021/ct900587b.

[233] A. D. MacKerell Jr, N. Banavali, and N. Foloppe, *Development and current status of the CHARMM force field for nucleic acids*, Biopolymers **56** (2000) 257, DOI: 10.1002/1097-0282(2000)56:4<257::AID-BIP10029>3.0.CO;2-W.

[234] A. Onufriev, D. Bashford, and D. A. Case, *Exploring protein native states and large-scale conformational changes with a modified generalized born model*, Proteins **55** (2004) 383, DOI: 10.1002/prot.20033.

[235] V. Ásgeirsson, C. A. Bauer, and S. Grimme, *Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules*, Chem. Sci. **8** (7 2017) 4879, DOI: 10.1039/C7SC00601B.

[236] P. M. Zimmerman, *Growing string method with interpolation and optimization in internal coordinates: Method and examples*, J. Chem. Phys. **138** (2013) 184102, DOI: 10.1063/1.4804162.

[237] P. Zimmerman, *Reliable transition state searches integrated with the growing string method*, J. Chem. Theory Comput. **9** (2013) 3043, DOI: 10.1021/ct400319w.

[238] S. Dohm, M. Bursch, A. Hansen, and S. Grimme, *Semiautomated Transition State Localization for Organometallic Complexes with Semiempirical Quantum Chemical Methods*, J. Chem. Theory Comput. **16** (2020) 2002, DOI: 10.1021/acs.jctc.9b01266.

[239] *Conformer-Rotamer Ensemble Sampling Tool based on the xtb Semiempirical Extended Tight-Binding Program Package* `crest`, https://github.com/grimme-lab/crest, Accessed: 2021-12-28.

[240] F. Bohle, M. Bursch, E. Caldeweyher, S. Dohm, S. Ehlert, J. Koopman, H. Neugebauer, P. Pracht, K. Schmitz, S. Schmitz, J. Seibert, S. Spicher, and S. Grimme., *User Guide to Semiempirical Tight Binding*, https://xtb-docs.readthedocs.io/en/latest/contents.html, Accessed: 2020-04-08.

[241] G. Brandl, *Sphinx Python Documentation Generator*, https://www.sphinx-doc.org/en/master/, 2020-04-09, 2007.

[242] A. Project, *Asciidoctor: A fast text processor & publishing toolchain for converting AsciiDoc to HTML5, DocBook & more*, https://asciidoctor.org/, 2020-04-09, 2002.

[243] R. Lindh, A. Bernhardsson, G. Karlström, and P.-Å. Malmqvist, *On the use of a Hessian model function in molecular geometry optimizations*, Chem. Phys. Lett. **241** (1995) 423, DOI: 10.1016/0009-2614(95)00646-L.

[244] F. Eckert, P. Pulay, and H.-J. Werner, *Ab initio geometry optimization for large molecules*, J. Comput. Chem. **18** (1997) 1473, DOI: 10.1002/(SICI)1096-987X(199709)18:12<1473::AID-JCC5>3.0.CO;2-G.

[245] C. G. Broyden, *The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations*, IMA J. Appl. Math. **6** (1970) 76, DOI: 10.1093/imamat/6.1.76.

[246] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Mathematical programming **45** (1989) 503, DOI: doi.org/10.1007/BF01589116.

[247] E. W. Dijkstra et al., *A note on two problems in connexion with graphs*, Numerische mathematik **1** (1959) 269, DOI: 10.1007/BF01386390.

[248] M. Bursch, A. Hansen, and S. Grimme, *Fast and Reasonable Geometry Optimization of Lanthanoid Complexes with an Extended Tight Binding Quantum Chemical Method*, Inorg. Chem. **56** (2017) 12485, DOI: 10.1021/acs.inorgchem.7b01950.

[249] S. Grimme and M. Steinmetz, *Effects of London dispersion correction in density functional theory on the structures of organic molecules in the gas phase*, Phys. Chem. Chem. Phys. **15** (38 2013) 16031, DOI: 10.1039/C3CP52293H.

[250] T. Risthaus, M. Steinmetz, and S. Grimme, *Implementation of nuclear gradients of range-separated hybrid density functionals and benchmarking on rotational constants for organic molecules*, J. Comput. Chem. **35** (2014) 1509, DOI: 10.1002/jcc.23649.

[251] J. J. P. Stewart, *Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements*, J. Mol. Model. **13** (2007) 1173, DOI: 10.1007/s00894-007-0233-4.

[252] J. Řezáč and P. Hobza, *Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods*, J. Chem. Theory Comput. **8** (2012) 141, DOI: 10.1021/ct200751e.

[253] A. D. Becke, *Density-functional thermochemistry. III. The role of exact exchange*, J. Chem. Phys. **98** (1993) 5648, DOI: 10.1063/1.464913.

[254] 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*, Phys. Chem. Chem. Phys. **7** (2005) 3297, DOI: 10.1039/B508541A.

[255] P. Jurečka, J. Sponer, J. Cerny, 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*, Phys. Chem. Chem. Phys. **8** (2006) 1985, DOI: 10.1039/B600027D.

[256] L. Gráfová, M. Pitoňák, J. Řezáč, and P. Hobza, *Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set*, J. Chem. Theory Comput. **6** (2010) 2365, DOI: 10.1021/ct1002253.

[257] J. Řezáč, K. E. Riley, and P. Hobza, *Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules*, J. Chem. Theory Comput. **8** (2012) 4285, DOI: 10.1021/ct300647k.

[258] 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*, J. Chem. Phys. **119** (2003) 12129, DOI: 10.1063/1.1626543.

[259] J. J. P. Stewart, *Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters*, J. Mol. Model. **19** (2012) 1, DOI: 10.1007/s00894-012-1667-x.

[260] Y. Cui, Y. Yue, G. Qian, and B. Chen, *Luminescent Functional Metal–Organic Frameworks*, Chemi. Rev. **112** (2012) 1126, DOI: 10.1021/cr200101d.

[261] D. A. Pantazis, X.-Y. Chen, C. R. Landis, and F. Neese, *All-Electron Scalar Relativistic Basis Sets for Third-Row Transition Metal Atoms*, J. Chem. Theory Comput. **4** (2008) 908, DOI: 10.1021/ct800047t.

[262] D. Aravena, F. Neese, and D. A. Pantazis, *Improved Segmented All-Electron Relativistically Contracted Basis Sets for the Lanthanides*, J. Chem. Theory Comput. **12** (2016) 1148, DOI: 10.1021/acs.jctc.5b01048.

[263] J. L. Banks, H. S. Beard, Y. Cao, A. E. Cho, W. Damm, R. Farid, A. K. Felts, T. A. Halgren, D. T. Mainz, J. R. Maple, et al., *Integrated modeling program, applied chemical theory (IMPACT)*, J. Comput. Chem. **26** (2005) 1752, DOI: 10.1002/jcc.20292.

[264] D. Q. McDonald and W. C. Still, *AMBER torsional parameters for the peptide backbone*, Tetrahedron Lett. **33** (1992) 7743, DOI: https://doi.org/10.1016/0040-4039(93)88034-G.

[265] D. M. Ferguson and P. A. Kollman, *Can the Lennard–Jones 6-12 function replace the 10-12 form in molecular mechanics calculations?*, J. Comput. Chem. **12** (1991) 620, DOI: 10.1002/jcc.540120512.

[266] S. Schmitz, J. Seibert, K. Ostermeir, A. Hansen, A. H. Göller, and S. Grimme, *Quantum Chemical Calculation of Molecular and Periodic Peptide and Protein Structures*, J. Phys. Chem. B **124** (2020) 3636, DOI: 10.1021/acs.jpcb.0c00549.

[267] M. Korth and W. Thiel, *Benchmarking Semiempirical Methods for Thermochemistry, Kinetics, and Noncovalent Interactions: OMx Methods Are Almost As Accurate and Robust As DFT-GGA Methods for Organic Molecules*, J. Chem. Theory Comput. **7** (2011), PMID: 26605482 2929, DOI: 10.1021/ct200434a, eprint: https://doi.org/10.1021/ct200434a.

[268] J. C. Kromann, C. Steinmann, and J. H. Jensen, *Improving solvation energy predictions using the SMD solvation method and semiempirical electronic structure methods*, J. Chem. Phys. **149** (2018) 104102.

[269] C. Riplinger, B. Sandhoefer, A. Hansen, and F. Neese, *Natural triple excitations in local coupled cluster calculations with pair natural orbitals*, J. Chem. Phys. **139** (2013) 134101.

[270] C. Riplinger, P. Pinski, U. Becker, E. F. Valeev, and F. Neese, *Sparse maps – A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory*, J. Chem. Phys. **144** (2016) 024109.

[271] H. Kruse, A. Mladek, K. Gkionis, A. Hansen, S. Grimme, and J. Sponer, *Quantum chemical benchmark study on 46 RNA backbone families using a dinucleotide unit*, J. Chem. Theory Comput. **11** (2015) 4972, DOI: 10.1021/acs.jctc.5b00515.

[272] M. Marianski, A. Supady, T. Ingram, M. Schneider, and C. Baldauf, *Assessing the Accuracy of Across-the-Scale Methods for Predicting Carbohydrate Conformational Energies for the Examples of Glucose and α-Maltose*, J. Chem. Theory Comput. **12** (2016) 6157, DOI: 10.1021/acs.jctc.6b00876.

[273] T. Weymuth, E. P. A. Couzijn, P. Chen, and M. Reiher, *New Benchmark Set of Transition-Metal Coordination Reactions for the Assessment of Density Functionals*, J. Chem. Theory Comput. **10** (2014) 3092, DOI: 10.1021/ct500248h.

[274] M. A. Iron and T. Janes, *Evaluating Transition Metal Barrier Heights with the Latest Density Functional Theory Exchange–Correlation Functionals: The MOBH35 Benchmark Database*, J. Phys. Chem. A **123** (2019) 3761, DOI: 10.1021/acs.jpca.9b01546.

[275] M. A. Iron and T. Janes, *Correction to "Evaluating Transition Metal Barrier Heights with the Latest Density Functional Theory Exchange–Correlation Functionals: The MOBH35 Benchmark Database"*, J. Phys. Chem. A **123** (2019) 6379, DOI: 10.1021/acs.jpca.9b06135.

[276] A. Schäfer, H. Horn, and R. Ahlrichs, *Fully optimized contracted Gaussian basis sets for atoms Li to Kr*, J. Chem. Phys. **97** (1992) 2571, DOI: 10.1063/1.463096.

[277] S. Grimme, *Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations*, J. Chem. Theory Comput **15** (2019) 2847, DOI: 10.1021/acs.jctc.9b00143.

[278] T. Betz, S. Zinn, and M. Schnell, *The shape of ibuprofen in the gas phase*, Phys. Chem. Chem. Phys. **17** (6 2015) 4538, DOI: 10.1039/C4CP05529B.

[279] F. Schubert, M. Rossi, C. Baldauf, K. Pagel, S. Warnke, G. von Helden, F. Filsinger, P. Kupser, G. Meijer, M. Salwiczek, B. Koksch, M. Scheffler, and V. Blum, *Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala19-Lys + H+vs. Ac-Lys-Ala19 + H+ and the current reach of DFT*, Phys. Chem. Chem. Phys. **17** (11 2015) 7373, DOI: 10.1039/C4CP05541A.

[280] V. Carta, S. H. M. Mehr, and M. J. MacLachlan, *Controlling Ligand Exchange through Macrocyclization*, Inorg. Chem. **57** (2018) 3243, DOI: 10.1021/acs.inorgchem.8b00031.

[281] G. P. Connor, N. Lease, A. Casuras, A. S. Goldman, P. L. Holland, and J. M. Mayer, *Protonation and electrochemical reduction of rhodium– and iridium–dinitrogen complexes in organic solution*, Dalton Trans. **46** (41 2017) 14325, DOI: 10.1039/C7DT03476H.

[282] P. Pracht, R. Wilcken, A. Udvarhelyi, S. Rodde, and S. Grimme, *High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pKa values in the context of the SAMPL6 challenge*, J. Comput.-Aided Mol. Des. **32** (2018) 1139, ISSN: 1573-4951, DOI: 10.1007/s10822-018-0145-7.

[283] S. Grimme, *Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules*, Angew. Chem. Int. Ed. **52** (2013) 6306, DOI: 10.1002/anie.201300158, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201300158.

[284] J. Koopman and S. Grimme, *Calculation of Electron Ionization Mass Spectra with Semiempirical GFNn-xTB Methods*, ACS Omega **4** (2019) 15120, DOI: 10.1021/acsomega.9b02011, eprint: https://doi.org/10.1021/acsomega.9b02011.

[285] C. A. Bauer and S. Grimme, *How to Compute Electron Ionization Mass Spectra from First Principles*, J. Phys. Chem. A **120** (2016), PMID: 27139033 3755, DOI: 10.1021/acs.jpca.6b02907, eprint: https://doi.org/10.1021/acs.jpca.6b02907.

[286] S. A. Katsyuba, E. E. Zvereva, and S. Grimme, *Fast Quantum Chemical Simulations of Infrared Spectra of Organic Compounds with the B97-3c Composite Method*, J. Phys. Chem. A **123** (2019), PMID: 30958005 3802, DOI: 10.1021/acs.jpca.9b01688, eprint: https://doi.org/10.1021/acs.jpca.9b01688.

[287] Y.-P. Li, A. T. Bell, and M. Head-Gordon, *Thermodynamics of Anharmonic Systems: Uncoupled Mode Approximations for Molecules*, J. Chem. Theroy Comput. **12** (2016), PMID: 27182658 2861, DOI: 10.1021/acs.jctc.5b01177, eprint: https://doi.org/10.1021/acs.jctc.5b01177.

[288] C. Hättig, W. Klopper, A. Köhn, and D. P. Tew, *Explicitly correlated electrons in molecules*, Chem. Rev. **112** (2012) 4, DOI: 10.1021/cr200168z.

[289] K. I. Assaf, M. Florea, J. Antony, N. M. Henriksen, J. Yin, A. Hansen, Z.-w. Qu, R. Sure, D. Klapstein, M. K. Gilson, S. Grimme, and W. M. Nau, *HYDROPHOBE Challenge: A Joint Experimental and Computational Study on the Host–Guest Binding of Hydrocarbons to Cucurbiturils, Allowing Explicit Evaluation of Guest Hydration Free-Energy Contributions*, J. Phys. Chem. B **121** (2017) 11144, DOI: 10.1021/acs.jpcb.7b09175.

[290] M. Steinmetz, A. Hansen, S. Ehrlich, T. Risthaus, and S. Grimme, *Accurate Thermochemistry for Large Molecules with Modern Density Functionals*, Top. Curr. Chem. **365** (2015), ed. by E. R. Johnson 1, DOI: 10.1007/128_2014_543.

[291] A. N. Bootsma and S. Wheeler, *Popular Integration Grids Can Result in Large Errors in DFT-Computed Free Energies*, ChemRxiv (2019), DOI: 10.26434/chemrxiv.8864204.

[292] D. J. Tantillo, J. Chen, and K. N. Houk, *Theozymes and compuzymes: Theoretical models for biological catalysis*, Curr. Opin. Chem. Biol. **2** (1998) 743, ISSN: 13675931, DOI: 10.1016/S1367-5931(98)80112-9.

[293] U. Ryde and P. Söderhjelm, *Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods*, Chem. Rev. **116** (2016) 5520, ISSN: 15206890, DOI: 10.1021/acs.chemrev.5b00630.

[294] C. Clementi, *Coarse-grained models of protein folding: toy models or predictive tools?*, Curr. Opin. Chem. Biol. **18** (2008) 10, ISSN: 0959440X, DOI: 10.1016/j.sbi.2007.10.005.

[295] R. D. Schaeffer, A. Fersht, and V. Daggett, *Combining experiment and simulation in protein folding: closing the gap for small model systems*, Curr. Opin. Chem. Biol. **18** (2008) 4, ISSN: 0959440X, DOI: 10.1016/j.sbi.2007.11.007.

[296] Q. Shao and W. Zhu, *Assessing AMBER force fields for protein folding in an implicit solvent*, Phys. Chem. Chem. Phys. **20** (2018) 7206, ISSN: 14639076, DOI: 10.1039/c7cp08010g.

[297] M. Zheng, M. Biczysko, Y. Xu, N. W. Moriarty, H. Kruse, A. Urzhumtsev, M. P. Waller, and P. V. Afonine, *Including crystallographic symmetry in quantum-based refinement: Q— R# 2*, Acta Crystrallogr. **D76** (2020) 41, DOI: 10.1107/S2059798319015122.

[298] J. G. Lees, A. J. Miles, F. Wien, and B. A. Wallace, *A reference database for circular dichroism spectroscopy covering fold and secondary structure space*, Bioinformatics **22** (2006) 1955, ISSN: 13674803, DOI: 10.1093/bioinformatics/btl327.

[299] C. Bannwarth, S. Grimme, J. Seibert, and S. Grimme, *Electronic Circular Dichroism of [16]Helicene With Simplified TD-DFT: Beyond the Single Structure Approach*, Chirality **28** (2016) 365, ISSN: 1520636X, DOI: 10.1002/chir.22594.

[300] M. Masnyk, A. Butkiewicz, M. Górecki, R. Luboradzki, C. Bannwarth, S. Grimme, and J. Frelek, *Synthesis and Comprehensive Structural and Chiroptical Characterization of Enones Derived from (–)-α-Santonin by Experiment and Theory*, J. Org. Chem. **81** (2016) 4588, DOI: 10.1021/acs.joc.6b00416.

[301] J. D. Queen, M. Bursch, J. Seibert, L. R. Maurer, B. D. Ellis, J. C. Fettinger, S. Grimme, and P. P. Power, *Isolation and Computational Studies of a Series of Terphenyl Substituted Diplumbynes with Ligand Dependent Lead-Lead Multiple-Bonding Character.*, J. Am. Chem. Soc. **141** (2019) 14370, ISSN: 0002-7863, DOI: 10.1021/jacs.9b07072.

[302] G. te Velde, F. M. Bickelhaupt, E. J. Baerends, C. Fonseca Guerra, S. J. A. van Gisbergen, J. G. Snijders, and T. Ziegler, *Chemistry with ADF*, J. Comput. Chem. **22** (2001) 931, DOI: 10.1002/jcc.1056.

[303] J. Hutter, M. Iannuzzi, F. Schiffmann, and J. VandeVondele, *CP2K: atomistic simulations of condensed matter systems*, WIRES Comput. Mol. Sci. **4** (2014) 15, DOI: 10.1002/wcms.1159.

[304] C. Baerlocher and L. B. McCusker, *Database of Zeolite Structures*, http://www.iza-structure.org/databases/, Accessed: 2020-05-20.

[305] C. Baerlocher, L. B. McCusker, and D. H. Olson, *Atlas of zeolite framework types*, Elsevier, 2007.

[306] A. J. Cruz-Cabeza, S. M. Reutzel-Edens, and J. Bernstein, *Facts and fictions about polymorphism*, Chem. Soc. Rev. **44** (2015) 8619, DOI: 10.1039/C5CS00227C.

[307] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, et al., *Functional materials discovery using energy–structure–function maps*, Nature **543** (2017) 657, DOI: 10.1038/nature21419.

[308] J. G. Brandenburg, E. Caldeweyher, and S. Grimme, *Screened exchange hybrid density functional for accurate and efficient structures and interaction energies*, Phys. Chem. Chem. Phys. **18** (2016) 15519, DOI: 10.1039/C6CP01697A.

[309] B. Aradi, B. Hourahine, and T. Frauenheim, *DFTB+, a sparse matrix-based implementation of the DFTB method*, J. Phys. Chem. A **111** (2007) 5678, DOI: 10.1021/jp070186p.

[310] S. Smeets, L. B. McCusker, C. Baerlocher, E. Mugnaioli, and U. Kolb, *Using FOCUS to solve zeolite structures from three-dimensional electron diffraction data*, J. Appl. Crystallogr. **46** (2013) 1017, DOI: 10.1107/S0021889813014817.

[311] L. Wilbraham, R. S. Sprick, K. E. Jelfs, and M. A. Zwijnenburg, *Mapping binary copolymer property space with neural networks*, Chem. Sci. **10** (2019) 4973, DOI: 10.1039/C8SC05710A.

[312] R. Baldwin and D. Baker, *Peptide Solvation and H-bonds*, Elsevier, 2006.

[313] U. Shmueli and A. Wilson, *Conformation in biology*, 1982.

[314] R. Göttlich, B. C. Kahrs, J. Krüger, and R. W. Hoffmann, *Open chain compounds with preferredconformations*, Chemical Communications (1997) 247.

[315] G. Tasi, R. Izsák, G. Matisz, A. G. Császár, M. Kállay, B. Ruscic, and J. F. Stanton, *The origin of systematic error in the standard enthalpies of formation of hydrocarbons computed via atomization schemes*, ChemPhysChem **7** (2006) 1664.

[316] K. Goossens and H. De Winter,
Molecular dynamics simulations of membrane proteins: An overview,
J. Chem. Inf. Model. **58** (2018) 2193.

[317] G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.-F. Truchon, and W. D. Cornell,
Comparison of topological, shape, and docking methods in virtual screening,
J. Chem. Inf. Model. **47** (2007) 1504.

[318] K. S. Pitzer,
Thermodynamics of Gaseous Hydrocarbons: Ethane, Ethylene, Propane, Propylene, n-Butane, Isobutane, 1-Butene, Cis and Trans 2-Butenes, Isobutene, and Neopentane (Tetramethylmethane),
J. Chem. Phys. **5** (1937) 473.

[319] W. Herrebout, B. Van der Veken, A. Wang, and J. Durig, *Enthalpy difference between conformers of n-butane and the potential function governing conformational interchange*, J. Chem. Phys. **99** (1995) 578, DOI: 10.1021/j100002a020.

[320] R. M. Balabin, *Enthalpy difference between conformations of normal alkanes: Raman spectroscopy study of n-pentane and n-butane*, J. Phys. Chem. A **113** (2009) 1012.

[321] A. Basu, M. Mookherjee, E. McMahan, B. Haberl, and R. Boehler,
Behavior of Long-Chain Hydrocarbons at High Pressures and Temperatures,
J. Phys. Chem. B (2022), DOI: 10.1021/acs.jpcb.1c10786.

[322] G. D. Smith and R. L. Jaffe, *Quantum chemistry study of conformational energies and rotational energy barriers in n-alkanes*, J. Phys. C **100** (1996) 18718.

[323] N. L. Allinger, J. T. Fermann, W. D. Allen, and H. F. Schaefer III,
The torsional conformations of butane: Definitive energetics from ab initio methods,
J. Chem. Phys. **106** (1997) 5143.

[324] A. Salam and M. Deleuze,
High-level theoretical study of the conformational equilibrium of n-pentane,
J. Chem. Phys. **116** (2002) 1296.

[325] J. M. Goodman,
What is the longest unbranched alkane with a linear global minimum conformation?,
J. Chem. Inf. Comp. Sci. **37** (1997) 876.

[326] N. O. Lüttschwager, T. N. Wassermann, R. A. Mata, and M. A. Suhm,
The last globally stable extended alkane, Angew. Chem. Int. Ed. **52** (2013) 463,
DOI: 10.1002/anie.201202894.

[327] J. N. Byrd, R. J. Bartlett, and J. A. Montgomery Jr,
At what chain length do unbranched alkanes prefer folded conformations?,
J. Phys. Chem. A **118** (2014) 1706, DOI: 10.1021/jp4121854.

[328] D. G. Liakos and F. Neese,
Domain based pair natural orbital coupled cluster studies on linear and folded alkane chains,
J. Chem. Theory Comput. **11** (2015) 2137, DOI: 10.1021/acs.jctc.5b00265.

[329] J. M. Martin, *What can we learn about dispersion from the conformer surface of n-pentane?*,
J. Phys. Chem. A **117** (2013) 3118.

[330] P. v. R. Schleyer, J. E. Williams Jr, and K. Blanchard,
Evaluation of strain in hydrocarbons. The strain in adamantane and its origin,
J. Am. Chem. Soc. **92** (1970) 2377.

[331] M. D. Wodrich, C. Corminboeuf, and P. v. R. Schleyer,
Systematic errors in computed alkane energies using B3LYP and other popular DFT functionals,
Organic letters **8** (2006) 3631.

[332] A. Karton, D. Gruzman, and J. M. Martin,
Benchmark Thermochemistry of the C n H2 n+ 2 Alkane Isomers (n= 2- 8) and Performance of DFT and Composite Ab Initio Methods for Dispersion-Driven Isomeric Equilibria,
J. Phys. Chem. A **113** (2009) 8434.

[333] S. Grimme, *Semiempirical hybrid density functional with perturbative second-order correlation*,
J. Chem. Phys. **124** (2006) 034108, DOI: 10.1063/1.2148954.

[334] U. R. Fogueri, S. Kozuch, A. Karton, and J. M. Martin,
The melatonin conformer space: Benchmark and assessment of wave function and DFT methods for a paradigmatic biological and pharmacological molecule,
J. Phys. Chem. A **117** (2013) 2269, DOI: 10.1021/jp312644t.

[335] S. Kozuch, S. M. Bachrach, and J. M. Martin, *Conformational equilibria in butane-1, 4-diol: a benchmark of a prototypical system with strong intramolecular H-bonds*,
J. Phys. Chem. A **118** (2014) 293, DOI: 10.1021/jp410723v.

[336] M. K. Kesharwani, A. Karton, and J. M. Martin,
Benchmark ab initio conformational energies for the proteinogenic amino acids through explicitly correlated methods. Assessment of density functional methods,
J. Chem. Theory Comput. **12** (2016) 444, DOI: 10.1021/acs.jctc.5b01066.

[337] S. Spicher, E. Caldeweyher, A. Hansen, and S. Grimme, *Benchmarking London dispersion corrected density functional theory for noncovalent ion–π interactions*,
Phys. Chem. Chem. Phys. **23** (2021) 11635, DOI: 10.1039/D1CP01333E.

[338] S. G. Balasubramani, G. P. Chen, S. Coriani, M. Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, et al., *TURBOMOLE: Modular program suite for ab initio quantum-chemical and condensed-matter simulations*,
J. Chem. Phys. **152** (2020) 184107.

[339] ORCA – an ab initio, density functional and semiempirical program package, V. 5.0.1, F. Neese, MPI für Kohlenforschung, Mülheim a. d. Ruhr (Germany), **2021**.

[340] ORCA – an ab initio, density functional and semiempirical program package, V. 4.2.1, F. Neese, MPI für Kohlenforschung, Mülheim a. d. Ruhr (Germany), **2020**.

[341] J. Rezac, C. Greenwell, and G. J. Beran, *Accurate noncovalent interactions via dispersion-corrected second-order Møller–Plesset perturbation theory*,
J. Chem. Theory Comput. **14** (2018) 4711, DOI: 10.1021/acs.jctc.8b00548.

[342] D. G. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, et al.,
PSI4 1.4: Open-source software for high-throughput quantum chemistry,
J. Chem. Phys. **152** (2020) 184108, DOI: 10.1063/5.0006002.

[343] J. Lee and M. Head-Gordon,
Regularized orbital-optimized second-order Møller–Plesset perturbation theory: A reliable fifth-order-scaling electron correlation model with orbital energy dependent regularizers,
J. Chem. Theory Comput. **14** (2018) 5203, DOI: 10.1021/acs.jctc.8b00731.

[344] M. Pitoňák, P. Neogrády, J. Černỳ, S. Grimme, and P. Hobza, *Scaled MP3 non-covalent interaction energies agree closely with accurate CCSD (T) benchmark data*,
ChemPhysChem **10** (2009) 282, DOI: 10.1002/cphc.200800718.

[345] O. Vahtras, J. Almlöf, and M. W. Feyereisen,
Integral approximations for LCAO-SCF calculations, Chem. Phys. Lett. **213** (1993) 514.

[346] R. A. Kendall and H. A. Früchtl, *The impact of the resolution of the identity approximate integral method on modern ab initio algorithm development*, Theor. Chem. Acc. **97** (1997) 158, DOI: 10.1007/s002140050249.

[347] 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*,
Phys. Chem. Chem. Phys. **7** (2005) 3297, DOI: 10.1039/b508541a.

[348] F. Weigend, *Accurate Coulomb-fitting basis sets for H to Rn*,
Phys. Chem. Chem. Phys. **8** (2006) 1057.

[349] Y. Guo, C. Riplinger, U. Becker, D. G. Liakos, Y. Minenkov, L. Cavallo, and F. Neese,
Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)], J. Chem. Phys. **148** (2018) 011101, DOI: 10.1063/1.5011798.

[350] R. A. Kendall, T. H. Dunning, and R. J. Harrison, J. Chem. Phys. **96** (1992) 6796.

[351] T. Helgaker, W. Klopper, H. Koch, and J. Noga,
Basis-set convergence of correlated calculations on water, J. Chem. Phys. **106** (1997) 9639, DOI: 10.1063/1.473863.

[352] "DFTB+ general package for performing fast atomistic simulations",
https://github.com/dftbplus/dftbplus. Accessed: 2021-05-03.

[353] M. Gaus, A. Goez, and M. Elstner,
Parametrization and Benchmark of DFTB3 for Organic Molecules,
J. Chem. Theory Comput. **9** (2013) 338, DOI: 10.1021/ct300849w.

[354] M. Gaus, X. Lu, M. Elstner, and Q. Cui, *Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications*,
J. Chem. Theory Comput. **10** (2014) 1518, DOI: 10.1021/ct401002w.

[355] X. Lu, M. Gaus, M. Elstner, and Q. Cui, *Parametrization of DFTB3/3OB for Magnesium and Zinc for Chemical and Biological Applications*, J. Phys. Chem. B **119** (2015) 1062, DOI: 10.1021/jp506557r.

[356] M. Kubillus, T. Kubar, M. Gaus, J. Rezac, and M. Elstner, *Parameterization of the DFTB3 method for Br, Ca, Cl, F, I, K, and Na in organic and biological systems*,
J. Chem. Theory Comput. **11** (2015) 332, DOI: 10.1021/ct5009137.

[357] T. A. Niehaus, M. Elstner, T. Frauenheim, and S. Suhai, *Application of an approximate density-functional method to sulfur containing compounds*, J. Mol. Struct. (Theochem) **541** (2001) 185, DOI: 10.1016/S0166-1280(00)00762-4.

[358] V. Q. Vuong, J. Akkarapattiakal Kuriappan, M. Kubillus, J. J. Kranz, T. Mast, T. A. Niehaus, S. Irle, and M. Elstner, *Parametrization and Benchmark of Long-Range Corrected DFTB2 for Organic Molecules*, J. Chem. Theory Comput. **14** (2018) 115, DOI: 10.1021/acs.jctc.7b00947.

[359] "Molecular Orbital PACkage", https://github.com/openmopac/mopac. Accessed: 2021-12-17.

[360] J. J. Stewart, *Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters*, Journal of molecular modeling **19** (2013) 1, DOI: 10.1007/s00894-012-1667-x.

[361] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, et al., *Escaping atom types in force fields using direct chemical perception*, J. Chem. Theory Comput. **14** (2018) 6076.

[362] D. Mobley, C. Bannan, J. Wagner, A. Rizzi, N. Lim, and M. Henry, *openforcefield/smirnoff99Frosst: Version 1.1.0*, version 1.1.0, 2019, DOI: 10.5281/zenodo.3351714.

[363] Y. Qiu, D. G. Smith, S. Boothroyd, J. Wagner, C. C. Bannan, T. Gokey, H. Jang, V. T. Lim, X. Lucas, B. Tjanaka, M. R. Shirts, M. K. Gilson, J. D. Chodera, C. I. Bayly, D. L. Mobley, and L.-P. Wang, *openforcefield/openforcefields: Version 1.0.0 "Parsley"*, version 1.0.0, 2019, DOI: 10.5281/zenodo.3483227.

[364] J. Wagner, M. Thompson, D. Dotson, hyejang, S. Boothroyd, and J. Rodríguez-Guerra, *openforcefield/openff-forcefields: Version 2.0.0 "Sage"*, version 2.0.0, 2021, DOI: 10.5281/zenodo.5214478.

[365] J. Wagner, M. Thompson, D. L. Mobley, J. Chodera, C. Bannan, A. Rizzi, trevorgokey, D. Dotson, J. Rodríguez-Guerra, Camila, P. Behara, J. A. Mitchell, C. Bayly, JoshHorton, N. M. Lim, V. Lim, S. Sasmal, L. Wang, A. Dalke, SimonBoothroyd, I. Pulido, D. Smith, J. Horton, L.-P. Wang, and Y. Zhao, *openforcefield/openff-toolkit: 0.10.1 Minor feature and bugfix release*, version 0.10.1, 2021, DOI: 10.5281/zenodo.5601736.

[366] G. Landrum, P. Tosco, B. Kelley, Ric, sriniker, gedeck, R. Vianello, NadineSchneider, E. Kawashima, A. Dalke, D. N, B. Cole, D. Cosgrove, M. Swain, S. Turk, AlexanderSavelyev, G. Jones, A. Vaucher, M. Wójcikowski, D. Probst, V. F. Scalfani, guillaume godin, A. Pahl, F. Berenger, JLVarjo, K. Ujihara, strets123, JP, DoliathGavid, and G. Sforna, *rdkit/rdkit: 2021_09_3 (Q3 2021) Release*, version Release_2021_09_3, 2021, DOI: 10.5281/zenodo.5773460.

[367] J.-D. Chai and M. Head-Gordon, *Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections*, Phys. Chem. Chem. Phys. **10** (2008) 6615, DOI: 10.1039/B810189B.

[368] D. G. Smith, L. A. Burns, K. Patkowski, and C. D. Sherrill, *Revised damping parameters for the D3 dispersion correction to density functional theory*, J. Phys. Chem. Lett. **7** (2016) 2197, DOI: 10.1021/acs.jpclett.6b00780.

[369] J. Witte, N. Mardirossian, J. B. Neaton, and M. Head-Gordon, *Assessing DFT-D3 damping functions across widely used density functionals: Can we do better?*, J. Chem. Theory Comput. **13** (2017) 2043, DOI: 10.1021/acs.jctc.7b00176.

[370] "Reimplementation of the DFT-D3 program", https://github.com/awvwgk/simple-dftd3. Accessed: 2022-03-16.

[371] "Generally Applicable Atomic-Charge Dependent London Dispersion Correction", https://github.com/grimme-lab/xtb. Accessed: 2021-12-17.

[372] J. P. Perdew, J. Tao, V. N. Staroverov, and G. E. Scuseria, *Meta-generalized gradient approximation: Explanation of a realistic nonempirical density functional*, J. Chem. Phys. **120** (2004) 6898, DOI: 10.1063/1.1665298.

[373] 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"*, J. Phys. Chem. Lett. **11** (2020) 9248, DOI: 10.1021/acs.jpclett.0c03077.

[374] 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*, J. Phys. C **98** (1994) 11623, DOI: 10.1021/j100096a001.

[375] Y. Zhao and D. G. Truhlar, *Design of density functionals that are broadly accurate for thermochemistry, thermochemical kinetics, and nonbonded interactions*, J. Phys. Chem. A **109** (2005) 5656, DOI: 10.1021/jp050536c.

[376] R. Peverati and D. G. Truhlar, *Screened-exchange density functionals with broad accuracy for chemistry and solid-state physics*, Phys. Chem. Chem. Phys. **14** (2012) 16187, DOI: 10.1039/C2CP42576A.

[377] L. Goerigk and S. Grimme, *Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals: Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions*, J. Chem. Theory Comput. **7** (2011) 291, DOI: 10.1021/ct100466k.

[378] G. Santra, N. Sylvetsky, and J. M. Martin, *Minimally empirical double-hybrid functionals trained against the GMTKN55 database: revDSD-PBEP86-D4, revDOD-PBE-D4, and DOD-SCAN-D4*, J. Phys. Chem. A **123** (2019) 5129, DOI: 10.1021/acs.jpca.9b03157.

[379] J. Shee, M. Loipersberger, A. Rettig, J. Lee, and M. Head-Gordon, *Regularized second-order Møller–Plesset theory: A more accurate alternative to conventional MP2 for noncovalent interactions and transition metal thermochemistry for the same computational cost*, J. Phys. Chem. Lett. **12** (2021) 12084, DOI: 10.1021/acs.jpclett.1c03468.

[380] K. Nalin de Silva and J. M. Goodman, *What is the smallest saturated acyclic alkane that cannot be made?*, Journal of chemical information and modeling **45** (2005) 81, DOI: 10.1021/ci0497657.

[381] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, and A. K. Wilson, *Basis-set convergence in correlated calculations on Ne, N2, and H2O*, Chemical Physics Letters **286** (1998) 243.

[382] P. Jurečka, J. Šponer, J. Černỳ, 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*, Phys. Chem. Chem. Phys. **8** (2006) 1985.

[383] J. M. Martin and G. de Oliveira, *Towards standard methods for benchmark quality ab initio thermochemistry—W1 and W2 theory*, J. Chem. Phys. **111** (1999) 1843.

[384] J. M. Martin and S. Parthiban, "W1 and W2 theories, and their variants: thermochemistry in the kJ/mol accuracy range", *Quantum-Mechanical Prediction of Thermochemical Data*, Springer, 2001 31.

[385] A. G. Császár, W. D. Allen, and H. F. Schaefer, *In pursuit of the ab initio limit for conformational energy prototypes*, J. Chem. Phys. **108** (1998) 9751, DOI: 10.1063/1.476449.

[386] 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*, J. Chem. Phys. **136** (2012) 04B613, DOI: 10.1063/1.3700154.

[387] J. G. Brandenburg, M. Alessio, B. Civalleri, M. F. Peintinger, T. Bredow, and S. Grimme, *Geometrical correction for the inter-and intramolecular basis set superposition error in periodic density functional theory calculations*, J. Phys. Chem. A **117** (2013) 9282, DOI: 10.1021/jp406658y.

[388] L. R. Maurer, M. Bursch, S. Grimme, and A. Hansen, *Assessing Density Functional Theory for Chemically Relevant Open-Shell Transition Metal Reactions*, J. Chem. Theory Comput. **17** (2021) 6134, DOI: 10.1021/acs.jctc.1c00659.

[389] P. Pracht and S. Grimme, *Calculation of absolute molecular entropies and heat capacities made simple*, Chemical science **12** (2021) 6551, DOI: 10.1039/D1SC00621E.

[390] M. Pitonak and A. Heßelmann, *Accurate intermolecular interaction energies from a combination of MP2 and TDDFT response theory*, J. Chem. Theory Comput. **6** (2010) 168.

[391] G. Santra, E. Semidalas, and J. M. Martin, *Surprisingly Good Performance of XYG3 Family Functionals Using a Scaled KS-MP3 Correlation*, J. Phys. Chem. Lett. **12** (2021) 9368, DOI: 10.1021/acs.jpclett.1c02838.

[392] W. L. Jorgensen and T. B. Nguyen, *Monte Carlo simulations of the hydration of substituted benzenes with OPLS potential functions*, J. Comput. Chem. **14** (1993) 195.

[393] D. L. Mobley and J. P. Guthrie, *FreeSolv: a database of experimental and calculated hydration free energies, with input files*, Journal of computer-aided molecular design **28** (2014) 711.

[394] C. F. Wong and J. A. McCammon, *Dynamics and design of enzymes and inhibitors*, J. Am. Chem. Soc. **108** (1986) 3830, DOI: 10.1021/ja00273a048.

[395] J. G. Kirkwood, *Statistical mechanics of fluid mixtures*, J. Chem. Phys. **3** (1935) 300.

[396] P. H. Berens, D. H. Mackay, G. M. White, and K. R. Wilson, *Thermodynamics and quantum corrections from molecular dynamics for liquid water*, J. Chem. Phys. **79** (1983) 2375.

[397] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, *The statistical-thermodynamic basis for computation of binding affinities: a critical review*, Biophys. J. **72** (1997) 1047.

[398] R. W. Zwanzig, *High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases*, J. Chem. Phys. **22** (1954) 1420, DOI: 10.1063/1.1740409.

[399] W. L. Jorgensen and C. Ravimohan, *Monte Carlo simulation of differences in free energies of hydration*, J. Chem. Phys. **83** (1985) 3050, DOI: 10.1063/1.449208.

[400] J. Wereszczynski and J. A. McCammon, *Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition*, Quarterly reviews of biophysics **45** (2012) 1.

[401] N. Hansen and W. F. Van Gunsteren, *Practical aspects of free-energy calculations: a review*, J. Chem. Theory Comput. **10** (2014) 2632, DOI: 10.1021/ct500161f.

[402] A. Laio and M. Parrinello, *Escaping free-energy minima*, Proc. Nat. Acad. Sci., USA **99** (2002) 12562, DOI: 10.1073/pnas.202427399.

[403] G. M. Torrie and J. P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling*, Journal of Computational Physics **23** (1977) 187, DOI: 10.1016/0021-9991(77)90121-8.

[404] U. H. Hansmann, *Parallel tempering algorithm for conformational studies of biological molecules*, Chem. Phys. Lett. **281** (1997) 140, DOI: 10.1016/S0009-2614(97)01198-6.

[405] Y. Sugita and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding*, Chem. Phys. Lett. **314** (1999) 141, DOI: 10.1016/S0009-2614(99)01123-9.

[406] D. L. Mobley and P. V. Klimovich, *Perspective: Alchemical free energy calculations for drug discovery*, J. Chem. Phys. **137** (2012) 230901, DOI: doi.org/10.1063/1.4769292.

[407] W. L. Jorgensen and J. Tirado–Rives, *Molecular modeling of organic and biomolecular systems using BOSS and MCPRO*, J. Comput. Chem. **26** (2005) 1689, DOI: 10.1002/jcc.20297.

[408] I. Cabeza de Vaca, Y. Qian, J. Z. Vilseck, J. Tirado-Rives, and W. L. Jorgensen, *Enhanced Monte Carlo methods for modeling proteins including computation of absolute free energies of binding*, J. Chem. Theory Comput. **14** (2018) 3279, DOI: 10.1021/acs.jctc.8b00031.

[409] J. Tomasi and M. Persico, *Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent*, Chem. Rev. **94** (1994) 2027, DOI: 10.1021/cr00031a013.

[410] B. Honig and A. Nicholls, *Classical electrostatics in biology and chemistry*, Science **268** (1995) 1144.

[411] B. Roux and T. Simonson, *Implicit solvent models*, Biophysical chemistry **78** (1999) 1, DOI: 10.1016/S0301-4622(98)00226-9.

[412] C. J. Cramer and D. G. Truhlar, *Implicit solvation models: equilibria, structure, spectra, and dynamics*, Chem. Rev. **99** (1999) 2161.

[413] M. Orozco and F. J. Luque, *Theoretical methods for the description of the solvent effect in biomolecular systems*, Chem. Rev. **100** (2000) 4187, DOI: 10.1021/cr990052a.

[414] J. Tomasi, B. Mennucci, and R. Cammi, *Quantum mechanical continuum solvation models*, Chem. Rev. **105** (2005) 2999.

[415] 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 (1993) 799, DOI: 10.1039/P29930000799.

[416] S. Miertuš, E. Scrocco, and J. Tomasi, *Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects*, Chemical Physics **55** (1981) 117, DOI: 10.1016/0301-0104(81)85090-2.

[417] J. B. Foresman, T. A. Keith, K. B. Wiberg, J. Snoonian, and M. J. Frisch, *Solvent effects. 5. Influence of cavity shape, truncation of electrostatics, and electron correlation on ab initio reaction field calculations*, J. Phys. C **100** (1996) 16098, DOI: 10.1021/jp960488j.

[418] V. Barone and M. Cossi, *Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model*, J. Phys. Chem. A **102** (1998) 1995, DOI: 10.1021/jp9716997.

[419] M. Cossi, N. Rega, G. Scalmani, and V. Barone, *Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model*, J. Comput. Chem. **24** (2003) 669, DOI: 10.1002/jcc.10189.

[420] A. Klamt, *Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena*, J. Chem. Phys. **99** (1995) 2224.

[421] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, *Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions*, J. Phys. Chem. B **113** (2009) 6378, DOI: 10.1021/jp810292n.

[422] S. Ten-no, F. Hirata, and S. Kato, *A hybrid approach for the solvent effect on the electronic structure of a solute based on the RISM and Hartree-Fock equations*, Chemical physics letters **214** (1993) 391, DOI: 10.1016/0009-2614(93)85655-8.

[423] A. Kovalenko and F. Hirata, *Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach*, Chem. Phys. Lett. **290** (1998) 237, DOI: 10.1016/S0009-2614(98)00471-0.

[424] J. Heil and S. M. Kast, *3D RISM theory with fast reciprocal-space electrostatics*, J. Chem. Phys. **142** (2015) 114107, DOI: 10.1063/1.4914321.

[425] F. Lipparini, B. Stamm, E. Cances, Y. Maday, and B. Mennucci, *Fast domain decomposition algorithm for continuum solvation models: Energy and first derivatives*, J. Chem. Theory Comput. **9** (2013) 3637.

[426] J. J. Stewart, *Application of the PM6 method to modeling proteins*, Journal of molecular modeling **15** (2009) 765, DOI: 10.1007/s00894-008-0420-y.

[427] K. Kříž and J. Řezáč, *Reparametrization of the COSMO solvent model for semiempirical methods PM6 and PM7*, J. Chem. Inf. Model. **59** (2019) 229, DOI: 10.1021/acs.jcim.8b00681.

[428] M. Born, *Volumen und hydratationswärme der ionen*, Zeitschrift für Physik **1** (1920) 45, DOI: 10.1007/BF01881023.

[429] S. C. Tucker and D. G. Truhlar, *Generalized Born fragment charge model for solvation effects as a function of reaction coordinate*, Chem. Phys. Lett. **157** (1989) 164, DOI: 10.1016/0009-2614(89)87227-6.

[430] A. V. Marenich, R. M. Olson, C. P. Kelly, C. J. Cramer, and D. G. Truhlar, *Self-consistent reaction field model for aqueous and nonaqueous solutions based on accurate polarized partial charges*, J. Chem. Theory Comput. **3** (2007) 2011, DOI: 10.1021/ct7001418.

[431] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, *Universal solvation model based on the generalized born approximation with asymmetric descreening*, J. Chem. Theory Comput. **5** (2009) 2447, DOI: 10.1021/ct900312z.

[432] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, *Generalized born solvation model SM12*, J. Chem. Theory Comput. **9** (2013) 609, DOI: 10.1021/ct300900e.

[433] A. Pecina, R. Meier, J. Fanfrlik, M. Lepšik, J. Řezáč, P. Hobza, and C. Baldauf, *The SQM/COSMO filter: reliable native pose identification based on the quantum-mechanical description of protein–ligand interactions and implicit COSMO solvation*, Chemical Communications **52** (2016) 3312, DOI: 10.1039/C5CC09499B.

[434] K. Kříž and J. Řezáč, *Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design*, J. Chem. Inf. Model. **60** (2020) 1453, DOI: 10.1021/acs.jcim.9b01171.

[435] G. Hou, X. Zhu, and Q. Cui, *An Implicit Solvent Model for SCC-DFTB with Charge-Dependent Radii*, J. Chem. Theory Comput. **6** (2010) 2303, DOI: 10.1021/ct1001818.

[436] V. Barone, I. Carnimeo, and G. Scalmani, *Computational Spectroscopy of Large Systems in Solution: The DFTB/PCM and TD-DFTB/PCM Approach*, Journal of chemical theory and computation **9** (2013) 2052, DOI: 10.1021/ct301050x.

[437] Y. Nishimoto, *DFTB/PCM applied to ground and excited state potential energy surfaces*, J. Phys. Chem. A **120** (2016) 771.

[438] V. I. Lebedev and D. N. Laikov, *A quadrature formula for the sphere of the 131st algebraic order of accuracy*, Doklady Mathematics **59** (3 1999) 477.

[439] W. Keesom, *The second viral coefficient for rigid spherical molecules, whose mutual attraction is equivalent to that of a quadruplet placed at their centre*, Proc. R. Acad. Sci **18** (1915) 636.

[440] A. D. Becke, *Density-functional exchange-energy approximation with correct asymptotic behavior*, Phys. Rev. B **38** (1988) 3098, DOI: 10.1103/physreva.38.3098.

[441] "Conformer-Rotamer Ensemble Sampling Tool", https://github.com/grimme-lab/crest. Accessed: 2021-05-03.

[442] "Commandline ENergetic SOrting of Conformer Rotamer Ensembles", https://github.com/grimme-lab/censo. Accessed: 2021-05-03.

[443] "Scripts to automate CREST and CENSO for calculating various properties", https://github.com/grimme-lab/crenso. Accessed: 2021-05-03.

[444] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Q. Appl. Math. **2** (1944) 164, DOI: 10.1090/qam/10666.

[445] D. W. Marquardt, *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*, J. Soc. Ind. Appl. Math. **11** (1963) 431, DOI: 10.1137/0111030.

[446] H. Nakai and A. Ishikawa, *Quantum chemical approach for condensed-phase thermochemistry: Proposal of a harmonic solvation model*, J. Chem. Phys. **141** (2014) 174106, DOI: 10.1063/1.4900629.

[447] Y.-i. Izato, A. Matsugi, M. Koshi, and A. Miyake, *A simple heuristic approach to estimate the thermochemistry of condensed-phase molecules based on the polarizable continuum model*, Phys. Chem. Chem. Phys. **21** (2019) 18920, DOI: 10.1039/C9CP03226F.

[448] W. J. Lyman, W. F. Reehl, and D. H. Rosenblatt, *Handbook of chemical property estimation methods*, Washington, DC (USA); American Chemical Society, 1990.

[449] M. Reddy, R. Yang, M. E. Andersen, and H. J. Clewell III, *Physiologically based pharmacokinetic modeling: science and applications*, John Wiley & Sons, 2005.

[450] S. Spicher and S. Grimme, *Efficient Computation of Free Energy Contributions for Association Reactions of Large Molecules*, J. Phys. Chem. Lett. **11** (2020) 6606, DOI: 10.1021/acs.jpclett.0c01930.

[451] K. R. Mann, J. Gordon, and H. B. Gray, *Characterization of oligomers of tetrakis (phenyl isocyanide) rhodium (I) in acetonitrile solution*, J. Am. Chem. Soc. **97** (1975) 3553, DOI: 10.1021/ja00845a065.

[452] S. Grimme and J.-P. Djukic, *Cation–Cation "Attraction": When London Dispersion Attraction Wins over Coulomb Repulsion*, Inorganic Chemistry **50** (2011) 2619, DOI: 10.1021/ic102489k.

List of Figures

List of Tables

Acknowledgements

I like to thank many people without whom this thesis would not have been possible. I thank Stefan Grimme for the supervision of my Ph.D. studies and giving me the opportunity to develop and explore my own ideas and scientific as well as financial support. Furthermore, I thank Thomas Bredow for being my second reviewer of this thesis and the helpful scientific discussion.

I thank all my colleagues at the Mulliken Center for the friendly and productive working atmosphere. Formost, thanks to my long-term office partner and good friend Hagen Neugebauer as well as Jeroen Koopman as valued friend and role model. Thanks to Vahideh Alizadeh, Benedikt Baedorf, Fabian Bohle, Markus Bursch, Sebastian Dohm, Thomas Froitzheim, Thomas Gasevic, Johannes Gorges, Michael Häfner, Andreas Hansen, Christian Hölzer, Julia Kohn, Joachim Laun, Sarah Löffelsender, Fabian März, Jan Mewes, Marcel Müller, Christoph Plett, Philipp Pracht, Thomas Rose, Sebastian Spicher, Julius Stückrath, and Paul Zaby for pleasant years, fruitful collaborations, and scientific and non-scientific discussions. Special thanks go to the talented students, Marvin Friede, Stefan Jütten, Albert Katbashev, and Marcel Stahn, I got the opportunity to mentor and continue to collaborate with in various exciting projects. I am grateful to my mentors Eike Caldeweyher and Christoph Bannwarth for the great opportunity to participate in their research and projects. Many thanks to Tobias Ehlert, Thomas Gasevic, Christian Hölzer, and Marcel Stahn for proof-reading this thesis.

Further, I like to thank my collaborators, Gerit Brandenburg, Gerhardt Erker, Ben Hourahine, John Perdew for fruitful exchange. Many thanks go to Bálint Aradi for opening a new world of sustainable software development to me and Lori Burns for helping me find my way into the package management ecosystem. I am grateful for the steady administrative support by Claudia Kronz.

Last but not least, I want to thank my family for supporting me unconditionally over all the years.