Dissertation

Robust Atomistic Modeling of Large Molecules by Efficient Force-Field and Tight-Binding Methods

Sebastian Spicher

UNIVERSITÄT BONN

Robust Atomistic Modeling of Large Molecules by Efficient Force-Field and Tight-Binding Methods

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

vorgelegt von
Sebastian Spicher
aus Engelskirchen
–2021–

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

Dekan:	Prof. Dr. Walter Witke
Erster Gutachter:	Prof. Dr. Stefan Grimme
Zweiter Gutachter:	Prof. Dr. Thomas Bredow
Tag der Disputation:	29.10.2021
Erscheinungsjahr:	2021

To my family

"You can never solve a problem on the level on which it was created."

– Albert Einstein –

Publications

Parts of this thesis (in order of appearance) are published in peer-reviewed journals:

1. Spicher, S.; Grimme, S. "Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems", *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673. (Very Important Paper)

2. Spicher, S.; Bursch, M.; Grimme, S. "Efficient Calculation of Small Molecule Binding in Metal-Organic Frameworks and Porous Organic Cages", *J. Phys. Chem. C* **2020**, *124*, 27529–27541.

3. Spicher, S.; Abdullin, D.; Grimme, S.; Schiemann, O. "Modeling of spin–spin distance distributions for nitroxide labeled biomacromolecules", *Phys. Chem. Chem. Phys.* **2020**, *22*, 24282–24290. (2020 PCCP HOT Article)

4. Spicher, S.; Caldeweyher, E.; Hansen, A.; Grimme, S. "Benchmarking London dispersion corrected density functional theory for noncovalent ion–π interactions", *Phys. Chem. Chem. Phys.* **2021**, *23*, 11635–11648.

5. Spicher, S.; Grimme, S. "Efficient Computation of Free Energy Contributions for Association Reactions of Large Molecules", *J. Phys. Chem. Lett.* **2020**, *11*, 6606–6611.

6. Spicher, S.; Grimme, S. "Single-Point Hessian Calculations for Improved Vibrational Frequencies and Rigid-Rotor-Harmonic-Oscillator Thermodynamics", *J. Chem. Theory Comput.* **2021**, *17*, 1701–1714.

7. Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. "Fast and Accurate Quantum Chemical Modeling of Infrared Spectra of Condensed-Phase Systems", *J. Phys. Chem. B* **2020**, *124*, 6664–6670.

Further publications:

8. Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. "Revisiting Conformations of Methyl Lactate in Water and Methanol", *J. Chem. Phys.* **2021**, *155*, 024507.

9. Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. "Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods", *J. Chem. Theory Comput.* **2021**, *17*, 4250–4261.

10. Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. "Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules", *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

11. Fleck, N.; Heubach, C.; Hett, T.; Spicher, S.; Grimme, S.; Schiemann, O. "Ox-SLIM: Synthesis of and Site-Specific Labelling with a Highly Hydrophilic Trityl Spin Label", *Chem. Eur. J.* **2021**, *27*, 5292–5297.

12. Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. "Extended tight-binding quantum chemistry methods", *WIREs Comput. Mol. Sci.* **2021**, *11*, e1493.

13. Abdullin, D.; Brehm, P.; Fleck, N.; Spicher, S.; Grimme, S.; Schiemann, O. "Pulsed EPR Dipolar Spectroscopy on Spin Pairs with one Highly Anisotropic Spin Center: The Low-Spin Fe^{III} Case", *Chem. Eur. J.* **2019**, *25*, 14388–14398.

14. Abdullin, D.; Matsuoka, H.; Yulikov, M.; Fleck, N.; Klein, C.; Spicher, S.; Hagelueken, G.; Grimme, S.; Lützen, A.; Schiemann, O. "Pulsed EPR Dipolar Spectroscopy under the Breakdown of the High-Field Approximation: The High-Spin Iron(III) Case", *Chem. Eur. J.* **2019**, *25*, 8820–8828.

15. Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. "A generally applicable atomic-charge dependent London dispersion correction", *J. Chem. Phys.* **2019**, *150*, 154122.

16. Abdullin, D.; Fleck, N.; Klein, C.; Brehm, P.; Spicher, S.; Lützen, A.; Grimme, S.; Schiemann O. "Synthesis of μ_2-Oxo-Bridged Iron(III) Tetraphenylporphyrin–Spacer–Nitroxide Dimers and their Structural and Dynamics Characterization by using EPR and MD Simulations", *Chem. Eur. J.* **2019**, *25*, 2586–2596.

Presentations:

1. Talk: "From Biomacromolecules to Metal-Organic Frameworks: Development and Application of GFN-FF", *University of Zurich*, **2021**, Zürich, Switzerland.

2. Talk: "GFN-FF: A generic Force-Field for the Full Periodic Table", *RWTH Aachen*, **2020**, Aachen, Germany.

3. Poster: "Quantum Cluster Growth: An Explicit Approach towards Solvation Free Energies", *55th Symposium on Theoretical Chemistry (STC)*, **2019**, Rostock, Germany.

4. Poster: "Quantum Cluster Growth: An Explicit Approach towards Solvation Free Energies", *European Summerschool in Quantum Chemistry (ESQC)*, **2019**, Sicily, Italy.

5. Poster: "Quantum Cluster Growth: An Explicit Approach towards Solvation Free Energies", *258th ACS National Meeting*, **2019**, San Diego, CA, USA.

6. Poster: "GFN2-xTB: An extended Tight-Binding Quantum Chemical Method for Structures, Frequencies and Noncovalent Interactions across the Periodic Table", *34th Winter School in Theoretical Chemistry*, **2018**, Helsinki, Finland.

7. Poster: "GFN2-xTB: An extended Tight-Binding Quantum Chemical Method for Structures, Frequencies and Noncovalent Interactions across the Periodic Table", *54th Symposium on Theoretical Chemistry (STC)*, **2018**, Halle (Saale), Germany.

Abstract

Modern chemistry has almost no boundaries in elemental composition and molecular size. State-of-the-art chemical systems reach tremendous dimensions containing thousands of atoms, due to significant progress in the fields of polymer and supramolecular chemistry, material science, and biochemistry. This is associated with an increasing demand for the performance of computational methods to compare experiments and simulations without restrictions in size and composition of the investigated system. Universal, fast, and yet accurate methods are therefore of increasing importance and related popularity in the field of theoretical chemistry. Thus, this thesis is devoted to the development and application of efficient force-field (FF) and tight-binding (TB) methods for the robust atomistic simulation of large molecules. In particular, a generic FF is introduced, as well as improvements to existing semiempirical extended tight-binding (xTB) methods, for the accurate calculation of Geometries, harmonic vibrational Frequencies (HVF), and Noncovalent interaction (NCI) energies, termed as GFN methods.

The main subjects of interest for experimental and theoretical comparisons are often molecular geometries, reaction free energies, and spectroscopic properties such as infrared (IR) spectra. Their description requires not only accurate energies, but also efficient gradients (first derivative), harmonic vibrational frequencies (second derivative), and corresponding solvation models. However, with increasing molecular size and complexity, the number of computational methods that are applicable for energies, geometries, and frequencies decreases rapidly. For *ab initio* electronic structure methods, therefore, the limit is reached for not much more than one hundred atoms and hence, a fully quantum mechanical (QM) description is not possible. The next consequent step towards higher computational efficiency is represented by semiempirical quantum mechanical (SQM) methods, even though they are often not generally applicable due to incomplete parameterizations or conceptual shortcomings. Recently, this changed by the development of the GFN*n*-xTB family of methods ($n = \{0, 1, 2\}$), which are parameterized for a major part of the periodic table up to radon. The underlying approximations extend the accessible atom size regime to ~1000 atoms. Yet, without massively parallel supercomputers, the description of larger systems remains denied and classical approaches such as FFs have to be applied. Although many different types of FFs exist, universally accurate variants still represent an almost blank space in the repertoire of theoretical methods. Therefore, the development of more accurate (polarizable) FFs is named as a "holy grail" for computational organic- and biochemistry.

The first part of this thesis presents a new generic force-field within the GFN framework. This method, termed GFN-FF, represents a unique, partially polarizable, universal FF for the accurate description of structures and dynamics of large molecules and is developed to combine FF speed with SQM accuracy. What distinguishes it from other FFs is a full periodic table ($Z \leq 86$) parameterization and a completely automated setup routine. To yield high accuracy for NCIs, a sophisticated charge

model based on electronegativity equilibration (EEQ) of Gaussian type charge densities is employed and the treatment of Pauli repulsion and London dispersion interactions is analogous to TB methods. Additionally, a novel hydrogen bond correction is introduced. In this thesis, a detailed description of the underlying theory is given followed by illustrative application examples. It is shown that for structures of metal-organic frameworks (MOF) and biomacromolecules (proteins) the GFN-FF optimized structures correspond well to the experimental crystal structures. GFN-FF is here in many cases the only applicable computational method. On established benchmark sets for conformational and NCI energies, GFN-FF often reaches an accuracy that is comparable to SQM methods or even more sophisticated GGA density functionals.

The next part of this thesis explores the new possibilities of GFN-FF in combination with the conformer-rotamer ensemble sample tool (CREST) in the context of conformational space exploration for large and complex structures, ranging from biomacromolecules to metal-organic frameworks. In a first application-based study, the gas storage of greenhouse gases and bio-fuels, such as carbon dioxide and methanol, in MOFs and porous organic cages (POCs) is investigated. Optimal binding sites are determined by the CREST algorithm at the GFN-FF level of theory and re-optimized by DFT. The association energies calculated by GFN*n*-xTB and GFN-FF show comparable accuracy to the good performing (meta-)GGAs. As a second study, spin–spin distance distributions for nitroxide labeled mutants of azurin and T4 lysozyme are modeled by molecular dynamics (MD) simulations at the GFN-FF level of theory and compared to experimental EPR results. With deviations to the experiment of less than 2 Å in the mean spin–spin distances, GFN-FF outperforms competitive methods.

In the last part, GFN methods are assessed for the calculation of HVF from which the thermostatistical contributions to the free energies are derived within the modified rigid-rotor-harmonic-oscillator (RRHO) approximation. The accuracy of GFN2-xTB and GFN-FF is benchmarked in comparison to DFT reference data. As an outlook for future applications, free association energies, also including solvation effects, are calculated for protein-drug complexes of almost 5000 atoms. In addition, a new method termed single-point hessian (SPH) is introduced for improved HVF of general non-equilibrium structures, in which the input geometry is retained by the application of a biasing potential. Thereby, the SPH approach enables the calculation of accurate thermodynamics on every point of the potential energy surface (PES). Significant improvements in thermostatistical contributions and IR spectra are obtained by the SPH approach at the SQM and FF level of theory, if, *e.g.*, DFT structures are provided as input. Finally, the effect of explicit solvation is investigated in the context of IR spectra. For the first time, a novel algorithm named quantum cluster growth (QCG) is applied, yielding results remarkably close to the experimental reference spectra.

Overall, the methods developed and evaluated in this work present a great leap forward in theoretical chemistry, bridging the gap between theory and experiment for large molecules. GFN-FF and SPH calculations are added to the portfolio of computational methods and represent valuable and versatile tools for theoretical pre-screening and modeling. From organometallic to biochemical systems, the unique combination of efficiency, generality, and accuracy of the GFN-FF and GFN*n*-xTB methods is promising for future applications in protein-drug design, gas storage, explicit solvation, free energy computations, and IR spectra interpretation.

Zusammenfassung

Die moderne Chemie kennt fast keine Grenzen in der elementaren Zusammensetzung und Größe der Moleküle. Chemische Systeme erreichen heutzutage enorme Dimensionen mit Tausenden von Atomen, aufgrund von bedeutenden Fortschritten auf den Gebieten der Polymer- und supramolekularen Chemie, der Materialwissenschaft und der Biochemie. Damit verbunden ist eine steigende Anforderung an die Leistungsfähigkeit von computergestützten Methoden, um Experiment und Simulation ohne Einschränkungen in Größe und Zusammensetzung des untersuchten Systems zu vergleichen. Allgemeine, schnelle und dennoch genaue Methoden sind deshalb von zunehmender Bedeutung und damit verbundener Popularität im Bereich der theoretischen Chemie. Diese Arbeit widmet sich daher der Entwicklung und Anwendung effizienter Kraftfeld und *tight-binding* (TB) Methoden für die robuste Simulation großer Moleküle auf atomarer Ebene. Insbesondere werden ein generisches Kraftfeld, sowie Verbesserungen bestehender semiempirischer *extended tight-binding* (xTB) Methoden für die akkurate Berechnung von Geometrien, harmonischen Schwingungsfrequenzen und nichtkovalenten Wechselwirkungsenergien vorgestellt. Im Folgenden werden diese Methoden mit dem Akronym GFN bezeichnet.

Von Interesse für experimentelle und theoretische Vergleiche sind oft Molekülgeometrien, freie Reaktionsenergien und spektroskopische Eigenschaften wie Infrarot- (IR) Absorption. Deren Beschreibung benötigt nicht nur akkurate Energien, sondern auch effiziente Gradienten (erste Ableitungen), harmonische Schwingungsfrequenzen (zweite Ableitungen) und entsprechende Solvatationsmodelle. Mit zunehmender Molekülgröße und damit verbundener Komplexität nimmt jedoch die Anzahl der computergestützten Methoden, die für Energien, Geometrien und Frequenzen anwendbar sind, rapide ab. Für *ab initio* Elektronenstrukturmethoden liegt die Grenze daher bei nicht viel mehr als hundert Atomen womit eine vollständig quantenmechanische (QM) Beschreibung nicht möglich ist. Den nächsten konsequenten Schritt zu höherer Recheneffizienz stellen semiempirische quantenmechanische (SQM) Methoden dar, auch wenn sie aufgrund unvollständiger Parametrisierungen oder konzeptioneller Defizite oft nicht allgemein anwendbar sind. Dies hat sich erst kürzlich durch die Entwicklung der GFN*n*-xTB Methoden geändert ($n = \{0, 1, 2\}$), die für einen Großteil des Periodensystems bis hin zu Radon parametrisiert sind. Die zugrundeliegenden Näherungen erweitern die zugängliche Molekülgröße auf bis zu eintausend Atome. Jedoch ohne massiv parallele Supercomputer bleibt die Beschreibung größerer Systeme verwehrt und man muss auf klassische Ansätze wie Kraftfelder zurückgreifen. Obwohl viele verschiedene Arten von Kraftfeldern existieren, stellen genaue und gleichzeitig allgemeine Varianten eine fast leere Stelle im Repertoire der theoretischen Methoden dar. Daher wird die Entwicklung genauerer (polarisierbarer) Kraftfelder als "heiliger Gral" für die computergestützte organische Chemie und Biochemie bezeichnet.

Der erste Teil dieser Arbeit präsentiert ein neues generischen Kraftfeldes im Rahmen der GFN Methoden. Diese als GFN-FF bezeichnete Methode stellt ein einzigartiges, teilweise polarisierbares, universelles Kraftfeld für die akkurate Beschreibung der Struktur und Dynamik großer Moleküle dar. GFN-FF verbindet dabei die Geschwindigkeit eines Kraftfeldes mit nahezu quantenmechanischer Genauigkeit. Was es von anderen etablierten Kraftfeldern unterscheidet, ist eine nahezu vollständige Parametrisierung des Periodensystems ($Z \leq 86$) und eine automatisierte Routine zum Aufsetzen der Topologie und zur Bestimmung der Kraftkonstanten. Um eine hohe Genauigkeit für NCIs zu erreichen, wird ein hochentwickeltes Ladungsmodell, das auf Elektronegativitätsausgleich von gaußförmigen Ladungsdichten basiert verwendet und die Behandlung von Pauli Repulsion und London Dispersionswechselwirkungen ist analog zu den TB Methoden. Zusätzlich wird eine neuartige Wasserstoffbrückenbindungskorrektur eingeführt. In dieser Arbeit wird eine detaillierte Beschreibung der zugrundeliegenden Theorie gegeben, gefolgt von illustrativen Anwendungsbeispielen. Es wird gezeigt, dass für Strukturen von metallorganischen Gerüsten und Biomakromolekülen die GFN-FF optimierten Strukturen gut mit den experimentellen Kristallstrukturen übereinstimmen. Des weiteren erreicht GFN-FF auf etablierten Benchmarksätzen für Konformations- und nichtkovalente Wechselwirkungsenergien oft eine Genauigkeit, die mit SQM Methoden oder sogar mit wesentlich anspruchsvolleren GGA Dichtefunktionalen vergleichbar ist.

Der nächste Teil dieser Arbeit erforscht neue Möglichkeiten, die GFN-FF für große und komplexe Systeme bietet. Der Fokus liegt hierbei auf der Kombination mit dem Konformer-Rotamer-Ensemble Generator (CREST) für die Erforschung des niederenergetischen molekularen Konformationsraums. Der Anwendungsbereich reicht hierbei von Proteinen bis hin zu metallorganischen Gerüsten. In einer ersten anwendungsbezogenen Studie, wird die Gasspeicherung von Treibhausgasen und Biotreibstoffen, wie Kohlendioxid und Methanol, in metallorganischen Gerüsten und porösen organischen Käfigen untersucht. Bevorzugte Bindungsstellen werden mit dem CREST-Algorithmus auf GFN-FF Theorieniveau bestimmt und mit DFT Methoden nach optimiert. Die mit GFN*n*-xTB und GFN-FF berechneten Assoziationsenergien zeigen eine vergleichbare Genauigkeit wie die gut funktionierenden (meta-)GGAs. Als zweite Studie werden Spin–Spin Abstandsverteilungen für Nitroxid markierte Mutanten von Azurin und T4 Lysozym durch MD Simulationen auf GFN-FF Niveau modelliert und mit experimentellen EPR Ergebnissen verglichen. Mit Abweichungen von weniger als 2 Å von den mittleren Spin–Spin Abständen zum Experiment übertrifft die Genauigkeit von GFN-FF alle konkurrierende Methoden.

Im letzten Teil dieser Arbeit werden GFN Methoden für die Berechnung von harmonischen Schwingungsfrequenzen getestet, aus denen der thermostatistische Beitrag zu den freien Energien innerhalb der modifizierten Rigiden-Rotor-Harmonischer-Oszillator (RRHO) Näherung hergeleitet wird. Die Genauigkeit von GFN2-xTB und GFN-FF wird im Vergleich zu DFT Referenzdaten bewertet. Als Ausblick auf zukünftige Anwendungen werden freie Assoziationsenergien, auch unter Einbeziehung von Solvationseffekten, für Protein-Wirkstoff Komplexe mit fast 5000 Atomen berechnet. Darüber hinaus wird eine neue Methode, *single-point Hessian* (SPH) genannt, für verbesserte harmonischen Schwingungsfrequenzen von allgemeinen Nicht-Gleichgewichtsstrukturen eingeführt, bei der die Eingangsgeometrie durch die Anwendung eines Bias-Potentials erhalten bleibt. Dadurch ermöglicht der SPH Ansatz die Berechnung von genauen thermodynamischen Beiträgen an jedem Punkt der Energiehyperfläche. Sig-

nifikante Verbesserungen der thermostatistischen Beiträge und IR Spektren werden durch den SPH Ansatz auf SQM und Kraftfeld Niveau erreicht, wenn z.B. DFT Strukturen als Startpunkt zur Verfügung gestellt werden. Abschließend wird der Effekt von expliziter Solvatation für die Berechnung von IR Spektren untersucht. Zum ersten Mal wird dazu ein neuartiger Algorithmus namens *quantum cluster growth* (QCG) angewendet. Die auf diese Weise berechneten Ergebnisse kommen den experimentellen Referenzspektren bemerkenswert nahe.

Insgesamt stellen die in dieser Dissertation entwickelten und evaluierten Methoden für die theoretischen Chemie einen großen Sprung nach vorn dar und schließen somit weiter die Lücke zwischen Theorie und Experiment in der Beschreibung von großen Molekülen. GFN-FF und SPH Berechnungen ergänzen das Portfolio der computergestützten Chemie und stellen wertvolle und vielseitige Werkzeuge für die theoretische Vorauswahl und Modellierung dar. Von metallorganischen bis hin zu biochemischen Systemen, die einzigartige Kombination aus Effizienz, Universalität und Genauigkeit der GFN-FF und GFN*n*-xTB Methoden ist vielversprechend für zukünftige Anwendungen im Bereich des Protein-Wirkstoff Designs, der Gasspeicherung, der expliziten Solvatation, der Berechnung von freien Energie und der Interpretation von IR Spektren.

Contents

Part I.

Introduction and Theoretical Background

1. Introduction

Theoretical chemistry has evolved over the past decades from a minority in science into an indispensable companion of the experiment and an important and independent part of chemistry.[1–3] With the underlying laws of physics, the practical programming in computer codes, and the application to chemical as well as biological problems, computational chemistry touches and combines many fields of science.[4–14] Besides the verification and interpretation of experimental findings, computer simulations can yield further insights into the investigated systems on an atomistic level and reveal energetical, structural, and dynamical effects.[15–18] From a philosophical point of view, the computer may be regarded as a the microscope of the 21st century.

In academia and industry, computational simulations have found their way into scientific workflows.[19–22] A general scheme thereof is depicted in Figure 1.1. The driving force for academic and industrial research is usually a chemical idea or problem. Here, some of the largest and most important challenges of this century arise from the reasearch fields of catalyst,[23–26] drug,[27–30] and material design,[31–33] energy storage,[34–37] and sustainable chemistry.[38–41] At the beginning of a scientific work, there is usually a large number of potential compounds or substances (candidates) that look promising in terms of solving the problem. A lot of time and large amounts of resources are spent investigating and screening all possible candidates to end up with a much smaller number of final products that show the desired properties. At this point, theoretical chemistry can intervene and help to accelerate this process by computing equilibrium geometries, molecular dynamics (MD), reaction free energies, and spectroscopic properties.[42,43] While experimental synthesis and analysis are time consuming and hence rather expensive, computational simulations require only a fraction of the time, are inexpensive, and do not generate chemical waste. Especially the latter aspect is important in the context of green chemistry and sustainability.

Chemical research often focuses on a fundamental question: Does a reaction (A + B) take place and which is the thermodynamically most stable product (C or D, see Figure 1.1)?[44–46] One of the most important quantities in this context is the difference in the Gibbs free energy ΔG between products and reactants, named after Josiah Willard Gibbs.[47,48] The free energy provides information on how reaction mechanisms proceed, which intermediates are formed, how high the reaction barriers are, and which products are populated.[49,50] The corresponding reaction kinetics may also be derived from ΔG and are accessible via transition state theory (TST), developed by Henry Eyring.[51,52] The accurate calculation of free energies for all types of atoms and molecules remains one of the largest tasks for theoretical chemistry today and in the near future.[2] This is because ΔG is a very complex quantity that consists of multiple contributions, *i.e.*, the molecular geometry, gas phase energy, thermodynamics, and environmental effects, each of which is part of current theoretical research. This thesis focuses on

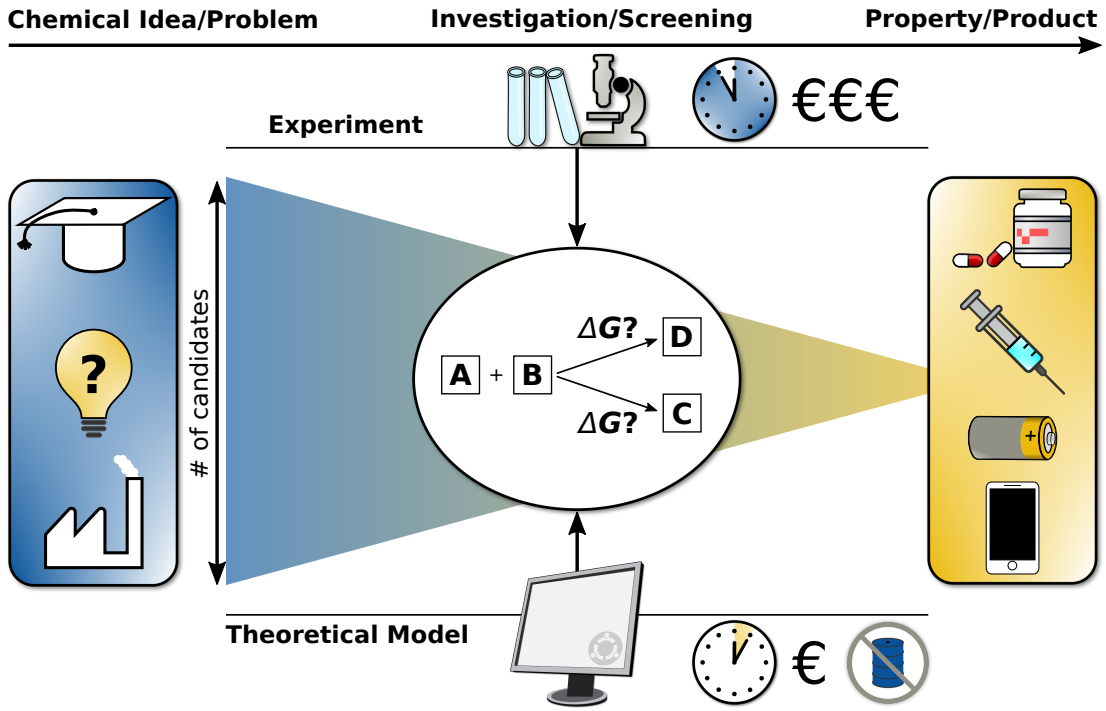


Figure 1.1.: General workflow of scientific research in academia and industry. To solve a chemical problem or to realize an idea, large amounts of possible compounds and substances (candidates) have to be investigated to end up with a small number of final products. Computational simulations can facilitate this process.

the improvement of ΔG and all of its components through the use of fast, accurate, and thus, efficient computational methods. In particular, large molecular systems (10^2–10^4 atoms) are investigated. For a general overview of condensed-phase systems under periodic boundary conditions (PBC), the reader is referred to Refs. 53,54. The individual contributions to the free energy are illustrated in Figure 1.2.

In the experimental laboratory, chemical reactions usually take place in solution at chosen temperature and pressure. Commonly 10^{20} molecules of a compound, which is less than 1 mmol, and additionally 10^{23} solvent molecules are present. In the initial theoretical model, on the contrary, often only a single molecule or a set of representative conformations is considered in vacuum at absolute zero (T = 0 K). The input geometry for the theoretical simulation is either generated manually,[55] or taken from crystallographic databases.[56,57] The corresponding gas phase energy E of this structure is calculated by molecular energy methods. To find the most stable atomic arrangement of the molecule, where the net inter-atomic force on each atom is zero, it is necessary to arrive at a local minimum on the potential energy surface (PES). Therefore, a geometry optimization is performed in which the molecular energy is minimized with respect to the nuclear coordinates $\mathbf{R}$ to yield the equilibrium geometry $\mathbf{R}_e$.[58] Note that an optimized geometry (stationary point on the PES) is always connected to the energy method that defines the PES. For flexible molecules, a single geometry optimization is usually not sufficient and multiple conformations need to be considered. Here, more sophisticated chemical space exploration algorithms have to be applied to determine the low-energy conformer ensemble.[59–61]

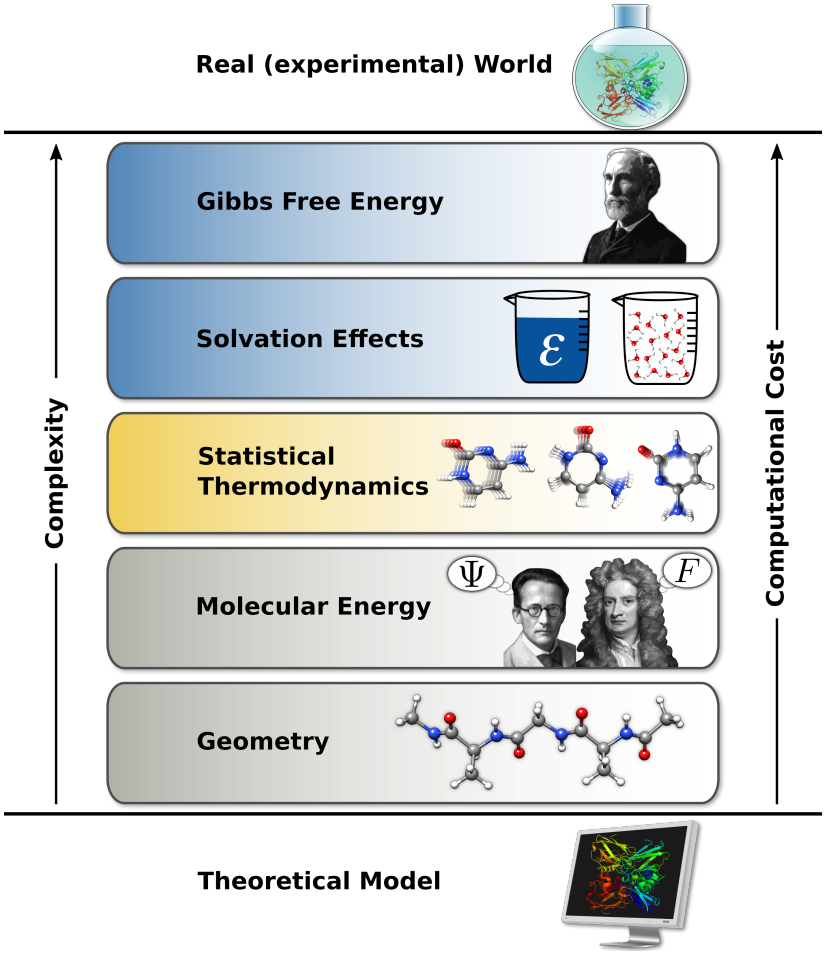


Figure 1.2.: Composition of the Gibbs free energy. The molecular energy is calculated on optimized geometries. Molecular degrees of freedom are added by statistical thermodynamics and the influence of the environment is accounted for by, *e.g.*, solvation effects.

The molecular energy can be calculated by molecular mechanics (MM) and by quantum mechanical (QM) methods. MM methods treat the dynamics of atoms classically, *i.e.*, by Newton's second law of motion, and are therefore often referred to as force-fields (FF).[62–65] In FFs, the electrons are no individual particles and the electronic structure is rather described by parametric functions. QM methods, on the other hand, treat the electrons as quantum mechanical particles within the Born–Oppenheimer approximation,[66] while the nuclei are still considered as classical particles. For the description of the electrons, the wave function Ψ is introduced, which is obtained by solving the electronic Schrödinger equation.[67] Hence, the respective QM methods are commonly referred to as electronic structure methods. A more detailed description of the underlying theory is given in Chapter 2. The molecular degrees of freedom (DOF) regarding translation, rotation, and vibration are added by statistical thermodynamics. The sum of these thermostatistical contributions, also including zero-point vibrational energy (ZPVE), enables the transition from the (electronic) gas phase energy (0 K) to the Gibbs free energy at a given temperature.[68,69] If the experiment is performed in solution, additional solvent effects must be taken

into account by a suitable solvation model. Here, the most common representatives are either implicit continuum models,[70,71] or explicit approaches that introduce more molecules and often increase the system size drastically.[72]

The underlying physical laws and mathematical equations necessary for the description of the molecular energy are exactly known, but the emerging dimensions are much too many and computationally demanding to be generally solvable.[73] Note that for ΔG at least the first and second order derivative of the molecular energy with respect to the nuclear displacements are required for equilibrium geometries and thermodynamics, which additionally increases the overall complexity. For more than a few atoms,[74] the full QM description is thus not possible and with Moore's law nearing its end, *i.e.*, silicon-based microprocessors reach their capability limits,[75,76] computer hardware is not the solution to the problem. Therefore, it becomes necessary to develop approximate, more efficient, and yet accurate QM and FF methods, which are applicable to large molecules for the calculation of energies, geometries, and thermodynamic quantities. An overview of existing theoretical methods is given in Figure 1.3, focusing on the degree of empiricism (approximations) and the applicable system size.

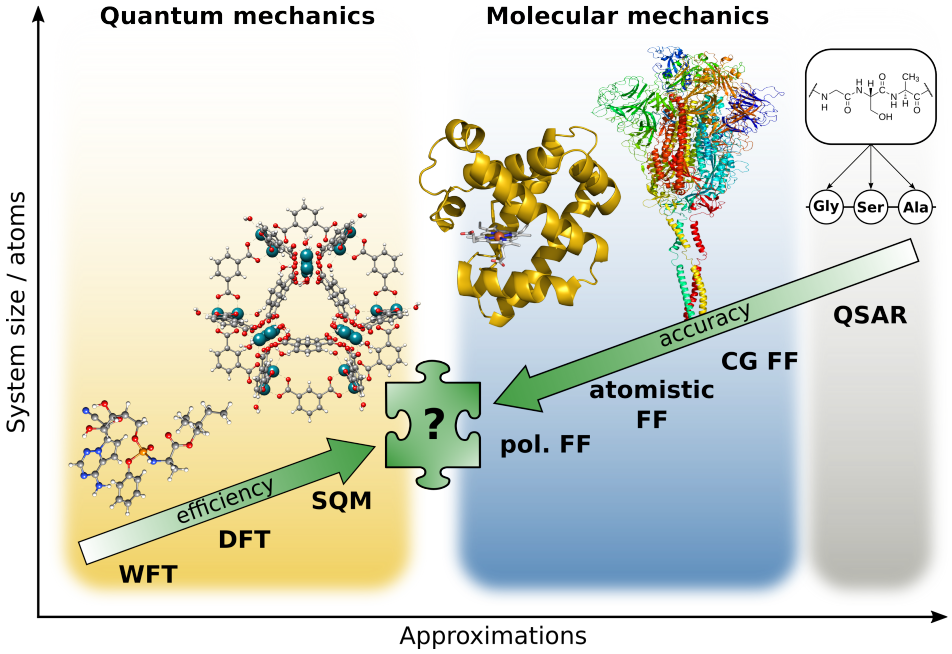


Figure 1.3.: Overview of computational methods according to their degree of empiricism and applicability. The shown molecules are remdesivir, a rhodium–organic cuboctahedra, zinc azurin, and the spike protein of SARS-CoV-2.

Wave function theory (WFT) is conceptually closest to the exact solution of the electronic Schrödinger equation and thus, often the most accurate method available.[17] In correlated WFT, the electrons are fully interacting, which leads to many-body problems and incurs high computational costs. Molecules described by different WFT methods are mostly limited to less than one hundred atoms (see remdesivir[77] in Figure 1.3), depending on the applied approximations. In the simplest form of WFT, which is called Hartree–Fock (HF) theory,[78,79] the exact many-body wave function of the system is approximated by

a single Slater determinant (see Section 2.1.2).[80] Kohn–Sham density functional theory (KS-DFT or simply DFT) draws the connection between the energy of a system and its electron density (see Section 2.1.5).[81,82] With the introduction of reasonable approximations regarding the electron exchange-correlation (XC), DFT methods can routinely treat systems with more than one hundred atoms.[83] Consequently, DFT is the most frequently used electronic structure method in computational chemistry and was rewarded with the Nobel price in 1998 to Walter Kohn.[84] Due to the low amount of empiricism in both, WFT and DFT, they are generally applicable to any atom type or molecule and show high accuracy. A more detailed description of WFT and DFT methods is given in Chapter 2. Semiempirical quantum mechanical (SQM) methods introduce further approximations to *ab initio* schemes (see Section 2.1.6). Formally, SQM methods are related to electronic structure methods because they are based on either HF or DFT frameworks. The replacement of various interaction terms by empirical potentials and the use of minimal basis sets accelerates the calculation times and increases the treatable system size to roughly one thousand atoms. This enables the routine description of metal-organic polyhedra such as the rhodium containing cuboctahedra in Figure 1.3.[85] A disadvantage of SQM methods is that parameters must be introduced for empirical approximations, limiting general applicability. Prominent examples for SQM methods are the widely used, NDDO-based PM*x*[86–88] methods derived from HF, as well as the DFT based density functional tight-binding[89–92] (DFTB) methods and the more recently developed extended tight-binding methods of the GFN*n*-xTB family.[93–96] An extensive overview of SQM methods can be found in Refs. 97,98. Neglecting the electronic structure of a molecule and replacing it by interatomic interaction potentials is the main approximation in atomistic FFs.[15] This simplification leads to a tremendous speedup compared to SQM methods and molecules beyond thousand atoms can be described. FFs are thus frequently applied for the simulation of proteins such as zinc azurin.[99] Nevertheless, the functional form of the FF potentials introduces a (large) set of atom and atom-pair specific parameters. Such parameters often only exist for a limited number of elements and structural motifs, requiring individual, laborious parametrizations. Here, a distinction is made between classical atomistic FFs with fixed, usually atom-centered, charges and polarizable FFs, which account for variations in the charge distribution within a dielectric environment.[100,101] Leaving the atomistic description, coarse-grained (CG) FFs represent the next level of empiricism.[102] The combination of several atoms into chemically related fragments allows the description of millions of atoms, such as in the spike protein of the SARS-CoV-2 virus shown in Figure 1.3.[103] Finally, chemoinformatic methods must also be mentioned, as they are often represented by quantitative structure-activity relationships (QSAR).[104–106] At the core of QSAR modeling are chemical descriptors reflecting various levels of chemical structure representation, which range from one-dimensional molecular formulas to two-dimensional structural formulas.[107] A second important class of chemoinformatics are neural networks, which aim at both, WFT accuracy and FF speed.[108,109] Machine learning (ML) potentials are not considered in this work, for more information see Refs. 110,111.

A large gap regarding efficiency, generality, and accuracy exists between FF and (S)QM methods. The speedup of SQM methods is limited by expensive mathematical operations,[112] such as matrix diagonalization, that show a cubic scaling with the system size. Accurate special purpose FFs exist, such as AMBER,[62] CHARMM,[64] and OPLS,[113] but lack generality and general FFs such as UFF,[114] DREI-

DING,[115] and ESFF[116] show insufficient accuracy. For this reason, accurate (polarizable) force-fields are named as a remaining holy grail in computational chemistry.[2] Within this thesis, a new generic FF is developed that exceeds general FFs in terms of accuracy performing more like a specialized FF, and thus, represents a missing piece in the "puzzle" of computational methods (see Figure 1.3). This method, termed GFN-FF,[117] is a completely automated, partially polarizable, generic FF designed in close accordance with the GFN*n*-xTB methods. Furthermore, all important terms that contribute to the Gibbs free energy, namely geometries, reaction energies, and thermodynamics, are extensively tested and improved by GFN methods (SQM and FF) for large molecules ranging from organometallic to biochemical systems.[96] Therefore, this thesis deals with the following three objectives:

1. Development of GFN-FF, a generic force-field for the full periodic table (Part II).
2. Application of efficient computational methods for organometallic and biochemical systems (Part III).
3. Improving vibrational frequencies and thermodynamics for association free energies of large molecules (Part IV).

Overall, this thesis presents the development, evaluation, and application of new efficient computational methods for the sophisticated simulation of large molecular systems. These methods are tested for their individual performance and in combination with other existing state-of-the-art procedures to enable new computational schemes and optimize chemical screening protocols. Thereby, various subdomains of theoretical chemistry including molecular dynamics, thermodynamics, free energies, and related properties such as infrared (IR) spectroscopy of organic and inorganic compounds are studied in detail. In Chapter 2 of Part I, an overview of the theoretical background of selected key methods relevant for this thesis is given. Further, in Part II, GFN-FF is introduced. First, a detailed description of the underlying theory is given followed by illustrative application examples. GFN-FF optimized structures are compared to the crystal structures of organometallic and biochemical systems. On established benchmark sets for conformational and noncovalent interaction energies, the performance of GFN-FF is compared to GFN*n*-xTB, PM*x*, and much more sophisticated density functionals. Part III of this thesis explores new possibilities of GFN-FF in combination with the conformer-rotamer ensemble sample tool (CREST). The gas storage of greenhouse gases and bio-fuels, such as carbon dioxide and methanol, in metal-organic frameworks (MOFs) and porous organic cages (POCs) is investigated. Further, spin-spin distance distributions for nitroxide labeled mutants of azurin and T4 lysozyme are modeled by MD simulations (Appendix A2.3) at the GFN-FF level of theory and compared to experimental EPR results. In Part IV, GFN methods are assessed for the calculation of the thermostatistical contributions to the free energies. The accuracy of GFN*n*-xTB and GFN-FF is benchmarked in comparison to DFT reference data. Free association energies, also including solvation effects, are calculated for protein-drug complexes reaching from more than 2500 to almost 5000 atoms. In addition, a new method termed single-point hessian (SPH) is introduced for improved thermodynamics and IR spectra of general non-equilibrium structures on every point of the PES. Also, the effect of explicit solvation is investigated in the context of IR spectra. For this purpose, a new algorithm termed quantum cluster growth (QCG) is developed and tested. Part V finally summarizes the achievements of this thesis and provides perspectives regarding their use in chemical research.

2. Theoretical Background

Chemistry is referred to as the science of molecular construction, transformation, and change.[118] Changes in free energies within a chemical reaction give insights into the underlying mechanisms and reveal possible products. Hence, relative free energies are a fundamental quantity in chemistry. The association free energy of molecules A and B (*e.g.*, host and guest) forming a noncovalently bound complex C is calculated in the supramolecular approach according to $\Delta G = G(C) - G(A) - G(B)$.[119,120] This notation is followed throughout this work and Δ always refers to the difference between product and reactant for a given quantity.

Thermodynamically, the Gibbs free energy of a reaction at temperature T is defined by the reaction enthalpy ΔH and entropy ΔS as given in Eq. 2.1.[15] The reaction enthalpy ΔH can be divided into a temperature dependent $\Delta H(0\,K \rightarrow T)$ and independent $\Delta H(0\,K)$ part, as well as volume work $p\Delta V$ (see Eq. 2.1a). Assuming a normal reference state ($1\,\mathrm{mol\,L^{-1}}$), the latter term is a constant for an ideal gas with a molar volume of $24.8\,\mathrm{L\,mol^{-1}}$ (1 bar, 298.15 K). The temperature dependency in ΔH (and also ΔS) results from molecular translations, rotations, and vibrations, which are themselves functions of T.

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) \tag{2.1}$$

$$= \overbrace{\Delta H(0\,K) + \Delta H(0\,K \rightarrow T) + p\Delta V}^{\Delta H(T)} - T\Delta S(T) \tag{2.1a}$$

$$= \overbrace{\Delta E + \Delta ZPVE}^{\Delta H(0\,K)} + \Delta H(0\,K \rightarrow T) + p\Delta V - T\Delta S(T) \tag{2.1b}$$

$$= \Delta E + \Delta G_{\mathrm{corr.}}(T) \tag{2.1c}$$

The temperature independent part of the reaction enthalpy consists of the difference in the molecular energy $\Delta E = E(C) - E(A) - E(B)$ and the ZPVE (see Eq. 2.1c), which is the residual energy arising from quantum mechanical motions at zero temperature. Finally, all terms besides the molecular energy can be summarized as the thermostatistical contribution $\Delta G_{\mathrm{corr.}}(T)$ to correct from a reaction energy ΔE to a reaction free energy ΔG. This separation is common in theoretical chemistry[119] and a large variety of QM and FF methods exists for the computation of ΔE. On the contrary, the calculation of $\Delta G_{\mathrm{corr.}}(T)$ introduces many difficulties that will be addressed in this thesis.

In this chapter, a general overview is given about the computational methods that are used throughout this work, following the separation *vide supra* in ΔE and its corrections to the Gibbs free energy. The focus lies on methodologies that are applied, but not explicitly developed in this thesis. First, molecular energy methods will be discussed. Electronic structure theories, such as Hartree–Fock, Kohn–Sham DFT and extended tight-binding methods, are introduced in Section 2.1 and classical force-fields in

Section 2.2. In the second part of this chapter (Section 2.3), the most important contributions to the free energy will be given, namely thermodynamics, solvation effects, and molecule conformations. The following theoretical background is mainly based on Refs. 15–17,96.

2.1. Electronic Energy Methods

2.1.1. Definition of the Electronic Hamiltonian

The object of study in electron structure theory is the movement of negatively charged electrons in the presence of positively charged nuclei in atoms or molecules.[17] Electrons are very light particles and they cannot be described correctly by classical mechanics. In the Schrödinger picture of quantum mechanics[121], the electronic system is described by the time-dependent (TD) wave function $\Psi_K(t)$ according to Eq. 2.2

$$\Psi_K(t) = \Psi_K e^{-iE_K t}. \tag{2.2}$$

Here, Ψ_K is the amplitude of the wave function in the stationary electronic state K and $e^{-iE_K t}$ refers to the time-dependent phase factor. The evolution of the electrons in time is given by the non-relativistic, time-dependent Schrödinger equation (SE)[67]

$$i\hbar\frac{\partial}{\partial t}\Psi_K(t) = \hat{H}\Psi_K(t)\,, \tag{2.3}$$

which describes the quantum mechanical connection between the TD wave function $\Psi_K(t)$ and the Hamiltonian $\hat{H}$ (*vide infra*), *i.e.*, the operator corresponding to the total energy. Because $\hat{H}$ in Eq. 2.3 is time-independent, the TD part of the wave function can be separated for stationary states. This way, the time-independent SE is obtained[15,16]

$$\hat{H}\,\Psi_K = E_K\Psi_K\,, \tag{2.4}$$

where E_K is the energy of the electronic state K and Ψ_K the corresponding wave function. Both are obtained as the respective eigenvalue and eigenfunction of the non-relativistic Hamiltonian. In the following, atomic units are used throughout.[16] $\hat{H}$ is composed of kinetic ($\hat{T}$) and potential energy operators ($\hat{V}$) for the electrons and the nuclei, indicated by e and n, respectively (Eq. 2.5)

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{V}_{ee} + \hat{V}_{ne} + \hat{V}_{nn}\,. \tag{2.5}$$

Within the Born–Oppenheimer approximation (BOA),[66] the wave functions of atomic nuclei and electrons are treated separately. This is based on the fact that the nuclei are much heavier than the electrons. [a] As a result of the BOA, the kinetic energy operator of the nuclei ($\hat{T}_n$) and the nuclei–nuclei repulsion operator ($\hat{V}_{nn}$) can be removed from Eq. 2.5, leading to the electronic Hamiltonian given

[a] The mass of an electron is $5.485{\cdot}10^{-4}$ u, whereas the mass of a proton is 1.007 u.[122] Hence, the electron is three orders of magnitude lighter than a proton.

in Eq. 2.6[17]

$$\hat{H} \equiv \hat{H}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{ne} = \hat{h}_e + \hat{V}_{ee} \tag{2.6a}$$

$$= -\frac{1}{2}\sum_{i}^{N} \hat{\nabla}_i^2 + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_{i}^{N}\sum_{A}^{O} \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} \,. \tag{2.6b}$$

The electronic Hamiltonian consists of the kinetic energy operator of the electrons $\hat{T}_e$, the electron–electron interaction operator $\hat{V}_{ee}$, and the Coulomb interaction between electrons and nuclei $\hat{V}_{ne}$ (*cf.* Eq. 2.6a). All single-electron components are often expressed by the one-electron operator $\hat{h}_e$, *i.e.*, $\hat{T}_e$ and $\hat{V}_{ne}$. In Eq. 2.6b, N is the total number of electrons and O the number of nuclei. Z corresponds to the respective nuclear charge of atom A, while $\mathbf{r}$ and $\mathbf{R}$ represent the position vectors of the electrons, denoted by the subscripts i and j, and the nuclei. Solving the electronic Schrödinger equation yields the electronic energy as the expectation value of the electronic Hamilton operator given in Dirac's bra–ket notation[123] (Eq. 2.7)

$$\int_{-\infty}^{+\infty} \Psi_K^* \hat{H} \Psi_K d\mathbf{a} \equiv \left\langle \Psi_K \left| \hat{H} \right| \Psi_K \right\rangle \equiv H_{KK} = E_K \left\langle \Psi_K | \Psi_K \right\rangle = E_K \,. \tag{2.7}$$

On the left hand side of Eq. 2.7 the integration is done over all variables $\mathbf{a}$. The overlap integral on the right hand side of Eq. 2.7 vanishes, since the wave function Ψ_K is orthonormalized according to Eq. 2.8

$$\langle \Psi_K | \Psi_L \rangle = \delta_{KL} = \begin{cases} 0 & \text{for } K \neq L \\ 1 & \text{for } K = L \end{cases} , \tag{2.8}$$

where δ_{KL} is the Kronecker delta. Since the exact wave function necessary to solve the electronic Schrödinger equation is *a priori* unknown, it is the objective of WFT to determine appropriate approximations. One of the most important approximations in this context was introduced by Hartree and Fock.[78,79] HF theory is a cornerstone in theoretical chemistry and forms the basis for more accurate WFT methods.

2.1.2. Hartree–Fock Theory

To generate approximate solutions to the exact N-body wave function Ψ_K, the variational principle is employed, which states that any approximate wave function has an energy $\tilde{E}_K$ above or equal to the exact energy E_K (Eq. 2.9)[124]

$$\tilde{E}_K = \frac{\left\langle \tilde{\Phi}_K \left| \hat{H} \right| \tilde{\Phi}_K \right\rangle}{\left\langle \tilde{\Phi}_K | \tilde{\Phi}_K \right\rangle} \geq \left\langle \Psi_K \left| \hat{H} \right| \Psi_K \right\rangle = E_K \,. \tag{2.9}$$

Therefore, a trial wave function $\tilde{\Phi}_K$ is constructed from a set of parameters, and its energy is minimized as a function of these parameters to determine the trial function that is closest to the exact Ψ_K.[15] In Hartree–Fock theory, the trial function of the exact N-body wave function is approximated by a single

Slater determinant (SD) for the electronic ground state Φ_0 (Eq. 2.10) containing N electrons

$$\Psi_K \approx \tilde{\Phi}_K \equiv \Phi_0(1,2,\cdots,\mathrm{N}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) & \cdots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \cdots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathrm{N}) & \phi_2(\mathrm{N}) & \cdots & \phi_N(\mathrm{N}) \end{vmatrix} . \tag{2.10}$$

$\phi_i(k)$ denotes the i^{th} molecular orbital (MO) occupied by the k^{th} electron $\in \{1,\cdots,N\}$. A molecular orbital is defined as the wave function of a single particle system and consists of the spin part $\sigma_i \in \{\alpha,\beta\}$, and a spatial MO $\psi_i(\mathbf{r}_1)$ that depends on the position of electron $\mathbf{r}_1$ in Cartesian space (Eq. 2.11)

$$\phi_i(1) = \sigma_i \psi(\mathbf{r}_1) . \tag{2.11}$$

The Pauli principle[125] states that the total electronic wave function must be antisymmetric with respect to the interchange of any two electrons (fermions) 1 and 2, *i.e.*, $\Psi(1,2) = -\Psi(2,1)$.[16] Hence, the antisymmetrized product of one-particle functions represented in an SD is a valid wave function that is exact for non-interacting particles.

Inserting Φ_0 into the time-independent SE (*cf.* Eq. 2.4) yields the energy of a single SD. This corresponds to the HF energy of the N-electron wave function given by Eq. 2.12[15,16]

$$E_{\mathrm{HF}} = \sum_{i=1}^{N} h_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(J_{ij} - K_{ij}\right) + V_{NN} \tag{2.12a}$$

$$= \sum_{i}^{N} \langle \phi_i | h_i | \phi_i \rangle + \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N}\left(\left\langle \phi_j \middle| \hat{J}_i \middle| \phi_j \right\rangle - \left\langle \phi_j \middle| \hat{K}_i \middle| \phi_j \right\rangle\right) + V_{NN} . \tag{2.12b}$$

$\hat{h}_i$ is the one-electron operator (*cf.* Eq. 2.6a). The J_{ij} matrix elements are the Coulomb integrals and represent the classical Coulomb interaction (repulsion) between two charge distributions $\phi_i^2(1)$ and $\phi_j^2(2)$. The K_{ij} matrix elements are termed as (Pauli) exchange integrals and have no classical analogy. The HF energy is further expressed in terms of Coulomb $\hat{J}$ and exchange $\hat{K}$ operators (*cf.* 2.12b), which are given by Eq. 2.13

$$\hat{J}_i \left|\phi_j(2)\right\rangle = \left\langle \phi_i(1) \left| \frac{1}{\mathbf{r}_1 - \mathbf{r}_2} \right| \phi_i(1) \right\rangle \left|\phi_j(2)\right\rangle \tag{2.13a}$$

$$\hat{K}_i \left|\phi_j(2)\right\rangle = \left\langle \phi_i(1) \left| \frac{1}{\mathbf{r}_1 - \mathbf{r}_2} \right| \phi_j(1) \right\rangle \left|\phi_i(2)\right\rangle . \tag{2.13b}$$

With this, a set of MOs that minimize the HF energy whilst remaining orthogonal and normalized has to be determined. This is achieved by Lagrange multipliers in a constrained optimization where the condition is that the Lagrangian is stationary with respect to a small variation in the orbitals (Eq. 2.14)[15]

$$L = E - \sum_{i}^{N} \sum_{j}^{N} \lambda_{ij} \left(\langle \phi_i | \phi_j \rangle - \delta_{ij} \right) \tag{2.14a}$$

$$\partial L = \partial E - \sum_{i}^{N} \sum_{j}^{N} \lambda_{ij} \left(\langle \delta\phi_i | \phi_j \rangle - \langle \phi_i | \delta\phi_j \rangle \right) = 0 \,. \tag{2.14b}$$

Variation of the orbitals subject to this constrain leads to an effective one-electron problem known as the canonical Hartree–Fock equations (Eq. 2.15)[16]

$$\hat{f}_i \tilde{\phi}_i = \epsilon_i \tilde{\phi}_i \,, \tag{2.15}$$

where $\tilde{\phi}_i$ is a special set of molecular orbitals called canonical MOs and $\hat{f}$ is the Fock operator (Eq. 2.16)

$$\hat{f}_i(\mathbf{r}_1) = \hat{h}_i(\mathbf{r}_1) + \sum_{j}^{N} \left(\hat{J}_j(\mathbf{r}_1) - \hat{K}_j(\mathbf{r}_1) \right) \,. \tag{2.16}$$

The Hartree–Fock equations form a set of pseudo-eigenvalue equations since the Fock operator depends on all the occupied MOs via the Coulomb and exchange operators (Eq. 2.16).[15] A specific Fock orbital can only be determined if all the other occupied orbitals are known. Therefore, the Hartree–Fock equations are solved iteratively in a self-consistent field (SCF) procedure. The Fock orbital energies are matrix elements of $\hat{f}_i$ with the canonical MOs (Eq. 2.17a). The tilde notation is omitted for simplicity

$$\epsilon_i = \left\langle \phi_i \left| \hat{f}_i \right| \phi_i \right\rangle = h_i + \sum_{j}^{N} \left(J_{ij} - K_{ij} \right) \tag{2.17a}$$

$$E_{\mathrm{HF}} = \sum_{i}^{N} \epsilon_i - \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} \left(J_{ij} - K_{ij} \right) + V_{NN} \,. \tag{2.17b}$$

The total energy can then be reformulated in terms of MO energies using the definition of the Fock operator. Because it counts the electron–electron repulsion twice, the total energy is not the sum of MO orbital energies. Also, note that for one electron the Coulomb J_{ii} and exchange K_{ii} terms exactly cancel. HF and related WFT methods are therefore self-interaction error (SIE) free, which is one of the major issues in DFT (*vide infra*). In HF, the electron–electron repulsion is accounted as an average effect. It is assumed that the spatial distribution between an electron and all the other electrons is described by one set of orbitals, due to the approximation of a single SD as the trial wave function. Hartree–Fock is therefore referred to as a mean-field theory, which implies that electron correlation (*vide infra*) is neglected. This drawback is present also in other mean-field theories such as DFT.

2.1.3. Basis Set Approximation

The HF equations (Eq. 2.15) represent a set of integro-differential equations and remain unsolvable since the required MOs are unknown *a priori*.[16] An efficient way to express the unknown MOs in terms

of a set of known functions was introduced by Roothaan and Hall.[126,127] Here, the spatial part ψ_i of the molecular orbitals is expanded as a linear combination of M atom-centered orbitals χ_μ (LCAO) (Eq. 2.18)[128]

$$\psi_i(\mathbf{r}_1) = \sum_{\mu}^{M} c_{\mu i}\chi_\mu(\mathbf{r}_1)\,. \tag{2.18}$$

Note that these orbitals are generally no solutions to the HF problem. In Eq. 2.18, the unknowns are the coefficients $c_{\mu i}$ instead of the functions themselves. Inserting the expansion in Eq. 2.15 yields

$$\hat{f}_i \sum_{\mu}^{M} c_{\mu i}\chi_\mu = \epsilon_i \sum_{\mu}^{M} c_{\mu i}\chi_\mu\,. \tag{2.19}$$

Multiplying from the left by χ_ν followed by integration results in the Roothaan–Hall equations, which are usually expressed in a single matrix equation (Eq. 2.20)[15]

$$\mathbf{FC} = \mathbf{SC}\epsilon\,. \tag{2.20}$$

$\mathbf{F}$ is the Fock matrix with elements $F_{\mu\nu} = \langle\chi_\mu|\hat{f}|\chi_\nu\rangle$, $\mathbf{C}$ is is the corresponding LCAO-MO coefficient matrix, and $\mathbf{S}$ is the atomic orbital (AO) overlap matrix with elements $S_{\mu\nu} = \langle\chi_\mu|\chi_\nu\rangle$. ϵ is a diagonal matrix and contains the MO eigenvalues (orbital energies). For a fixed set of AO functions located on the atoms (basis set), the Roothaan–Hall formalism reduces the solution of the HF equations to a variational optimization of the LCAO-MO coefficients.[17] The elements of the Fock matrix are in general given by Eq. 2.21

$$F_{\mu\nu} = \underbrace{\langle\chi_\mu|\hat{h}|\chi_\nu\rangle}_{H_{\mu\nu}} + \underbrace{\sum_{\lambda}^{M}\sum_{\sigma}^{M} P_{\lambda\sigma}\left[\langle\chi_\mu\chi_\nu|\chi_\lambda\chi_\sigma\rangle - \frac{1}{2}\langle\chi_\mu\chi_\nu|\chi_\sigma\chi_\lambda\rangle\right]}_{G_{\mu\nu}}\,, \tag{2.21}$$

where $H_{\mu\nu}$ is the one-electron part and $G_{\mu\nu}$ is the two-electron part. The evaluation of the two-electron integrals $G_{\mu\nu}$ is the most time consuming part in the Roothaan–Hall formalism and the computational cost of mean-field procedures formally scales to $O(M^4)$ with respect to M atomic orbitals. Further in Eq. 2.21, the density matrix $P_{\mu\nu}$ is introduced, where the matrix elements are given by Eq. 2.22

$$P_{\mu\nu} = \sum_{i}^{M_{\text{occ.}}} n_i C^*_{\mu i} C_{\nu i}\,. \tag{2.22}$$

Here, the sum runs over all occupied MOs ($M_{\text{occ.}}$) and n_i is the respective occupation number. Due to the dependence of the Fock matrix elements on the coefficients $\mathbf{C}$ and the density matrix $\mathbf{P}$, the Roothaan–Hall equations are solved iterative in a SCF procedure until the HF (or KS-DFT) energy (Eq. 2.23) is converged[16]

$$E_{\text{HF}} = \frac{1}{2}\sum_{\mu}^{M}\sum_{\nu}^{M} P_{\mu\nu}\left(H_{\mu\nu} + F_{\mu\nu}\right)\,. \tag{2.23}$$

The Roothaan–Hall approach has become the standard way to solve the SCF, not just for HF, but also for KS-DFT.

Employing basis sets is common practice in quantum chemistry and many pre-compiled sets of AOs exist for each element. Some of the most prominent basis sets are Pople-style basis sets,[129] Dunning's correlation-consistent basis sets,[130–132] and Ahlrichs' basis sets.[133–135] The AO basis sets are given by normalized linear combinations (contractions) of primitive Gaussian-type orbitals (GTO),[136] due to the analytic form and efficient calculation of the involved integrals. GTOs can be written in Cartesian coordinates as

$$\chi(x, y, z) = N x^{l_x} y^{l_y} z^{l_z} e^{-\zeta r^2} , \tag{2.24}$$

where the sum of l_x, l_y, and l_z determines the type of orbital, *e.g.*, $l_x + l_y + l_z = 1$ is a p-orbital. The cardinal number ζ of the basis set is defined by the number of contracted GTOs per AO shell, *i.e.*, the occupied orbital shells of the free atom (s, p, d, f). In an infinitely large and complete basis set (CBS) of AOs, the Roothaan–Hall formalism is equivalent to the exact HF solution. However, this is not practical, and smaller but yet sufficient basis sets (triple-, or quadruple-ζ) are employed. Calculations using finite basis sets result in the so-called basis set superposition error (BSSE) and the basis set incompleteness error (BSIE). For too small basis sets, both can be severe and limit the achievable accuracy.[137,138]

2.1.4. Electron Correlation Methods

In Hartree–Fock theory, the exact electron–electron interaction is replaced by a mean-field, in which N non-interacting electrons experience the average interaction of the remaining $(N - 1)$ electrons.[16] Based on a single Slater determinant, the HF wave function accounts for 99 % of the total energy in a sufficiently large basis set. Nevertheless, the remaining 1 % is often decisive for the correct description of chemical interactions. The difference in energy between the fully interacting system in the given basis set and the HF energy is called the electron correlation energy (Eq. 2.25)[139,140]

$$E_{\text{corr}} = E - E_{\text{HF}} . \tag{2.25}$$

Physically, E_{corr} describes the correlated movement of the electrons avoiding each other. This influences their mutual distance that is on average larger than described by a single SD wave function. In HF, there is no intraorbital correlation from electron pairs with the same spin because the SD in HF is antisymmetric and obeys the Pauli principle. This so-called Fermi correlation is thus inherent in HF theory. The opposite spin correlation is called the Coulomb correlation and is not included in the HF wave function. It has therefore the largest contribution to E_{corr}.[15]

Within a given basis set, the HF method determines the energetically best one-determinant trial wave function. To improve on HF results and include electron correlation, the trial wave function must consist of more than one SD (Φ). A general multi-determinant approach can be written as[16]

$$\Psi = a_0 \Phi_{\text{HF}} + \sum_i a_i \Phi_i . \tag{2.26}$$

Because the HF solution already recovers about 99 % of the total electronic energy, Φ_{HF} is usually a good starting point and a_0 is close to one. All other determinants Φ_i are generated by replacing occupied MOs in the HF determinant with virtual (unoccupied) MOs, resulting in SDs that are singly, doubly, triply, quadruply, *etc.* excited relative to the HF determinant. The maximum number of excitations is N, the

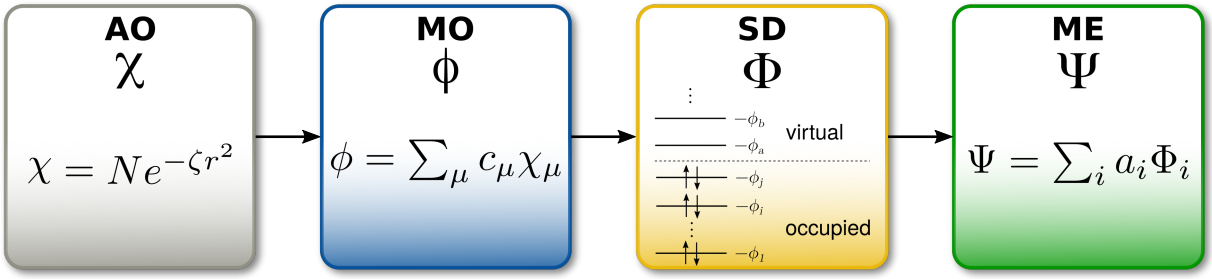


Figure 2.1.: Progression from atomic orbitals (AO) (basis functions), to molecular orbitals (MO), to Slater determinants (SD), and to a many-electron (ME) wave function composed of SDs. Figure adapted from Ref. 15.

number of electrons. The corresponding determinants are often referred to as singles (S), doubles (D), triples (T), quadruples (Q), *etc.* While the basis set (AOs) determines the size of the one-electron basis and limits the description of the MOs, it is the number of included SD that determines the size of the many-electron basis in a multi-determinant wave function. The number of included excitations (SDs) limits the description of electron correlation.[15] A general overview thereof is given in Figure 2.1. The objective for electron correlation methods is now to determine the excitations that contribute the most to $E_{\mathrm{corr.}}$, and to calculate the coefficients in front of the (excited) SDs.

The two most widely known WFT methods that follow this approach are configuration interaction[141,142] (CI) and coupled cluster (CC) theory.[143,144] First, an excitation operator $\hat{\mathbf{T}}$ is defined as $\hat{\mathbf{T}} = \hat{\mathbf{T}}_1 + \hat{\mathbf{T}}_2 + \hat{\mathbf{T}}_3 + \cdots + \hat{\mathbf{T}}_N$, where N is the number of electrons. The $\hat{\mathbf{T}}_i$ operator acts on the HF ground state wave function ($\Phi_{\mathrm{HF}} \equiv \Phi_0$) and generates all i^{th} exited SDs. In the CI formalism, the excitation operator acts linear on the HF wave function, with λ being the order of excitation (Eq. 2.27a)

$$\Psi_{\mathrm{CI}} = \left(1 + \hat{\mathbf{T}}\right)\Phi_0 = \left(1 + \sum_{\lambda}^{N} \hat{\mathbf{T}}_\lambda\right)\Phi_0 \tag{2.27a}$$

$$\Psi_{\mathrm{CC}} = e^{\hat{\mathbf{T}}}\Phi_0 = \sum_{k=0}^{\infty} \frac{1}{k!}\hat{\mathbf{T}}^k\Phi_0\,. \tag{2.27b}$$

The corresponding coupled cluster wave function is defined in Eq. 2.27b. The difference in CC methods is that $\hat{\mathbf{T}}$ is applied as an exponential expanded in a Taylor series and optimized according to the expansion coefficients (excitation amplitudes). This includes all corrections of a given type to infinite order $k = \infty$. In practice, however, the expansion is truncated to finite values of k. The CC wave function contains at each excitation level additional terms arising from products of excitations, so-called disconnected excitations, which are missing in the CI wave function. The main advantage of

CC over CI results from these disconnected excitations, as they preserve the size consistency (and size extensivity) of CC methods. CI, on the other hand, is not size-consistent (extensive). With increasing basis set size, the virtual orbital space grows tremendously, and hence, CI and CC methods show significantly higher computational costs than HF. The incorporation of all possible excited determinants, which corresponds in a given basis set to the exact solution of the electronic SE, is called full configuration interaction (FCI).[145] However, FCI is only applicable for a few light atoms.[145,146] Instead CCSD(T), that is CC with single and double excitations and triple excitations taken from perturbation theory (*vide infra*), is often regarded as the "gold standard" in quantum chemistry.[144] The formal scaling of CCSD(T) $\left(O(M^7)\right)$ is further reduced by modern variants such as DLPNO-CCSD(T) whilst maintaining the overall accuracy.[147–150] Throughout this work, it is therefore often used as the theoretical reference level.

Electron correlation effects may also be added to HF by the concept of perturbation theory (PT).[151] Here, the full Hamiltonian $\hat{H}$ is expressed as

$$\hat{H} = \hat{H}_0 + \lambda \hat{H}' \,, \tag{2.28}$$

where $\hat{H}_0$ refers to an effective zeroth order Hamiltonian, $\hat{H}'$ is the perturbation operator, and λ is the perturbation parameter. In Møller–Plesset (MP) perturbation theory, $\hat{H}_0$ is chosen to be the sum over Fock operators.[152] This approach counts electron–electron repulsion twice and the zeroth-order energy equals the sum over orbital energies. The first-order energy correction fixes the double counting of the electron–electron repulsion and reproduces the HF energy. The inclusion of correlation energy starts at the second order in Møller–Plesset perturbation theory.[152] The corresponding MP2 correlation energy expression is given by Eq. 2.29[15]

$$E_{\mathrm{MP2}} = \sum_{i<j}^{\mathrm{occ.}} \sum_{a<b}^{\mathrm{virt.}} \frac{\left[\langle \phi_i \phi_j | \phi_a \phi_b \rangle - \langle \phi_i \phi_j | \phi_b \phi_a \rangle\right]^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} \,, \tag{2.29}$$

with i and j indicating occupied and a and b virtual orbitals. The MP2 approach recovers a significant portion of the correlation energy but also has well known weaknesses. For small HOMO-LUMO gaps, as they occur, *e.g.*, for metal complexes, the MP2 energy approaches $-\infty$. The advantage of MP2 is its non-iterative nature and its relatively small scaling with the AO basis set size of $O\left(M^5\right)$, which is why MP2 correction is frequently applied to compute the correlation energy.[153]

2.1.5. Kohn–Sham Density Functional Theory

Density functional theory draws the connection between the electron density of a system and its molecular energy and properties.[154] DFT originates from the Thomas–Fermi model[155,156] for the electronic structure of materials, but was initially implemented in a theoretical framework by Walter Kohn and Pierre Hohenberg formulating the two Hohenberg–Kohn theorems.[81,82] The first Hohenberg–Kohn theorem states that there is a one-to-one connection between the exact energy and an (unknown) functional $F[\rho(\mathbf{r})]$ of the electron density $\rho(\mathbf{r})$ (see Figure 2.2). This functional is universal and identical

for any system. Following, the second Hohenberg–Kohn theorem introduces the variational principle,

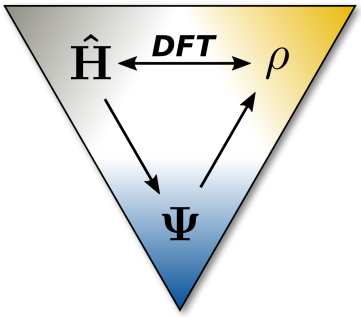


Figure 2.2.: DFT draws the connection between the electron density of the ground state and the energy of the system, as stated by the first Hohenberg–Kohn theorem.

$E[\tilde{\rho}(\mathbf{r})] \geq E_0$, stating that an energy for any trial density $\tilde{\rho}(\mathbf{r})$ is always higher than the energy calulated from the exact density.[157]

The Hohenberg–Kohn theorems form the basis for density functional theory but they lead to no further applicable approximations on how to construct the exact density functional. In analogy to the electronic Hamiltonian (Eq. 2.6), an electron density dependent energy can be formulated as[154,158]

$$E_{\mathrm{DFT}}[\rho] = T_e[\rho] + V_{ne}[\rho] + V_{ee}[\rho] \tag{2.30}$$

$$= T_e[\rho] + V_{ne}[\rho] + (J[\rho] + K[\rho]) \,, \tag{2.30a}$$

where $T_e[\rho]$ and $V_{ne}[\rho]$ are the kinetic energy and the nuclear–electron interaction energy, respectively, associated with a given electron density. The density dependent electron–electron interaction $V_{ee}[\rho]$ is further divided into a Coulomb part $J[\rho]$ and an exchange part $K[\rho]$ (Eq. 2.30a). This basic DFT approach is inherently orbital free and the density $\rho(\mathbf{r})$ depends only on three spatial variables, instead of $3N$ as in WFT. Nevertheless, the exact description of $T_e[\rho]$, in particular, remains problematic, and the Thomas–Fermi model was not suitable for chemical reactions, as it did not yield bound molecules.

The problem was solved by Kohn and Sham who introduced a fictitious reference system of non-interacting electrons, but with the same (exact) density as the real system.[82] In KS-DFT, the kinetic energy is calculated in analogy to HF theory from a Slater determinant approach ($T_{\mathrm{SD}}[\rho]$) and hence, orbitals are re-introduced.[154] The approximate trial density is constructed from a set of auxiliary orbitals ϕ_i (basis set) and given by the summed probability densities of the individual orbitals[159]

$$\rho(\mathbf{r}) = \sum_{\sigma=\alpha,\beta} \rho_\sigma(\mathbf{r}) = \sum_{\sigma=\alpha,\beta} \sum_{i}^{N_\sigma} |\phi_i(\mathbf{r})|^2 \,. \tag{2.31}$$

In Eq. 2.31 the density is divided into the different spin components, and N_σ is the total number of electrons with spin σ ($N_\sigma = N_\alpha + N_\beta$).

The difference in between the exact and approximated kinetic energy, as well as the correlation and

exchange energy, are combined in the exchange-correlation functional $E_{\mathrm{XC}}[\rho]$ (Eq. 2.32)

$$E_{\mathrm{XC}}[\rho] = (T[\rho] - T_{\mathrm{SD}}[\rho]) + (V_{ee}[\rho] - J[\rho]) \tag{2.32}$$
$$= E_{\mathrm{X}}[\rho] + E_{\mathrm{C}}[\rho] \tag{2.32a}$$
$$= \int \varepsilon_{\mathrm{XC}} \left[\hat{\nabla}^a \rho(\mathbf{r})\right] \rho(\mathbf{r}) d\mathbf{r} \,. \tag{2.32b}$$

E_{XC} is often further decomposed into an exchange E_{X} and correlation E_{C} part (*cf.* Eq. 2.32a) that depends on the energy density $\varepsilon_{\mathrm{XC}}$. This latter term is a functional of the different electron density derivatives $\hat{\nabla}^a \rho(\mathbf{r})$ (Eq. 2.32b). Thus, Eq. 2.30a can be reformulated in terms of Kohn–Sham DFT

$$E_{\mathrm{KS}}[\rho] = T_{\mathrm{SD}}[\rho] + V_{ne}[\rho] + J[\rho] + E_{\mathrm{XC}}[\rho] \tag{2.32}$$
$$= \left[\sum_{i=1}^{N} \left\langle \phi_i \left| -\frac{1}{2}\nabla^2 \right| \phi_i \right\rangle\right] + V_{ne}[\rho] + J[\rho] + (E_{\mathrm{X}}[\rho] + E_{\mathrm{C}}[\rho]) \,. \tag{2.33a}$$

Minimizing E_{KS} with respect to the orbitals results in the Kohn–Sham equations[15] (Eq. 2.34)

$$\hat{f}_i^{\mathrm{KS}}[\rho]\phi_i = \epsilon_i \phi_i \tag{2.34}$$
$$= \left[\hat{h}_i[\rho] + \sum_{j}^{N} \left(\hat{J}_{ij}[\rho] + v_{\mathrm{XC}}[\rho]\right)\right] \phi_i \,, \tag{2.34a}$$

where the Kohn–Sham operator $\hat{f}^{\mathrm{KS}}$ is given in Eq. 2.34a and the exchange-correlation potential $v_{\mathrm{XC}}[\rho]$ is given by Eq. 2.35

$$v_{\mathrm{XC}}[\rho] = \frac{\delta E_{\mathrm{XC}}[\rho]}{\delta \rho} \,. \tag{2.35}$$

In analogy to HF (Eq. 2.15), the Kohn–Sham equations (Eq. 2.34) are solved iteratively in an SCF procedure, employing the same methodologies such as the Roothaan–Hall formalism.

The exact density functional would provide the correct exchange and correlation energy. Thus, KS-DFT offers an appealing approach to include electron correlation effects within a mean-field theory and improves upon HF. Nevertheless, approximations to the density functional must be employed since the exact functional of an arbitrary density is unknown.[159] Note that in Eq. 2.34a $v_{\mathrm{XC}}[\rho]$ is approximated and hence, does not cancel the Coulomb interaction J_{ii} of an electron with itself. This leads to the SIE in local and semi-local density functionals, whereas HF is SIE free (*cf.* Eq. 2.17).[160,161] Over the last decades, theoreticians have developed an entire "zoo" of different density functional approximations[162] (DFA), ranging from purely empirical to physically motivated approaches. Perdew and Schmidt introduced the metaphor of "Jacob's ladder" to rank different DFAs according to the underlying approximations in $v_{\mathrm{XC}}[\rho]$.[154,159,163] A general overview is given in Figure 2.3.

The idea of Jacob's ladder is that DFAs at higher levels build on those below them, always adding a new level of complexity to get closer to the exact functional. The lowest runk, and thus the basis of all DFAs, is occupied by local (spin) density approximations (LSDA) derived from the uniform electron gas (UEG).[164,165] In the UEG, there are no variations in the density and thus, E_{XC} only depends on

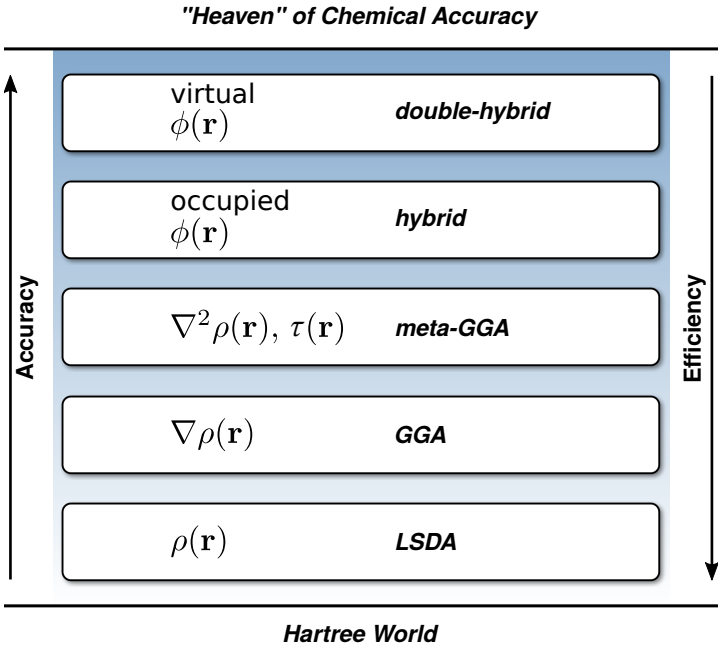


Figure 2.3.: Illustration of DFA categorization according to the "Jacob's ladder." ρ is the electron density, τ the kinetic energy density, ∇ the Nabla operator, and ϕ a molecular orbital.

$\rho(\mathbf{r})$.[164,165] LSDAs are improved by the general gradient approximation (GGA) type of functionals, which additionally take into account the density derivative $\nabla\rho(\mathbf{r})$.[166–168] Higher order derivatives $\nabla^2\rho(\mathbf{r})$ or the kinetic energy density[169] τ are included by meta-GGA functionals.[170,171] In hybrid DFAs, the (semi-)local exchange functional is modified by a fraction of non-local exact exchange from HF, the so-called Fock exchange.[172,173] Introducing Fock exchange reduces the SIE in hybrid functionals compared to semi-local DFAs. Double-hybrid (DH) DFAs represent the highest rung on Jacob's ladder and include virtual orbital information, *e.g.*, by means of (modified) Møller–Plesset[152] perturbation theory as in WFT.[174,175] For more information about the different rungs of DFAs, see Appendix A2.1. In the following DFT always refers to KS-DFT, if not stated differently.

2.1.6. Semi-Empirical Quantum Mechanical Methods

Tight-Binding Methods

Semiempirical QM methods are empirical extensions to *ab initio* methods.[89] Formally, they belong to the class of electronic structure methods but introduce empirical potentials and approximations to various terms to accelerate calculations. Usually, minimal basis sets are employed as well. SQM methods derived from KS-DFT are called tight-binding (TB) methods. The most prominent representatives thereof are the density functional based tight-binding (DFTB) methods, developed by Elstner and co-workers,[89–92,176–179] and the more recently introduced extended tight-binding (xTB) methods of the GFN*n*-xTB family of methods.[93–96] In the following, the connection of the xTB methods to DFT and DFTB is outlined. The equations are adapted from Refs. 93,94,96.

The energy expression for extended TB methods is derived from a general DFA that includes a non-local (NL) correlation contribution. The total Kohn-Sham energy expression is then given by[93] (Eq. 2.36)

$$E_{\text{tot}} = \sum_{i}^{M} f_i \int \phi_i^*(\mathbf{r}) \left[\hat{T}(\mathbf{r}) + V_n(\mathbf{r}) + \varepsilon_{\text{XC}}^{\text{LDA}}[\rho(\mathbf{r})] + \frac{1}{2} \int \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \Phi_{\text{C}}^{\text{NL}}(\mathbf{r}, \mathbf{r}') \right) \rho(\mathbf{r}') \mathrm{d}\mathbf{r}' \right] \phi_i(\mathbf{r}) \mathrm{d}\mathbf{r} + V_{NN} \,. \tag{2.36}$$

Here, ψ_i are molecular spatial orbitals, M their total number, and f_i their respective occupation. $\hat{T}(\mathbf{r})$ is the kinetic energy operator, $\hat{V}_n(\mathbf{r})$ is the Coulomb operator, and $\varepsilon_{\text{XC}}^{\text{LDA}}[\rho(\mathbf{r})]$ is the semilocal exchange-correlation (XC) energy per particle. The inner integral contains the interelectronic Coulomb and nonlocal correlation *via* the kernel $\Phi_{\text{C}}^{\text{NL}}(\mathbf{r}, \mathbf{r}')$. The latter term includes dispersion interactions (see Appendix A2.1) and establishes the connection between tight-binding and intermolecular force-fields (*vide infra*). The total energy can then be reformulated in terms of a reference density ρ_0 composed of a superposition of spherical, neutral atomic reference densities (SADs) $\rho_0 = \sum_A \rho_0^A$, and a density difference $\Delta\rho$ with $\rho = \rho_0 + \Delta\rho$ (Eq. 2.37)

$$E_{\text{tot}} = E_0^{\text{H}} + \Delta E^{\text{H}} + E_{\text{XC}}^{\text{LDA}}[\rho] + E_{\text{C}}^{\text{NL}}[\rho, \rho'] \,. \tag{2.37}$$

Here, the last two terms on the right hand side are non-separable exchange-correlation (local and non-local) energies. Eq. 2.37 is equivalent to Eq. 2.36, just reformulated in terms of the difference density $\Delta\rho$. The energy at the reference density E_0^{H} and the energy difference ΔE^{H} are given by

$$E_0^{\text{H}} = E_{nn} + \sum_{i}^{M} f_{0,i} \int \phi_i^*(\mathbf{r}) \left[\hat{T}(\mathbf{r}) + V_n(\mathbf{r}) + \frac{1}{2} \int \frac{1}{|\mathbf{r} - \mathbf{r}'|} \rho_0(\mathbf{r}') \mathrm{d}\mathbf{r}' \right] \phi_i(\mathbf{r}) \mathrm{d}\mathbf{r} \,, \tag{2.38}$$

$$\Delta E^{\text{H}} = + \sum_{i}^{M} \Delta f_i \int \phi_i^*(\mathbf{r}) \left[\hat{T}(\mathbf{r}) + V_0(\mathbf{r}) + \frac{1}{2} \int \frac{1}{|\mathbf{r} - \mathbf{r}'|} \Delta\rho(\mathbf{r}') \mathrm{d}\mathbf{r}' \right] \phi_i(\mathbf{r}) \mathrm{d}\mathbf{r} \,, \tag{2.39}$$

with the reference potential $\hat{V}_0(\mathbf{r})$ given as

$$V_0(\mathbf{r}) = \sum_{A}^{O} \left(\int \frac{1}{|\mathbf{r} - \mathbf{r}'|} \rho_0^A(\mathbf{r}') \mathrm{d}\mathbf{r}' - \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \right) . \tag{2.40}$$

In Eq. 2.40, M is the number of orbitals (Eqs. 2.38, 2.39) and O the number of nuclei. In DFTB methods, the total energy is expanded in a Taylor series around $\Delta\rho = 0$ (Eq. 2.41), where the density fluctuations are typically restricted to the valence orbitals

$$E[\rho] = E^{(0)}[\rho_0] + E^{(1)}[\rho_0, \delta\rho] + E^{(2)}\left[\rho_0, (\delta\rho)^2\right] + E^{(3)}\left[\rho_0, (\delta\rho)^3\right] + \cdots \,. \tag{2.41}$$

The most sophisticated variants truncate this expansion after the third order term,[92,178] just like the GFN1-xTB[94] and GFN2-xTB[93] methods. GFN0-xTB[95] corresponds to truncation after the first order term and GFN-FF truncates the expansion after the zeroth order term (*vide infra*). The zeroth order

energy term only depends on neutral atomic reference densities as all terms but E_0^{H} cancel. The energy is computed as the sum of non-interacting atomic total energies (E_A), as well as interatomic repulsion and London dispersion interaction terms (Eq. 2.42)

$$E^{(0)}\left[\rho_0\right] = \sum_A^O E_A\left[\rho_0^A\right] + \frac{1}{2}\sum_A^O\sum_B^O\left(E_{\mathrm{rep}}\left[\rho_0^A,\rho_0^B\right] + E_{\mathrm{disp}}\left[\rho_0^A,\rho_0^B\right]\right) \tag{2.42a}$$

$$= \sum_A^O E_A\left[\rho_0^A\right] + \underbrace{\left(E_{\mathrm{rep}}^{(0)} + E_{\mathrm{disp}}^{(0)}\right)}_{\text{Lennard-Jones/Buckingham-type potential}} . \tag{2.42b}$$

In Eq.2.42, E_{disp} arises from the long-range correlation effects and E_{rep} results from overlap of the atomic reference densities ρ_0^A. This establishes the connection between TB methods and intermolecular force-field potentials of Lennard-Jones[180] or Buckingham[181] type, which is important for the course of this thesis.

At first order, density fluctuations enable changes in the energy but not in the electrostatic potential and therefore, Coulomb interactions are yet missing. The energy term is given by Eq. 2.43

$$E^{(1)}\left[\rho_0,\delta\rho\right] = \Delta E^{\mathrm{H}}\left[\Delta f_i = \delta f_i, \Delta\rho = 0\right] + \frac{1}{\partial\rho}\left(\partial E_{\mathrm{XC}}^{\mathrm{LDA}}[\rho_0] + \partial E_{\mathrm{C}}^{\mathrm{NL}}[\rho_0,\rho_0']\right)\delta\rho \tag{2.43a}$$

$$\approx E_{\mathrm{EHT}}^{(1)} + E_{\mathrm{disp}}^{(1)} . \tag{2.43b}$$

First order terms include the description of covalent bonds in TB theories. The ΔE^{H} term is equivalent to extended Hückel theory[182,183] (EHT) and is given as the sum of the changes in the atomic valence energies. The major savings of TB methods in computation time result from the EHT treatment, for which only one-electron integrals are evaluated according to Eq. 2.44

$$E_{\mathrm{EHT}}^{(1)} = \sum_i^M n_i\left\langle\Psi_i\middle|\hat{H}_0\middle|\Psi_i\right\rangle = \sum_{\mu,\nu}^M\sum_i^M n_i C_{\mu i}C_{\nu i}\left\langle\phi_\mu\middle|\hat{H}_0\middle|\phi_\nu\right\rangle \equiv \sum_{\mu,\nu}^M P_{\mu,\nu}H_{\mu,\nu} . \tag{2.44}$$

Here, $\hat{H}_0$ is formally a one-electron operator and the corresponding off-diagonal matrix elements $H_{\mu,\nu}$ are approximated (see Appendix A2.2). The computationally demanding two-electron terms are thus neglected or implicitly compensated by the empirical approximations to higher order Coulomb and XC terms. TB methods are effectively 2–3 orders of magnitude faster than (semi-) local DFAs.[184]

At second order, interatomic electrostatic, one-center exchange-correlation, and NL correlation terms occur (Eq. 2.45)

$$E^{(2)}\left[\rho_0,\delta\rho\right] = \Delta E^{\mathrm{H}}\left[\Delta f_i = \delta f_i, \Delta\rho = \delta\rho\right] + \frac{1}{\partial\rho\,\partial\rho'}\left(\partial^2 E_{\mathrm{XC}}^{\mathrm{LDA}}[\rho_0] + \partial^2 E_{\mathrm{C}}^{\mathrm{NL}}[\rho_0,\rho_0']\right)\delta\rho\,\delta\rho' \tag{2.45a}$$

$$\approx E_{\mathrm{ES+XC}}^{(2)} + E_{\mathrm{disp}}^{(2)} \tag{2.45b}$$

$$\approx \frac{1}{2}\sum_A^O\sum_B^O q_A q_B \gamma_{AB}^h . \tag{2.45c}$$

In GFN*n*-xTB schemes, these are generally condensed in a Mataga–Nishimoto–Ohno–Klopman damped Coulomb interaction γ_{AB}^{h} between atomic or shell charge monopoles $q_{A,B}$.[185–187] With second or higher order terms, the tight-binding energies require a self-consistent solution, because the charges q entering the TB Hamiltonian (Eq. 2.45c) are obtained from a Mulliken population analysis.[128] Thus, they depend on the overlap integral $S_{\mu,\nu}$ and density matrix $P_{\mu,\nu}$. TB methods are therefore often referred to as self-consistent charge (SCC) procedures instead of SCF.

At third and higher orders, no contributions from ΔE^{H} occur, and only the local and non-local XC terms lead to energy according to (Eq. 2.46)

$$E^{(3)}\left[\rho_0, \delta\rho\right] = \frac{1}{\partial\rho\,\partial\rho'\,\partial\rho''}\left(\partial^3 E_{\mathrm{XC}}^{\mathrm{LDA}}[\rho_0] + \partial^3 E_{\mathrm{C}}^{\mathrm{NL}}[\rho_0, \rho_0']\right)\delta\rho\,\delta\rho'\,\delta\rho'' \tag{2.46a}$$

$$\approx E_{\mathrm{XC}}^{(3)} + E_{\mathrm{disp}}^{(3)}\,. \tag{2.46b}$$

These (semi-)local effects are included in DFTB3, as well as in GFN1- and GFN2-xTB Hamiltonians to stabilize relatively high charged atoms. Introducing empirical potentials inevitably leads to the use of element (pair-wise) parameters,[89,92] which are usually adapted from (or fitted to) pre-computed *ab initio* values. Parameterizations are often available only for a few, mostly main group element pairs, while a full periodic table parameterization is missing. In the context of general applicability, there is no real alternative to KS-DFT. Recently, this changed with the development of the extended tight-binding methods.

2.1.7. The GFN Family of Methods

The GFN*n*-xTB methods ($n = \{0, 1, 2\}$) represent a trilogy of general applicable extended tight-binding methods with special focus on geometries, harmonic vibrational frequencies, and noncovalent interaction energies.[93–96] The evolution of GFN*n*-xTB methods further inspired the development of a generic force-field, termed GFN-FF,[117] which is the first non-electronic GFN variant. An overview of all GFN methods and their mutual connection is given in Figure 2.4. What all GFN methods have in common, is the coverage of a full periodic table parameterization for elements up to radon ($Z \leq 86$). A mostly global and element-specific fitting strategy is applied, thereby avoiding element-pair-specific parameters. The molecule training set is versatile and covers thousands of structures reaching from small hydrides or oxides of the respective elements to large transition metal complexes. The description of van-der-Waals interactions represents a second parallel between the GFN methods, treating London dispersion and Pauli-exchange repulsion almost identically. The latest DFT-D dispersion correction is always employed in combination with a classical, exponentially damped $1/R$, pair-wise repulsion term. Further, all GFN methods are equipped with robust and efficient implicit solvation models.[96] Two variants are currently available. One model is based on the analytical linearized Poisson–Boltzmann (ALPB) model,[188] the other is a generalized Born and surface area model (GBSA).[189] All electronic GFN methods employ predefined minimal basis sets of the STO–mG type[190] for the valence electrons, where Slater-type orbitals (STO) are approximated by combinations of Gaussian-type atomic orbitals. The application, evaluation, and improvement of the existing GFN*n*-xTB methods is a major part of this

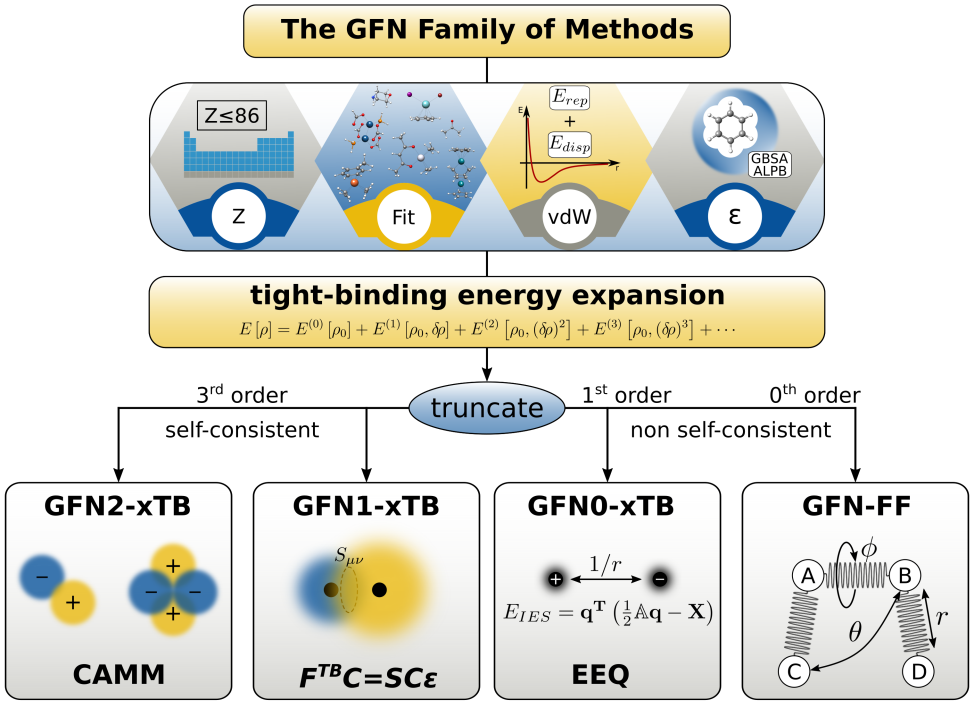


Figure 2.4.: The GFN family of methods. GFN1- and GFN2-xTB truncate the tight-binding energy expansion at third order. Truncation after the third order leads to the non-selfconsistent GFN0-xTB method. The first non-electronic GFN variant is GFN-FF, a formally zeroth-order-only TB scheme classified as a generic force-field.

thesis. For more details about the underlying theory, see the original publications Refs. 93–96 or Appendix A2.2. The GFN methods can be sorted according to the order in which the tight-binding energy expansion (Eq. 2.41) is truncated. Both, GFN1- and GFN2-xTB include up to third order terms. GFN1-xTB[94] represents the first generation of xTB methods. It features zeroth order Coulomb interactions described by the empirical repulsion term E_{rep}, the well-known D3 dispersion correction, and a halogen bond correction E_{XB}. Covalent bonds are described at first order by the Hückel term E_{EHT}. At second order, the isotropic monopole Coulomb interactions are described by Eq. 2.45c.

The essential novelty in the GFN2-xTB[93] method is the inclusion of anisotropic second order density fluctuation effects for the electrostatic E_{AES} and exchange-correlation E_{AXC} energies *via* short-range damped interactions of cumulative atomic multipole moments (CAMM).[191–193] As a result, the halogen bond correction is obsolete in GFN2-xTB. The dispersion energy is described by a modified, self-consistently solved D4 dispersion model, where the atomic partial charges are taken from the Mulliken population. GFN2-xTB is currently the most sophisticated TB method.

Truncation of the TB expansion (Eq. 2.41) after the first order leads to GFN0-xTB.[95] Covalent bonds are described by an extended Hückel term, which is improved by a classical short-range bond correction E_{SRB}. The classical repulsion term E_{rep} and non-selfconsistent D4 dispersion energy are added. The essential innovation in GFN0-xTB is the incorporation of semi-classical atomic charges, determined variationally from a classical electronegativity charge equilibrium (EEQ) model.[194,195] Hence, GFN0-

xTB is a non-selfconsistent method (no SCC required) and diagonalizes the Hamiltonian only once. Computationally it is the fastest, but also the most empirical GFN*n*-xTB method.

GFN-FF[117] is the first non-electronic GFN variant and can formally be regarded as a zeroth-order-only TB scheme. The missing first order terms, which are responsible for covalent bonding, are replaced by classical molecular mechanics for bonds, angles, and torsions. Therefore, GFN-FF is classified as a general force-field. The development, application, and evaluation of GFN-FF is a major part of this thesis. Detailed information about the underlying theory are given in Chapter 3 of Part II. A general overview of FF methods is given in Section 2.2.

2.2. Force-Field Methods

The QM calculation of the electronic energy for a given nuclear configuration is non-trivial and beyond a certain system size rather impossible. Force-field methods circumvent the description of the electronic structure by formulating the molecular energy as a parametric function of only the nuclear coordinates. [15] The main actors in force-field methods are thus the atoms and electrons are not considered as individual particles. Consequently, the bonding information must be provided by suitable functions, instead of emerging from the solution of the electronic Schrödinger equation.[15] The corresponding potential energy terms introduce parameters, which must be fitted to experimental or first principle computational data. Often, many atom-pair-specific parameters are needed, making it difficult to develop broadly applicable FF methods.

A general FF energy can be expressed as a sum of terms that describe the energy required for specific distortions of a molecule

$$E_{\mathrm{FF}} = \underbrace{E_{\mathrm{bond}} + E_{\mathrm{bend}} + E_{\mathrm{tors}}}_{\text{covalent}} + \underbrace{E_{\mathrm{vdW}} + E_{\mathrm{es}}}_{\text{noncovalent}} . \tag{2.47}$$

Here, the energy of bond stretching and compression is given by E_{bond}, E_{bend} is the energy required for angle bending, and the torsional energy for rotation around a bond is E_{tors}. E_{vdW} is the van-der-Waals energy, *i.e.*, the sum of repulsion and dispersion energy. E_{es} is the electrostatic energy. The first three terms describe the interaction of covalently bound atoms, while the latter two terms include the noncovalent atom–atom interactions. The energy terms are illustrated in Figure 2.5.

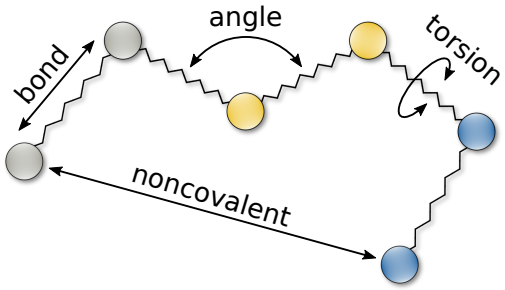


Figure 2.5.: Illustration of the fundamental force-field energy terms. Figure adapted from Ref. 15.

Changes in the bond distance R_{AB} between two atoms A and B are described by the energy function

E_{bond}. The expansion in a Taylor series around an equilibrium bond length R^0, which is then terminated at second order, leads to Eq. 2.48.[15]

$$E_{\text{bond}}\left(R_{AB} - R^0_{AB}\right) = E(0) + \frac{dE}{dR}\left(R_{AB} - R^0_{AB}\right) + \frac{1}{2}\frac{d^2E}{dR^2}\left(R_{AB} - R^0_{AB}\right)^2 . \tag{2.48}$$

Here, E_0 is the zero point for the energy scale and thus set to zero. Since the derivatives are evaluated at $R = R^0$, the second term in the expansion is also equal to zero. The first non-zero term arises at second order and Eq. 2.48 can be re-written as Eq. 2.49

$$E_{\text{bond}}\left(R_{AB} - R^0_{AB}\right) = \sum_{\text{bonds}} k_{AB}\left(R_{AB} - R^0_{AB}\right) = \sum_{\text{bonds}} k_{AB}\left(\Delta R_{AB}\right)^2 , \tag{2.49}$$

where the sum is over all bonds A–B and k_{AB} is the respective force constant (FC). The potential in Eq. 2.49 is quadratic in the displacement Δ from the equilibrium geometry and has the form of a harmonic oscillator. Simple harmonic bond potentials are widely used in common FFs[62,64,113] as they are sufficient for determining most equilibrium geometries. Nevertheless, the harmonic potential does not allow bond dissociation and yields unphysical values for strongly distorted structures. For a bond stretched to infinity, the energy should convergence towards the dissociation energy. This criterion is satisfied by the Morse potential (Eq. 2.50a)[196]

$$E_{\text{Morse}}(\Delta R) = D\left(1 - e^{-\alpha \Delta R}\right)^2 \tag{2.50a}$$

$$\alpha = \sqrt{\frac{k}{2D}}, \tag{2.50b}$$

where D is the dissociation energy, α controls the width of the potential, and k is the related FC at the minimum of the potential well. The Morse function reproduces the correct dissociation behavior over a wide range of distances and is frequently used in dissociative FFs.[197]

Bending an angle θ_{ABC} formed by three atoms A–B–C, where B is the central atom, requires the energy E_{bend}. In analogy to the bond energy, E_{bend} is usually expanded as a Taylor series around a reference bond angle θ^0_{ABC} and terminated at second order. In this harmonic approximation, the bending energy is given by Eq.2.51 as the sum over all angles

$$E_{\text{bend}}\left(\theta_{ACB} - \theta^0_{ABC}\right) = \sum_{\text{angles}} k_{ABC}\left(\theta_{ABC} - \theta^0_{ABC}\right)^2 , \tag{2.51}$$

where k_{ABC} is the bending force constant. If the central atom B is sp^2-hybridized and in a trigonal planar arrangement with atoms A, C, and D, then there is a significant energy penalty associated with the pyramidalization of B.[15] Therefore, an out-of-plane energy bend term (E_{oop}) is added, while the in-plane angles (θ_{ABC}, θ_{ABD}, and θ_{CBD}) are treated as in the general case above. E_{oop} may also be written as a quadratic function (Eq. 2.52)

$$E_{\text{oop}}(d) = \sum_{B} k_B d^2, \tag{2.52}$$

where d is the out-of-plane distortion and k_B the respective force constant.

E_{tors} is the energy required for the rotation around the B–C bond in a four-atom sequence A–B–C–D. The torsion angle, also known as a dihedral angle, is formed by three consecutive bonds in a molecule (A–B, B–C, C–D) and defined by the angle created between the two outer bonds (*cf.* Figure 2.5). The torsional energy function is periodic in the angle ω and can be expressed as a Fourier series given in Eq. 2.53

$$E_{\text{tors}}(\omega) = \sum_{\text{torsins}} \sum_{n=1} V_n \cos(n\omega)\,. \tag{2.53}$$

V_n is the force constant that determines the magnitude of the rotation barrier around the B–C bond. For $n = 1$, the rotation is periodic by 360°, for $n = 2$ by 180°, for $n = 3$ by 120°, *etc.* The sum is in Eq. 2.53 is over all respective torsion angles.

The noncovalent contributions to E_{FF} consist of the van-der-Waals energy and the electrostatic interactions.[62,198] E_{vdw} is a distance-dependent interaction between atoms or molecules, which is zero at larger interatomic distances. At short distances, E_{vdW} becomes very repulsive as a result of the Pauli exclusion principle that prevents the collapse of molecules. At intermediate distances, there is moderate but highly important interatomic attraction from induced multipole interactions, due to electron correlation effects. Thereof, the leading term is the induced dipole–dipole interaction, which varies as the inverse sixth power of the distance between atoms A and B ($1/R^6$). Other contributions from induced dipole–quadrupole ($1/R^8$) , quadrupole–quadrupole ($1/R^{10}$), *etc.*, interactions exist as well. The asymptotic behavior at long distances is dominated by the ($1/R^6$) dependence and the force associated with this interactions is well-know as the London dispersion force[199,200] (see Chapter 6 in Part III or Appendix A2.1). A general functional form that fulfills the above mentioned criteria for E_{vdW} is the Lennard–Jones (LJ) potential[180] in Eq. 2.54, which is often applied in FFs

$$E_{\text{LJ}}(R_{AB}) = \sum_{A=1}^{O-1} \sum_{B=A+1}^{O} \frac{C_{12}}{(R_{AB})^{12}} - \frac{C_6}{(R_{AB})^6} \tag{2.54a}$$

$$= \sum_{A=1}^{O-1} \sum_{B=A+1}^{O} \varepsilon \left[\left(\frac{R^0_{AB}}{R_{AB}} \right)^{12} - 2 \left(\frac{R^0_{AB}}{R_{AB}} \right)^{6} \right]. \tag{2.54b}$$

In Eq. 2.54a, the repulsive part is given by an ($1/R^{12}$) dependence and C_{12} and C_6 are the respective constants. O is the total number of atoms. A more widely used form of the LJ potential is given in Eq. 2.54b, where R^0_{AB} is the minimum energy distance and ε the depth of the minimum. Note that it is not possible to derive the functional form of the repulsive interaction and the exponent in the repulsive part is an empirical parameter. A generalized Lennard–Jones type potential is applied, *e.g.*, in the Merck Molecular Force-Field (MMFF).[201]

Atomic partial charges in molecules result from differences in the electronegativities of the respective atoms. The electrostatic interaction between point charges is given by the Coulomb potential

$$E_{\text{es}}(R_{AB}) = \sum_{A=1}^{O-1} \sum_{B=A+1}^{O} \frac{Q_A\, Q_B}{\varepsilon R_{AB}}\,, \tag{2.55}$$

where Q_A and Q_B are the atomic partial charges of atoms A and B, R_{AB} is their mutual distance, and ε the dielectric constant. The MM2 and MM3 force-fields use a bond dipole description instead of atomic partial charges, but only little difference in the performance was observed.[202,203] Note that due to the sum over atom pairs, the noncovalent contributions (E_{vdW}, E_{es}) scale quadratically with the system size, while the covalent terms scale linearly. Atomic partial charges are assigned by fitting to the molecular electrostatic potential (MEP) from *ab initio* electronic structure methods.[204] The MEP ϕ_{es} at the point $\mathbf{r}$ is defined by the nuclear charges Z and the electronic wave function Ψ (Eq. 2.56)[15]

$$\phi_{\text{es}}(\mathbf{r}) = \sum_{A}^{O} \frac{Z_A}{|\mathbf{R}_a - \mathbf{r}|} - \int \frac{\Psi^2(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} d\mathbf{r}' \,. \tag{2.56}$$

In the fitting procedure, an error function of the form shown in Eq. 2.57 is minimized

$$ErrF(\mathbf{Q}) = \sum_{r}^{N_{\text{points}}} \left(\phi_{\text{es}}(\mathbf{r}) - \sum_{A}^{O} \frac{Q_A(\mathbf{R}_A)}{|\mathbf{R}_A - \mathbf{r}|} \right)^2 , \tag{2.57}$$

constraining the sum of the partial charges Q_i to equal the total charge of the molecule. Usually thousands of points (N_{points}) are sampled near around the molecule. Restrained MEP fitting techniques are used, *e.g.*, in the AMBER force-field.[62]

2.3. Free Energy Contributions

At the very beginning of Chapter 2 in Eq. 2.1c, the Gibbs free energy for a reaction was decomposed in two parts, the molecular energy ΔE and the sum of all corrections to free energy $\Delta G_{\text{corr.}}(T)$. In Sections 2.1 and 2.2, QM, TB, and FF methods were introduce to calculate the molecular energy. In this section, the focus will lie on the correction term to free energies. Following an established thermostatistical protocol,[119] $\Delta G_{\text{corr.}}(T)$ is separated in two parts (Eq. 2.58)

$$\Delta G_{\text{corr.}}(T) = \Delta G_{\text{trv}}(T) + \Delta\delta G_{\text{solv}}(T) \,. \tag{2.58}$$

Here G_{trv} is the thermostatistical contribution from energy to free energy at absolute temperature T, accounting for translation, rotation, and vibration degrees of freedom (DOF), also including the zero-point vibrational and volume work terms. $\Delta\delta G_{\text{solv.}}(T)$ is the temperature dependent solvation free energy, which has to be considered when the reaction takes place in solution. First, an overview of statistical thermodynamics is given in Section 2.3.1, followed by an introduction to solvation effects in Section 2.3.2. Conformational degrees of freedom are considered in Section 2.3.3.

2.3.1. Statistical Thermodynamics

Statistical thermodynamics provides the connection between the partition function Q for a microscopic system and external macroscopic properties.[68] Fundamental thermodynamic functions, such as the

Gibbs free energy G, enthalpy H, and the entropy S can be derived directly from Q according to[15]

$$G = H - TS = kT\,V\left(\frac{\partial lnQ}{\partial V}\right)_T - kT\,lnQ \tag{2.59a}$$

$$H = kT^2\left(\frac{\partial lnQ}{\partial T}\right)_V + kT\,V\left(\frac{\partial lnQ}{\partial V}\right)_T \tag{2.59b}$$

$$S = kT\left(\frac{\partial lnQ}{\partial T}\right)_V + k\,lnQ\,. \tag{2.59c}$$

The symbols used in Eq. 2.59 and throughout this section are listed in Table 2.1. The calculation of Q

Table 2.1.: Symbols used in this section for the derivation of thermostatistical quantities.

Q	partition function
g_i	degeneracy of state i
ϵ_i	energy of state i
k (k_B)	Boltzmann's constant
T	the temperature
M	mass of the molecule
h	Planck's constant
V	volume of molecule confined in a cubic box (molar volume of an ideal gas)
R	the gas constant
N_A	Avogadro's number
I_A, I_B, I_C	moment of inertia around principle axes of rotation A, B, and C
ν_i	the vibrational frequency for mode i
σ	rotational symmetry number

requires the knowledge of all possible quantum states of a given system. For this, the electronic and the nuclear Schrödinger equation have to be solved, which is practically impossible for more than three atoms.[15] Instead, the energy levels for an isolated molecule in a single conformation are often calculated within the rigid-rotor harmonic-oscillator (RRHO) approximation.[68,119] Therein, the electronic, translation, vibration, and rotation DOFs are assumed to be separable and the total energy can be approximated as a sum of the respective terms. Applying the RRHO approximation, the partition function can be calculated exactly for an ideal gas (isolated molecule). Q_{tot} is then written as the product of the individual partition functions (Eq. 2.60)

$$Q_{\text{tot}} = Q_{\text{elec}}Q_{\text{trans}}Q_{\text{rot}}Q_{\text{vib}}\,, \tag{2.60}$$

where the individual terms are given in Eq. 2.61

$$Q_{\text{elec}} = \sum_i g_i e^{-\epsilon_i/kT} \tag{2.61a}$$

$$Q_{\text{trans}} = (2\pi MkT)^{\frac{3}{2}}\,h^{-3}V \tag{2.61b}$$

$$Q_{\text{rot}} = \frac{8\pi^2}{\sigma h^3}\,(2\pi kT)^{\frac{3}{2}}\,\sqrt{I_A\,I_B\,I_C} \tag{2.61c}$$

$$Q_{\text{vib}} = \prod_i^{\text{modes}} \frac{e^{-h\nu_i/2kT}}{1 - e^{-h\nu_i/kT}}\,. \tag{2.61d}$$

The enthalpy and entropy contributions depend on the logarithm of Q, and hence, the respective quantities can be formulated as the sums of their DOFs

$$H_{\text{tot}} = H_{\text{elec}} + H_{\text{trans}} + H_{\text{rot}} + H_{\text{vib}} \tag{2.62a}$$

$$S_{\text{tot}} = S_{\text{elec}} + S_{\text{trans}} + S_{\text{rot}} + S_{\text{vib}} \,. \tag{2.62b}$$

The electronic partition function involves a sum over electronic quantum states obtained by solving the electronic Schrödinger equation. Because the energy difference between the ground state and the first excited state is usually much larger than kT, only the ground state is further considered. With the partition functions at hand (*cf.* Eq. 2.61), the enthalpy and entropy terms can be calculated by performing the differentiation in Eq. 2.59. The results for one mole of non-linear molecules are given in Eqs. 2.63 and 2.64

$$H_{\text{trans}} = \frac{5}{2}RT \tag{2.63a}$$

$$H_{\text{rot}} = \frac{3}{2}RT \tag{2.63b}$$

$$H_{\text{vib}} = RT \sum_{i}^{\text{modes}} \left(\frac{h\nu_i}{kT}\right) \frac{e^{-h\nu_i/kT}}{1 - e^{-h\nu_i/kT}} \,, \tag{2.63c}$$

$$S_{\text{trans}} = \frac{5}{2}R + R\ln\left(\frac{V}{N_A}\left(\frac{2\pi MkT}{h^2}\right)^{\frac{3}{2}}\right) \tag{2.64a}$$

$$S_{\text{rot}} = R\left[\frac{3}{2} + \ln\left(\frac{\sqrt{\pi}}{\sigma}\left(\frac{8\pi^2 kT}{h^2}\right)^{\frac{3}{2}}\sqrt{I_A I_B I_C}\right)\right] \tag{2.64b}$$

$$S_{\text{vib}} = R\sum_{i}^{\text{modes}}\left[\left(\frac{h\nu_i}{kT}\frac{e^{-h\nu_i/kT}}{1 - e^{-h\nu_i/kT}}\right) - \ln\left(1 - e^{-h\nu_i/kT}\right)\right] \,. \tag{2.64c}$$

Note that the first part of the vibrational enthalpy (Eq. 2.63c), which is the sum of $1/2\,h\nu$ contributions, gives the zero-point energies. Finally, G_{trv} can be calculated according to Eq. 2.59, where the enthalpy and entropy terms are expressed as the sums of their translation, rotation, and vibration DOFs according to Eq. 2.65

$$G_{\text{trv}} = (H_{\text{trans}} + H_{\text{rot}} + H_{\text{vib}}) - T\,(S_{\text{trans}} + S_{\text{rot}} + S_{\text{vib}}) \,. \tag{2.65}$$

2.3.2. Solvation Effects

To model chemical systems in solution realistically, solvent effects must be taken into account. This can be done either by the explicit inclusion of solvent molecules and dynamical sampling,[205,206] or by a parameterized implicit solvent model.[207–210] Due to its much higher computational efficiency, an implicit treatment of solvent effects is desirable in combination with the GFN methods, for the application to large molecules. Two kinds of implicit solvation models are available in the *xtb* program, a recently implemented analytical linearized Poisson–Boltzmann (ALPB) model[211] and a well established

generalized Born (GB) model. During the course of this thesis, the GB model was applied and the theoretical details will be limited to this model. The equations are adapted from Ref. 96.

In the GB model, a molecule is considered as a continuous region with a dielectric constant ϵ_{in}, which is surrounded by an infinite solvent with a dielectric constant ϵ_{out}.[212] The electrostatic interaction in the presence of a polarized solvent within the GB model is given by Eq. 2.66[189]

$$\delta G_{\text{GB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{out}}}\right)\sum_{A=1}^{O}\sum_{B=1}^{O}\frac{q_A q_B}{\left(R_{AB}^2 + a_A a_B \exp\left[-\frac{R_{AB}^2}{4a_A a_B}\right]\right)^{\frac{1}{2}}}, \tag{2.66}$$

where a_A, a_B are the effective Born radii of the atoms A and B, and O is the total number of atoms. The denominator on the right-hand side of Eq. 2.66 corresponds to the canonical interaction kernel proposed by Still.[189] The Born radii are evaluated by an Onufriev–Bashford–Case (OBC) corrected pairwise approximation to the molecular volume according to

$$\frac{1}{a_A} = \frac{1}{a_{\text{scale}}}\left(\frac{1}{R_A^{\text{cov}} - R_{\text{offset}}} - \frac{1}{R_A^{\text{cov}}} \cdot \tanh\left[b\eta_A - c\eta_A^2 + d\eta_A^3\right]\right), \tag{2.67}$$

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.[213] η_A is the pairwise approximation to the volume integral given by

$$\eta_A = \frac{R_A^{\text{cov}} - R_{\text{offset}}}{2}\sum_{B}\Omega(R_{AB}, R_A^{\text{cov}}, s_B R_B^{\text{cov}}), \tag{2.68}$$

where Ω is the pairwise function used to approximate the volume integral. The covalent radius of atom B is scaled by the element-specific descreening value s_B, which compensates for the inherent volume overestimation. In addition to the polar term, a non-polar surface area contribution is added, which depends on the solvent accessible surface area (SASA)

$$\delta G_{\text{SASA}} = \sum_{A=1}^{N}\gamma_A\sigma_A\,. \tag{2.69}$$

Here, γ_A is the surface tension, and σ_A is the SASA of atom A. The SASA is calucalted by integration on an angular Lebedev grid.[214] Combining the GB and the SASA term leads to a generalized Born surface area (GBSA) model,[189] for which the total solvation free energy is given by Eq. 2.70

$$\delta G_{\text{solv}} = \delta G_{\text{GB}} + \delta G_{\text{SASA}} + \delta G_{\text{shift}}\,. \tag{2.70}$$

δG_{shift} is included as an additional shift depending on the chosen reference state of the solution. The GBSA solvation free energy includes four global parameters that are fitted to reproduce COSMO-RS16 solvation free energies.[215–218] For further details about the solvation models in the *xtb* program, see Refs. 96,188. Leaving out the explicit atomistic description of solvent comes at the cost of an insufficient description of local solute–solvent interactions, *i.e.*, hydrogen bonds, ion distributions, and salt

bridges.[219] Therefore, in Chapter 9 of Part IV, an explicit solvation model, termed quantum cluster growth (QCG), is introduced to investigate the effect of explicit solvent molecules in the context of IR-spectroscopy.

2.3.3. Conformations

For rigid molecules, the partitioning in Eq. 2.1c and Eq.2.58 is sufficiently accurate. Nevertheless, for flexible systems several problems arise[15] and the concept of a single molecular structure is replaced by that of an ensemble of structures, which are in equilibrium with each other.[59] Under the assumption that all DOFs (electronic, translation, rotation, vibration, conformation, solvent) are separable, the free energy of the structure ensemble is obtained as (Eq. 2.71)

$$G_{\text{ensemble}} = \bar{G} + G_{\text{conf}} \,. \tag{2.71}$$

Here, G_{conf} is the conformational free energy part consisting of C distinguishable conformers in the ensemble. The average free energy $\bar{G}$ is then given by Eq. 2.72

$$\bar{G} = \sum_{i}^{C} p_i G_i \,, \tag{2.72}$$

where the Boltzmann population p_i is calculated according to Eq. 2.73

$$p_i = \frac{e^{-\frac{G_i}{kT}}}{\sum_i^C e^{\frac{G_i}{kT}}} \,. \tag{2.73}$$

k is the Boltzmann constant and G_i is the molecular free energy of the ensemble member i obtained as (*cf.* Eqs. 2.1c, 2.58)

$$G = E_{\text{gas}} + G_{\text{trv}}(T) + G_{\text{solv}}(T) \,. \tag{2.74}$$

Taking into account an entire structure ensemble of a flexible molecule, Eq. 2.74 is rewritten as

$$\Delta G = \Delta \bar{E} + \Delta \bar{G}_{\text{mRRHO}}(T) + \Delta\delta \bar{G}_{\text{solv.}}(T) + \Delta G_{\text{conf.}}(T) \,, \tag{2.75}$$

where the overlined quantities are computed in analogy to Eq. 2.72. Prerequisite for the accurate calculation of G_{conf} is the knowledge of the complete structure ensemble, wherefore the exploration of the PES and the detection of global minima is necessary. Therefore, the conformer-rotamer ensemble sampling tool, abbreviated CREST, is applied several times throughout this thesis. A short overview is given in Appendix A2.3. For a detailed description, see the original publication Ref. 60. A description on how to compute G_{conf} using the *crest* program is given in Ref. 61.

Part II.

Development of a Generic Force-Field for the Full Periodic Table

Experimental chemistry usually deals with billions of atoms simultaneously. Already an invisible drop, like a micromole of water (~10^{-7} g) contains 10^{17} atoms. As a matter of fact, the description of such large systems by high accuracy DFT will probably never be possible and instead, classical force-fields must be employed. Therefore, the development of new FFs with improved accuracy is an important part of research in theoretical chemistry.[2]

Currently, there is an imbalance between highly specialized FFs, which are tailor-made for a certain type of molecules, and general FFs, whose parameterization covers a larger amount of elements (see Figure 2.6). Prominent and widely applied specialized FFs, such as AMBER,[62] CHARMM,[64] or OPLS,[113] can yield high accuracy but are mostly limited to the systems and properties they were designed for. The universal force-field[114] (UFF) was introduced in 1992 and remains one of the few truly general FFs to date. The challenge is to design an efficient FF that pairs computational speed, accuracy, and general applicability. In addition, an efficient FF must be backward compatible and describe not only large but also small systems with comparable accuracy. This is especially important in the context of chemical space exploration,[60,220] where not the molecular size is the time determining step, but rather the sheer number of calculations that need to be performed due to the enormous combinatorial complexity of the problem.[2] A force-field that combines all requirements of accuracy, speed, and universal applicability is yet missing.

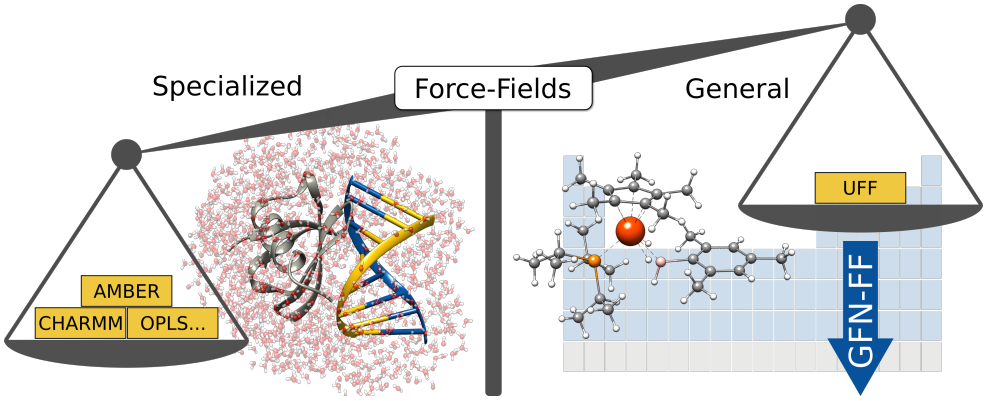


Figure 2.6.: Imbalance between specialized and general force-fields. With GFN-FF, an efficient FF is introduced that combines accuracy, computational speed, and full periodic table parameterization.

Therefore, Part II (Chapter 3) of this thesis is devoted to the development and initial testing of a partially polarizable, generic FF for the accurate description of geometries, harmonic frequencies, and noncovalent interaction energies.[117] This method, termed GFN-FF, is parameterized for all elements up to radon ($Z \leq 86$) and employs only element-specific and a few global fitting parameters, which reduces the total amount of empirical data compared to other FFs drastically. GFN-FF is designed to describe small and medium sized transition metal complexes equally well as large metal-organic frameworks or biomacromolecules.

Section 3.1 outlines existing simplified methods and their underlying approximations briefly. The theoretical background of GFN-FF is given in Section 3.2. Here, the connection between GFN-FF and

the closely related GFN*n*-xTB methods is drawn and common "ingredients" are pointed out. Also, the automatized topology assignment is outlined, which is a key aspect in the construction of a robust force-field. Next, the total energy of GFN-FF is presented and broken down into its components, whereof each term is discussed individually.

With this new FF method at hand, the computational modeling of a variety of chemical systems involving transition metals and biochemical compounds becomes possible. Section 3.3 deals with the geometry optimization of large metal-organic polyhedra and cutouts from periodic frameworks in comparison to experimental crystal structure cutouts. The quality of GFN-FF for transition metal complex structures is evaluated on the TMG145 benchmark set.[221] Furthermore, the structures of (metallo-)proteins are investigated. Among others, the structure of hemoglobin (~10.000 atoms) is fully optimized and an MD simulation is conducted for mutants of myoglobin. GFN-FF is compared to other established force-fields, such as OPLS2005[65] and AMBER*,[222,223] on a set of 70 organic peptide and protein structures.[224] Finally, the accuracy of GFN-FF for the calculation of various interaction energies is tested on standard benchmark sets, such as the GMTKN55[162] and S30L,[120] in comparison to high-level QM reference data.

Overall, it is found that GFN-FF optimized structures are in excellent agreement with the experiment, and in terms of interaction energies, almost the accuracy of much more sophisticated (S)QM methods is reached. Thereby, GFN-FF is two orders of magnitude faster than established TB methods, and six orders of magnitude faster than DFT methods, enabling the use on standard desktop computers.

3. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems

Sebastian Spicher[a] and Stefan Grimme[a]

Received: 23 March 2020, Published online: 18 May 2020

Selected as Very Important Paper

Reprinted (adapted) with permission[†] from
Spicher, S.; Grimme, S. *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.

DOI 10.1002/anie.202004239

Own manuscript contribution

- Major contributions to the development of GFN-FF
- Implementation of GFN-FF into the *xtb* code
- Performing all the calculations
- Interpretation of the results
- Writing the manuscript

[a]Mulliken Center for Theoretical Chemistry, Institute of Physical and Theoretical Chemistry, University of Bonn, Beringstraße 4, 53115 Bonn, Germany
[†]Permission requests to reuse material from this chapter should be directed to WILEY-VCH.

Abstract Modern chemistry seems to be unlimited in molecular size and elemental composition. Metal-organic frameworks or biological macromolecules involve complex architectures and a large variety of elements. Yet a general and broadly applicable theoretical method to describe the structures and interactions of molecules beyond the 1000 atom size regime semi-quantitatively is not self-evident. For this purpose, a generic force-field named GFN-FF is presented, which is completely newly developed to enable fast structure optimizations and molecular dynamic simulations for basically any chemical structure consisting of elements up to radon. The freely available computer program requires only starting coordinates and elemental composition as input from which, fully automatically, all potential energy terms are constructed. GFN-FF outperforms other force-fields in terms of generality and accuracy, approaching in many cases the performance of much more elaborate quantum mechanical methods.

3.1. Introduction

Concepts for designing molecules with desired (bio)chemical activities or physical properties have become state-of-the-art in experimental chemistry.[225,226] Molecular size and complexity has no boundaries and the elemental composition is versatile[227]. Within the last decades, the field of theoretical chemistry has evolved into an indispensable part of chemistry and is proven to be an important companion of the experiment.[1] Computational chemistry is able to explore the chemical space and provide experimentalists with useful information in order to circumvent resource demanding trial and error procedures.[228,229] In a cleaner and greener future for chemistry, theory is an essential tool supporting the experiment and increasing economic and environmental sustainability.[230] The constantly growing diversity of chemical compound space requires the development of new methods that can be applied in the analysis and prediction of complex molecular systems. Yet, a universal, fast, and easy to use method is missing that is capable of providing qualitatively correct molecular models beyond the size of a thousand atoms with arbitrary elemental composition.[2]

Even though today's ensemble of theoretical methods is quite versatile, it is limited in application. On the basis of wave function theory (WFT), methods have been developed that can provide highly accurate total energies and potential energy surfaces (PES) for small to medium sized molecules in the gas phase.[147] Kohn-Sham density functional theory (KS-DFT or simply DFT) draws the connection between the energy of a system and its electron density. With the introduction of reasonable approximations, DFT methods can provide accurate PES for systems with up to a few hundred atoms routinely.[231] Latest developments in the field of semiempirical quantum mechanical (SQM)[98,232] methods have further extended the treatable molecular size with special attention regarding the computation of Geometries, Frequencies, and Noncovalent interactions (GFN).[93,94] Within the extended tight-binding (xTB) theoretical framework, equilibrium structure optimizations and molecular dynamics (MD) simulations are feasible for large molecular systems, aiming at a comparable accuracy as DFT.[233] Still, the routine handling of several thousands of atoms is beyond the scope for the aforementioned methods and it is, therefore, necessary to apply more drastic but still physically reasonable approximations to reduce computational demands.[3]

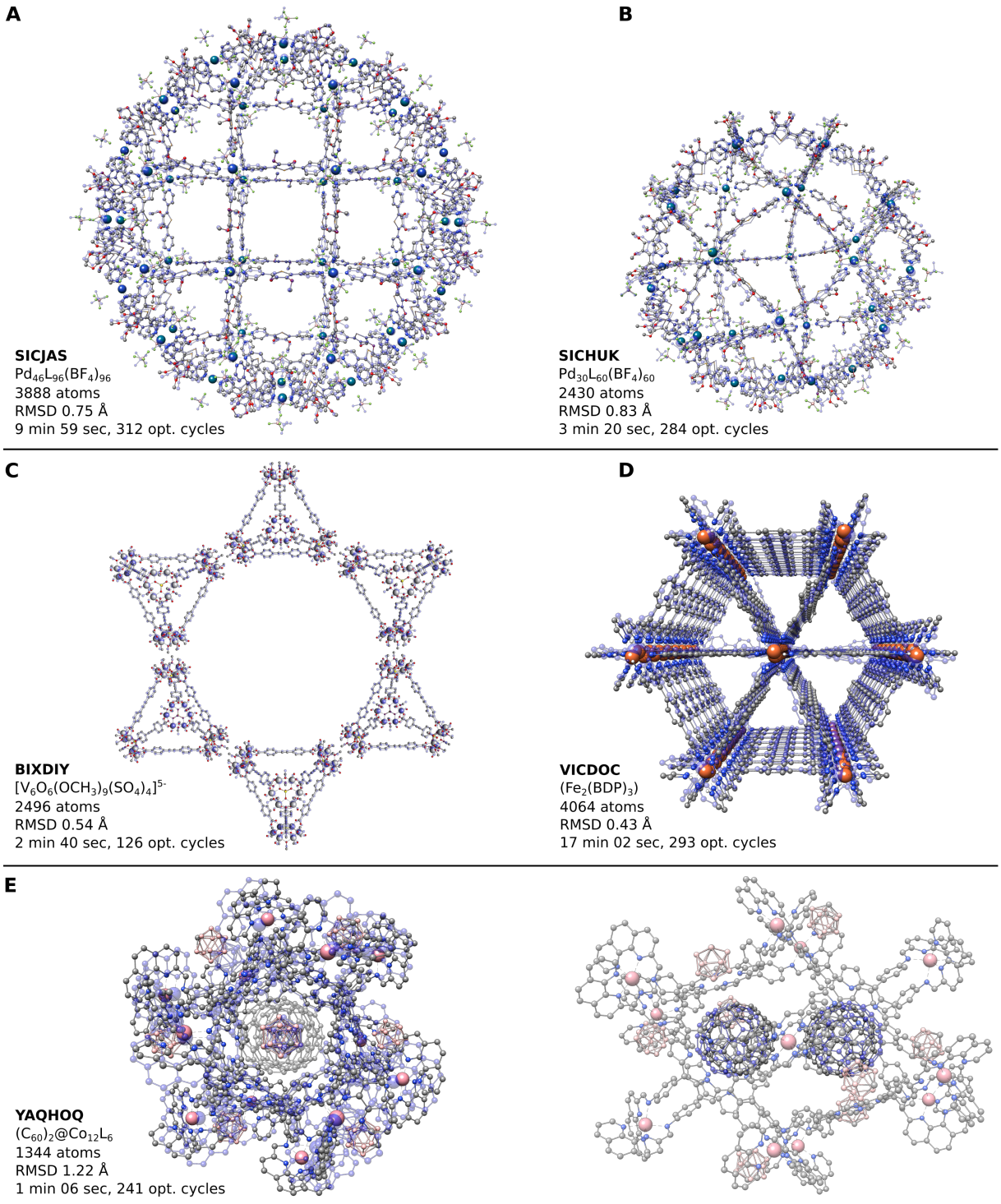


Figure 3.1.: Modeling of metal-organic porous materials with GFN-FF. (A-E) RMSD minimized structure overlay between the optimized GFN-FF geometries (transparent blue) and crystal structures of five systems. The CSD identifier are given as well as heavy atom RMSD values, total computation wall-times, and the required number of geometry optimization cycles.

Neglecting the electronic structure of a molecule and replacing it with classical interatomic interaction potentials is the main approximation in classical, atomistic force-fields (FFs). Their great benefit is to leave out the costly and difficult description of the electronic structure and substituting it by chemical-knowledge-motivated classical energy expressions. Force-fields specialized for the accurate description of a certain class of chemical systems exist for various application fields. Organics are well described by GAFF and MM3,[203] while CHARMM,[234,235] Amber,[62] and OPLS[113] focus on the description of proteins. In material science, DREIDING[115] and MOF-FF[236] are widely used. Limitations of those special purpose force-fields are manifold as they are not suited for interdisciplinary use, given the fact that parameters only exist for a limited amount of elements and structural motifs. Until now, only a single general applicable FF covering a full periodic table parameterization exists. This universal force-field (UFF)[114] was first introduced in 1992 and ever since advancements on this subject could not prevail, requiring most of the time individual, laborious parameterizations. Within this work, the idea of a general, easy to use force-field is revived within the GFN framework. The presented method, named GFN-FF, represents a generic, fully automated potential for the accurate description of an unlimited variety of molecular systems. GFN-FF is designed to combine high force-field speed with the accuracy of QM methods at unsurpassed robustness. For a manifold of systems, GFN-FF is currently the only applicable atomistic method to provide reasonable theoretical molecular structures. As examples for the diverse possible applications, a selected set of five porous metal-organic materials is shown in Figure 3.1. Their discussion will follow in the results Section 3.3 below.

3.2. Methodology

3.2.1. The GFN Force-Field

The idea of a general GFN-type FF is inspired by the latest developments in the field of SQM methods, namely the evolution of GFN1-, GFN2, and especially GNF0-xTB[95] methods, where the latest key ingredient was the introduction of a classical electronegativity equilibrium (EEQ) atomic charge model[194,195] for the description of pair–wise interatomic electrostatic interactions. This allowed truncating the fundamental expansion of the DFT energy $E[\rho]$ in terms of electron density fluctuations $\delta\rho$ after the first-order term, leading to a non-self-consistent method that employs classical atomic charges. GFN-FF introduces approximations to the remaining quantum mechanical terms in GFN0-xTB by replacing most of the extended Hückel type theory (EHT) for covalent bonding by classical bond, angle, and torsion terms. To highlight the ancestry from the xTB methods, the similarities and differences between FF and QM methods are illustrated in Figure 3.2.

All GFN methods cover a full periodic table parameterization for elements up to radon ($Z \leq 86$). This broad coverage is not self-evident even for SQM methods. To yield accurate results, the empirical FF parameters are fitted to reproduce DFT (B97-3c[237]) equilibrium geometries and frequencies as well as theoretical reference noncovalent interactions energies. A mostly global and element specific fitting strategy is applied thereby mostly avoiding element pair specific parameters. The molecule training set is versatile and covers currently about 8.000 structures reaching from small hydrides or oxides of

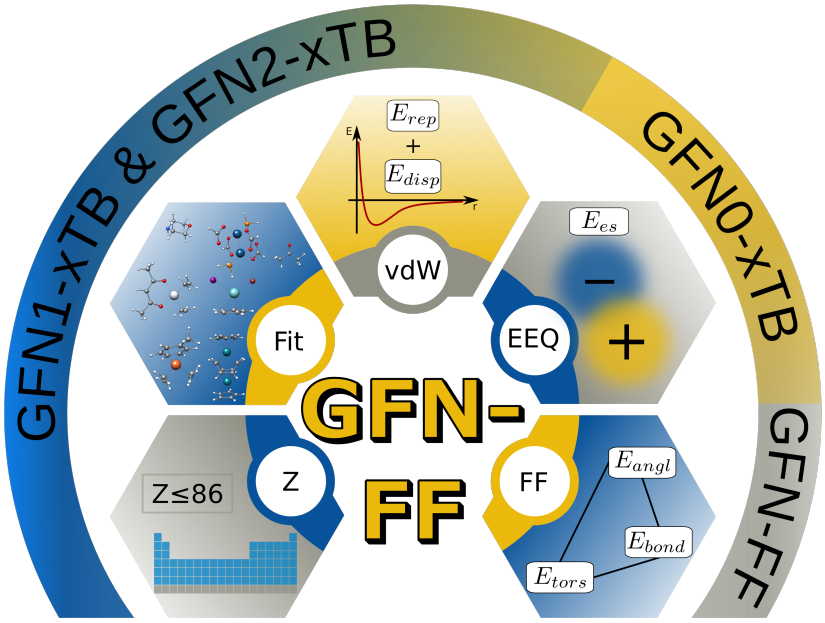


Figure 3.2.: The GFN-family of methods. The graphic shows the ingredients to the GFN-FF potential energy and the connection to the family members.

the respective elements to large transition-metal complexes. This approach is a unique feature of all GFN methods and differs strongly from the parameterization strategies of other force-fields.[205,235,238,239] The potential energy terms in GFN-FF are physically based and more sophisticated than simple, often used harmonic functions. Due to this well-defined basis, parameters arise naturally from the potential energy terms and their number is rather small. With only 18 specific parameters per element, GFN-FF is constructed upon a framework flexible enough to describe a vast majority of chemical systems. The quality and complexity of the potential functions determine the accuracy of GFN-FF rather than the sheer amount of parameters. This contradicts a current trend in theoretical chemistry to solve complicated many-body problems with a huge number of parameters, as it is done in, *e.g.*, machine learning and neural networks.[240,241] The description of van-der-Waals interactions represents another parallel between the GFN methods, treating London-dispersion and Pauli-exchange repulsion almost identically. To accurately treat the important π-conjugated systems such as aromatic hydrocarbons or graphenic materials, GFN-FF retains an iterative Hückel QM scheme for a selected set of atoms. From the resulting bond orders, the force constants and other energy relevant terms are derived, leading to high accuracy for conjugated molecules. The assignment of parameters and setup of force constants is key to the performance of any FF. A simplified flow-chart of the automatic setup is given in Figure 3.3. The automation of the force-field setup is a unique feature of GFN-FF. As input, only Cartesian coordinates and the elemental composition are required, from which the topological covalent bonding information, as well as atomic charges and bond orders, are generated fully automatically. With this information at hand, all potential energy terms are constructed.

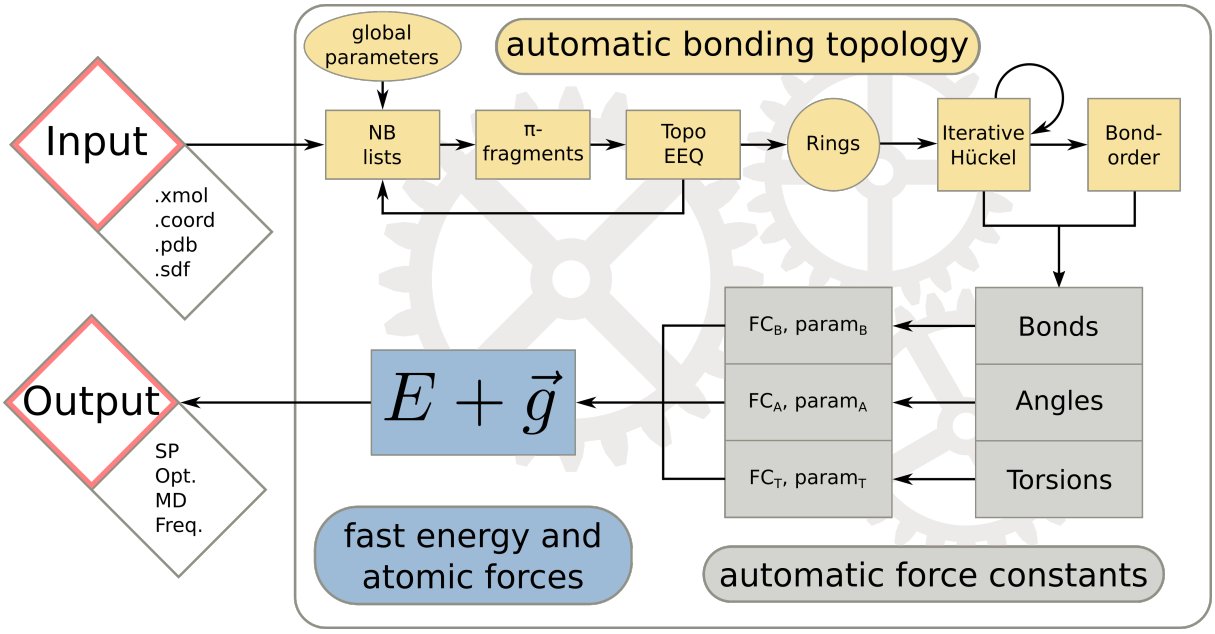


Figure 3.3.: GFN-FF internal flow chart. The implementation in the *xtb* program provides fully automated force-field setup, which is the generation of the topology and derivation of the force constants and other energy term related parameters.

3.2.2. The GFN-FF Total Energy

The total GFN-FF energy expression is given by

$$E_{\text{GFN-FF}} = E_{\text{cov}} + E_{\text{NCI}}\,, \tag{3.1}$$

where E_{cov} refers to the bonded FF energy and E_{NCI} describes the intra- and intermolecular noncovalent interactions (NCI). In the covalent part interactions are described by asymptotically correct bonding, angular and torsional terms. For the bonding term a new Gaussian-type potential is proposed that allows bond cleavage thus turning GFN-FF into a dissociative force-field. Repulsive terms are added for bonded and non-bonded interactions separately. Additionally, a new three-body bonding correction that extends beyond the sum of pair–wise interactions is included yielding

$$E_{\text{cov}} = E_{\text{bond}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{rep}}^{\text{bond}} + E_{\text{abc}}^{\text{bond}}\,. \tag{3.2}$$

In the noncovalent part, electrostatic interactions are described by the EEQ model. It is employed to calculate the entire electrostatic energy and isotropic atomic partial charges, which goes beyond the fixed charge model used in many other FFs. Overall GFN-FF uses two sets of EEQ charges. One set depends on the actual molecular geometry, whereas another set of charges is exclusively bond topology based, introducing further polarizability to the FF and leading to large simplifications for the gradient computations. Dispersion interactions are taken into account by a simplified version of the established D4 scheme,[242] which is the most accurate dispersion correction available and superior to the corre-

sponding description in standard FFs. Without detailed QM information, the accurate description of important noncovalent hydrogen (and halogen) bonds (HB/XB) is challenging. Therefore, newly developed charge-dependent HB/XB corrections are applied. These unique potentials include information about the location of electron lone-pairs, via an exclusion principle over neighboring atoms. The noncovalent energy expression is given by

$$E_{\mathrm{NCI}} = E_{\mathrm{IES}} + E_{\mathrm{disp}} + E_{\mathrm{HB}} + E_{\mathrm{XB}} + E_{\mathrm{rep}}^{\mathrm{NCI}} \,. \tag{3.3}$$

Despite its complexity, GFN-FF reaches quadratic scaling $O(N^2)$ in terms of energy and gradient calculation time with respect to system size at a moderate prefactor and is thus, almost on par with established force-fields in terms of computational speed.

Covalent topology

In GFN-FF the covalent topology is implemented in neighbor lists. In a first step, neighbor lists are generated according to highly selective inter-atomic distance criteria, where the actual distance between two atoms R_{AB}, is compared to a pre-computed reference value R_{AB}^0, describing a common bonding distance between those two atoms in their actual environment. It is calculated according to

$$R_{AB}^0 = \left(R_A^0 + R_B^0 + R_{sft}\right)\,\left(1 - c_1|\Delta EN| - c_2|\Delta EN|^2\right)\,, \tag{3.4}$$

where R_A^0 and R_B^0 are modified CN dependent D3 damping radii and c_1 and c_2 are period-specific parameters. The parameters are fitted to reproduce PBEh-3c[231] (a few) and B97-3c (mainly) equilibrium bond lengths, using about 15–20 reference molecules per element. For the atomic electronegativity (EN), Pauling values are used. R_{sft} is an empirically determined element specific shift to the interatomic distance. For $R_{AB} < f_R(q)R_{AB}^0$ a covalent bond is assigned between atoms A and B, where $f_R(q)$ is a charge-dependent scaling function that is unknown at this point and therefore set to unity. Depending on the number of neighbors, a hybridization state is assigned to each atom. With the information of neighboring atoms and their formal hybridization state at hand, π-conjugated fragments can be assigned with a corresponding number of π-electrons to separated parts of the entire system. Based on this initial topology, the first set of topology based EEQ charges $\mathbf{q_t}$ is derived. Having atomic charges at hand, the procedure is iterated once (see Figure 3.3), applying $\mathbf{q_t}$ to generate the function $f_R(q)$. This results in an improved description of the topology that is propagated in a better π-fragment assignment and an improved set of topological charges within only one iteration step. Applying more than one iteration did not lead to further improvements in the results. Subsequently, ring systems are identified automatically and this information is included up to a ring size of six atoms.

The proper description of π-conjugated systems is one of the key ingredients for an accurate force-field. In GFN-FF, Hückel theory is applied to describe conjugation for the elements B, C, N, O, F, P, and S. This is done in an iterative fashion, where the off-diagonal elements of the Hückel matrix depend on the density matrix P_{AB} in order to avoid over-delocalization. For more details, the reader is referred to the *xtb* source code.[243] The former π-assignment is used and for each fragment a Hückel matrix is set

up. Matrix diagonalization is performed in parallel, returning a set of eigenvalues ϵ_i and eigenvectors $\mathbf{c_i}$, which are then used to construct the ground state density matrix according to

$$P_{AB} = \sum_i n_i c_{iA} c_{iB} \,. \tag{3.5}$$

The Fermi smearing technique[244] is used to allow fractional orbital occupations n_i throughout this routine. This ensures that electronically degenerate and (poly)radical situations are treated automatically. Hückel theory introduces quantum mechanics to GFN-FF and provides important bonding information as used previously in MMFF94.[201] This comes at the cost of cubic scaling with regard to the number of π-electrons in the Hückel subsystems and thus, the formal scaling of the GFN-FF setup is for large π-conjugated molecules $O(N^3)$. However, due to the fact that normally only small π-fragments are taken into account and the entire setup procedure is called just once, the overall computational costs remain low. Nevertheless, the program has been successfully tested for molecule adsorption problems on huge graphene flakes with about 5000 atoms. With a complete topology assignment and knowledge about atomic charges, ring- and π-systems, in combination with the set of element specific fitted parameters, all force constants (FC), and relevant equilibrium values for the bond, bend, and torsion potentials can be derived. Along with the other energy terms as discussed below, the total energy E and its analytical gradient $\mathbf{g}$ are calculated. In GFN-FF the entire setup routine and calculation of energy and gradient, as formulated above (see Figure 3.3), is executed fully automatic.

Bond term

For a covalent bond between two atoms A and B, the energy E_{bond} is calculated by a newly developed Gaussian type function

$$E_{\text{bond}} = \sum_{\text{bonds}} -k_{str} \cdot exp\left[-\eta_{str}\left(1 + k_{EN}|\Delta EN(AB)|^2\right) \cdot \left(R_{AB} - R^0_{AB}\right)^2\right] , \tag{3.6}$$

which allows smooth dissociation of the AB bond into the separated atoms. This is a reversible process allowing broken bonds to reform. However, the formation of new bonds is not possible. In Eq. 3.6, R_{AB} is the actual interatomic distance, R^0_{AB} is the equilibrium value (Eq. 3.4) and η_{str} and k_{EN} are global fitting parameters. ΔEN is the difference between the EN of the elements $\Delta EN = EN_A - EN_B$. k_{str} is the bonding force constant, which is determined as a function of the coordination number (f_{CN}), atomic charges (f_{qq}), π-bond character (f_π), element type (main group metal or heavy element f_{hvy}), and whether the atoms are part of a ring system f_{rng})

$$k_{str}(AB) = f_{CN} \cdot f_{qq} \cdot f_\pi \cdot f_{hvy} \cdot f_{rng} \cdot k_b(A) k_b(B) \,. \tag{3.7}$$

k_b represent element specific bond parameters determined from the fit to reference data. In general, an element specific parameter that represent a function of global parameter and atomic properties will be termed with an f. In terms of FCs, element specific parameter that were determined by a fit to reference data will be termed with a k. If a covalent bond is of OH or NH type and the O/N atoms are involved

in hydrogen bonding, the exponent in Eq. 3.6 is multiplied by the function f_{CN_H}

$$f_{CN_H} = (1.0 - 0.1\, mCN_H)\ , \tag{3.8}$$

where mCN_H is the modified D3 coordination number of the hydrogen atom that is part of the bond and also the central atom within a hydrogen bond. The mCN is discussed in the electrostatic interaction part. For bonds involved in hydrogen bonding motifs, this modification of Eq. 3.6 models the population of the anti-bonding σ^* orbital by the hydrogen bond acceptor atom and the resulting weakening/elongation of the respective bond. As a result, GFN-FF reproduces the splitting of, *e.g.*, the OH stretching vibrational band in the spectrum of the water dimer.

Bending term

For the bending term involving three atoms A-B-C, where A is the central atom, the following energy expression is employed for angles θ_{ABC}

$$E_{\text{bend}} = \sum_{\text{angles}} f_{dmp} k_{bnd} \begin{cases} \left(\theta_{ABC} - \theta^0_{ABC}\right)^2 & \theta^0_{ABC} \approx \pi \\ \left(cos(\theta_{ABC}) - cos(\theta^0_{ABC})\right)^2 & \text{else.} \end{cases} \tag{3.9}$$

Here, a case distinction is made between equilibrium angles close to linearity ($\theta^0_{ABC} \approx \pi$) and all other cases, which apply a double-minimum function that allows inversion. The FCs k_a represent element specific parameters. Here, k_{bnd} is the bending FC, which is determined as a function of the atomic charges (f_{qq}) also including a correction for small angles f_{sml} and metals as the central atom of the angle f_{mtl}

$$k_{bnd}(ABC) = f_{qq} \cdot f_{sml} \cdot f_{mtl} \cdot k_a(A) k_a(B) k_a(C)\,. \tag{3.10}$$

k_a represent element specific angular bending parameters determined from the fit to reference data. $f_{dmp} = f_{dmp}(AB) f_{dmp}(AC)$ is the product of the distance dependent damping functions (modified from Ref. 245) for AB and AC that are given by

$$f_{dmp}(AB, R) = \frac{1}{1 + k_{dmp}\left(\frac{R_{AB}}{R^{cov}_{AB}}\right)^4}\,, \tag{3.11}$$

with the covalent distance for the pair $R^{cov}_{AB} = R^{cov}_A + R^{cov}_B$ as the sum of covalent atomic radii[246] and a global parameter k_{dmp}, chosen such that the potential vanishes approximately at twice the covalent distance. This damping of bending as well as torsional terms (*vide infra*) for long distances allows proper dissociation of the molecule into atoms.

Torsion term

The energy expression E_{tors} for a rotation around a bond AB with atom C connected to A and atom D to B is given by

$$E_{\text{tors}} = \sum_{\text{torsions}} f_{dmp} k_{tor} \left[1 + cos\left(n(\psi - \psi_0) + \pi\right)\right] , \tag{3.12}$$

where f_{dmp} is the triple damping product in analogy with the bending term (for CA, AB, BD) and ψ_0 is the equilibrium torsion angle. The appropriate multiplicity number n describes a rotation that is periodic by 360° for $n = 1$, the $n = 2$ term is periodic by 180°, the $n = 3$ term is periodic by 120° and so on. k_{tor} is the torsion force constant, which determines the size of the rotational barrier around the A-B bond. It is determined as a function of the σ- (f_σ) and π- (f_π) character of the A-B bond and of the atomic partial charges (f_{qq})

$$k_{tor} = f_\sigma \cdot f_\pi \cdot f_{qq} \cdot k_t(AB) k_t(CD) . \tag{3.13}$$

k_t represent element specific torsion parameters determined from the fit to reference data. Out-of-plane or improper torsion terms are constructed in the same manner, taking into account all three-fold coordinated sp^2-hybridized atoms, π-atoms, and additionally three-coordinated nitrogen atoms. This choice yields physically correct double-minimum potentials and proper inversion barriers. For out-of-plane and improper torsion, a different force constant k'_{tor} is applied.

Repulsion term

Unlike other force-fields, where harmonic potentials, generalizations of Lennard-Jones[245] or Morse-type potentials are applied to describe bond stretching, in GFN-FF bond elongation and compression are considered separately. As a novelty in GFN-FF, repulsive forces arising upon bond compression are treated in analogy to classical nuclear repulsion terms as in Ref. 94

$$E_{\text{rep}}^{\text{bond}} = \sum_A \sum_B \eta_{\text{rep}}^{\text{bond}} \frac{Z_A^{\text{eff}} Z_B^{\text{eff}}}{R_{AB}} exp\left(-\sqrt{\alpha_A \alpha_B R_{AB}^3}\right) . \tag{3.14}$$

Here Z_A^{eff} and Z_B^{eff} are effective nuclear valence charges of atoms A and B, which, like α_A and α_B are element specific parameters. $\eta_{\text{rep}}^{\text{bond}}$ is a global scaling parameter of the repulsion energy especially for covalently bond atoms.

Bonded three-body term

For the description of three-body effects, the well-known ATM term[247,248] taken from the theory of dispersion forces is adopted here for the covalent regime. Therefore, the atoms A-B and B-C must be covalently connected. Such a term has never been used before in a force-field. In GFN-FF, the following expression is used to describe a correction to the interaction between three covalently bond atoms A, B, and C

$$E_{abc}^{\text{bond}} = \sum_{ABC} C_{abc} \frac{(3\, cos\theta_a\, cos\theta_b\, cos\theta_c + 1)}{(R_{AB} R_{AC} R_{BC})^3} . \tag{3.15}$$

Here, θ_a, θ_b and θ_c are the internal angles of the triangle formed by R_{AB}, R_{BC} and R_{AC}, respectively. In comparison to the description of three-body effects in the context of dispersion interaction, the triple-dipole constant C_9^{ABC} is replaced by the triple-nuclear constant C_{abc} given by

$$C_{abc} = f_{q,a} f_{q,b} f_{q,c}\, \eta_{abc}\, (Z_A Z_B Z_C)^{1/3}\,, \tag{3.16}$$

where f_q are linear functions of the respective atomic charge. η_{abc} is a global parameter and Z is the scaled nuclear charge. E_{abc}^{bond} may be considered as a non-additive three-body correction to covalent binding which appears in QM methods as many center nuclear-electron attraction and electron-electron repulsion integrals.

Electrostatic energy

For the description of isotropic electrostatic (IES) energy, a classical charge model based on electronegativity equilibration (EEQ) of Gaussian type model charge densities is used, as previously applied in the DFT-D4 scheme.[242,249] This EEQ model allows charge distribution on the whole system in an optimal way, includes penetration effects, and thus can describe neutral, as well as charged systems. The IES energy expression, which is obtained variationally, is given by

$$E_{\text{IES}} = \sum_A \left[\chi_A q_A + \frac{1}{2}\left(J_{AA} + \frac{2\gamma_{AA}}{\sqrt{\pi}} \right) q_A^2 \right] + \sum_{A>B} q_A q_B \frac{erf(\gamma_{AB} R_{AB})}{R_{AB}}\,, \tag{3.17}$$

or within matrix notation as

$$E_{\text{IES}} = \mathbf{q}^{\mathrm{T}} \left(\frac{1}{2} \mathbb{A}\mathbf{q} - \mathbf{X} \right)\,, \tag{3.18}$$

where elements of the $\mathbf{X}$ vector and elements of the $\mathbb{A}$ matrix are given by

$$X_A = \Omega_A - EN_A\,, \tag{3.19}$$

with

$$A_{AB} = \begin{cases} J_{AA} + \frac{2\gamma_{AA}}{\sqrt{\pi}} & \text{for A = B,} \\ \frac{erf(\gamma_{AB} R_{AB})}{R_{AB}} & \text{otherwise.} \end{cases} \tag{3.20}$$

Here, $\gamma_{AB} = \frac{1}{\sqrt{a_A^2 + a_B^2}}$ with the atomic radii a_A and J_{AA} represents element dependent atomic hardness. The modified electronegativity X_A is computed from the fitted atomic electonegativities EN_A and a scaled logarithmic coordination number (CN) Ω_A

$$\Omega_A = \kappa_A \sqrt{log\left(\frac{1 + exp(CN_{\text{max}})}{1 + exp(CN_{\text{max}} - mCN_A)} \right)}\,, \tag{3.21}$$

where κ_A is an element dependent scaling factor, CN_{max} the maximum CN and mCN_A the modified D3 CN according to

$$mCN_A = \frac{1}{2} \sum_{B \neq A} \left[1 + erf\left(-7.5 \left(\frac{R_{AB}}{R_{AB}^{cov}} - 1 \right) \right) \right]\,. \tag{3.22}$$

The logarithmic function applied has the following properties. For small values, it is identical to the modified CN in Eq. 3.22, but for larger values, it approaches asymptotically the highest chemically reasonable value for a CN in the molecular case (CN_{max}). This behavior of Ω_A prevents the modified electronegativity X_A to adopt nonphysical values in highly coordinated systems. In molecular systems the atomic charges have to sum up to the total charge of the system $\sum_A q_A = q_{tot}$. Adding this constrains in terms of Lagrange multipliers to Eq. 3.18 leads to the modified linear system of equations

$$\begin{pmatrix} \mathbb{A} & \mathbf{1} \\ \mathbf{1}^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ q_{\mathrm{tot}} \end{pmatrix} . \tag{3.23}$$

Solving this set of linear equations leads to the geometry dependent charges q_A. This classical charge model introduces four further empirical parameters (J_{AA} , a_a , EN_A and κ_A) per element to GFN-FF. The formal scaling of the procedure is $O(N^2)$ with the number of atoms N and therefore GFN-FF scales with this same order of magnitude. The analytical derivative of the partial atomic charges with respect to nuclear displacements $\frac{\partial \mathbf{q}}{\partial R_i}$ scales cubic $O(N^3)$ and thus it is impractical for a fast and general force-field. Thus, in GFN-FF a second set of purely topology dependent charges $\mathbf{q_t}$ is introduced as mentioned already above in the setup procedure. The Floyd–Warshall algorithm[250] is employed to determine the shortest covalent path between all pairs of atoms. Replacing the interatomic distance R_{AB} in Eq. 3.17 by the sum of all covalent radii that lie on the shortest path between atom A and B leads to geometry independent but purely topology based atomic partial charges

$$R_{AB}^{\mathrm{topo}} = \sum_{i}^{A \to B} \eta_{\mathrm{topo}} \, R_i^{'cov} . \tag{3.24}$$

Here, η_{topo} is a global scaling parameter and the sum of covalent radii $R_{AB}^{'cov} = R_A^{cov} + R_B^{cov}$.[251] This topological approach eliminates the costly analytical derivative of these charges with respect to nuclear displacements but still includes the environmental effects of neighboring atoms. In addition to the dependency of the charges on the system geometry in the EEQ model, the topology based GFN-FF charges (which are used to derive various potential energy terms) take further polarization effects into account. GFN-FF is thus a partially polarizable but still efficient force-field with environment dependent atomic charges.

Dispersion energy

In GFN-FF long-range dispersion is treated by applying a modified version of the DFT-D4 scheme. Here, in analogy to the well established D3 scheme,[252,253] the pair–wise dipole-dipole dispersion coefficients are calculated from a Casimir–Polder integration over atomic dynamic polarizabilities $\alpha(i\omega)$.

$$C_6^{AB,D3} = \frac{3}{\pi} \int_0^{\infty} d\omega \, \alpha^A(i\omega) \alpha^B(i\omega) . \tag{3.25}$$

The resulting C_6^{D3} coefficients are multiplied by a charge-scaling function ζ

$$C_6^{AB} = \zeta^A \zeta^B C_6^{AB,D3}, \tag{3.26}$$

with the atom specific charge function ζ^A defined as

$$\zeta^A \left(z^A, z^{A,ref}\right) = exp\left[\beta_1 \left\{1 - exp\left[\gamma^A \left(1 - \frac{z^{A,ref}}{z^A}\right)\right]\right\}\right], \tag{3.27}$$

where the chemical hardness γ_A is taken from Ref. 254 and determines as an element specific parameter the steepness of the scaling function while β_1 is a global parameter. The effective nuclear charge z^A is defined as the sum of the nuclear charge of atom A and the topological atomic partial charges q_t^A

$$z^A = Z^A + q_t^a. \tag{3.28}$$

$z^{A,ref}$ are effective nuclear charges for element specific reference systems employing modified nuclear charges for all elements beyond krypton. In contrast to the D4 model, where atomic reference polarizabilities $\alpha(i\omega)$ are scaled by ζ, in GFN-FF the charge scaling is applied to pair–wise C_6^{AB} coefficients. C_8^{AB} coefficients are computed recursively from the C_6^{AB} coefficients. The dispersion energy is calculated according to

$$E_{\text{disp}}^{(6,8)} = -\sum_{A,B}\sum_{n=6,8} s_n \frac{C_{(n)}^{AB}}{R_{AB}^{(n)}} f_{damp}^{(n)}\left(R_{AB}\right), \tag{3.29}$$

where s_n scales the individual multipolar contributions and $f_{damp}^{(n)}$ denotes the rational Becke–Johnson (BJ) damping function.[255] Treating only two-body dispersion contributions in GFN-FF leads to a scaling of $O(N^2)$ for the energy and gradient computation.

Hydrogen bond correction

For atom pairs involving the electronegative elements C, N, O, F, Si, P, S, Cl, As, Se, Br, Sb, Te, and I, an additional hydrogen bonding correction is applied when as third atom hydrogen is involved. A case distinction is made between a classical hydrogen bonding motif, where the hydrogen atom is covalently bound to the donor atom A and noncovalently interacting with a hydrogen acceptor atom B. In the second case, the H atom is noncovalently interacting with both, the donor and the acceptor atom (*i.e.*, it is covalently bonded neither to A nor to B). The total hydrogen bonding energy is given by

$$E_{\text{HB}} = E_{\text{HB}}^{A-H\cdots B} + E_{\text{HB}}^{A\cdots H\cdots B}. \tag{3.30}$$

Here, the sum is taken over specific atom triples AHB (donor–hydrogen–acceptor) as

$$E_{\text{HB}}^{A-H\cdots B} = -\sum_{AHB} f_{dmp}^{srt} f_{dmp}^{lng} \Upsilon_{dmp}^{out} \chi_{AHB}^{\alpha} \left(\frac{\omega_{AB}}{R_{AB}^3} + \frac{\omega_{BH}}{R_{BH}^3}\right), \tag{3.31}$$

where ω_{AB} and ω_{BH} are global parameters that determine the weighting of the distance dependency between donor–acceptor (R_{AB}) and acceptor–hydrogen (R_{BH}). The dependence on the inverse third power of the distance is chosen due to the strong polarization and dipole-dipole character of hydrogen bonding, which differs from the electrostatic Coulomb interaction. To smoothly interpolate between donor/acceptor character, the interaction strength is modified to ensure that the hydrogen bonding correction vanishes for weak donor basicity and weak acceptor acidity character

$$\chi^{\alpha}_{AHB} = c^{A}_{a}\rho^{A}_{q}\, c^{B}_{b}\rho^{B}_{q}\, \rho^{H}_{q}\, \eta_{hb}\,. \tag{3.32}$$

Here, η_{hb} is a global scaling parameter and c^{A}_{a} and c^{B}_{b} are global fitting parameters describing the acidity of atom A and the basicity of atom B. Charge dependencies are included for atoms A, B and H by employing the charge functions ρ_q, which are based on the topology dependent charges $\mathbf{q_t}$ and the global parameters k_{q1} and k_{q1}

$$\rho^{A,B}_{q} = \frac{exp\left(-k_{q1}\, q^{A,B}_{t}\right)}{exp\left(-k_{q1}\, q^{A,B}_{t}\right) + k_{q2}}\,, \tag{3.33}$$

and

$$\rho^{H}_{q} = \frac{exp\left(k_{q1}\, q^{H}_{t}\right)}{exp\left(k_{q1}\, q^{H}_{t}\right) + k_{q2}}\,. \tag{3.34}$$

Here, f^{srt}_{dmp} is a damping function reducing the H-bond contribution to zero for short distances between donor and acceptor

$$f^{srt}_{dmp} = \frac{1}{1 + \left(\frac{\eta_{srt} R'^{cov}_{AB}}{R^{2}_{AB}}\right)^{\gamma_{srt}}}\,, \tag{3.35}$$

where η_{srt} and γ_{srt} are global parameters and $R'^{cov}_{AB} = R'^{cov}_{A} + R'^{cov}_{B}$. For long distances between donor and acceptor a corresponding damping function f^{lng}_{damp} is applied with η_{srt} and γ_{srt} as global parameters

$$f^{lng}_{dmp} = \frac{1}{1 + \left(\frac{R^{2}_{AB}}{\eta_{lng}}\right)^{\gamma_{lng}}}\,. \tag{3.36}$$

The strength of a hydrogen bond correlates with the angle formed by the three atoms AHB involved. In order to introduce an angular dependency the out-of-line damping function Υ^{out}_{dmp} is introduced

$$\Upsilon^{out}_{dmp} = f^{out}_{H} \prod_{i}^{\Delta_B} f^{out}_{i}\,. \tag{3.37}$$

Here, the first out-of-line term f^{out}_{H} is applied so that the hydrogen bond contribution vanishes for nonlinear arrangements

$$f^{out}_{H} = \frac{2}{1 + exp\left[\frac{\eta^{H}_{out}}{R'^{cov}_{A} + R'^{cov}_{B}}\left(\frac{R_{AH}+R_{BH}}{R_{AB}}\right) - 1\right]}\,, \tag{3.38}$$

where η_{out}^{H} is a global parameter. For the correct description of hydrogen bonding motifs, knowledge about the presence and location of paired electrons in lone-pair orbitals (LP) is crucial. Because of the absence of information on electrons and occupied orbitals in a force-field, this dependence is nontrivial to realize. GFN-FF addresses the problem of how to locate the LP by a simple geometrical model. The LP is assumed to be located at maximum distance to all covalently bound neighbors of the hydrogen accepting atom, obeying the chemical VSEPR principle. The second term in Eq. 3.37 is the product of damping functions f_i^{out} over the number of covalently bound neighbors of the hydrogen acceptor Δ_B

$$f_i^{out} = \frac{2}{1 + exp\left[-\frac{\eta_{out}^{\Delta}}{R_A^{cov}+R_B^{cov}}\left(\frac{R_{Ai}+R_{Bi}}{R_{AB}}\right) - 1\right]} - 1\,, \tag{3.39}$$

where η_{out}^{Δ} is a global parameter. This damping function can be regarded as the inverse of f_H^{out} and it is applied so that the hydrogen bond contribution vanishes for linear arrangements of the neighboring atoms of the acceptor atom in between the donor–acceptor bond. To the authors knowledge, this is the first attempt to introduce LPs to force-fields via an exclusion principle over neighbor lists.

For carbonyl- and nitro-groups involved as a lone-pair donor within a hydrogen bond, a torsion and bending potential is applied for the noncovalent interaction between donor and acceptor. The hydrogen bond potential is modified by,

$$E_{\mathrm{HB}}^{A-H\cdots B} = -\sum_{AHB} f_{dmp}^{srt} f_{dmp}^{lng} \Upsilon_{dmp}^{out} \chi_{AHB}^{\alpha} \left(\frac{\omega_{AB}}{R_{AB}^3} + \frac{\omega_{BH}}{R_{BH}^3}\right) f_{tors}(\psi)\, f_{bend}(\theta)\,, \tag{3.40}$$

where $f_{tors}(\psi)$ is a torsion function and $f_{bend}(\theta)$ a bending function, derived from the potentials used for covalent bending and torsion respectively (*vide supra*). The angles ψ and θ are defined in Figure 3.4. By this refinement of the HB potential, the position of the hydrogen atom is favored within the plane of the carbonyl group at an angle θ of 120° or 240° respectively, which are the positions of the lone-pairs at the carbonyl oxygen atom.

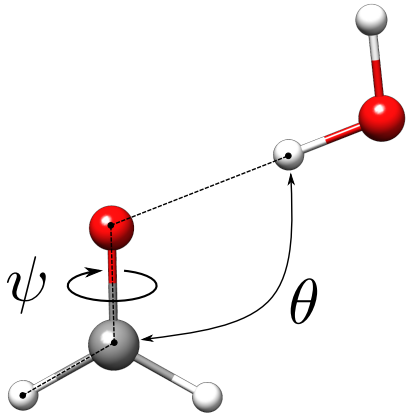


Figure 3.4.: Torsion and bend angle of the noncovalent interaction of water and formaldehyde. Bending and torsion potentials are chosen to place the hydrogen atom of the H-bond in the carbonyl plane at an angle of 120° or 240°.

The H-bond term for the second case is given by a sum over specific atom triples AHB (donor/acceptor–hydrogen–donor/acceptor) as

$$E_{\mathrm{HB}}^{A\cdots H\cdots B} = -\sum_{AHB} f_{dmp}^{srt} f_{dmp}^{lng} f_{H}^{out} \frac{\chi_{AHB}^{\beta}}{R_{AB}^{3}} \,. \tag{3.41}$$

Here the same short- and long-range damping functions are used as before. For the out-of-line damping only the linearity relation regarding the hydrogen atom is applied and neighbors are not further included. For symmetric H-bonds the interaction strength is modified by

$$\chi_{AHB}^{\beta} = \frac{c_a^A R_{BH}^4 + c_a^B R_{AH}^4}{R_{AH}^4 + R_{BH}^4} \cdot \frac{c_b^A \rho_q^A R_{BH}^4 + c_b^B \rho_q^B R_{AH}^4}{R_{AH}^4 + R_{BH}^4} \cdot q_t^H \,, \tag{3.42}$$

where the first term is a measure of the acidity of hydrogen donor and acceptor and the second term the corresponding basicity scaled by the topological charge of the hydrogen atom. Again, for weak donor basicity and acceptor acidity character the hydrogen bonding correction vanishes.

Halogen bond correction

The electron acceptor property of the halogens is based on the so-called sigma hole model.[256] To describe this effect in GFN-FF, a similar potential as for the hydrogen bond correction is used for the halogen bonding situation D-X-Y where atom D is any donor atom, X is the halogen or other sigma-hole containing elements[257] (P, S, Cl, As, Se, Br, Sb, Te, and I) and Y can be any acceptor atom from group 15–17. The interaction is treated similar to hydrogen bonding as a sum over all atom triples DXY

$$E_{\mathrm{XB}} = -\sum_{DXY} f_{dmp}^{srt} f_{dmp}^{lng} f_{X}^{out} \frac{\chi_{DXY}^{\alpha}}{R_{XY}^{3}} \,, \tag{3.43}$$

where again short- and long-range damping functions are applied and f_X^{out} is the out-of-line damping function for atom X. The interaction strength χ_{DXY}^{α} in this version depends only on the halogen atom and its acceptor

$$\chi_{DXY}^{\alpha} = c_b^Y \rho_q^Y \, c_b^X \rho_q^X \,, \tag{3.44}$$

where c_b are the corresponding global basicity parameters and ρ_q the charge dependent scaling functions.

3.3. Results and Discussion

To conduct chemically sensible atomistic modeling, knowledge about molecular structure, binding motifs, and structural dynamics is essential. Accurate molecular geometries give insights into the composition and functionality of the investigated system. GFN-FF, as implemented in the free *xtb* program, is equipped with a highly sophisticated quasi-Newton geometry optimization engine,[96] which is also used by QM methods. Implementation of a fragmented Hessian scheme provides the necessary speedup

to be practical also at the FF level. All results discussed in the following examples are given in detail in Appendix A3.

The class of organic-inorganic hybrid crystalline porous materials, referred to as metal-organic frameworks (MOF), attracts much attention due to their potential application in gas storage, chemical separation, drug transport, and catalysis.[258] For their theoretical description only a few specialized methods are in principle available, as for instance UFF, UFF4MOF,[259] and MOF-FF. Since the latter is only parameterized for certain metal-organic binding motifs, UFF is the only true competitor for GFN-FF as a general black-box FF. However, UFF shows deficiencies in the description of conjugated systems and η-metal-coordinated bonding motifs. For a selected set of five metal-organic polyhedra (MOP) and MOF cut-outs depicted in Figure 3.1, GFN-FF is to the authors' knowledge the only method capable of performing geometry optimizations whilst keeping the initial structure intact. Fujita *et al.*[260] synthesized the world's largest metal-organic Goldberg polyhedra through a self-assembly reaction. The structure consists of 3888 atoms in total and is made of 46 Pd^{2+} ions coordinated square planar by 96 organic ligands. Charge neutrality is conserved by two BF_4^- molecules per palladium ion. With GFN-FF the structure is optimized properly. An overlay with the crystal structure is shown in Figure 3.1A. The heavy-atom root mean square deviation (RMSD) of only 0.75 Å, indicates an excellent agreement between experiment and theory. For the second largest MOP shown in Figure 3.1B, a similarly accurate result is obtained. Gong *et al.*[261] performed a bottom-up construction of supramolecular organic polyhedra with tetrahedral symmetries. In a noncovalently bound three-dimensional mesh of tetrahedra, each tetrahedral corner consists of vanadium-oxide clusters. For the geometry optimization, the structural motif was truncated in a star-like shape to end up at 2496 atoms in total shown in Figure 3.1C. With an RMSD of 0.54 Å, the GFN-FF optimized structure corresponds very well with the experiment. A triangular channel framework constructed of $Fe^{III}_2(BDP)_3$ units (BDP^{2-} = 1,4-benzenedipyrazolate) was synthesized by Herm *et al.*[262] A cutout of 4064 atoms, shown in Figure 3.1D, is chosen and optimized by GFN-FF, yielding an RMSD of only 0.43 Å. With 1344 atoms the cuboctahedron $Co^{II}_{12}L_6$ hosting two C_{60} fullerene molecules, synthesized by Rizzuto *et al.*[263] and shown in Figure 3.1E, is small enough to be described at the SQM level of theory. However, the electronic structure of twelve cobalt ions and eight negatively charged borate-clusters, leading to an overall molecular charge of +16, is too difficult, and hence, the iterative self-consistent field calculations fail to converge. On contrary, GFN-FF is able to describe the structure with an RMSD of 1.22 Å compared to the experimental crystal structure. This somewhat larger RMSD is mainly caused by a movement of the noncovalently bound borate-clusters due to the absence of confining crystal packing effects in the molecular calculation.

To further show the excellent performance of GFN-FF, the molecular structure of hemoglobin is optimized starting from a molecular crystal structure cut-out. The theoretical treatment is challenging here because of the co-existence of a large bio-organic framework (overall $\approx$ 9000 atoms) and the chemically complicated organometallic heme groups. The overlay of the crystal structure (yellow) and the GFN-FF optimized geometry (blue) of hemoglobin is shown in Figure 3.5A. The calculation was conducted with an implicit Generalized Born (GB) solvation model augmented with the solvent accessible surface area (SA). This GBSA solvation model is implemented in the *xtb* program and available for GFN-FF. The inclusion of solvation effects is essential for the accurate modeling of bio-macromolecules or porous

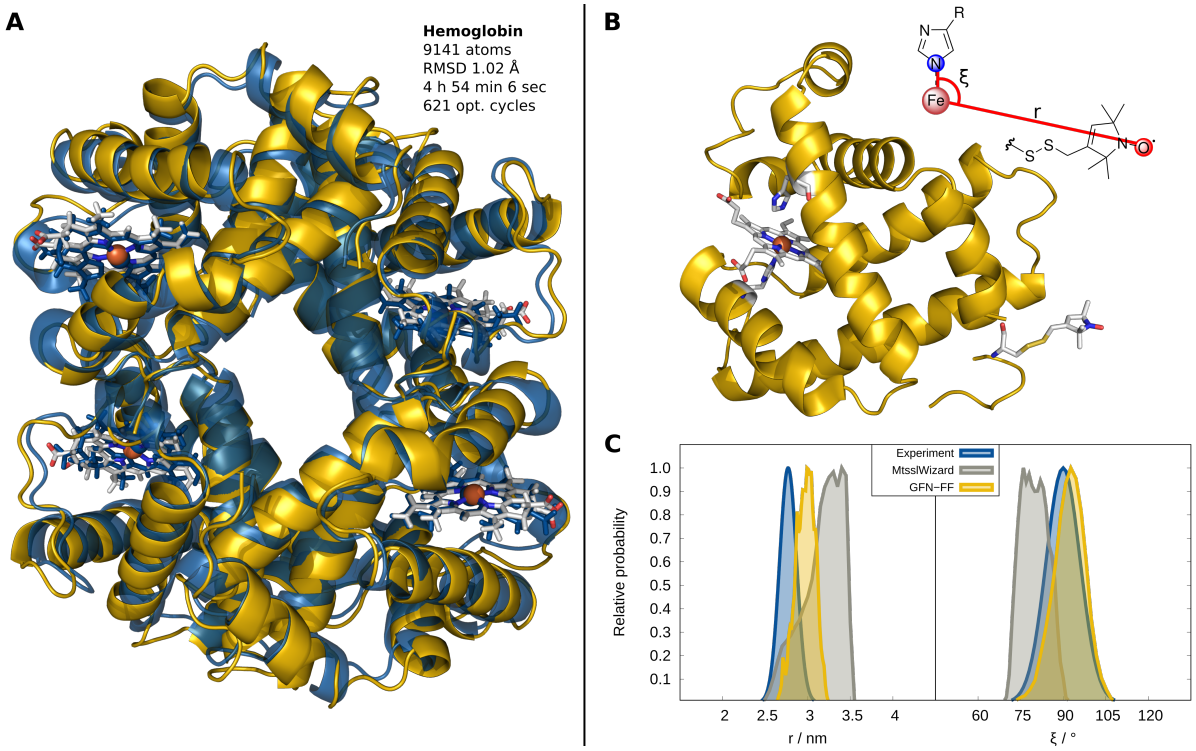


Figure 3.5.: Structures and dynamics of metallo-proteins. (A) Geometry optimization of the hemoglobin structure at the GFN-FF/GBSA(H_2O) level of theory. (B) Myoglobin mutant Q8R1 with an open shell iron (III) and a nitroxide spin label covalently attached. (C) Comparison of $P(r)$ and $P(\xi)$ to experimental EPR data and previous MtsslWizzard results.

materials in order to prevent structures from a "gas phase collapse". Within only 5 hours and 621 optimization cycles on four Intel© Xeon E5-2660 v4 @ 2.00 GHz CPUs, a stationary point on the PES was found. The heavy atom RMSD between the experimental and the GFN-FF structure is only 1.02 Å, which is an excellent result for such a comparison. It shows that GFN-FF is an efficient and technically robust FF with a physically reasonable potential that is capable of describing amino acids and metal containing heme groups similarly well. Due to the current occurrences, COVID-19, the illness caused by the new coronavirus, is the focus of clinical research. To show the utility of GFN-FF, a successful geometry optimization was performed on the COVID-19 main protease in complex with an inhibitor N3 starting from the crystal structure.[264] Again, a small RMSD between the theoretical and experimental structure of 0.95 Å is found (for details see Appendix A3).

One of the prime application of force-field methods is MD simulations. From the obtained (space-time) trajectory of the atoms, geometrical and molecular properties are derived that can be directly compared to the experiment. Abdullin *et al.*[265] conducted electronic paramagnetic resonance (EPR) measurements on a met-myoglobin mutant (**Q8R1**) shown in Figure 3.5B. EPR spectroscopy was used to measure the average distance r between a high-spin Fe^{3+} ion and a nitroxide spin label termed MTSSL, as well as the angle ξ formed by the iron ion, its nitrogen ligand of a histidine amino acid, and the nitroxide group. To obtain the radial distribution $P(r)$ and angular distribution $P(\xi)$ for **Q8R1**,

the structure and dynamics were determined *in silico* by GFN-FF and compared to experimental EPR values and previous theoretical estimates performed with MtsslWizard.[266] MtsslWizard is a program that searches for possible MTSSL conformations, which do not clash with a static model of the protein. Such a simple model is often the method of choice since established protein force-fields do neither provide parameters for metals, nor the chosen nitroxide spin-label. With GFN-FF, MD simulations were carried out for 1 ns at 298 K employing the GBSA(H_2O) solvation model (see Figure 3.5C). Compared to the maximum in the EPR measured distance distribution, GFN-FF shows a deviation in the maximum of only 3 Å, which is within the experimental error. For the angular distribution, the difference is only 2 ° and hence, the GFN-FF dynamical structure average is in almost perfect agreement with the experiment. This again demonstrates the accuracy of the presented FF and proves the reliability of the therewith calculated PES also for non-equilibrium situations. The results with MtsslWizard (deviations for maxima of $P(r)$ of 9 Å and 15° for $P(\xi)$, respectively) are clearly worse. A related study on the B1 immunoglobulin-binding domain of a protein termed GB1 using a tailored FF for Cu-containing metallo-proteins appeared recently.[267] The laborious "hand-made" parameterization described could have been completely avoided by using GFN-FF, which is available not only for copper but also all other transition metals.

The performance of GFN-FF compared to other general as well as highly specialized FFs is depicted in Figure 3.6 for a benchmark set of 70 organic peptide and protein structures,[224] where geometry optimizations with OPLS2005,[65] AMBER*,[222,223] UFF, and GFN2-xTB were conducted. Figure 3.6A shows the average deviations of four dihedral angles in degree and the C_α and heavy atoms RMSD in Å with respect to the X-Ray structures. A structural example is shown in Figure 3.6B. The deviations of the angles ϕ, ψ, χ, and ω are soft descriptors regarding local displacements of the protein backbone and are shown in detail in Appendix A3. For the angles ϕ, ψ, and χ, GFN-FF yields about the same or even better accuracy as the special-purpose method OPLS2005 and is clearly more accurate than AMBER*. The larger deviations for ω indicate that the barrier for rotation around the peptide C-N bond seems to be underestimated by GFN-FF, which is only observed for larger proteins but not for smaller peptides. The RMSD for C_α and heavy atoms are comparable to the much more elaborate GFN2-xTB QM method. For a general FF that has not been specifically developed for proteins, GFN-FF performs overall excellently on the tested protein structures. On the other hand, UFF provides large deviations for the angles ϕ, ψ and ω as well as for the C_α and heavy atom RMSD, indicating that protein structures are not well described by UFF.

For transition-metal complexes, the quality of theoretical structures is tested on the challenging TMG145 benchmark set[221] (for structural examples see Figure 3.6E). The performance of GFN-FF is compared to UFF and GFN2-xTB with reference to high-quality DFT optimized structures (TPSSh-D3(BJ)-ATM/def2-TZVPP[134,252,268] level). In Figure 3.6C, 941 bond lengths d(M-A), including mainly the transition-metal ligands, are shown. With a mean absolute deviation (MAD) of 9.7 pm, GFN-FF performs just as well as the GFN2-xTB QM method with a MAD of 8.3 pm. Furthermore, the reproduction of 2846 bond angles around the transition-metal center ∠(A-M-B) is compared in Figure 3.1D. Again GFN-FF (MAD = 5.7 °) performs similar as GFN2-xTB (MAD = 3.9 °). For angles and bond lengths, GFN-FF clearly outperforms UFF, which yields a MAD of 14.6 pm, and 8.4 °, respectively. All 145 GFN-

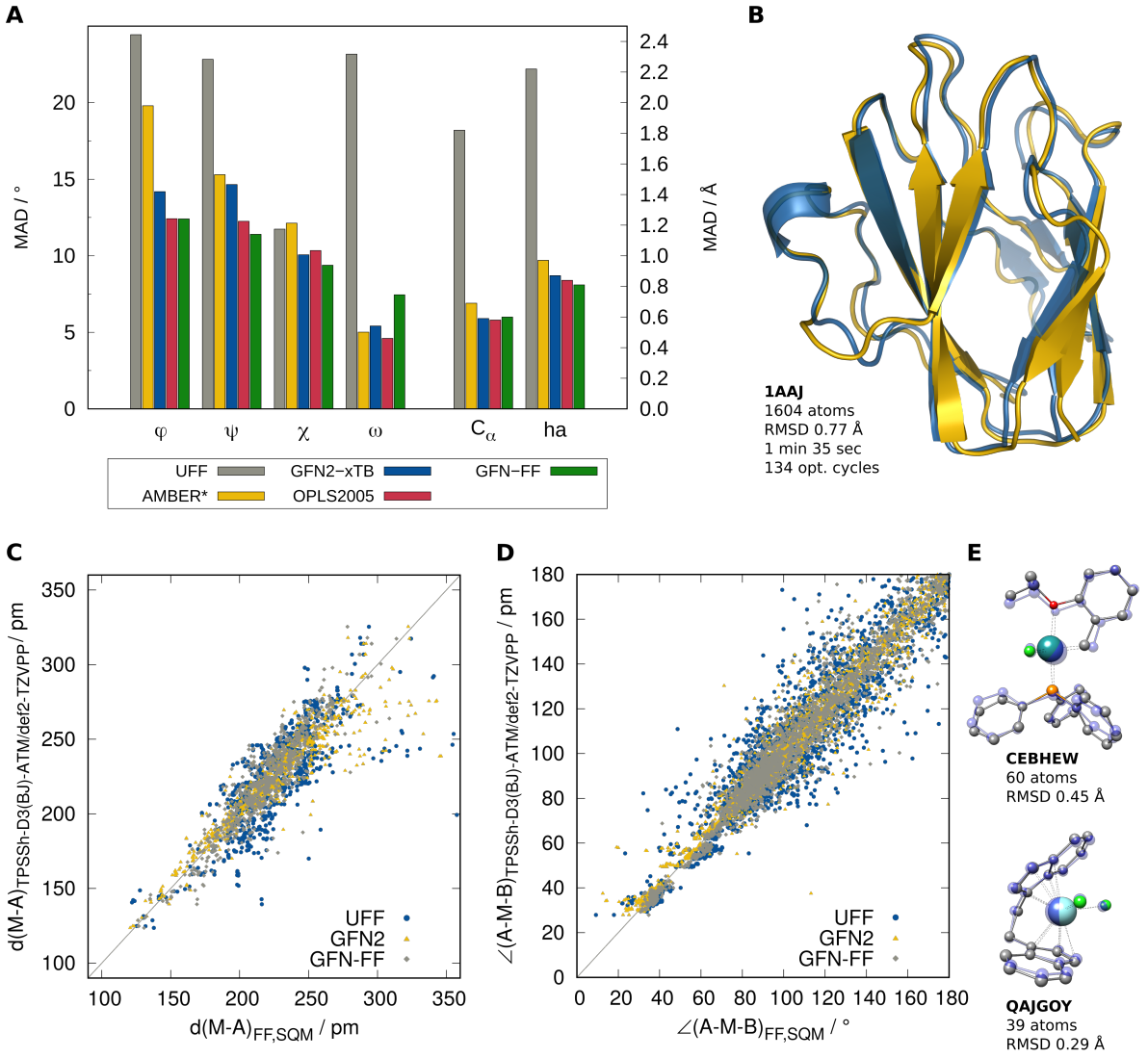


Figure 3.6.: Comparison of GFN-FF to established theoretical methods. (A) Average deviations of four types of dihedral angles (in degree) for 70 protein structures with respect to the crystal structure as well as average C_α and heavy atoms RMSD (in Å) with example (B). (C-D) Performance for computed bond lengths and angles of the TMG145 benchmark set. Correlation plots for bond lengths and angles obtained with GFN2-xTB, GFN-FF and UFF with reference to DFT structures. An example complex shown in (E).

FF optimized structures fulfill the previous chemical correctness criteria used to identify structures that are chemically transformed, dissociated, or critically deformed during optimization, while UFF produces 75 out of 145 structures completely wrong. The performance and robustness of GFN-FF for transition-metal complexes are outstanding and unmatched by its direct competitor UFF.

As the previous results and comparisons suggest, GFN-FF provides almost as accurate results for equilibrium structures as sophisticated QM methods. To investigate this further for interaction energies, GFN-FF results are compared to QM ones on various interaction energy benchmark sets as shown in Figure 3.7. Intermolecular noncovalent interactions are investigated on supramolecular host–guest

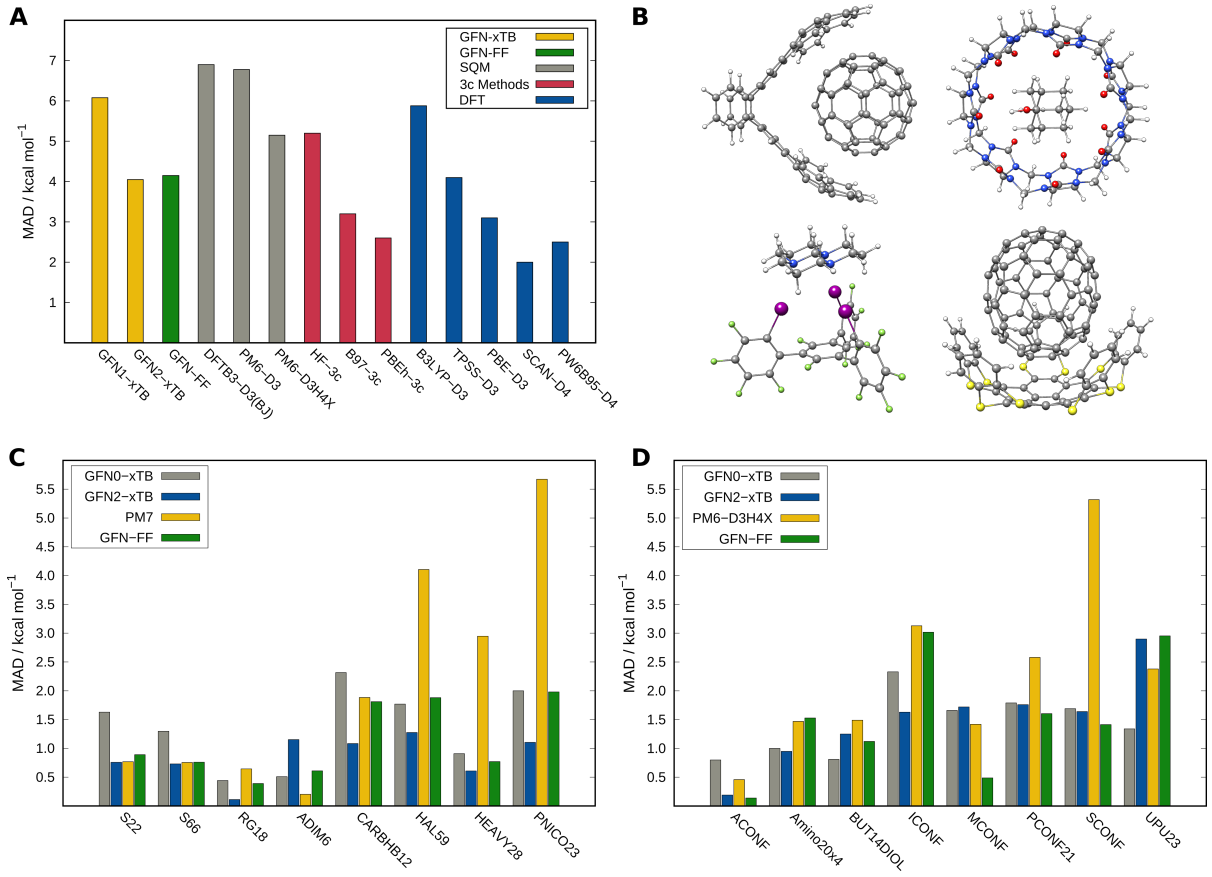


Figure 3.7.: Comparing GFN-FF performance for established benchmark sets. (A) Association energies averaged over all systems in the S30L set computed with different methods. The reference energies are obtained at the DLPNO-CCSD(T)/CBS level. Four examples of structures are shown in (B). In (C) MAD values are given for intermolecular noncovalent interaction sets and several conformational benchmark sets in (D), including the two xTB methods, GFN-FF, PM7 and PM6-D3H4X.

systems taken from the S30L benchmark.[120] In Figure 3.7A, the performance of GFN-FF and other QM methods is shown compared to DLPNO-CCSD(T)/CBS reference values.[237] The MAD of the association energy for the entire test set is given and Figure 3.7B shows four example complexes. With an overall MAD of 4.15 kcal mol^{-1}, GFN-FF outperforms most of the SQM methods and is even on par with some dispersion corrected DFT methods. In Figure 3.7C, the performance of GFN-FF is shown for all NCI energy subsets of the huge GMTKN55 database.[162] With an overall MAD of 1.13 kcal mol^{-1} the accuracy of GFN-FF is comparable to that of GFN1-xTB and outperforming the SQM method PM7. The GMTKN55 subsets dealing with conformational energies are shown in Figure 3.7D. Again, GFN-FF is just as good as the GFN*n*-xTB QM methods with a MAD of 1.53 kcal mol^{-1}. Here, the excellent performance for the relative energies of alkane (ACONF) and melatonin (MCONF) conformers is noted.

3.4. Conclusion

The development of accurate polarizable force-fields is named as one of the remaining holy grails for computational chemistry.[2] With GFN-FF, a generic partially polarizable force-field is presented that is unique in its universality and accuracy, meeting, to a large extent, the specified requirements in Ref. 2. It is a robust and fully automated black-box method for the modeling and design of materials, organometallic and biochemical systems. The performance is tested on various examples and many established benchmark sets. For biological macro-molecules, it is shown that GFN-FF is able to simulate the dynamics of a met-myoglobin mutant and reproduces the experimental EPR distance measurements par excellence. For a test set of 70 protein equilibrium structures, GFN-FF has proven to perform similarly or even slightly better than highly specialized protein force-fields. For metal-organic materials, GFN-FF is, in many cases, the only currently applicable method, and for a highly complex test set of transition-metal complexes, it exceeds the only real competitor UFF by far in terms of accuracy, robustness, and efficiency. For structures and energies, GFN-FF is approaching the accuracy of semiempirical QM methods, in some cases reaching even DFT accuracy. The main limitation of GFN-FF (similar to all other non-reactive force-fields) is that the input structure must be reasonably close to a "normal" chemical bonding situation such that the initial topology analysis works properly. If this is not the case, a few pre-optimization steps with a GFN*n*-xTB QM method to adjust the covalent bonding network may be applied. In this work, a new quality standard is set for general force-fields, providing high universality paired with almost QM accuracy at still high computational speed. An easy-to-use and freely available computer program implementing GFN-FF can be downloaded for extended applications in physical- and biochemistry.[243]

Acknowledgments

This work was supported by the DFG in the framework of the "Gottfried-Wilhelm-Leibniz" prize to Stefan Grimme. Sebastian Spicher thanks the "Fond der Chemischen Industrie (FCI)" for financial support. The authors thank M. Bursch, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert and the users of the *xtb* program for helpful discussions and many initial tests of the method.

Part III.

Application of Efficient Computational Methods for Organometallic and Biochemical Systems

Force-fields enable new computational possibilities regarding molecular size and complexity. Two of the most important fields of application are materials science and biochemistry. In material science, force-fields are often the method of choice for predicting properties of, *e.g.*, metal-organic frameworks, due to the difficult electronic structure and sheer size of the systems.[269] In the modeling of biochemical systems, MD simulations at the force-field level represent a valuable tool for characterizing the structural dynamics of DNA, (m)RNA, and (metallo-)proteins.[270]

Part III of this thesis represents the application and evaluation of GFN-FF, and the GFN*n*-xTB methods in general, to fundamental scientific questions regarding gas adsorption and storage in porous materials and the structural investigation of (metallo-)proteins.[271,272] The efficient storage of gases and fuels in porous materials is one of the most important fields of research in the context of sustainable chemistry, energy conservation, and greenhouse gas reduction. Therefore, in Chapter 4, the adsorption of climate gases, bio-fuels, and anti-cancer drugs in metal-organic frameworks and porous organic cages is modeled by a combination scheme of GFN-FF and the NCI-iMTD algorithm of CREST (see Figure 3.8).[273] The determined binding sites are further re-optimized at the DFT level to conduct a large and versatile benchmark set of 117 different gas–framework combinations. The accuracy of the GFN methods is compared to highly accurate DFT reference association energies (PBE0-D4/def2-TZVP) and the fully automatic determined binding sites are compared to other studies in the literature.

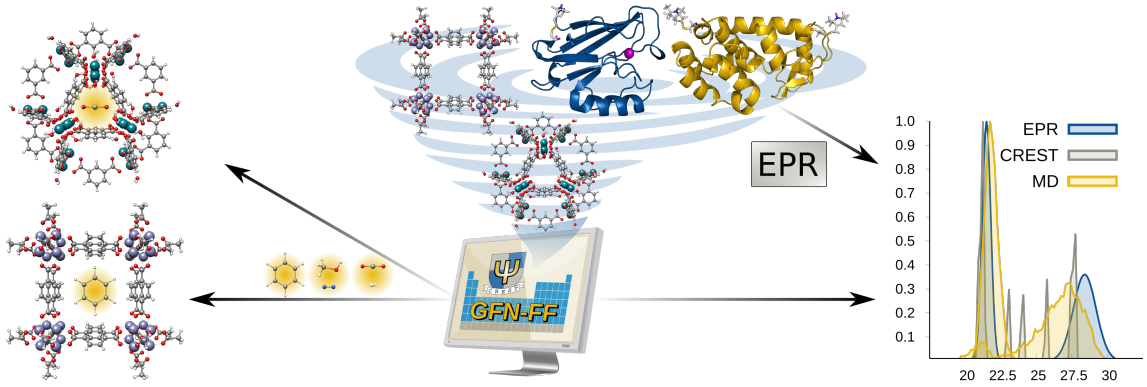


Figure 3.8.: Various applications of GFN-FF. On the left, GFN-FF is employed in combination with CREST to determine automatically optimal binding sites of gases in metal-organic frameworks. On the right, GFN-FF MD simulations are performed to model distance distributions of spin-labeled (metallo-)proteins.

In Chapter 5, a new approach termed CREST/MD is developed and tested for the calculation of distance distributions in spin-labeled (metallo-)proteins. The elucidation of the three-dimensional protein structures is inevitable for the understanding of their biological function. Theoretical approaches are desirable to facilitate this complex task and yield further insights at the microscopic level. In CREST/MD, GFN-FF MD simulations are performed for the most populated spin-label conformations, obtained by the CREST algorithm at the GFN-FF level (see Figure 3.8).[274] For realistic modeling of protein dynamics the incorporation of solvation effects is essential and the implicit GBSA solvation model is employed within all GFN-FF calculations. Experimental reference distributions were provided by

means of electron paramagnetic resonance (EPR) spectroscopy for spin-labeled mutants of azurin and T4 lysozyme. Since the EPR measurements were performed in a highly diluted glassy frozen solution, a non-periodic description with an implicit solvation model is well suited for the comparison between experiment and theory. The new CREST/MD approach is tested in comparison to previously existing theoretical approaches of modeling spin–spin distance distributions.

Chapter 6 presents a benchmark set, termed IONPI19, which is compiled for inter- and intramolecular ion–π interactions.[184] Ion-π interactions have implications for protein structure and function.[275] Thereby, the interaction of charged amino groups with aromatic side chains is a common binding motif.[276] The focus in this chapter lies on accurate dispersion corrected DFT methods in comparison to highly accurate coupled cluster reference values. The applied and evaluated methods represent state-of-the-art density functionals from different rungs of Jacob's ladder (*cf.* Section 2.1.5). With an average molecule size of 31 atoms, the IONPI19 set is also an excellent benchmark to test the backward compatibility of the GFN methods for smaller but electronically difficult systems. Besides DFT, GFN*n*-xTB, PM6-D3H4X, PM7, and GFN-FF are tested.

Regarding the variety and complexity of the investigated systems, ranging from organometallic, over highly charged, to biochemical systems, Part III displays the accuracy, robustness, and versatility of the GFN-FF and GFN*n*-xTB methods for the description of geometries and noncovalent interactions.

4. Efficient Calculation of Small Molecule Binding in Metal-Organic Frameworks and Porous Organic Cages

Sebastian Spicher,[a] Markus Bursch,[a] and Stefan Grimme[a]

Received: 22 September 2020, Published online: 3 December 2020

Reprinted (adapted) with permission[†] from
Spicher, S.; Bursch, M.; Grimme, S. *J. Phys. Chem. C* **2020**, *124*, 27529–27541.

DOI 10.1021/acs.jpcc.0c08617

Own manuscript contribution

- Conception of the study and choice of methods
- Performing all the calculations
- Interpretation of the results
- Writing the manuscript

[a]Mulliken Center for Theoretical Chemistry, Institute of Physical and Theoretical Chemistry, University of Bonn, Beringstraße 4, 53115 Bonn, Germany
[†]Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

Abstract The activation, storage, and separation of gases and fuels are closely related to the reduction of greenhouse gas emissions, the widespread use of renewable energies, and the application of industrial gases. Metal-organic frameworks (MOF) and porous organic cages (POC) are an emerging class of crystalline porous materials that show promising characteristics in this field. Yet, their accurate theoretical description poses a challenge to existing methods due to the sheer size of the pores and cages, as well as their often complex structure. In this work, the performance of generally applicable density functional approximations (DFAs), semiempirical quantum mechanical (SQM) methods, and force-fields (FFs) for the calculation of binding energies of various gases in molecular cutouts of MOFs and POCs is tested with reference to high-level PBE0-D4/def2-TZVP hybrid DFT energies. Therefore, favorable binding sites for greenhouse gases (CO_2), energy-related gases (H_2, methanol, and benzene), and industrial gases (N_2) are determined by an efficient conformer search algorithm (CREST). The resulting structures are further optimized by DFT (B97-3c), semiempirical (GFN2-xTB), and force-field (GFN-FF) methods, to yield the binding sites and corresponding energies. With mean absolute deviations ranging from 1.1 to 1.4 kcal mol^{-1} for all tested systems, the considered GFN methods reach an accuracy remarkably close to the DFT reference, justifying their application for efficient binding site screening. In comparison, the widely used PMx methods show on average 1.0 kcal mol^{-1} larger deviations. Furthermore, the application of single-point, multi-level approaches and the parallelism of potential energy surfaces are discussed.

4.1. Introduction

The efficient storage and targeted use of gases and fuels are one of the large challenges of this century.[34,277,278] Carbon dioxide separation is crucial for the alleviation of the global greenhouse effect[279] and the storage of, *e.g.*, hydrogen or methanol is indispensable for the further development of new biological fuels and concepts of renewable energy applications.[280–283] The storage and application of gases, such as nitrogen and oxygen, are important for environmental[284] as well as industrial processes and the synthesis of industrial chemicals.[285,286]

A widespread and promising class of compounds for future developments are metal-organic frameworks (MOF). They represent an aspiring class of crystalline porous materials consisting of metal-organic nodes coordinated by mostly organic linker ligands to form one-, two-, or three-dimensional structures.[287–290] MOFs are characterized by an open framework leading to high porosity and a large surface area, which is well suited for application in gas storage, chemical separation, drug transport and catalysis.[37,291–297] Unlike infinite MOFs, metal-organic polyhedra (MOP) form a subgroup of discrete self-assembled supramolecules, composed of edge-sharing molecular polygons.[298–300] A second emerging class of such materials are porous organic cages (POC), which are discrete, three-dimensional molecular assemblies.[301–303] POCs show comparable characteristics in terms of fuel/gas adsorption as their metal-organic relatives.[304–306] The exploration of advanced porous materials and their possible applications has become an intense subject of research at the border between organic and inorganic chemistry.[307]

In order to gain an atomistic understanding of molecular structures and interactions, theoretical modeling of MOFs and POCs serves as a valuable tool, and hence, the accurate description of porous frameworks is a crucial goal in computational chemistry.[308–312] One of the main challenges for *ab initio* methods such as density functional theory (DFT) is the sheer size of the systems of interest. For periodic MOFs it is common practice to choose saturated molecular cutouts that represent a repeating unit within the infinite framework.[313,314] For MOPs and POCs this is not necessary, since those are already discrete molecular systems. Thus, the size of the pore determines the minimal chemically reasonable molecular size that can be truncated to, varying from 200 to far over 1000 atoms.[260] This restricts the range of applicable DFT methods drastically, and even low-cost DFT methods like the composite methods B97-3c[237] and PBEh-3c[231] reach their practicable limits for the larger systems.

The next consequent step regarding better efficiency represent semiempircal quantum mechanical (SQM) methods, which are routinely applicable to systems with 500–1000 atoms.[96] Nevertheless, most of these methods lack a full parameterization for transition-metals and are thus not generally applicable in this field of chemistry. The two most notable exceptions are the widely used, NDDO based PM*x*[87,88,315] methods and the more recently developed extended tight-binding methods of the GFN*n*-xTB[93,94,96] family. The latter proved to be more robust and accurate for structure optimization and thermochemistry of transition-metal complexes.[221] Even more commonly used, due to further decreased computational costs compared to SQM and DFT methods, are force-fields (FFs), which are established as a common choice for the description of porous materials.[269,311,316,317] As FFs describe the atoms within a molecule only by classical potentials, they are applicable to much larger (> 1000 atoms) systems than SQM methods. Nevertheless, the functional form of the FF potentials introduces a set of atom specific parameters, which is unique for each FF. Such parameters often only exist for a limited number of elements and structural motifs, requiring individual, laborious parameterizations for the description of MOFs.[311,318–320] Recently, a partially polarizable generic FF has been introduced, termed GFN-FF,[117] which is a promising, generally applicable candidate for the description of porous materials due to its full periodic table parameterization ($Z \leq 86$).

Special method adaptions and specific method developments were used in many studies of gas adsorption in porous (metal-)organic systems. FF re-parameterization[321–323] and neural network training[324] represent some of the most common, even though inconvenient, techniques. A comparison of general, efficient out-of-the-box applicable methods on a versatile test set for small molecule binding to large (metal-)organic structures like MOPs or POCs is yet missing. Thus, in this work, a benchmark study for binding energies of gases and fuels at various metal-organic frameworks, polyhedra, and porous organic cages is presented. Generally applicable low-cost DFT, SQM and FF methods, namely B97-3c, PBEh-3c, GFN1-xTB,[94] GFN2-xTB,[93] PM6-D3H4,[86] PM7,[88] GFN-FF, and UFF[114] are tested for their performance compared to hybrid DFT gas phase binding energies (PBE0-D4[242,325]/def2-TZVP[326]//B97-3c). Other popular density functional approximations (DFAs) such as the range-separated hybrid ωB97X-V/def2-TZVP,[327] TPSS-D4/def2-TZVP,[328] and PBE-D4/def2-TZVP[166] are tested as well for comparison. Further, a workflow for the determination of binding sites by the efficient conformational search algorithm CREST[60] in combination with GFN-FF is presented, serving as the starting point for subsequent higher-level structure optimizations.

4.2. Molecular Test Set

The presented test set consists of one saturated metal-organic framework cutout (MOF-5, **1a**: H, C, O, Zn), two metal-organic polyhedra (Rh-MOP, **1b**: H, C, O, Rh; Pd-MOP$^+$, **1c**: H, C, N, Pd), where one is highly positively charged, and three different porous organic cages (Q-POC, **2a**: H, C, N, O; B-POC, **2b**: H, C, B, O; F-POC, **2c**: H, C, N, F) with sizes between 144 and 480 atoms. All structures are illustrated in Figure 4.1.

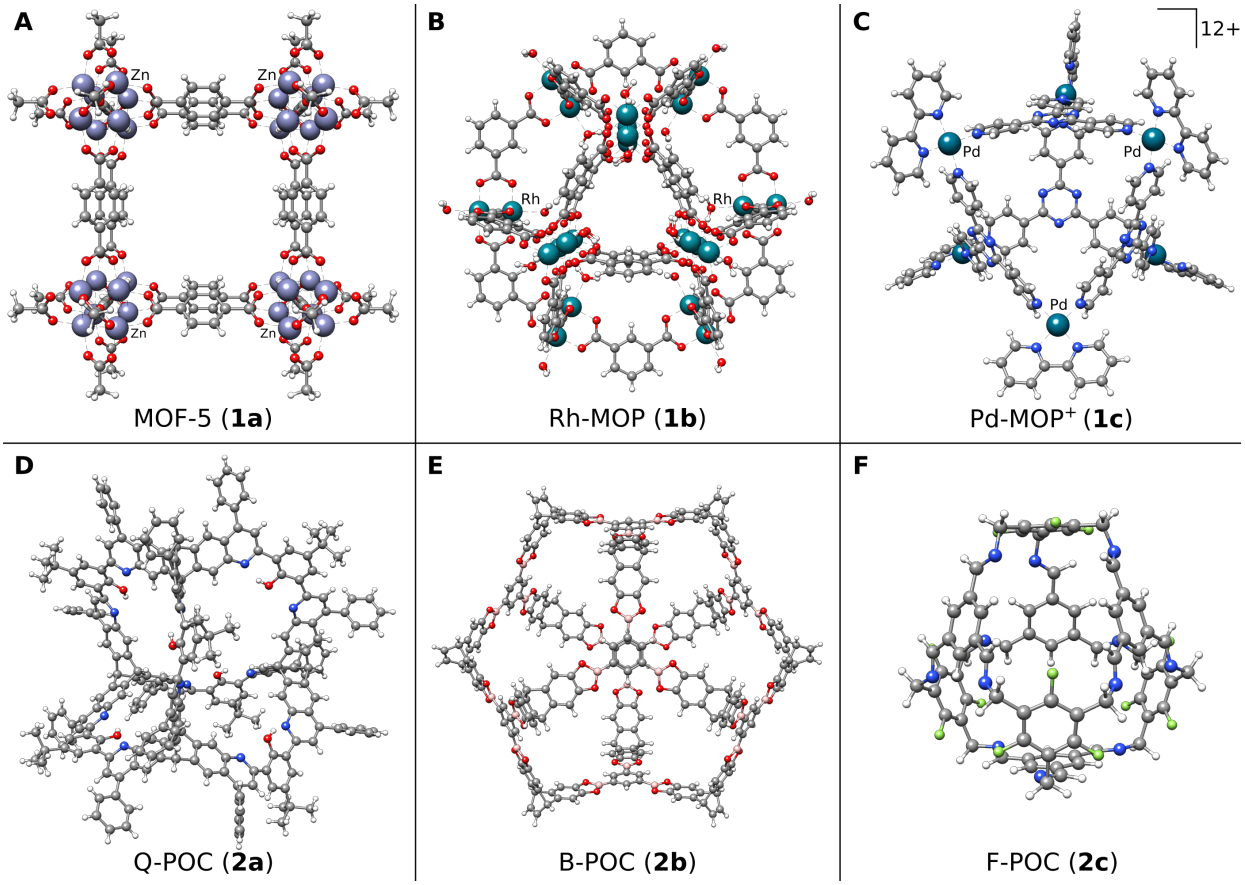


Figure 4.1.: Molecular structures of the MOFs, MOPs and POCs investigated in this work. (A) Molecular cutout of the cavity of MOF-5 (**1a**). (B) Rhodium containing metal-organic polyhedra (Rh-MOP, **1b**) saturated with water. (C) Positively charged ocatahedral Palladium containing MOP (Pd-MOP$^+$, **1c**). (D) A porous organic quinoline cage (Q-POC, **2a**). (E) Cuboctahedral [12+8] boron dotted cage (B-POC, **2b**). (F) A porous fluorinated organic [4+4] imine cage showing CO_2 and H_2 adsorption (F-POC, **2c**).

Figure 4.1A shows the molecular cutout of one cavity in the $Zn_4(O)(BDC)_3$, MOF-5, framework[329] (BDC=benzene-dicarboxylate). Eight clusters constitute a unit cell and enclose a large cavity. Figure 4.1B shows a rhodium-based metal-organic cuboctahedra, $[Rh_2(bdc)_2(solv)_2]_{12}\cdot(solv)$.[85] It is constructed from 12 units of binuclear paddlewheel complexes (M_2) with 24 dicarboxylate BDC bridging linkers. The metal centers are saturated with water molecules. The octahedral coordination cage in Figure 4.1C is self-assembled from four electron-deficient panel ligands (2,4,6-tris(4-pyridyl)-1,3,5-triazine) and six

metal corners (cis-endcapped Pd(II) complexes).[330] The overall net charge of the MOP accounts to +12 and it resembles an example for charged systems. In Figure 4.1D a robust porous organic quinoline cage is shown, transformed of a [4+6] salicylimine cage by Povarov cyclization.[331] In a one-step 48-fold condensation reaction of twelve molecules of triptycene tetraol with eight molecules of triboronic acid, the cuboctahedral [12+8] boron dotted cage in Figure 4.1E is formed.[332] Finally, Figure 4.1F shows a porous organic [4+4] imine cage containing perfluorinated aromatic panels.[333] The electronic structure of these systems is closed-shell with significant HOMO-LUMO gaps, *i.e.*, no to very little bi- or polyradical character is present. As guest molecules, CO_2, H_2, N_2, benzene (C_6H_6), and methanol (CH_3OH) are bound to the six hosts. Overall, 117 different binding sites of non-covalently bound host-guest complexes are evaluated. The mean binding energy at the reference DFT level is -6.1 kcal mol^{-1}.

4.3. Methods

A chemically reasonable binding site determination is the starting point for the successful calculation of host-guest binding energies. As geometry optimization algorithms may just find the next local minimum on the potential energy surface (PES), much more elaborate PES screening techniques are required to identify various possible binding sites. Recently, CREST has been introduced as an efficient automated scheme for the *in silico* sampling of the low-energy molecular chemical space by semiempirical tight-binding methods combined with the meta-dynamics (MTD) driven search algorithm iMTD.[220] In this work CREST is employed in combination with GFN-FF as the underlying level of theory. The computational efficiency of a FF method is mandatory in the context of (metal-)organic cages, due to the investigated system sizes of up to 500 atoms. The general workflow applied in this work is illustrated in Figure 4.2. The molecular input in Cartesian coordinates is generated starting from the crystal structure database[56] (CSD) entry. For MOFs, molecular pore cutouts are saturated by hydrogen atoms accordingly. Guest molecules are placed manually at the center of mass (COM) of the respective molecular pores/cages (host). To screen for different binding sites, the non-covalent interaction (NCI)/iMTD algorithm in CREST is employed. This is a special run-type in which a constraining ellipsoidal shaped potential is added in the MTD simulations. The additional potential prevents dissociation or transformation of the noncovalently interacting (NCI) complexes. Since possible binding sites of the guest molecule inside the host are in the focus, the biasing root-mean-square deviation (RMSD) potential in the MTD is only applied to the guest molecule throughout. The host is free of any constraints, allowing structural relaxation and adaption to the guest. From the CREST calculation, a structure ensemble of NCI complexes within a 6 kcal mol^{-1} energy window is obtained. The energetic ranking of such an ensemble is depicted in Figure 4.2. The selection of binding sites consists of the following steps. The energetically lowest conformation on the GFN-FF PES is always selected as the first binding site. From there on, the conformational energy ladder is ascended and further structurally different binding sites are determined manually, based on chemical intuition. All selected binding sites are post-processed by full geometry optimizations on different levels of theory including B97-3c as the reference structure level of theory. Note, that an accurate account of London dispersion corrections[334] in DFT as well as in the low-level methods is mandatory for an accurate description of the here

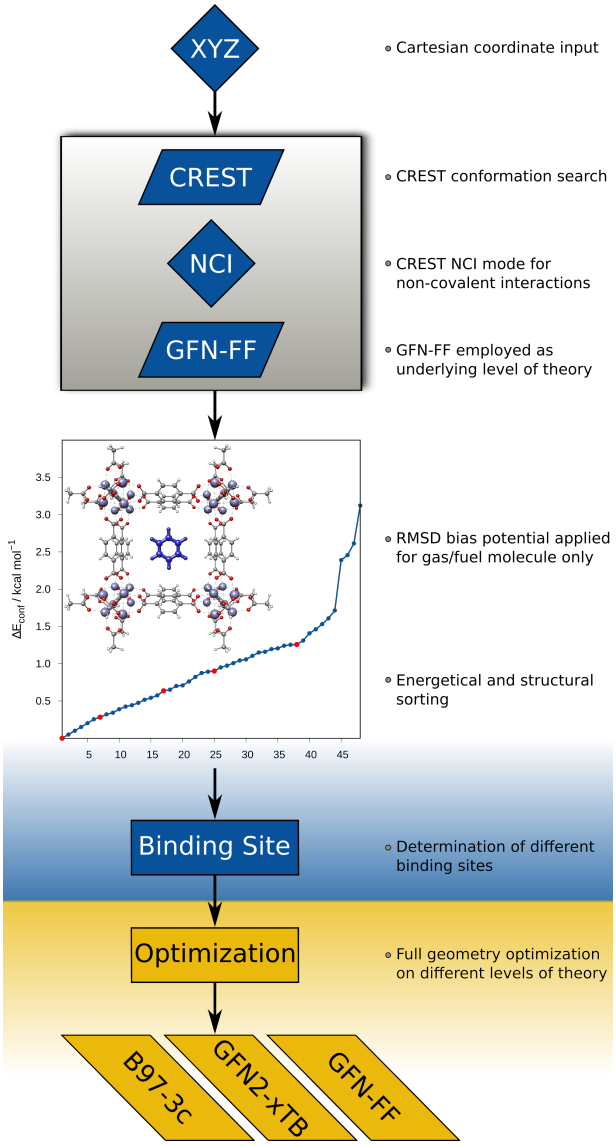


Figure 4.2.: Workflow for the calculation of optimal binding sites and binding energies for small molecules. Binding sites are determined by the NCI/iMTD algorithm in CREST employing GFN-FF as the underlying method. The resulting structures are post-processed on different levels of theory including B97-3c as the reference structure level.

considered NCIs and adsorption phenomena. This, in particular, holds for MOF/POC which contains highly polarizable atoms and functional groups. Without dispersion correction, *e.g.*, in DFT calculations, the binding energies decrease significantly and many of the investigated association complexes are not bound ($\Delta E_{\text{bind}} \geq 0$). This is reflected by the mean binding energy, which absolutely decreases from -6.1 kcal mol^{-1} (PBE0-D4/def2-TZVP) to 1.6 kcal mol^{-1} for PBE0/def2-TZVP without dispersion correction (at B97-3c optimized geometries).

4.3.1. Computational Details

All quantum mechanical calculations were performed with the TURBOMOLE 7.2.1 or 7.5[335,336] (DFT, UFF[114]), xtb 6.3.2[337] (GFN1-xTB,[94] GFN2-xTB,[93], GFN-FF[117]), and MOPAC 2016[338] (PM6-D3H4,[86] PM7[88]) program packages. The ANCOpt optimizer was applied for GFN*n*-xTB, GFN-FF and PM*x* optimizations as implemented in the xtb 6.3.2 program package for comparability with default convergence criteria 10^{-7} E_h for energies and 10^{-5} E_h/bohr for gradients. All reference binding energies were calculated applying the PBE0[325] hybrid functional with the def2-TZVP[326] basis set with default convergence criteria for energies and gradients as implemented in TURBOMOLE. The D4[242,249,339] London dispersion correction was applied throughout. Default effective core potentials (ECPs) ECP-28 and ECP-60[340] were used for all elements with atomic numbers larger than 36 (Kr) to take into account scalar relativistic effects in the DFT calculations. The resolution-of-identity (RI) approximation for the Coulomb integrals was generally used to speed up the DFT calculations by using matching default auxiliary basis sets.[341,342] For the integration of the exchange-correlation contribution, the numerical quadrature grid *m*4 was employed. For UFF the partial charges needed for the electrostatic terms were calculated with the Charge Equilibration model (QEq) from Rappé[343] as implemented in TURBOMOLE's version of UFF. All calculations were performed on Intel© Xeon E5-2660 v4 @ 2.00 GHz machines.

4.4. Results and Discussion

4.4.1. Structure Optimization

First, the following question is addressed: how accurate are the chosen low-cost methods for the description of the molecular structures of MOFs, MOPs, and POCs?[344] Therefore, B97-3c, GFN2-xTB, PM6-D3H4, UFF, and GFN-FF fully optimized gas phase geometries were compared to the initial molecular crystal structures cutouts (without guest molecules). As a measure for structural correlation, the heavy atom (excluding hydrogen) root-mean-square deviation (hRMSD[345] in Å) from either the X-ray structure cutout or the B97-3c optimized structure was calculated. The results are listed in Table 4.1 and illustrated in Figure 4.3A.

Table 4.1.: hRMSD between the molecular crystal structure cutout and the structures optimized with B97-3c, GFN2-xTB, PM6-D3H4, UFF, and GFN-FF. The CSD identifiers, if available, are given in bold font. The hRMSD (in Å) in comparison to the B97-3c optimized structure is given in parentheses.

hRMSD / Å	MOF-5 (**SAHYIK**)	Rh-MOP (**DALTES**)	Pd-MOP$^+$ (**HUMJIL**)	Q-POC (**WUTGOK**)	B-POC (**ZIRCIO**)	F-POC	Avg.
B97-3c	0.04	0.18	0.53	0.46	0.18	0.03	0.24
PM6-D3H4	0.07 (0.03)	0.54 (0.57)	0.51 (0.15)	0.59 (0.43)	0.26 (0.13)	0.62 (0.61)	0.43 (0.32)
GFN2-xTB	0.23 (0.27)	0.20 (0.05)	0.47 (0.08)	0.44 (0.16)	0.31 (0.18)	0.10 (0.11)	0.29 (0.14)
GFN-FF	0.08 (0.11)	0.33 (0.26)	0.52 (0.09)	0.41 (0.24)	0.24 (0.14)	0.50 (0.50)	0.35 (0.22)
UFF	2.19 (2.21)	0.64 (0.62)	1.34 (1.22)	0.48 (0.53)	0.46 (0.39)	0.44 (0.44)	0.93 (0.90)

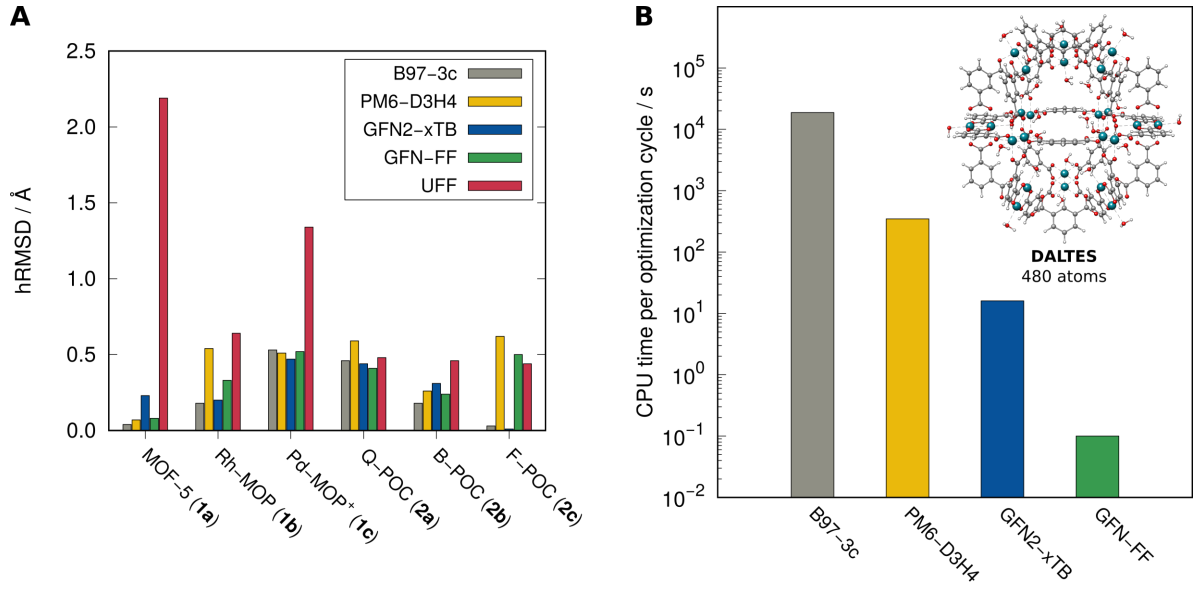


Figure 4.3.: (A) hRMSD between the molecular crystal structure cutout and the structures optimized with B97-3c, GFN2-xTB, PM6-D3H4, UFF, and GFN-FF. (B) CPU time per optimization cycle in seconds on one Intel© Xeon E5-2660 v4 @ 2.00 GHz CPU for B97-3c, PM6-D3H4, GFN2-xTB, and GFN-FF using the example of **1b**.

On average, all methods, except UFF, produce reasonable hRMSD of less than 0.5 Å. For the comparison of molecular crystal structure cutouts and gas phase optimized structures, and considering the neglect of secondary solvent and crystal packing effects, the agreement between theory and experiment is remarkable. The largest deviations occur for Pd-MOP$^+$ **1c**, which is to be expected, due to the high net charge and the absence of counter ions in our treatment. For the fluorinated POC **2c**, PM6-D3H4 shows a noticeable deviation, which is mainly due to wrong torsion angles at the imine bonds. Also, the saturated Rh-MOP **1b** shows larger deviations from the crystal structure cutout, arising from the re-organization of the water ligands at the metal centers. B97-3c produces good structures, showing the lowest average RMSD of 0.24 Å. It is noticeable that going from a force-field method to much more sophisticated SQM methods and even beyond high-level DFT calculations, the changes in accuracy are rather small. For example, GFN-FF, which is 5-6 orders of magnitude faster than B97-3c, is capable of providing structures at (almost) DFT quality. Computational timings for a single geometry optimization cycle of each respective method for **1b** illustrate the distinct differences in computational cost (Figure 4.3B). The good structural agreement with experimental and full DFT optimized structures justifies the usage of GFN-FF in combination with CREST for the generation of host–guest binding complexes as binding site candidates, which is computationally not feasible at any other theoretical level (except for GFN1-, GFN2-xTB but at much higher computational cost).

4.4.2. Binding Sites and Energies

Combining six hosts and five guests at three to five different binding sites leads to a benchmark set of 117 structures in total. Binding energies ΔE_{bind} in kcal mol^{-1} are calculated as the differences in

total energies E between the fully optimized complex and the respective hosts and guests (Eq. 4.1, *i.e.*, the geometry relaxation induced by the binding is included. The calculation of Gibbs free energies by adding standard thermostatistical contributions is easily possible and has recently been considered in detail[346] but is not pursued here.

$$\Delta E_{\text{bind}} = E(\text{Complex}) - E(\text{Host}) - E(\text{Guest}) \tag{4.1}$$

According to our years of experience with the calculation of NCIs,[120,162,334,347] the residual basis set superposition error for the large AO basis sets used is small (< 5% of ΔE_{bind}) or even practically negligible by construction of the methods (for PBEh-3c, B97-3c and GFN methods). The performance of all tested methods is evaluated in terms of the mean absolute deviation (MAD) from the reference PBE0-D4/def2-TZVP//B97-3c values. MADs are calculated separately for each molecular host (**1a**–**2c**). As the geometry optimization of **2a** at the DFT level showed convergence difficulties due to the rotatable tert-butyl and phenyl-residues, for the energy study a truncated version of **2a** was used, where those residues were replaced by hydrogen atoms. The correlation of binding energies calculated by the tested low-cost methods and the reference is further evaluated by the Pearson correlation coefficient r_p. Conformational ordering is evaluated by the Spearman rank correlation coefficient r_s. For r_p and r_s values close to unity are desired, indicating high correlation and parallelism of the corresponding PESs. Combined MAD_{mean}, $r_{p,\text{mean}}$, and $r_{s,\text{mean}}$ values are calculated as the arithmetic mean for all six host systems. The results for the entire test set are illustrated in Figure 4.4. MAD values for each

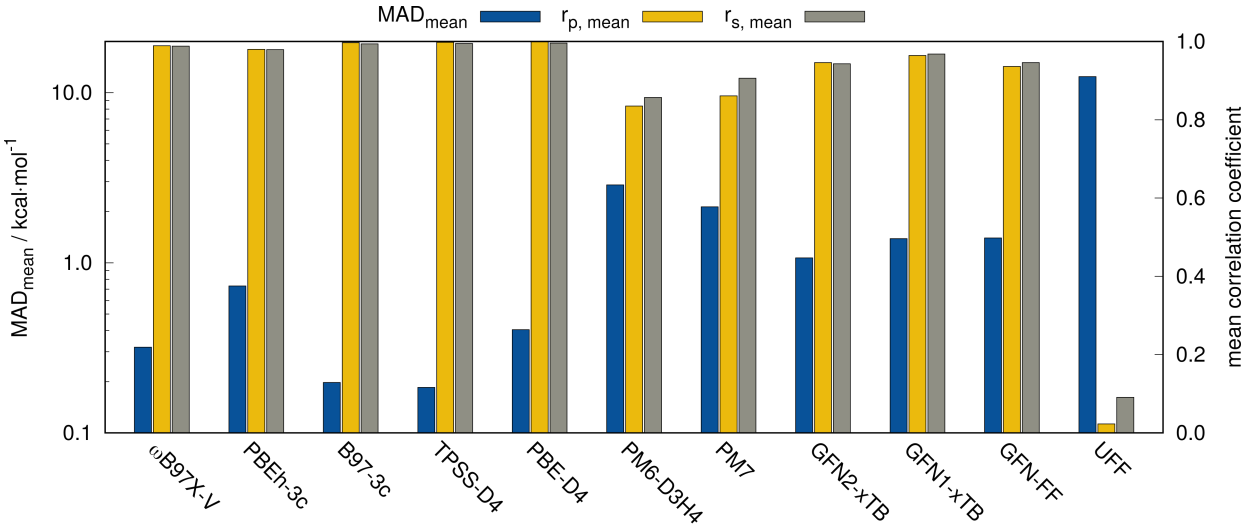


Figure 4.4.: Statistical evaluation of MAD_{mean}, $r_{p,\text{mean}}$, and $r_{s,\text{mean}}$ coefficients with respect to PBE0-D4/def2-TZVP reference values. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

system and the averaged mean values (MAD_{mean}) are additionally listed in Table 4.2. The binding energy level diagrams for each host-guest combination are shown in Appendix A4. Employing different DFAs serves as a cross-check for the chosen PBE0-D4/def2-TZVP reference. Except for PBEh-3c, all

DFT methods behave as expected and follow Jacob's ladder[159,163] scheme for DFA classification. B97-3c and TPSS-D4/def2-TZVP perform almost equally good with a MAD_{mean} of 0.2 kcal mol^{-1} compared to the reference, followed by PBE-D4/def2-TZVP with a MAD_{mean} of 0.4 kcal mol^{-1}. The range-separate hybrid ωB97X-V is the only applied functional with nonlocal correlation (VV10[348]), whereas all other DFAs are combined with the semi-classical DFT-D correction.[334] With a MAD_{mean} of 0.3 kcal mol^{-1}, the performance is similar to the best performing (meta-)GGAs with reference to PBE0-D4. For the accurate description of NCI, a sufficiently large basis set is mandatory. The built-in double-ζ basis set in PBEh-3c may be too small compared to the triple-ζ basis employed for the other density functionals which probably explains the MAD_{mean} of 0.7 kcal mol^{-1}. To validate this assumption, PBE-D4/def2-SVP single-point calculations were additionally performed for the entire test set, yielding a MAD_{mean} value of 2.2 kcal mol^{-1}. This is an increase in the mean absolute deviation of 1.8 kcal mol^{-1} compared to the triple-ζ basis set calculation (MAD_{mean}(PBE-D4/def2-TZVP) = 0.4 kcal mol^{-1}). All DFAs yield high correlation coefficients r_p and r_s, with excellent values $\geq$ 0.98 throughout the entire test set (detailed r_p and r_s values are given in Appendix A4).

Table 4.2.: MAD between PBE0-D4/def2-TZVP (reference) and all other tested methods.

MAD / kcal mol^{-1}	MOF-5 (**1a**)	Rh-MOP (**1b**)	Pd-MOP$^+$ (**1c**)	Q-POC (**2a**)	B-POC (**2b**)	F-POC (**2c**)	MAD_{mean}
PBEh-3c	0.75	1.02	0.74	0.44	0.56	0.88	0.73
B97-3c	0.16	0.25	0.15	0.17	0.23	0.24	0.20
TPSS-D4	0.16	0.27	0.14	0.19	0.14	0.22	0.19
PBE-D4	0.22	1.01	0.21	0.33	0.29	0.36	0.40
PM6-D3H4	0.85	10.05	1.31	0.78	0.96	3.27	2.87
PM7	1.05	2.13	2.20	1.26	0.94	5.12	2.31
GFN2-xTB	0.75	1.47	1.64	0.99	0.44	1.13	1.07
GFN1-xTB	0.84	2.79	0.77	1.26	1.20	1.48	1.39
GFN-FF	0.77	2.13	1.07	1.30	0.77	2.36	1.40
UFF	12.16	15.21	3.23	32.32	2.78	8.79	12.4

In the SQM class of methods, both GFN*n*-xTB variants perform only slightly worse compared to the DFT methods, yielding small MAD_{mean} of 1.1 kcal mol^{-1} and 1.4 kcal mol^{-1} for GFN2-xTB and GFN1-xTB, respectively (see Figure 4.4). This is further indicated by the high correlation coefficients around 0.95 clearly outperforming the tested PM*x* methods, which yield worse MAD_{mean} of 2.9 and 2.3 kcal mol^{-1} for PM6-D3H4 and PM7, respectively. PM6-D3H4 specifically fails for **1b**, strongly overestimating the binding energies for benzene and CO_2 (see Table 4.2). The GFN-FF force-field reaches almost the quality of GFN1-xTB with a MAD_{mean} value of 1.4 kcal mol^{-1} of and mean correlation coefficients around 0.95, whilst being two orders of magnitude faster. A correlation diagram comparing all GFN methods to the DFT reference is shown in Figure 4.5. The two outliers observed for GFN2-xTB are H_2 and N_2 coordinating to the Pd^{2+} metal center of the charged Pd-MOP$^+$ (**1c**). With GFN2-xTB this interaction is overestimated compared to GFN1-xTB and GFN-FF whereby the differences arise from the electrostatic energy term. UFF is included in this study as the only available generally applicable FF alternative to GFN-FF. UFF provides a large MAD_{mean} value >10 kcal mol^{-1} and correlation coefficients

<0.1. Thus, regarding its poor performance, UFF should be avoided for such applications.

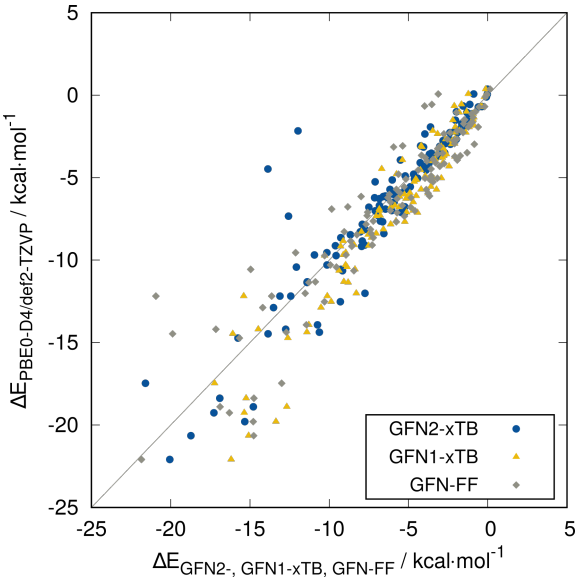


Figure 4.5.: Correlation plots for binding energies obtained with GFN2-xTB, GFN1-xTB, and GFN-FF with reference to the PBE0-D4/def2-TZVP energies. All binding energies are calculated on B97-3c geometries.

The accurate calculation of binding properties requires chemically reasonable binding motifs. Hence, the binding site generation is a major part of this and related application works. Therefore, the quality of the fully automated binding structure ensembles, which were used for all previous calculations, was exemplary tested. Optimal binding sites of CO_2 molecules in MOF-5 have previously been investigated by ONIOM (MP2/6-31G**:HF/6-31G**) calculations.[349] Pianwanit *et al.* used clusters of four metal corners and three connecting linkers as a model for MOF-5. Guest molecules were manually assigned to lie parallel and perpendicular to linker and corner domains of each cluster. The optimal binding site and orientation of CO_2 in the cavity of MOF-5 were predicted to be perpendicular to the metal corner at 5 Å distance with corresponding binding energy of −2.2 kcal mol^{-1}. The CO_2@MOF-5 system is used to compare the automatically generated complexes with previous results. Contrary to Ref. 349, all binding sites in this work were obtained fully automatically by the conformational search procedure in CREST and the whole unit cell was taken into account (see Figure 4.1A), rather than a truncated cluster model. Binding energies were calculated on the fully B97-3c optimized structures and larger triple-ζ basis sets were employed for all other DFT calculations in contrast to the double-ζ basis used in Ref. 349. The results for CO_2@MOF-5 are depicted in Figure 4.6.

With our fully automated approach, the same binding sites as in Ref. 349 are found and the same trends for binding energies are obtained. The two orientations of CO_2 perpendicular to corner and linker mark the two weakest bound motifs. The two most strongly bound motifs show CO_2 orientated parallel to the corner. We obtain distances of 4.5 Å between the carbon atom in CO_2 and the central oxygen atom of the respective MOF-5 corner. Binding energies range from −5 to −6 kcal mol^{-1} on the PBE0-D4/def2-TZVP level of theory. The absolute larger binding energies compared to Ref. 349

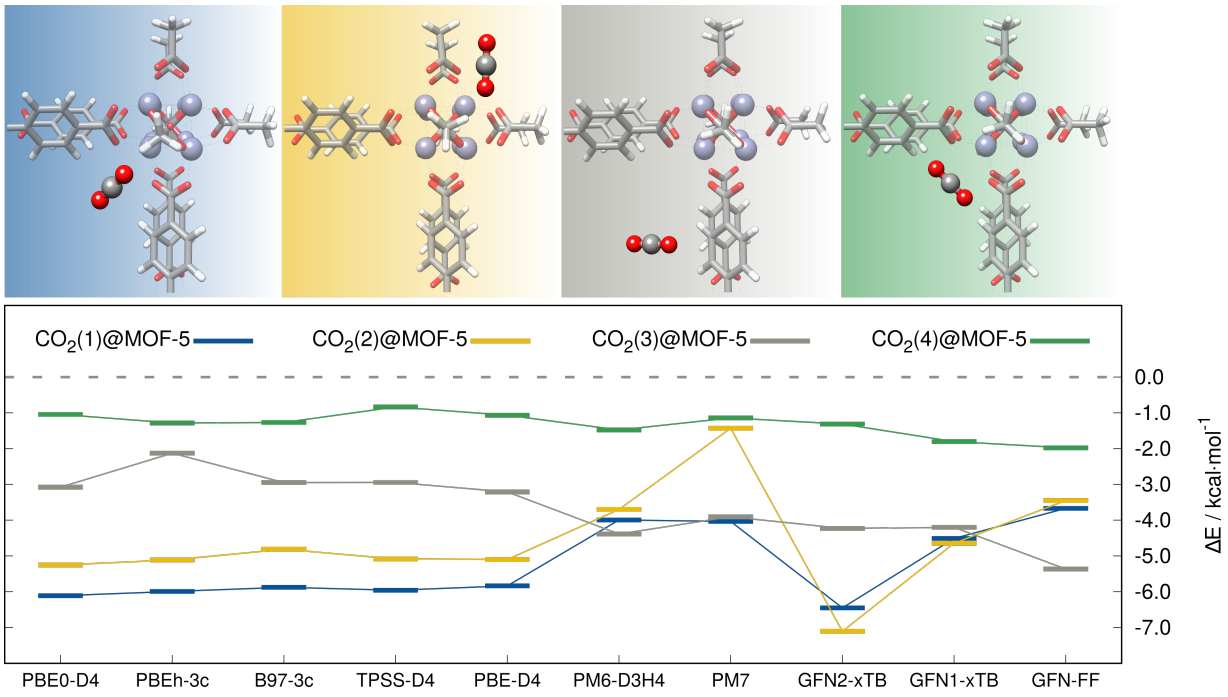


Figure 4.6.: Binding motifs and binding energies for CO_2@MOF-5 (MOF-5). On top, the four different binding motifs (1–4) of CO_2@MOF-5 are shown. The ordering corresponds to the respective PBE0-D4/def2-TZVP energies. For all DFAs, beside the composite methods, a def2-TZVP basis set is employed. Colored backgrounds indicate the matching structural binding motifs.

may result from the larger triple-ζ basis set and the underlying fully relaxed DFT geometries applied in this study. All tested DFAs reproduce the energetic ordering of the binding motifs correctly, while the GFN*n*-xTB methods overestimate the binding energy in motif 2 (yellow), which is on the other hand underestimated by the PM*x* methods and GFN-FF. GFN-FF further overestimates the binding energy in motif 3 (grey). Motif 4 (green) is correctly identified as the weakest one by all methods. Overall, although none of the SQM/FF methods is able to predict the correct energetic ordering in all cases, qualitatively reasonable orderings and absolute interaction energies are obtained at very low computational effort.

4.4.3. Structure Energy Correlation

A common practice in theoretical studies is a so-called multi-level approach (see *e.g.*, Ref. 120) in which structures are optimized on a more efficient, lower theoretical level compared to the final single-point energy (SPE) computation. Looking at Figure 4.3, the comparison of CPU times per optimization cycle, reveals the urgency for this approach. While B97-3c geometry optimizations take up to five hours per optimization step for the largest structures of the test set, GFN2-xTB and GFN-FF require just a few seconds. To validate this approach, all structures obtained from the CREST conformational search were additionally optimized on the GFN2-xTB and GFN-FF level of theory. Based on GFN2-xTB and GFN-FF structures, binding energies were calculated by the same methods as before. In the follow-

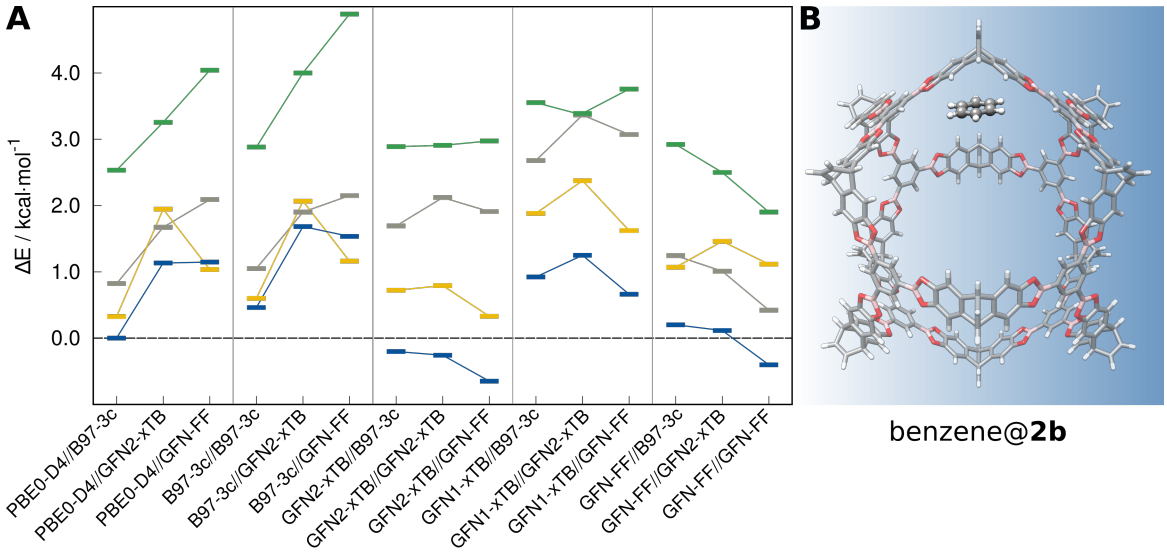


Figure 4.7.: **A** Energy level diagram of the four conformations of benzene@**2b**. Binding energies are shown for PBE0-D4/def2-TZVP, B97-3c, GFN1-xTB, GFN2-xTB, and GFN-FF calculated on DFT, SQM and FF optimized structures. **B** Energetically lowest binding site of benzene on **2b** at the PBE0-D4/def2-TZVP level of theory.

ing, binding energies of each method calculated on the SQM or FF structure (method//GFN2-xTB, method//GFN-FF) are compared with the binding energy obtained by the same method on the DFT structure (method//B97-3c). To ensure a meaningful comparison, it must be guaranteed that the optimized motifs do not change significantly upon geometry optimizations by the different methods. B-POC (**2b**) fulfills these criteria for all 23 different binding motifs and is thus well suited for this comparison. The correlation of binding energies and the underlying structures is explicitly shown for the benzene guest in **2b**. Figure 4.7 depicts the energy level diagram of all four motifs found for benzene@**2b** (Figure 4.7A) and the B97-3c structure of benzene(1)@**2b** (Figure 4.7B). Three connected energy levels belong to the binding energy calculated by the same method, but on differently optimized structures. In this comparison, PBE0-D4 and B97-3c are chosen as representative DFT methods. It is noticeable that all DFT binding energies calculated on SQM or FF structures decrease absolutely compared to those calculated on the DFT reference structure. This has an influence on the ordering of the conformational energies and some re-ranking is observed. For GFN*n*-xTB and GFN-FF methods, this effect is less pronounced, and moving from a DFT to a force-field structure causes only small changes. Hence, GFN1-xTB, GFN2-xTB, and GFN-FF binding energies are quite similar on all tested geometries and the energetic ordering remains unchanged compared to those computed on B97-3c structures. These findings also hold for the other guest molecules bound to B-POC. The results for all 23 complexes are quantified in Figure 4.8. Detailed values are listed in Appendix A4. The MAD_{mean} for DFT methods calculated on GFN structures (DFT/GFN-xTB and DFT/GFN-FF)) ranges from 1.0 to 1.5 kcal mol^{-1} with reference to the same DFT method calculated on the B97-3c optimized structures (DFT//B97-3c). This deviation is significant in comparison to the mean reference binding energy of −6.2 kcal mol^{-1} and is also reflected in smaller correlation coefficients, with values of about 0.90 for r_p and r_s, respectively. For the tested GFN methods

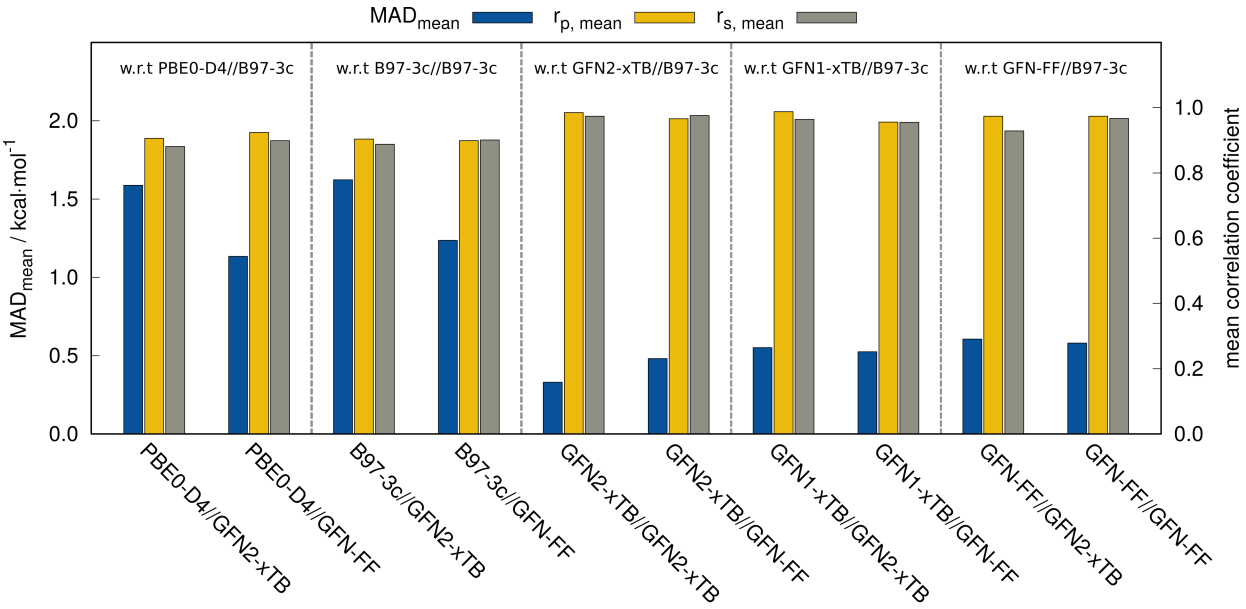


Figure 4.8.: MAD_{mean}, $r_{p,mean}$, and $r_{s,mean}$ calculated for 23 different guest binding motifs in **2b** on geometries optimized at different theoretical levels. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

on GFN structures (GFN//GFN) the binding energy deviations from the respective GFN//B97-3c level amount to an average of just 0.5 kcal mol^{-1} and all correlation coefficients lie above 0.95. Some caution is therefore required in multi-level approaches where DFT energies are calculated on GFN structures, not only because binding energies decrease, but especially because conformational ranks may change. Nevertheless, GFN2-xTB energies calculated on GFN-FF structures, as it would be done in applications beyond the 1000 atom regime, seem to yield rather good results, indicating higher parallelism of the corresponding PES.

4.4.4. Highly Charged Species

To increase the diversity in terms of strength and nature for the interactions considered, binding energies are further investigated for anionic guests bound to cationic hosts. Such systems are specifically interesting in the context of anion templated synthesis.[350] Therefore, binding positions were assessed for the sulfate (SO_4^{2-}) and tetrafluoroborate (BF_4^-) within a Zn variant of a cationic urea-functionalized M_4L_6 host[351] (**3a**) shown in Figure 4.9. The strong electrostatic interactions in the gas phase lead to high mean interaction energies of −675 kcal mol^{-1} for the sulfate anion and −269 kcal mol^{-1} for the tetrafluoroborate anion, respectively. Thus it is convenient to also consider the relative MAD_{mean} within the statistical evaluation for better comparison to previous results. For the calculation of binding energies, the same methods as before are applied. The results in comparison to the PBE0-D4/def2-TZVP reference are given in Figure 4.10. Reference binding energies are best reproduced by the (meta-)GGA functionals TPSS and PBE deviating by only 1.3 and 1.1 kcal mol^{-1} respectively, which is a relative error of just 0.2 %. The composite low-cost DFT composite methods PBEh-3c and B97-3c perform slightly worse

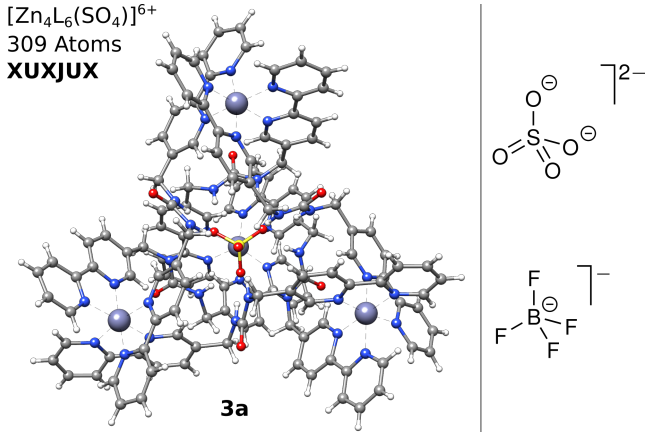


Figure 4.9.: Sulfate (SO_4^{2-}) and tetrafluoroborate (BF_4^-) anions binding to $[Zn_4L_6]^{8+}$ forming, *i.e.*, complex $[Zn_4L_6(SO_4)]^{6+}$ **3a**. **XUXJUX** was modified, replacing Ni by Zn to avoid SCF convergence issues in the DFT calculations.

with deviations of 3.7 to 6.6 kcal mol^{-1} (1.0-1.4 %). The range-separate hybrid ωB97X-V with non-local correlation yields comparable deviations of 1.4 % with respect to the PBE0 reference interaction energies. In the SQM class of methods, GFN1-xTB performs only slightly worse than the DFT methods, yielding small MAD_{mean} of 6.3 kcal mol^{-1} (1.4 %). PM6-D3H4 and GFN2-xTB follow with deviations of 10.1 and 13.8 kcal mol^{-1}, which represent errors of 2.7 and 3.5 % respectively. PM7 fails for the highly charged systems considered here, underestimating the binding energies for BF_4^- and SO_4^{2-} by more than 37 %. GFN-FF applied with default settings strongly overestimates the interaction energy by 95 % due to the artificial charge transfer of the EEQ model. Thus, employing charge constraints for the molecular fragments (host and guest) either manually or via GFN*n*-xTB charges is recommended. This reduces the MAD_{mean} value drastically to 8.9 kcal mol^{-1} (relative MAD_{mean} = 2.5 %), whilst remaining two orders of magnitude faster than all SQM methods. This approach was also recommended in the context of supramolecular association reactions and protein optimizations in Ref. 117. UFF, which applies an unconstrained charge equilibration model, deviates by more than 60 % from the reference values and should thus be avoided for highly charged systems.

4.4.5. Drug Binding

The molecular guests considered so far were limited to small organic molecules only. As an outlook and future perspective, the complexity is increased by a transition metal complex as a guest. Ruthenium arene PTA (RAPTA) represents a class of experimental cancer drugs,[352,353] consisting of a central ruthenium(II) ion coordinated by an arene ligand, chlorides, and 1,3,5-triaza-7-phosphaadamantane (PTA). One derivative is [Ru(η^6-p-cymene)Cl_2(PTA)] (RATPA-C) (Figure 4.11A). RAPTA-C was bound to the POCs **2a** and **2b**, focusing on the most favorable binding site at the GFN-FF level. Therefore, the same procedure as before was applied, but only the energetically lowest CREST conformation was post-processed by DFT. B97-3c optimized structures of RAPTA-C@**2a** and RAPTA-C@**2b** are shown in Figure 4.11B and 4.11C. Binding energies were calculated by the same DFT, SQM, and FF methods as before.

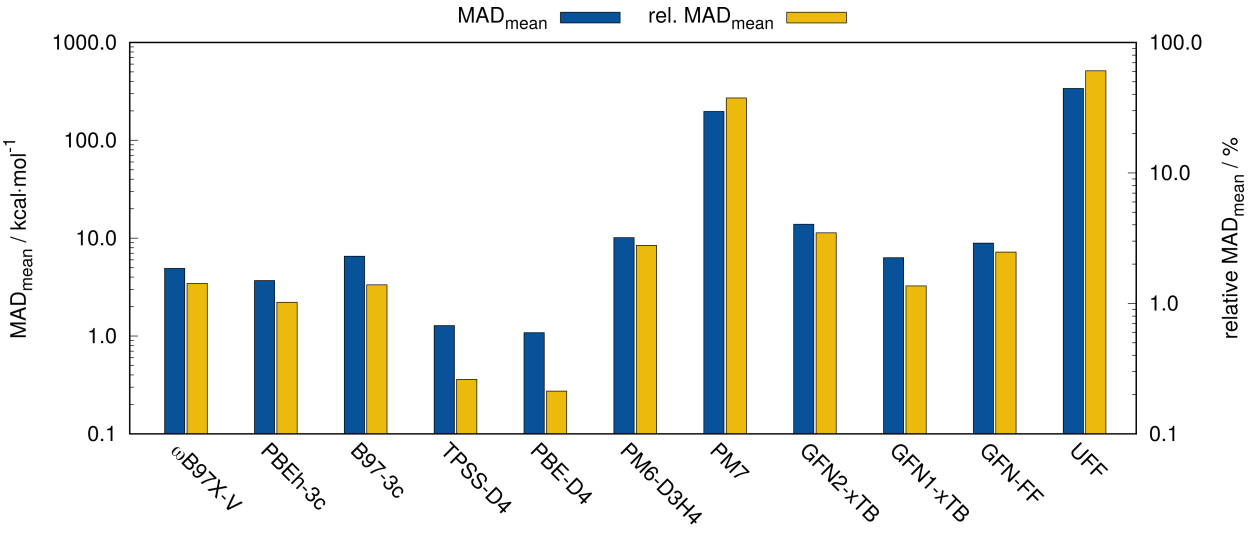


Figure 4.10.: Statistical evaluation of the MAD_{mean} and relative MAD_{mean} with respect to PBE0-D4/def2-TZVP reference values for BF_4^- and SO_4^{2-} bound to **3a**. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

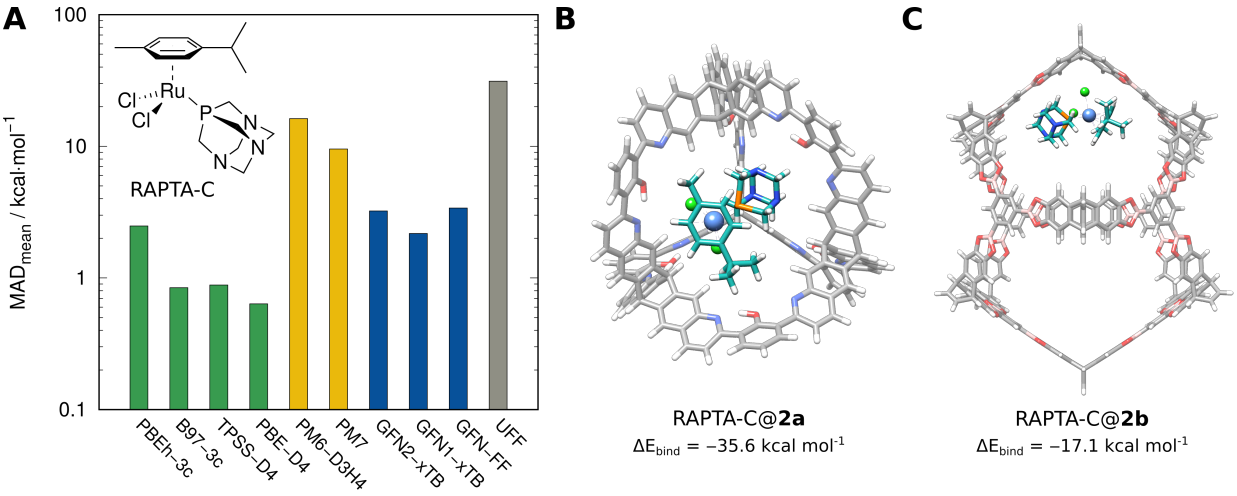


Figure 4.11.: **A** MAD_{mean} for the binding energies of RAPTA-C to **2a** and **2b** with reference to PBE0-D4/def2-TZVP reference data. **B** B97-3c optimized structure of RAPTA-C bound to **2a**. **C** B97-3c optimized structure of RAPTA-C bound to **2b**. Also, the respective binding energies on the PBE0-D4/def2-TZVP level are given.

PBE0-D4/def2-TZVP reference energies for RAPTA-C@**2a** and RAPTA-C@**2b** amount to −35.6 and −17.1 kcal mol^{-1} respectively. Although the deviations are generally larger, the results shown in Figure 4.11A (note the logarithmic scale) are comparable to those discussed *vide supra*. All (meta-)GGA functionals reproduce the hybrid reference values within a deviation of less than 0.9 kcal mol^{-1}. Detailed values are given in Appendix A4. As noted before, PBEh-3c again does not achieve this accuracy. GFN2-xTB and GFN-FF show an acceptable degree of accuracy with MAD_{mean} values of about 3 kcal mol^{-1}, whereas GFN1-xTB is the best performing GFN method with a deviation of just 2.2 kcal mol^{-1}. As for

the other systems, neither of the PM*x* methods nor UFF are capable of providing an adequate description of transition metal complexes bound to porous organic cages.

4.5. Conclusion

In this work, small molecule binding in metal-organic constructs and porous organic cages by DFT and low-cost (mainly quantum mechanical) methods was investigated. An efficient, fully automated procedure for the generation of preferred binding sites was shown, combining the conformer search algorithm CREST with GFN-FF. The obtained binding motifs at MOPs, MOFs, and POCs were further optimized on the B97-3c DFT level of theory and reference binding energies were calculated at the PBE0-D4/def2-TZVP hybrid level for 117 different host-guest complexes.

The resulting binding motifs were compared to those from previous theoretical studies for CO_2@MOF5 and a good mutual agreement was observed. For the entire benchmark set, the accuracy for binding energies and structures calculated by SQM and FF methods was evaluated compared to DFT results. Of all SQM approaches tested, the GFN methods performed well, yielding small deviations in the range from 1.1 to 1.4 kcal mol^{-1} throughout. The GFN-FF force-field yields particularly good results, regarding its very low computation times compared to the SQM methods. PM6-D3H4 and PM7 worked well in some cases but showed a large deviation for others. With MAD_{mean} values of 2.3 to 2.9 kcal mol^{-1}, the accuracy is clearly lower than with GFN methods. UFF on the other hand, which was tested as a second general FF, was not able to adequately describe any of the occurring noncovalent binding motifs.

In the context of non-covalent binding to MOFs/MOPs and POCs a substitution of both, the energy calculation and the geometry optimization, by any of the GFN*n*-xTB methods is justified if the corresponding geometry optimization does not indicate significant basic issues in reproducing the electronic structure of the investigated host-guest complex. If the energy calculation is performed with a DFT method, non-parallel PES between DFT and the SQM/FF methods applied for the geometry optimization may also increase the deviations in binding energy compared to the fully DFT calculated values. We recommend to investigate this effect on a case-by-case basis. The complexity of the systems was increased by binding two anions (BF_4^- and SO_4^{2-}) and the ruthenium containing cancer drug RAPTA-C on different cages, where only GFN methods produced results comparable to the DFT reference. Even though this work focuses on gas phase binding energies, the performance and computational efficiency of the GFN methods also enables the computation of Gibbs free association energies as recently shown in Ref. 346, which is not practicable at the DFT level of theory due to the high computational cost. Such calculations, also including solvation effects by the available continuum models, open up completely new possibilities for the comparison between theory and experiment in this field of supramolecular chemistry.

Acknowledgments

This work was supported by the DFG in the framework of the priority program 1807 “Control of London dispersion interactions in molecular chemistry”. Sebastian Spicher thanks the “Fond der Chemischen Industrie (FCI)” for financial support. The authors thank A. Hansen and P. Pracht for fruitful discussions.

5. Modeling of Spin–Spin Distance Distributions for Nitroxide Labeled Biomacromolecules

Sebastian Spicher,[a] Dinar Abdullin,[b] Stefan Grimme,[a] and Olav Schiemann[b]

Received: 17 September 2020, Published online: 21 October 2020

Selected as 2020 PCCP Hot Article

Reprinted (adapted) with permission[†] from
Spicher, S.; Abdullin, D.; Grimme, S.; Schiemann, O. *Phys. Chem. Chem. Phys.* **2020**, *22*, 24282–24290.

DOI 10.1039/D0CP04920D

Own manuscript contribution

- Conception of the study and choice of methods
- Developing the workflow
- Performing all the calculations
- Interpretation of the results
- Writing the manuscript

[a]Mulliken Center for Theoretical Chemistry, Institute of Physical and Theoretical Chemistry, University of Bonn, Beringstraße 4, 53115 Bonn, Germany
[b]Institute of Physical and Theoretical Chemistry, University of Bonn, Wegelerstr. 12, 53115 Bonn, Germany
[†]Permission requests to reuse material from this chapter should be directed to the Royal Society of Chemistry.

Abstract Electron Paramagnetic Resonance (EPR) spectroscopy is a powerful method for unraveling the structures and dynamics of biomolecules. Out of the EPR toolbox, Pulsed Electron–Electron Double Resonance spectroscopy (PELDOR or DEER) enables one to resolve such structures by providing distances between spin centers on the nanometer scale. Most commonly, both spin centers are spin-labels or one is a spin-label and the other is a paramagnetic metal ion, cluster, amino acid, or cofactor radical. Often, the translation of the measured distances into structures is complicated by the long and flexible linker connecting the spin center of the spin-label with the biomolecule. Nowadays, this challenge is overcome by computational methods but the currently available approaches have a rather large mean error of roughly 2-3 Å. Here, the new GFN-FF general force-field is combined with the fully automated Conformer-Rotamer Ensemble Sample Tool (CREST) [Pracht *et al., Phys. Chem. Chem. Phys.*, 2020, **22**, 7169-7192] to generate conformer ensembles of the R1 side chain (methanthiosulfonate spin-label (MTSL) covalently bound to a cysteine) in several cysteine mutants of azurin and T4 lysozyme. In order to determine the Cu^{2+}-R1 and R1-R1 distance distributions, GFN-FF based MD simulations were carried out starting from the most probable R1 conformers found by CREST. The deviation between theory and experiment in mean inter-spin distances was 0.98 Å on average for the mutants of azurin (1.84 Å for T4 lysozyme) and the right modality was obtained. The error of the most probable distances for each mode was only 0.76 Å in the case of azurin. This CREST/MD procedure does thus enable precise distance-to-structure translations and provides a means to disentangle label from protein conformers.

5.1. Introduction

Knowledge of the structure and dynamics of biomacromolecules is important for an understanding of their function.[354] In this regard techniques such as X-ray crystallography,[355] cryogenic electron microscopy (cryo-EM),[356–358] Förster resonance energy transfer (FRET),[359–361] nuclear magnetic resonance (NMR) spectroscopy,[362,363] and electron paramagnetic resonance (EPR) spectroscopy[364] have delivered important insights. The latter one offers in the form of Pulsed Electron–Electron Double Resonance spectroscopy (PELDOR or DEER)[365,366] a method that provides distances between spin centers in the range of 1.6 to 16 nm. Such spin centers can be paramagnetic metal ions[367] or clusters,[368] as well as amino acid or cofactor radicals.[369] If diamagnetic, spin centers can be introduced by means of site-directed spin-labeling (SDSL) with nitroxides.[370,371] The most widely used spin-label is the methanethiosulfonate spin-label (MTSL),[372] which binds covalently to cysteines, forming the so-called R1 side chain (Figure 5.1). In this side chain, the electron spin is localized on the N-O group of the nitroxide and the PELDOR experiment provides the distance to this spin center in form of distance distributions. However, for the structure of the biomolecule, the distance to the C_α atom, which the R1 side chain is connected to, is of interest. Thus, in order to translate the measured spin-spin distance into a structure, the intrinsic flexibility and length of the R1 side chain have to be taken into account. Yet, there is a non-trivial correlation between the C_α-C_α and the spin-spin distances, which depend on the conformational ensemble of the R1 side chain in its macromolecular environment.

In recent years, the field of theoretical chemistry has evolved into an indispensable part of natural science and is proven to be an important companion to the experiment.[1] Theory is able to provide

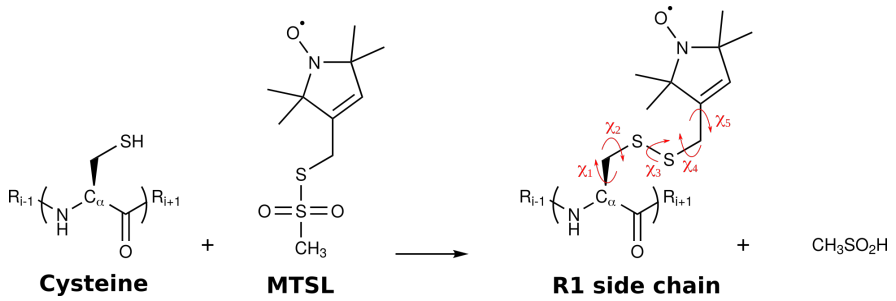


Figure 5.1.: The chemical reaction of a cysteine residue and MTSL forming the so-called R1 side chain. The spin is localized to more than 95 % between the N and O atoms. The five dihedral angles χ_1-χ_5 corresponding to the rotatable bonds of R1 are shown by red arrows.

accurate equilibrium structures by geometry optimizations as well as atomistically resolved classical motions by molecular dynamics (MD) simulations.[373] It is thus well suited for applications in structural biology.[374] However, the sheer size of typical biomacromolecules consisting of thousands of atoms excludes basically all *ab initio* quantum mechanical (QM) methods and also most of the semiempirical QM methods as potential candidates for the computation of dynamics for the entire system.[97,375–377] Therefore, atomistic force-fields (FF) are established as the standard level of theory in the description of biomacromolecules[239,378–380] and used routinely by common programs such as Amber,[62] CHARMM,[235] or OPLS.[113] However, specialized biomolecular FFs are not suited for interdisciplinary use, because their parameters only exist for a limited amount of chemical elements and structural motifs. For highly coordinated metal centers, which are present in metalloproteins, laborious "hand-made" parameterization is often required.[267] For consistent routine application, this approach is not practical and, hence, experimentalists are in urgent need of reliable routine theoretical support.[381] For this reason, it is often necessary to use much simpler methods such as mtsslWizard,[382,383] PRONOX,[384] Rosetta,[385] or MMM[386,387] that are based on an accessible volume or an energy-weighted library approach, respectively. Alternatively, MD-based approaches that employ simplified FFs, like the one used in the MDDS approach by Islam *et. al.*,[388] have been proposed and were shown to be slightly better than the more simple methods above when used to predict inter-spin distance distributions.

Recently, a completely automated, partially polarizable generic FF has been developed for the accurate description of geometries, frequencies and non-covalent interactions (GFN-FF),[117,346] covering a full periodic table parameterization ($Z \leq 86$). This universal method combines the high computational speed of FFs with an accuracy approaching that of low-level QM methods. Furthermore, a new conformer-rotamer ensemble sample tool (CREST)[60] has recently been published, which is an efficient scheme for the sampling of the low-energy chemical space by GFN methods[96] combined with a metadynamics (MTD)[220] driven search algorithm. All these computational methods are coded efficiently and are freely available for general use.[337]

In this work, we propose an automated composite scheme for the calculation of inter-spin distance distributions of MTSL spin-labeled proteins. For the first time, the combination of CREST and MD, both employing GFN-FF as the underlying theoretical model, is applied to large biomolecules. This CREST/MD approach is advantageous compared to conventional MD based treatments of the R1 con-

formers.[389–394] Purely MD based methods are problematic for the prediction of the conformer ensemble, because conformational barriers may not be overcome in unbiased simulations at ambient temperature. This problem was further addressed by Sezer *et al.*[395,396] The potential of GFN-FF based MD simulations for the prediction of distance distributions between paramagnetic metal ions and spin-labels has been briefly described in the original publication.[117] Here, the more thorough CREST/MD approach is described and extensively tested for several single-labeled azurin mutants and double-labeled T4 lysozyme (T4L) mutants (Figure 5.2).

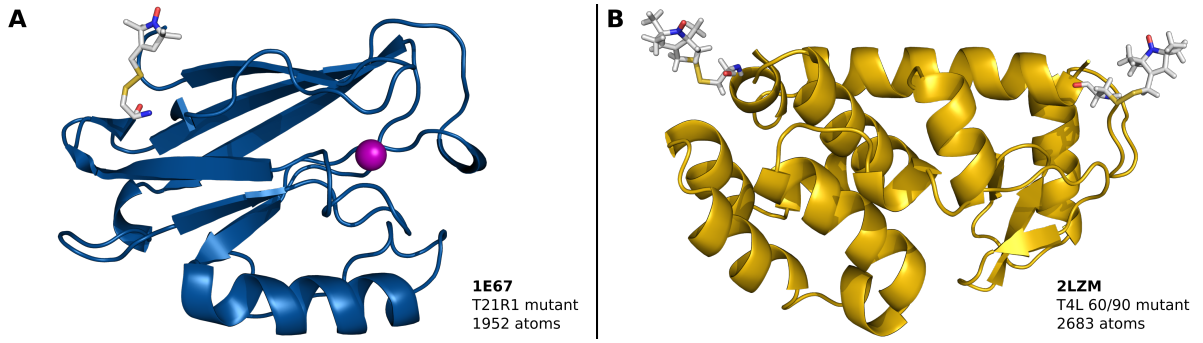


Figure 5.2.: Graphic illustration of the two investigated biomacromolecules. (A) Crystal structure of azurin with MTSL attached at position T21. (B) Crystal structure of T4 lysozyme with two MTSL attached at positions 60 and 90. The PDB identifiers are given in bold.

5.2. Method

The general workflow from the input structure to the final distance distribution, including CREST and MD at the GFN-FF level of theory, is illustrated in Fig 5.3. The main idea is to perform a constrained conformational search, followed by completely unconstrained independent MD simulations for different conformers to obtain a thermally equilibrated (classically averaged) structure. The inter-spin distance distributions are then calculated from the MD trajectory. The procedure can be divided into four steps

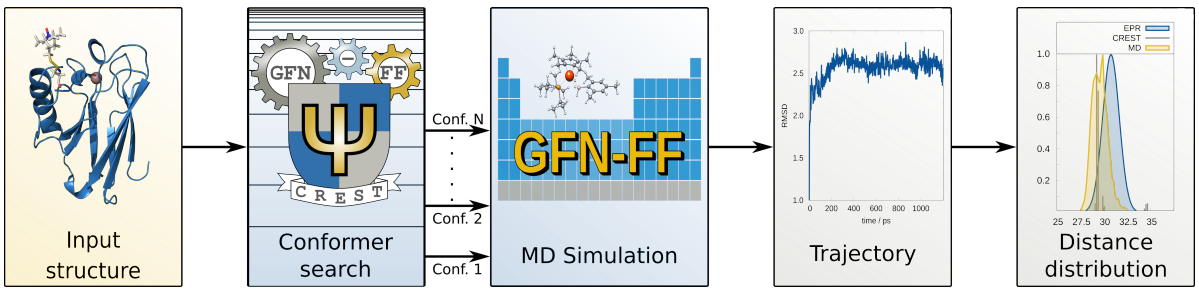


Figure 5.3.: The proposed general scheme for the calculation of spin-label distance distributions. Conformational analysis is performed on the input structure by the CREST algorithm, employing GFN-FF for the underlying potential energy surface. MD simulations on the same level of theory are performed on the highly populated conformers. The resulting trajectories are evaluated to obtain the distance distribution function per conformer. Their Boltzmann weighted sum yields the overall distance distribution.

employing four different programs which are all non-commercial and freely available:

1) Input structure preparation: *e.g.*, *mtsslWizard*[382,383]
2) CREST conformer search: *crest*[60]
3) GFN-FF MD simulation: *xtb*[337]
4) Trajectroy analysis: *TRAVIS*[397]

First, the input structure was generated from the PDB entry of the protein by attaching a conformer of the R1 side chain at each position needed. Various alternative computer methods for this step are available,[55,382,383,385,386,398,399] here mtsslWizard was used. Hydrogen atoms were added automatically. This pre-processed input structure resembles a random conformation of the spin-label in the protein.

In the second step, low-energy spin-label conformations were generated via CREST at the GFN-FF level of theory, employing an implicit GBSA(H_2O)[96] solvation model. A full CREST conformational search for the entire proteins is not feasible, even at the GFN-FF level of theory. Therefore, it was assumed that the protein backbone is only weakly affected by the presence of the label.[381] The main focus is on the conformations of the R1 side chain, but the protein core is allowed to relax, *i.e.*, it is *not* geometrically frozen. In the CREST calculation, a biasing potential was added for all atoms of the spin-label. More specifically, the 31 atoms (62 atoms for T4 lysozyme) of the R1 side chains were included in the RMSD criteria of the MTD simulations. See the original Ref. 220 for smaller examples of this selective heating procedure. An additional constraint via harmonic pairwise potentials was applied for all other atoms of the protein structure. Therefore, a weak constraining force constant of 0.01 E_h/bohr was chosen. This reduction of the number of active degrees of freedom is necessary to decrease the size of the conformation space and, consequently to make the CREST run practical. In general, a weak constraining force is favorable, being physically closer to the real system and allowing structural relaxation of the protein upon the conformational change induced by the spin-label. On the downside, this leads to an increased computation time. The resulting fully optimized conformer ensemble of the R1 side chain attached to the protein consisted of 20 to 200 structures. The distance distribution of an ensemble was obtained as the sum of the inter-spin distances from all conformers weighted by their Boltzmann factor based on GFN-FF relative energies at the respective temperature. If multiple conformers were found, which showed a very similar distance between the spin centers (<1 Å), they were considered as one R1 conformer cluster. For more details about the conformer ensemble see Appendix A5.

In the third step, unconstrained GFN-FF MD simulations of the full system were performed for all conformers that had a relative Boltzmann weight larger than 50 % compared to the energetically lowest conformer (whose population is set to 100 %). In addition, only the conformer with the highest Boltzmann weight in a R1 conformer cluster was taken as the starting point for the MD simulation. The threshold of 0.5 was found to yield good results in comparison to the experiment at reasonable computational costs. Yet, this is an empirical parameter, which might have to be varied for other systems. Again, an implicit GBSA(H_2O) solvation model was applied. Subsequent MD simulations were performed in order to reach thermal equilibrium for the unconstrained system. Note, that the MD run

was not applied to overcome barriers and find structurally new conformations since the conformational space exploration step has already been conducted by CREST. The GFN-FF MD simulation adds temperature effects to the optimized equilibrium structures and removes all applied constraints introduced above. For this purpose, a simulation time of 1 ns after an equilibration phase of 200 ps was found to be sufficient. Shorter and longer simulation times were tested but did not lead to significant changes in the distance distributions, for more details see Appendix A5. Nevertheless, for systems with flexible domains longer simulation times might be necessary. The applied temperature in the MD was set to the freezing point of the solvent because it was assumed, that the conformational ensemble examined in the PELDOR experiment corresponds to the one conserved at the freezing point of the solvent. The effect of the glass transition temperature on the outcome of the MD simulations will be a point for future investigations, as it amounts to roughly 2/3 of the freezing point of the solvent.[400,401] Detailed investigation of temperature effects is given below which sheds some light on the reliability of this scenario.

The last step in this protocol was the analysis and evaluation of the obtained MD trajectories. The trajectory analyzer and visualizer *TRAVIS*[397] was chosen to calculate the radial distribution functions. The overall distance distribution was calculated as the Boltzmann weighted sum of distributions obtained from the MD trajectories of each evaluated conformer.

5.3. Results and Discussion

5.3.1. Azurin

First, we investigated six spin-labeled mutants of the metallo-protein azurin,[402–404] namely D69R1, S100R1, T21R1, T30R1, T61R1, and T91R1. Their synthesis and spin-labeling as well as the corresponding PELDOR experiments are described in Ref. 403. As a starting structure of azurin, its Zn-bound crystal structure (PDB ID 1E67) was taken due to the higher resolution as compared to Cu-bound azurin. Here, GFN-FF can show its full strength, as it is able to describe both metals without any additional parameterization. After full geometry optimization, the heavy atom root mean square deviation (hRMSD) between copper and zinc azurin was only 0.1 Å, which is negligible (see Appendix A5 for more details). For the azurin mutants, a water-ethylene glycol mixture (50 v/v %) was used as a buffer,[403] which has a freezing temperature of roughly 237 K. Thus, all MD simulations were performed at this temperature. The calculated Cu^{2+}-NO distance distributions of the six mutants are shown in Figure 5.4. Note that the distances measured in the PELDOR experiment are the distances between the spin densities ρ distributed over several atoms (Figure 5.4, top). For the R1 side chain, the spin density is equally distributed between the nitrogen and oxygen atoms ($\rho_N = \rho_O = 0.5$). Yet for the metal center, only 35 % of the spin density is on the Cu^{2+} ion, whereas 60 % are on the neighboring sulfur atom[403] of residue C112. This distribution of spin densities was taken into account in the evaluation of the MD trajectories (see Appendix A5).

To quantify the results, the obtained mean distance $\bar{r}$, the most probable distance r_p, and the distributions width by standard deviations σ_r (Table 5.1) were compared to the experimental distances

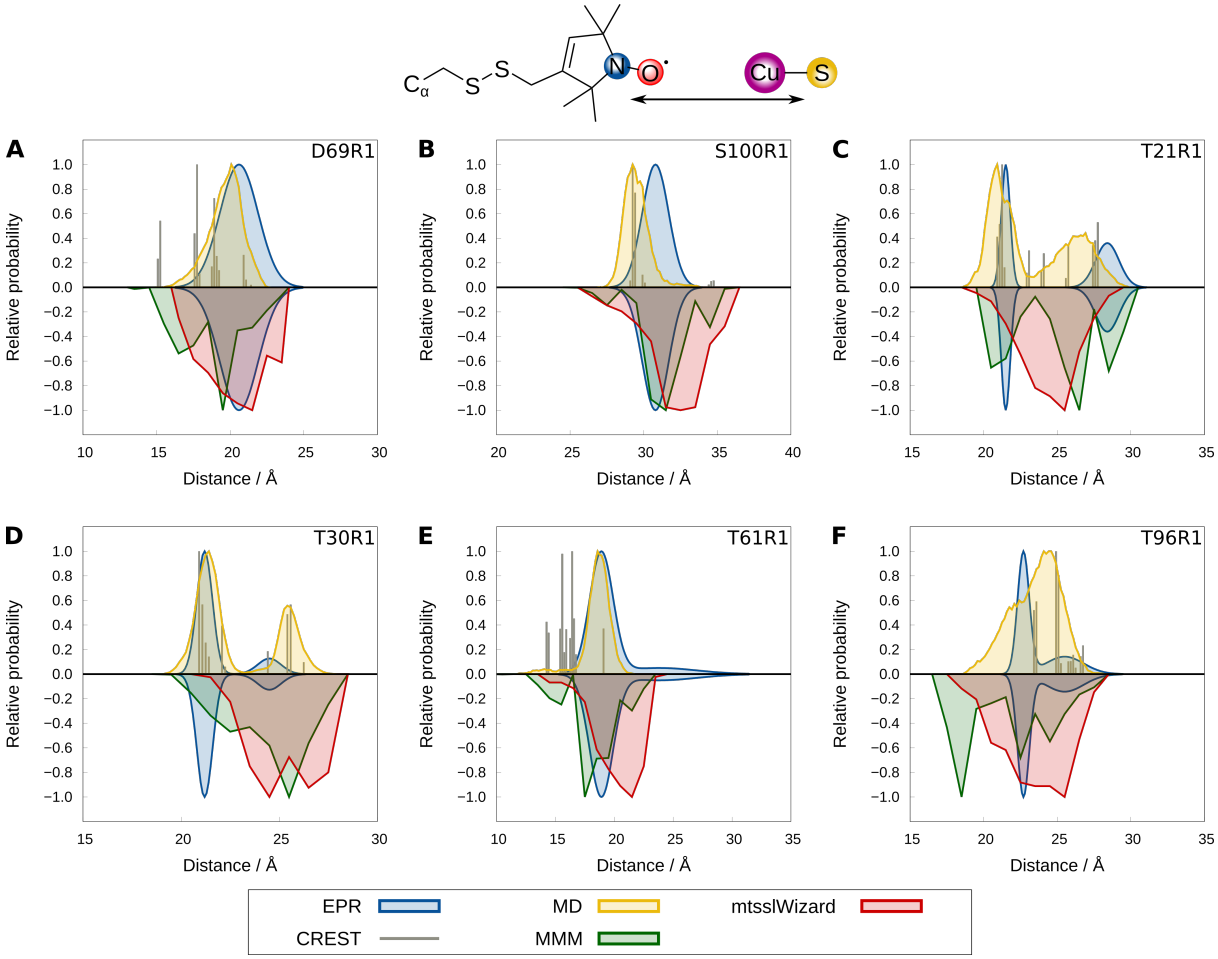


Figure 5.4.: Spin-label distance distributions for the six azurin mutants D69R1, S100R1, T21R1, T30R1, T61R1, and T96R1 (A-F). Top half of each plot: The results from the CREST conformations without any MD averaging (gray) overlayed with the distribution obtained by the GFN-FF MD simulation (yellow) and the experimental PELDOR data (blue). Bottom half of each plot: MMM (green) and mtsslWizard (red) distance distributions and the experimental PELDOR data (blue). The spin centers between which the distances were measured are illustrated on top.

in terms of mean absolute deviation (MAD), standard deviation (SD), and relative MAD (*rel*MAD). For all six mutants, the agreement between the CREST/MD approach and the experiment for the most probable distances r_p is remarkable with a MAD of 0.83 Å, SD of 1.14 Å and a relative error of 3.5 %. On the other hand mtsslWizard and MMM show MAD values ranging from 2.55 to 2.78 Å, which are relative deviations of more than 10 %. For the mean distances $\bar{r}$, almost equally good results are obtained for CREST/MD with a MAD of 0.97 Å, SD of 0.92 Å and a relative error of 4.3 %. mtsslWizard and MMM yield MAD values of 1.03 and 1.47 Å respectively. These relative deviations of 4–7 % for $\bar{r}$ are significantly smaller than the errors in r_p. The distribution widths σ_r of the experimental distance distributions were best reproduced by the CREST/MD approach with a MAD of just 0.29 Å, being on average only 12.2 % broader. MtsslWizard and MMM show on average 35–45 % broader distributions with absolute deviations of 1.19 and 1.72 Å respectively. In addition to the correspondence of

Table 5.1.: Statistical evaluation (in Å) of the mean spin-label distances $\bar{r}$, the most probable distances r_p, and the standard deviation σ_r for all distance distributions obtained *in silico* with CREST/MD, MMM, mtsslWizard and by PELDOR (Exp.) for the six azurin mutants.

Mutant	Exp.			CREST/MD			MMM			mtsslWizard		
	$\bar{r}$	r_p	σ_r	$\bar{r}$	r_p	σ_r	$\bar{r}$	r_p	σ_r	$\bar{r}$	r_p	σ_r
D69R1	20.6	20.6	2.0	19.7	20.1	2.1	19.0	19.5	3.3	19.9	21.5	2.5
S100R1	30.8	30.8	1.4	29.5	29.3	1.5	32.6	32.6	3.9	31.2	31.2	3.6
T21R1	24.8	21.5	2.8	23.4	20.9	3.1	23.9	26.5	3.6	24.2	25.5	3.6
T30R1	21.7	21.2	1.7	22.7	21.4	2.3	24.9	26.5	3.3	23.7	24.5	2.7
T61R1	19.6	18.9	2.6	18.5	18.6	2.4	19.5	18.5	4.5	19.9	20.5	3.6
T96R1	23.6	22.7	1.9	23.5	24.6	2.5	24.2	26.5	4.2	23.7	25.5	3.6
MAD				0.97	0.83	0.29	1.47	2.78	1.72	1.03	2.55	1.19
SD				0.92	1.14	0.30	1.84	3.55	0.62	1.51	1.11	0.63
*rel*MAD				4.3 %	3.5 %	12.2 %	6.7 %	12.4 %	44.8 %	4.1 %	10.3 %	35.5 %

the mean and most probable distances, the determination of the correct number of modes (modality) is of importance. Here, it should be highlighted that the presented CREST/MD procedure was the only theoretical model that predicted the multi-modal distance distributions correctly. MMM also yielded multi-modal distributions but differed significantly from the experiment. No multi-modality was observed for mtsslWizard. Thus, the good agreement of mtsslWizard with the experiment in terms of $\bar{r}$ loses its value, as the modality was not reproduced.

As the CREST/MD procedure was able to reproduce the experimentally found modality, it was further possible to compare the most probable distances per mode. For MMM and mtsslWizard this evaluation is not possible, due to missing or wrong modalities. The results are listed in Table 5.2. For the most probable distances r_p per mode, the deviation between the CREST/MD approach and the experiment is very small with a MAD of 0.76 Å, SD of 0.78 Å and a relative error of 3.0 %.

Table 5.2.: Statistical evaluation (in Å) of the most probable spin-label distances r_p obtained *in silico* with CREST/MD and by EPR for each mode of the five azurin mutants.

Mutant	Exp. r_p	CREST/MD r_p
D69R1	20.6	20.1
S100R1	30.8	29.3
T21R1	21.5	20.9
	28.4	26.9
T30R1	21.2	21.4
	24.5	25.4
T61R1	18.9	18.6
T96R1	22.7	22.4
	25.5	24.6
MAD		0.76
SD		0.78
*rel*MAD		3.0 %

For mutants T21R1 and T30R1, the importance of multiple MD simulations starting from different conformers for obtaining the experimentally observed bimodal distributions was revealed. While in both cases, the energetically lowest conformer yields the main peak at shorter distance, the second peak at longer distances is generated only by the second most probable conformation. For the chosen MD conditions, the energy barrier can not be overcome and, hence, the correct distribution is only obtained by using starting conformations from different R1 conformer clusters. For the mutants D69R1, S100R1, and T61R1, the picture is different. In all three cases, a similar mean distance is obtained from all starting conformations, implying low energy barriers between the conformers. The individual distributions obtained for each conformer are shown in Appendix A5. Mutant T96R1 is the only example where the relative probabilities of the R1 conformations are mismatched, indicating wrong conformational energies at the GFN-FF level. At this point, it should be mentioned that the performance of the whole procedure depends heavily on the accuracy of GFN-FF/GBSA for the potential energy surface and, in particular, the conformational energies. Even though this was tested thoroughly in Ref. 117 for small to medium sized systems, the extension to large biomolecules can not be assumed *per se*.

5.3.2. Temperature Effects

As mentioned above, all MD simulations were performed at the freezing point of the solvent, 237 K (freezing point of water with 50 v/v % ethylene glycol). To validate this approximation, temperature effects were investigated for the bimodal distributions of T21R1 and T30R1. Therefore, MD simulations were performed from 237 K to 298 K. Distance distributions obtained for all temperatures are given in Appendix A5. In Figure 5.5, the distributions are shown for the freezing point of the solvent (237 K) and at room temperature (298 K). In Figure 5.5A, two conformations that were found *in silico* are exemplary

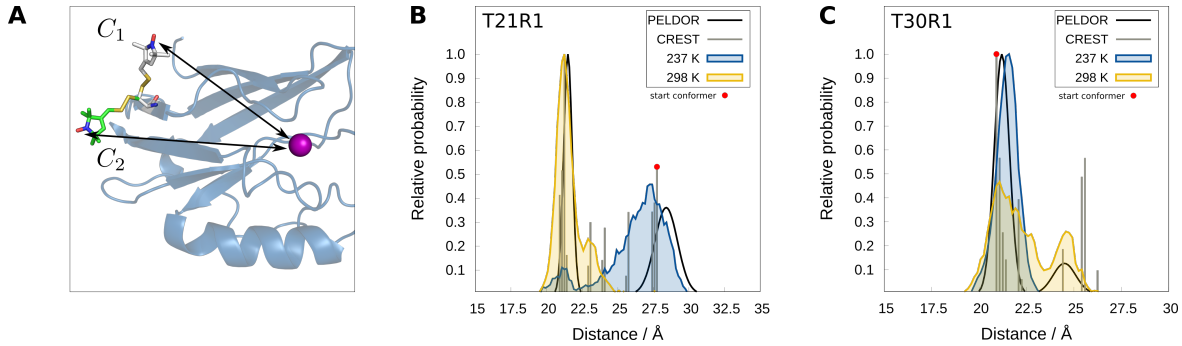


Figure 5.5.: Investigation of temperature effects in the MD simulation. In (A), two conformations with shorter (C_1) and longer (C_2) inter-spin distances of T21R1 are illustrated. Computations are performed at the freezing point of the solvent (237 K) and at room temperature (298 K), respectively, for the azurin mutants T21R1 [(B) and T30R1 (C)]. The starting conformation for the MD simulation is indicated by a red dot.

shown for T21R1. In the following, the conformation with the shorter inter-spin distance is called C_1, while the conformation with the larger inter-spin distance is called C_2. The main conformation C_1 of the R1 side chain in T21R1 is independent of temperature (see Appendix A5), indicating that the relevant conformational barriers are rather high (about 5–10 kcal mol^{-1}). However, when starting from

the C_2 conformation (Figure 5.5B, indicated by the red dot), significant changes were observed as the temperature was increased. The low-temperature MD indicates a small, but visible relative population of C_1. For the high-temperature MD, C_2 completely vanishes and converts to C_1. This highlights the importance of the chosen temperature for the MD simulation step. Importantly, the experimentally observed bimodality of T21R1 is obtained theoretically only for simulations close to the freezing point of the solvent. For T30R1 (Figure 5.5C), temperature effects were investigated as well. Here, the C1 conformer was again associated with the shorter Cu^{2+}-R1 distances and the C2 conformer with the longer distances. When starting the MD simulations from the conformation C1, only the C1 conformation is obtained at low temperature (237 K), whereas the conformation C1 partially converts into conformation C2 at 298 K. When compared to the PELDOR distance distribution, a slight overpopulation of the conformation C2 is observed, indicating a small error in the compared GFN-FF conformational energies. The conformation C_2, on the other hand, is independent of temperature (see Appendix A5).

5.3.3. T4 Lysozyme

The second investigated system consists of 19 mutants of T4 lysozyme.[405,406] Since there is no metal center present in T4L, two MTSL labels were attached in order to perform PELDOR measurements. Computing the distance distributions between two flexible R1 side chains rather than between one label and a fixed metal ion makes the conformational search considerably more complicated. The combined CREST/MD protocol was applied to all 19 structures. For the MD simulations, the temperature was set to 273 K (water with 20 v/v % glycerol). PELDOR distance distributions were taken from Ref. 405, where mean distances $\bar{r}$ for all mutants are tabulated. The *in silico* computed results are shown in Figure 5.6. The calculated values for $\bar{r}$, r_p, and σ_r are listed in Table 5.3.

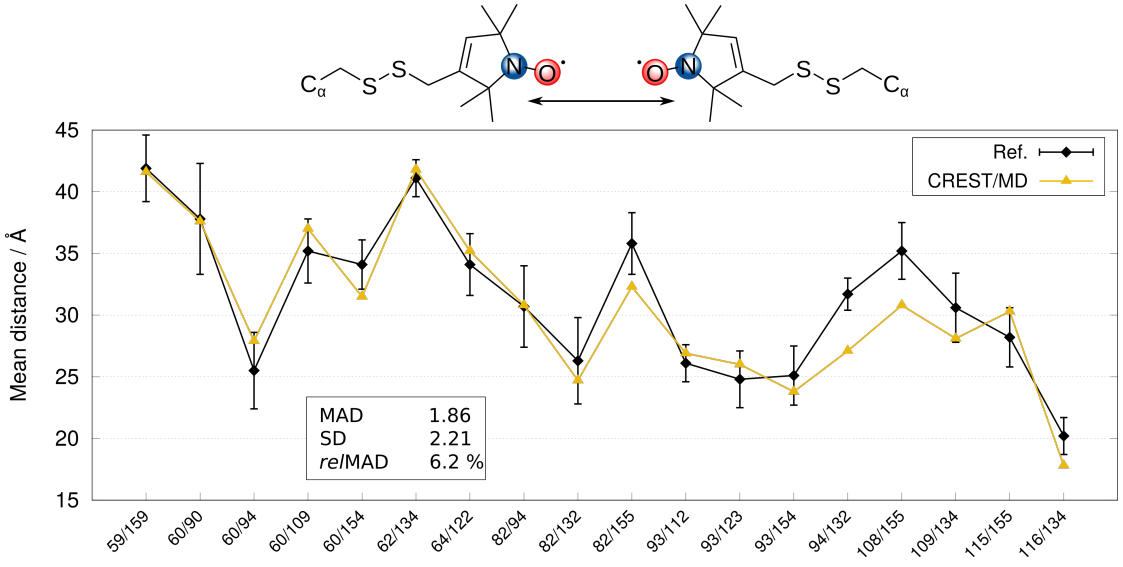


Figure 5.6.: Comparison of theory and experiment for mean spin-label distances of 19 T4L mutants. The experimental values are given in black and the distribution width is indicated by error bars. Mean distances obtained from GFN-FF MD simulations are shown in yellow. An illustration of the nitroxide-nitroxide distance is given on top.

Table 5.3.: Statistical evaluation (in Å) of the mean spin-label distances $\bar{r}$, the most probable distances r_p, and the standard deviation σ_r for all distance distributions obtained *in silico* with CREST/MD, MMM, mtsslWizard and by PELDOR (Exp.) for the 18 T4L mutants.

Mutant	Exp.			CREST/MD			MMM			mtsslWizard			MDDS		
	$\bar{r}$	r_p	σ_r	$\bar{r}$	r_p	σ_r	$\bar{r}$	r_p	σ_r	$\bar{r}$	r_p	σ_r	$\bar{r}$	r_p	σ_r
59/159	41.9	42.4	4.1	41.6	40.8	4.0	39.1	40.5	5.4	42.1	43.5	4.8	39.2	41.0	7.1
60/90	37.8	38.5	7.2	37.6	37.1	5.0	44.1	45.0	5.1	45.5	46.5	4.8	40.2	40.0	5.3
60/94	25.5	26.2	4.6	27.9	27.3	3.6	29.8	29.5	7.3	32.1	32.5	5.9	26.6	28.0	7.4
60/109	35.2	33.2	3.3	37.0	35.8	3.4	37.0	38.5	6.8	38.3	39.5	5.3	31.2	35.0	7.4
60/154	34.1	34.9	2.8	31.5	30.5	2.5	36.0	36.5	7.0	38.6	39.5	5.1	34.9	35.0	5.6
62/134	41.1	40.6	2.0	41.8	41.4	1.8	45.8	47.0	3.4	48.5	48.5	2.2	42.2	43.0	5.1
64/122	34.1	32.5	2.5	35.2	35.1	3.5	35.9	36.0	6.0	38.3	38.5	3.9	32.6	33.0	5.6
82/94	30.7	32.1	4.7	30.8	31.3	2.4	30.2	29.5	5.1	32.2	33.5	4.8	29.9	30.0	5.9
82/132	26.3	29.1	5.2	24.7	25.6	4.2	28.0	28.5	6.0	26.9	28.5	5.1	24.7	25.0	6.5
82/155	35.8	38.5	3.7	32.3	34.4	3.7	35.8	37.0	4.5	37.2	38.5	5.3	35.4	35.0	5.3
93/112	26.1	26.2	1.7	26.9	26.8	3.8	28.4	26.5	4.7	32.3	33.5	3.9	25.4	25.0	5.4
93/123	24.8	24.6	2.6	26.0	25.9	2.0	24.8	25.0	5.6	25.4	26.5	5.9	28.2	28.0	4.5
93/154	25.1	27.5	2.7	23.8	23.9	2.2	26.0	27.0	4.7	25.6	26.5	4.8	27.2	28.0	3.6
94/132	31.7	32.4	1.6	27.1	27.1	1.9	31.7	33.5	4.4	32.0	32.5	2.7	30.0	31.0	4.8
108/155	35.2	36.2	3.3	30.8	30.9	2.5	35.8	35.0	3.7	36.8	36.5	3.0	32.5	34.0	4.7
109/134	30.6	32.3	2.8	28.1	29.3	3.2	29.9	30.0	4.4	31.0	32.5	4.5	27.9	30.0	5.6
115/155	28.2	27.6	2.9	30.3	29.1	4.4	30.9	31.0	4.1	32.4	33.5	3.9	30.4	33.0	6.8
116/134	20.2	20.6	2.2	17.8	17.8	2.1	21.0	25.0	5.4	18.5	18.5	4.5	21.5	22.0	6.5
MAD				1.86	2.57	0.83	1.88	2.60	2.10	2.92	3.39	1.44	1.84	2.06	2.60
SD				2.21	2.69	1.13	2.15	3.00	1.53	2.85	3.47	1.32	2.13	2.50	1.48
relMAD				6.2 %	8.2 %	26.9 %	5.9 %	8.2 %	61.8 %	9.4 %	10.9 %	45.8 %	6.0 %	6.7 %	96.6 %

With a MAD of 1.86 Å, SD of 2.27 Å and a relative error of 6 %, the computed mean distances $\bar{r}$ show again a very good agreement with the experiment, although the error is roughly twice as large as for the azurin mutants. This is in line with the expectation, that the error per spin-label amounts to approximately 1.0 Å and is roughly additive. On the same test set, MMM and mtsslWizard yield MAD values of 1.88 and 2.92 Å respectively.[266] T4L mutants were also calculated in Ref. 388 with the MDDS approach, yielding a comparable MAD of 1.84 Å. r_p values for all tested methods are also listed in Table. 5.3. Due to the resolution limit of 0.5 Å for MMM and 1.0 Å for MDDS, the methods are subject to an uncertainty of ± 0.25 and ± 0.5 Å respectively.

While the performance in terms of $\bar{r}$ is quite similar for CREST/MD, MMM, and MDDS for the mutants of T4L, CREST/MD is the only method that is capable of reproducing the experimental distribution widths. The broadening amounts to 27 % with CREST/MD, while mtsslWizard and MMM predict 46%–62 % wider distributions. For MDDS, the distance distributions are almost twice as wide as the experiment. Distance distributions for all 19 mutants are shown in Appendix A5. For multiple mutants, a multi-modal distribution was found in the PELDOR experiment and by the CREST/MD protocol. While for most of them, the most intensive peak corresponds nicely with the experiment (*e.g.*, 60/90, 60/94, 60/109, 82/132), for some mutants differences in the modality occur (*e.g.*, 64/122, 60/94). For mutant 83/123 the experimentally found mean distance was not reproduced with CREST/MD. Comparison of the MD snapshots with the initial crystal structure revealed non-covalent interactions between the two spin-labels and, even though the tertiary structure was similar on a global view, significant differences in the positions of individual loops and helices (see Appendix A5). These conformational changes caused the short mean distance of only 5.8 Å, which is out of range for PELDOR measurements, where

the lower limit is about 16 Å.[407] Nevertheless, the strongly reduced modulation depth shown in Ref. 405 is an indication, that the predicated conformation was also present in the experiment. For this reason, mutant 83/123 was excluded from the statistical evaluation here. The overall agreement between theory and experiment can be regarded as remarkably good.

5.4. Conclusion

In this work, an automated workflow for the computation of structural (conformational) ensembles of nitroxide spin-labeled (metallo-) protein mutants was developed and tested. The procedure combines the conformer search algorithm CREST with the generic GFN force-field. The key ingredient in the protocol was the constrained generation of the conformer ensemble with CREST at the GFN-FF level of theory followed by a completely unconstrained GFN-FF MD simulation at the freezing point of the solvent to obtain thermally equilibrated structures. An implicit GBSA solvation model was applied throughout and necessary for realistic protein simulations. With this CREST/MD composite scheme, experimental distance distributions of azurin and T4 lysozyme mutants were successfully reproduced with a mean absolute deviation from the experiment of less than 2 Å (5–10 %) for $\bar{r}$. Compared to previous theoretical methods (see Ref. 404), this is the highest accuracy ever reached for the mutants of azurin. For T4 lysozyme, CREST/MD outperforms MMM and mtsslWizard and is on par with MDDS for $\bar{r}$. In terms of distribution width σ_r, CREST/MD best reproduces the experiment throughout all tested systems. A further advantage of this newly developed scheme is the obtained information about the modality of the distance distribution. Because no general assumptions about the structure of the biomolecule and spin-label were made in this work, the conclusions can be generalized to basically arbitrary labels, including those based on metal ions, as well as arbitrary biomolecules. The CREST/MD method is intended to serve as an accurate and robust tool for the comparison and validation of experimental findings, such as distance distributions measured by PELDOR. It may also help to translate data from paramagnetic relaxation enhancement (PRE) measurements in NMR or to predict/analyze site-specific effects in Dynamic Nuclear Polarization (DNP). Finally, it should be mentioned that the introduction of the spin-label was only needed for the PELDOR experiment, yet, the combination of CREST and GFN-FF may also be applied straight forward to other biological systems without any mutations or modifications.

Acknowledgments

This work was supported by the DFG in the framework of the “Gottfried-Wilhelm-Leibniz” prize to Stefan Grimme. Sebastian Spicher thanks the “Fond der Chemischen Industrie (FCI)” for financial support. The authors thank M. Bursch, A. Hansen and P. Pracht for helpful discussions.

6. Benchmarking London Dispersion Corrected Density Functional Theory for Noncovalent Ion–π Interactions

Sebastian Spicher,[a] Eike Caldeweyher,[a] Andreas Hansen,[a] and Stefan Grimme[a]

Received: 26 March 2021, Published online: 05 May 2021

Reprinted (adapted) with permission[†] from
Spicher, S.; Caldeweyher, C.; Hansen, A.; Grimme, S. *Phys. Chem. Chem. Phys.* **2021**, *23*, 11635–11648.

DOI 10.1039/D1CP01333E

Own manuscript contribution

- Compiling the benchmark set
- Conception of the study and choice of methods
- Performing all the DFT calculations
- Interpretation of the results
- Writing the manuscript

[a]Mulliken Center for Theoretical Chemistry, Institute of Physical and Theoretical Chemistry, University of Bonn, Beringstraße 4, 53115 Bonn, Germany
[†]Permission requests to reuse material from this chapter should be directed to the Royal Society of Chemistry.

Abstract The strongly attractive noncovalent interactions of charged atoms or molecules with π-systems are important binding motifs in many chemical and biological systems. These so-called ion–π interactions play a major role in enzymes, molecular recognition, and the structure of proteins. In this work, a molecular test set termed IONPI19 is compiled for inter- and intramolecular ion–π interactions, which is well balanced between anionic and cationic systems. The IONPI19 set includes interaction energies of significantly larger molecules (up to 133 atoms) than in other ion–π test sets and covers a broad range of binding motifs. Accurate (local) coupled cluster values are provided as references. Overall, 19 density functional approximations, including seven (meta-)GGAs, eight hybrid functionals, and four double-hybrid functionals combined with three different London dispersion corrections, are benchmarked for interaction energies. DFT results are further compared to wave function based methods such as MP2 and dispersion corrected Hartree–Fock. Also, the performance of semiempirical QM methods such as the GFN*n*-xTB and PM*x* family of methods is tested. It is shown that dispersion-uncorrected DFT underestimates ion–π interactions significantly, even though electrostatic interactions dominate the overall binding. Accordingly, the new charge dependent D4 dispersion model is found to be consistently better than the standard D3 correction. Furthermore, the functional performance trend along Jacob's ladder[159,163] is generally obeyed and the reduction of the self-interaction error leads to an improvement of (double) hybrid functionals over (meta-)GGAs, even though the effect of the SIE is smaller than expected. Overall, the double-hybrids PWPB95-D4/QZ and revDSD-PBEP86-D4/QZ turned out to be the most reliable among all assessed methods for the description of ion–π interactions, which opens up new perspectives for systems where coupled cluster calculations are no longer computationally feasible.

6.1. Introduction

Ion–π interactions refer to strongly attractive noncovalent interactions (NCI) between ions and mostly organic π-systems.[408,409] They are of crucial importance for many processes in chemistry and biology, such as controlling the regio- and stereoselectivity in organic reactions,[410,411] enabling important biological processes,[412–416] and determining the structures of molecules and proteins.[417–420] The application of quantum mechanical (QM) methods in the description of such ion–π systems is desirable for an in-depth understanding. Kohn–Sham Density Functional Theory (DFT), with its vast number of density functional approximations (DFA), is one of the most promising electronic structure methods for this purpose regarding the accuracy and computational efficiency.[3,421] Nevertheless, DFT methods have well-known weaknesses, like the one- and many-electron self-interaction error (SIE),[160,161,422] errors due to nondynamical correlation effects,[423,424] and the lack of long-range electronic correlation, so-called London dispersion (LD) interactions.[334] The effect on ion–π interactions is shown in this work.

The SIE affects even modern DFAs and may lead to severe SCF convergence problems,[425] artificial charge-transfer (CT),[426,427] and inaccurate NCI energies for larger inter-fragment distances. This is in contrast to Hartree–Fock (HF) theory and second-order Møller–Plesset perturbation theory (MP2), which are SIE free because the exchange integrals exactly cancel the self-interaction contributions

from the Coulomb integrals. This behavior is exploited by hybrid DFAs, where a fraction of exact exchange (also called Fock exchange) is mixed in, partially canceling the SIE. While large amounts of Fock exchange reduce the SIE, the resulting hybrid DFAs also inherit general shortcomings of HF, *e.g.*, a lacking description of Coulomb interactions by overestimating ionic contributions in the wave function. For a more general discussion on the one-electron SIE in DFT see, *e.g.*, Ref. 428, and for the related many-electron SIE see, *e.g.*, Ref. 422.

Mean-field electronic structure methods like HF do not describe long-range electronic correlation effects and hence cannot account for LD interactions. This drawback of HF is also present in DFT. The absence of LD interactions is long known[429] and various solutions have been developed in the context of LD-corrected DFT methods.[334,430–433] One strategy to fix the dispersion problem of conventional DFT has been the development of additive corrections. A popular and frequently used additive scheme is the "DFT-D3" correction, where the majority of the missing dispersion energy is accounted for by summing up the dispersion contributions of each atom pair.[252,253,334] Another approach is to add the nonlocal (NL) correlation energy effects as a function of the electron density to standard exchange-correlation DFAs, which is known as van der Waals density functional theory (vdW-DFT),[434] or approximations thereof (VV10).[348]

For the development and testing of state-of-the-art DFT methods, ion–π interactions as a class of NCIs are of special interest. Symmetry adapted perturbation theory[435–437] (SAPT) studies, which allow the separation into different energy components namely electrostatics, Pauli repulsion, induction, and LD, revealed that ion–π systems incorporate strong electrostatic and inductive components.[438] For highly polarizable systems, however, also LDs were identified as a crucial part of the ion–π interaction.[439] For this reason, ion–π interactions present a challenge for the density functional itself as well as for the added dispersion correction. In this work, we introduce a benchmark set composed of 19 molecules with strong ion–π interactions. It is termed IONPI19 and contains significantly larger molecules than in existing compilations and is well balanced between anionic and cationic systems. Various common "real-life" binding situations are covered as they occur in protein structures, molecular recognition, and supramolecular receptors. Intramolecular ion–π interactions are included as well. Hence, the IONPI19 set is an interesting test case for DFT and important addition to the pool of available benchmark sets.

As has been shown in previous studies,[440,441] well performing LD-corrected DFAs are able to reproduce coupled cluster reference interaction energies for cation-anion complexes, representing the building blocks of ionic liquids. These chemically often rather saturated systems are less prone to the SIE. Good results for small charged systems could also be obtained with DFT-D methods in Refs. 437 and 162, even though the benchmark sets discussed in these studies are composed of rather small ion–π systems. In this work, we want to find out if previous trends and findings also hold true for the IONPI19 set and we want to investigate the general importance of LD corrections for prototypical systems composed of cations/anions and π-systems. The common belief is that ion–π systems are dominated by electrostatic and inductive interactions[439] but little attention has been paid so far to the importance of LD in this context. In the present work, we will put a particular focus on the latter in the framework of LD-corrected DFT. Also MP2 (see *e.g.*, Refs. 442,443) and variants thereof[444] are common methods for

modeling ion–π interactions, although there are severe and well-known problems such as the overestimation of NCIs involving π-systems, particularly for π-π interactions.[445–449] Furthermore, similar to other post-HF correlation methods, MP2 is highly susceptible to the basis set superposition error[450,451] (BSSE), which leads to systematic overbinding with small and medium sized atom-centered basis sets. Due to persisting popularity of MP2 in NCI studies (see *e.g.*, Refs. 452,453), it is evaluated here as a competitor method.

To evaluate the performance of the methods mentioned above, reliable reference values of high accuracy are needed. For small to medium sized systems (up to about 30 atoms) explicitly correlated coupled cluster composite schemes such as the Weizmann protocols[454] (W1-F12 and W2-F12) have proven to yield highly accurate reference values. Yet, the respective computational cost is considerable. For larger systems (up to about 150 atoms), domain based local pair natural orbital coupled cluster theory (DLPNO-CCSD(T))[147,148] is still computational feasible and was already successfully applied,[150,237] even though the high accuracy of the W*n*-F12 protocols cannot be fully achieved. To reduce the additional errors due to the local (DLPNO) approximations, very tight threshold settings have to be applied[149] in addition to a proper complete basis set (CBS) extrapolation, which in turn also makes these calculations quite computationally demanding. The high-level reference values calculated in this work can also be useful in the development and validation of low-cost methods, *e.g.*, of special force-fields (FF),[455,456] since hardly any reliable coupled cluster reference values for ion–π interactions energies of larger molecules exist so far,[457] and neither have they been calculated with such an accurate setup.[458] The development of such FFs and respective workflows is an emerging field of research, especially with respect to the efficient description of ion–π interactions in proteins.[459,460]

First, a brief survey of the employed semi-classical LD-correction schemes is given followed by a description of the compiled IONPI19 benchmark set. Further, the results for this test set are presented and discussed for all employed methods. An energy decomposition analysis (EDA) is performed for the dissociation of an ion–π complex to investigate the effect of the SIE for GGA and hybrid DFAs. Due to significant increases in efficiency, accuracy, and related popularity, semiempirical QM (SQM) and FF methods are additionally tested and evaluated on the IONPI19 set. Timings are compared for all different types of methods with regard to their accuracy. Finally, general conclusions and method recommendations will be given.

6.2. Semi-Classical London Dispersion Corrections

To account for the missing LD interactions in the framework of DFT (and also HF), we apply two closely related semi-classical LD-correction schemes. First, the widely used DFT-D3 method with two-body contributions (only $E_{\text{disp}}^{(2)}$) with the standard Becke–Johnson (BJ) rational damping.[461,462] Second, we consider the default version of the recently introduced DFT-D4 scheme[242,249,339] including also three-body Axilrod–Teller–Muto[247,248] (ATM) contributions, where the dispersion energy is given by

$$E_{\text{disp}}^{\text{DFT-D4}} = E_{\text{disp}}^{(2)} + E_{\text{disp}}^{\text{ATM}}. \tag{6.1}$$

The basic formula for two-body dispersion interactions is the same in the D3 and D4 model, where the BJ rational damping form for the interatomic pair sum is employed,

$$E_{\text{disp}}^{(2)} = -\sum_{\text{AB}} \left[\frac{C_6^{\text{AB}}}{R_{\text{AB}}^6 + f\left(R_{\text{AB}}^0\right)^6} + s_8 \frac{C_8^{\text{AB}}}{R_{\text{AB}}^8 + f\left(R_{\text{AB}}^0\right)^8} \right], \tag{6.2}$$

with the three fitted damping and scaling parameters a_1, a_2, and s_8. Here, AB labels atom pairs, and $f(R_{\text{AB}}^0) = a_1 R_{\text{AB}}^0 + a_2$ is the BJ damping function with appropriate covalent radii.[253]

In both methods, the C_6 (and C_8) coefficients are obtained from precalculated frequency-dependent time-dependent DFT dipole polarizabilities.[463] In addition to the coordination number dependence in DFT-D3, classical atomic partial charges are included in DFT-D4, which are calculated by a charge model based on electronegativity equilibration of Gaussian type charge densities (EEQ).[464] According to many tests on neutral organic systems, DFT-D3 and DFT-D4 methods provide both accurate asymptotic dispersion energies of roughly coupled cluster accuracy[334] while D4 is somewhat superior for ionic or metallic cases.[242,465] Whether this also holds true for the important class of ion–π complexes is one of the main questions of the present work. Due to the high computational efficiency of the additive DFT-D schemes, they are also suitable for low-cost methods including force-fields.[117,273,274,346,466]

Other popular LD correction schemes exist, *e.g.*, the exchange-hole dipole method,[467–470] the many-body dispersion model,[430,471] the van der Waals family of density functionals.[472], or the non-local electron density dependent dispersion correction termed VV10 or DFT-NL.[348,473,474] For comparison, the latter is also tested in this work. For an in-depth analysis of other LD corrections and a more general discussion on the importance of LD effects for chemical bonding, see, *e.g.*, Refs. 334,433.

6.3. Description of the Molecular Test Set

The composition of the test set aims at both, smaller model systems as well as experimentally investigated ion–π systems. We arrived (after considering more than 30 candidate structures) at a statistically balanced set containing 19 exemplary systems featuring typical ion–π binding motifs. The average system size is about 32 atoms per molecule with the largest system consisting of 133 atoms. The test set is divided into smaller ($\leq$ 30 atoms) and larger ($>$ 30 atoms) systems, of which the first subset is shown in Figure 6.1 and the latter in Figure 6.2. For the ten cationic and nine anionic systems the mean interaction energy is -20.9 kcal mol^{-1}. Reference energies and estimated errors for each system as well as the corresponding computational reference level of theory are listed in Table 6.1.

Figure 6.1A shows the first seven systems of the IONPI19 set which are all cationic. Systems **1**–**3** show Li^+, Na^+, and K^+ bound to benzene. The three alkali-benzene complexes were taken from the CHB6 benchmark set[162,437] with the original reference interaction energies obtained at the CCSD(T)/CBS level of theory. Cation–π interactions are of particular interest for structural biology as the DNA bases are also able to participate therein. Systems **4**–**6** show Na^+ in complex with cytosine, Li^+ coordinated to the five-membered ring of adenine, and Na^+ in complex with the five-membered ring of guanine. The systems were taken from Ref. 475, where reference binding energies were also computed at the

CCSD(T)/CBS level of theory. System **7** was newly added to this benchmark and consists of anthracene and the cyclopropenyl cation ($C_3H_3^+$). The reference interaction energy of the rigid monomers was computed at the W1-F12 level.

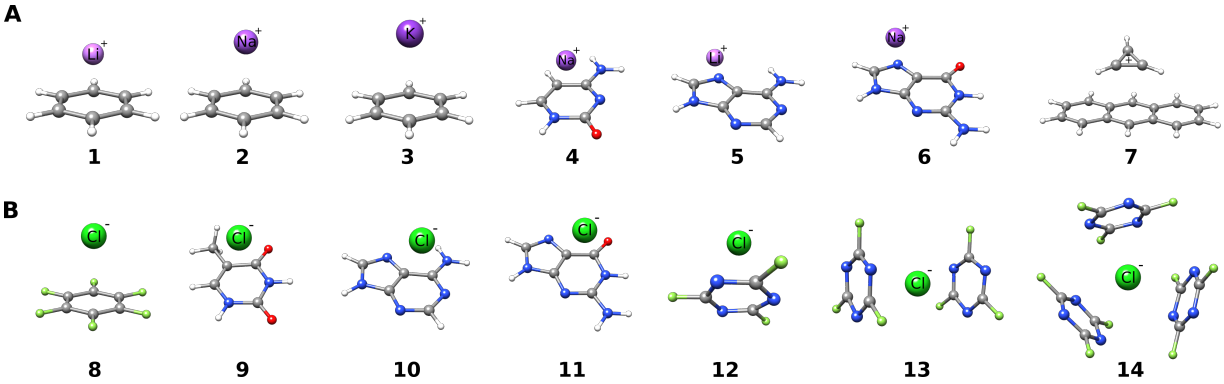


Figure 6.1.: Subset of the IONPI19 benchmark set containing small molecules (≤ 30 atoms). (A) Systems **1**–**7** include cation–π interactions. (B) Systems **8**–**14** include anion–π interactions.

The anionic systems of the small molecule subset are shown in Figure 6.1B. For **8** the anion–π interaction of hexafluorobenzene (C_6F_6) and chloride (Cl^-) is achieved by placing strong electron withdrawing substituents along the π-system. The reference interaction energy was computed at the W1-F12 level. Systems **9**–**11** show chlorine anions in complex with the six-membered ring of thymine, adenine, and guanine. These systems and their respective reference interaction energies were taken from Ref. 475. The test systems **12**–**14** are newly added to this benchmark and were taken from a study on designing receptors for molecular recognition. There, the additivity of anion–π interactions for 1:1, 1:2, and 1:3 (anion:π) complexes of trifluoro-1,3,5-triazine ($C_3F_3N_3$) with Cl^- ions[476] was investigated. All three systems were newly compiled for the IONPI19 set. Reference energies for the 1:1 complex were calculated at the W2-F12 level, whereas the 1:2, and 1:3 complex were computed at the DLPNO-CCSD(T1) / *VeryTightPNO* / CBS(aug-cc-pVTZ/aug-cc-pVQZ) level of theory.

The subset of larger molecules shown in Figure 6.2 was newly compiled for this work. **15** and **16** show the electron-deficit and cavity self-tunable macrocyclic host tetraoxacalix[2]arene[2]triazine forming 1:1 complexes with small anions (NO_3^-, SCN^-) as revealed by Wang and co-workers.[477] In complex, the two opposing triazine rings of tetraoxacalix[2]arene[2]triazine act as a pair of tweezers to interact with the included anions through cooperative anion–π and lone-pair electron-π interactions. The supramolecular cyclophane host-guest complex **17** is another interesting test system of practical relevance. This complex is able to catalyze N-alkylation to form cationic products *via* the Menschutkin reaction,[478] where it is assumed that the cation–π interaction plays a central role in catalysis. Hence, it is of relevance for the understanding of several biological methylation reactions.[479] For intramolecular ion–π interactions two test cases were chosen in which cation–π interactions contribute significantly to the stability of conformations.[150] **18** is based on a study of Dougherty *et al.*,[480] who proposed that the neurotransmitter acetylcholine can bind to acetylcholinesterase through cation–π interactions. A simplified system is taken from Ref. 481, where the folded ester conformation is proposed to be more stable than the unfolded one. For the isosteric 3,3-dimethylbutyl indole-3-acetate (*i.e.*, replaced ammo-

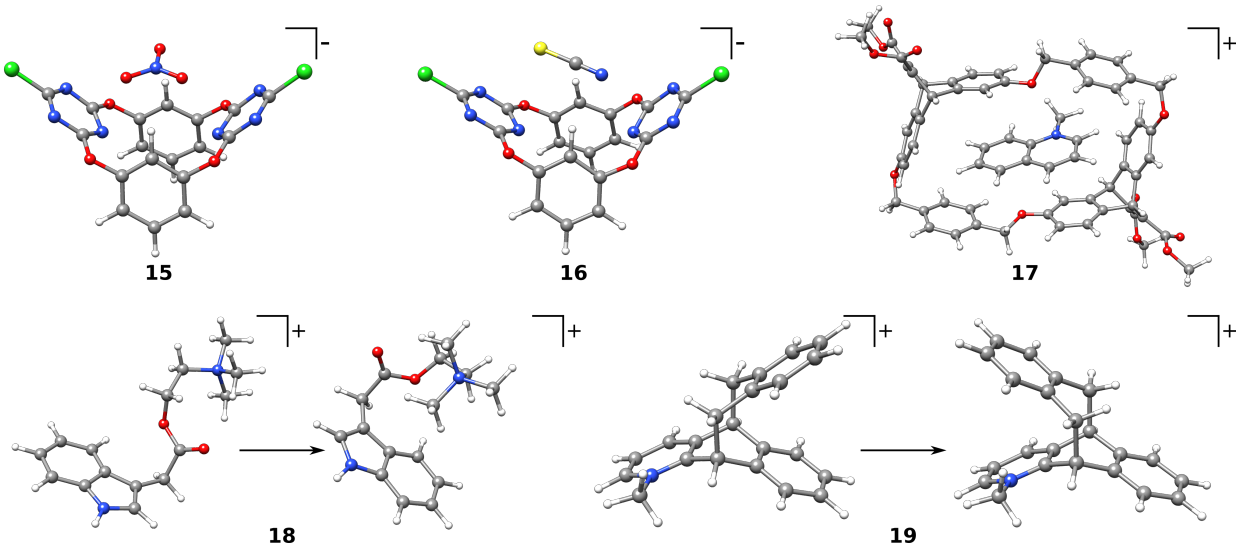


Figure 6.2.: Subset of the IONPI19 benchmark containing large molecules (> 30 atoms). Systems **15**–**17** show intermolecular ion–π interaction, whereas **18** and **19** are examples for intramolecular ion–π interactions.

nium nitrogen with carbon), an analogous folding is not observed. This implies that the cationic nature of the quaternary trimethylammonium group is responsible for this preferable association with the indole ring through cation–π interactions. Folded and unfolded conformations were generated with the recently published CREST algorithm.[60] **19** contains multiple interaction motifs that are able to compete with each other. This seesaw balance[482] adopts two distinct conformations that are either stabilized by cation–π or by π–π interactions. Experimental ^{1}H-NMR studies[453,483] in solution proposed that the cation–π bound conformer is stabilized by about 1.5 kcal mol^{-1}.

We are aware that intramolecular ion–π interactions introduce difficulties for fragment based methods such as SAPT. Therefore, full statistics are also given in Appendix A6 (Table A6.11, A6.12) for the IONPI17 set, where the intramolecular test cases (**18** and **19**) are excluded.

6.4. Computational Details

Typical DFAs from different classes of Jacob's ladder combined with the large def2-QZVPP basis set[134,135] were evaluated for the IONPI19 benchmark set. The DFA selection is based on results for previous benchmark studies[162] and on their popularity in the computational chemistry community.[484] All DFAs were assessed with the D3 and D4 London dispersion correction in the Becke–Johnson scheme and/or the nonlocal density-dependent NL (VV10) treatment in a non self-consistent form. For M06-L[485] and M06-2X[486] D3 was applied with zero damping.[252] D3 and D4 dispersion corrections were calculated with the *dftd3* and *dftd4* standalone programs.[487] The ORCA[488] implementation was used for ωB97X-D3(BJ) and to calculate all NL corrections. A list of the tested DFAs and dispersion correction combinations is given in Table 6.2.

All composite ("3c") DFT and r^2SCAN calculations were performed using the TURBOMOLE 7.5.1

Table 6.1.: Summary of the reference interaction and association energies ΔE and methods for the IONPI19 benchmark set. Values are given in kcal mol^{-1}.

System	$\Delta E_{\text{Ref.}}$	Estimated error	Reference
1	-39.1	± 0.8 (2.0 %)	437[a)]
2	-25.6	± 0.5 (2.0 %)	437[a)]
3	-19.9	± 0.8 (4.0 %)	437[a)]
4	-14.8	± 0.2 (1.5 %)	475[b)]
5	-25.7	± 0.4 (1.5 %)	475[b)]
6	-19.7	± 0.3 (1.5 %)	475[b)]
7	-21.5	± 0.2 (1.0 %)	this work[c)]
8	-14.6	± 0.2 (1.0 %)	this work[c)]
9	-10.4	± 0.1 (1.0 %)	475[d)]
10	-1.9	< 0.1 (1.0 %)	475[d)]
11	-5.7	± 0.1 (1.0 %)	475[d)]
12	-18.6	± 0.1 (0.5 %)	this work[e)]
13	-33.7	± 0.8 (2.5 %)	this work[f)]
14	-45.0	± 1.1 (2.5 %)	this work[f)]
15	-29.4	± 0.6 (2.0 %)	this work[g)]
16	-26.3	± 0.5 (2.0 %)	this work[g)]
17	-37.2	± 1.9 (5.0 %)	this work[h)]
18	-5.0	± 0.1 (2.5 %)	this work[f)]
19	-2.4	± 0.1 (2.5 %)	this work[f)]
mean	−20.9	± 0.5 (2.2 %)	

[a)] CCSD(T)/δCBS + counterpoise correction (details: see original publication).
[b)] CCSD(T)/δCBS + counterpoise correction + modified frozen core approximation, *i.e.*, $Li^+ = 1s^2$ (no core) and $Na^+ = [He]2s^22p^6$ ([He] core) (details: see original publication).
[c)] W1-F12.
[d)] CCSD(T)/deltaCBS + counterpoise correction (details: see original publication).
[e)] W2-F12.
[f)] DLPNO-CCSD(T1)/*VeryTightPNO*/CBS(aug-cc-pVTZ/aug-cc-pVQZ).
[g)] DLPNO-CCSD(T1)/*VeryTightPNO*/CBS(aug-cc-pVTZ/aug-cc-pVQZ) + counterpoise correction and deformation energy.
[h)] DLPNO-CCSD('T1')/*VeryTightPNO*/'CBS(def2-TZVPP/def2-QZVPP)' + counterpoise correction and deformation energy.

program package.[335,501] Computations of energies and geometry optimizations were conducted using the *ridft* and *jobex* programs of TURBOMOLE, respectively. The resolution-of-identity (RI) approximation for the Coulomb integrals was always applied using matching default auxiliary basis sets.[341,342] For the integration of the exchange-correlation contribution, the numerical quadrature grid *m*4 was employed. The default convergence criteria (10^{-7} E_h for energies and 10^{-5} E_h/bohr for gradients) were used throughout.

All other DFT, HF, MP2, and local coupled cluster calculations were carried out with the ORCA 4.2.1 program package.[488,502] The frozen core and RI approximations for the correlation part as well as *TightSCF* convergence criteria for the HF energy was employed for all double-hybrids, MP2, and CC methods. The domain based pair natural orbital local coupled cluster method[147] in its sparse maps[148] iterative triples[503] implementation (DLPNO-CCSD(T1)) employing *VeryTightPNO*[149] threshold settings was applied. An aug-cc-pVTZ/aug-cc-pVQZ[504] and def2-TZVPP/def2-QZVPP CBS extrapolation according to the schemes proposed by Helgaker/Klopper[505] (aug-cc basis sets) or Neese/Valeev[506] (def2 basis sets) was carried out for DLPNO-CCSD(T1). Matching auxiliary basis sets were applied for the

Table 6.2.: Tested DFAs and dispersion correction combinations.

Functional		D3	D4	NL	Reference
composite (3c)					
	PBEh-3c	✓	×	×	231
	B97-3c	✓	×	×	237
	r^2SCAN-3c	×	✓	×	489
(meta-)GGA					
	PBE	✓	✓	✓	166
	M06-L	✓	✓	×	485
	TPSS	✓	✓	✓	171
	r^2SCAN	✓	✓	✓	490,491
	B97M	✓	✓	✓	492
hybrid					
	M06-2X	✓	×	×	486
	MN15	×	×	×	493
	PBE0	✓	✓	✓	494
	PW6B95	✓	✓	✓	495
	B3LYP	✓	✓	✓	496,497
	ωB97M	✓	✓	✓	498
	ωB97X	✓	✓	✓	327
double-hybrid					
	B2PLYP	✓	✓	✓	174
	revDSD-PBEP86	×	✓	×	499
	revDSD-BLYP	×	✓	×	499
	PWPB95	✓	✓	✓	500

density fitting.[507] DLPNO-CCSD(T1)/*VeryTightPNO*/CBS(aug-cc-pVTZ/aug-cc-pVQZ) values were in general counterpoise (CP) corrected unless for the 1:2 (**13**) and 1:3 (**14**) complexes of $C_3F_3N_3$ with Cl^-, as the corresponding calculations with and without CP correction for the 1:1 complex (**12**) revealed a negligible residual BSSE of only 0.01 kcal mol^{-1}. Detailed information about the reference calculation for each system is shown in Table 6.1. To validate the accuracy of the DLPNO-CCSD(T1) reference values and to determine whether some of the systems show multireference character, a *T1* diagnostic according to Ref. 508 was performed (see Appendix A6 Table A6.10). Empirically, a value larger than 0.02 may indicate a significant nondynamical correlation.[509] This is not the case for the systems in the IONPI19 set, where the largest value is 0.017 for adenine in system **4**. The small to moderate values of the maximum *T2* amplitudes (largest value 0.067 for anthracene in **7**) further indicates that no problematic systems in terms of nondynamical correlation are included in the IONPI19 benchmark.[510]

CBS extrapolation was also performed for MP2. The MP2/CBS schemes correspond to the DLPNO-CCSD(T1) extrapolations for the individual systems. For **1**, **2**, and **6** MP2/CBS(aug-cc-pCVTZ/aug-cc-pCVQZ) was employed without RI since the AutoAux[507] basis showed linear dependencies. For **3** RI-MP2/CBS(aug-cc-pVTZ/aug-cc-pVQZ) was employed for C_6H_6 and RI-MP2/CBS(def2-TZVPPD/def2-QZVPPD) for K^+ as no non-relativistic "aug basis" was available. Corresponding auxiliary basis sets were employed. RI-MP2/CBS(aug-cc-pCVTZ/aug-cc-pCVQZ) (AutoAux[507] option in ORCA 4.2.1) was employed for **4** and **5**. For all systems **1**–**6**, the Boys-Bernadi CP correction was applied. The interaction energy of systems **7**–**16** was calculated without CP correction by RI-MP2/CBS(aug-cc-pVTZ/aug-cc-

pVQZ) including the corresponding auxiliary basis sets. For **17**, RI-MP2/CBS(def2-TZVPP/def2-QZVPP) was employed including the corresponding auxiliary basis sets, the CP correction, and deformation energy. RI-MP2/CBS(aug-cc-pVTZ/aug-cc-pVQZ) with the corresponding aux basis sets but without CP correction was employed for **18** and **19**. In general, the CP correction was applied only when basis set size or CBS extrapolation was not sufficient enough to minimize the BSSE.

For the largest test system **17**, a slightly more approximate CBS extrapolation scheme was employed for the local coupled cluster correlation energy since the full def2-QZVPP calculation was computationally unfeasible with the latter method. It is labelled as CBS/'def2-TZVPP/def2-QZVPP' in the following and refers to a multiplicative scaling of the DLPNO-CCSD('T1') correlation energy by the quotient of the respective CBS(def2-TZVPP/def2-QZVPP) and def2-TZVPP MP2 correlation energies. A similar CBS protocol was already successfully employed in Ref. 150. Note that the iterative correction to the triples correlation energy was calculated with the def2-TZVPP[134] basis set (labeled as 'T1') for this test system. The fact that the iterative triples could only be calculated with the def2-TZVPP basis set introduces a small additional error in the difference between iterative and non-iterative triples, but this is not significant since this correction amounts to 0.5 kcal mol^{-1} only and given that the estimated total error of the total association energy is ~1.9 kcal mol^{-1}. The high-level composite explicitly correlated coupled cluster protocols W1-F12 and W2-F12[454] were applied with the Molpro program package V. 2015.1.[511,512]

For systems **1**–**6** the Boys-Bernardi counterpoise correction[513] was applied for HF and all non-"3c" DFAs, which do not feature a fixed, composite basis set. This was because the alkali metal ion complexes showed a systematic overestimation of the interaction energy for all assessed DFT methods, which was not observed for the other 13 systems of the IONPI19 set. The CP correction reduced this BSSE on average by 0.3 kcal mol^{-1}. The CP uncorrected values are given in Appendix A6 in Table A6.9. Structures that were newly generated for the IONPI19 benchmark set (**7**, **8**, **12**–**19**) were all optimized at the PBEh-3c[231] level of theory. The lowest energy molecular conformers were obtained from the advanced conformer rotamer ensemble sampling tool[60,514] (CREST) in its default settings at the GFN2-xTB[93,96] level followed by DFT geometry re-optimizations at the PBEh-3c level of theory. All SQM and FF calculations were performed with the xtb 6.3.3[515] (GFN1-xTB,[94] GFN2-xTB, GFN-FF[117]), and MOPAC 2016[338] (PM6-D3H4X,[86] PM7[88]) program packages.

6.5. Results and Discussion

In Section 6.5.1 the performance of all tested DFAs and WFT methods for the IONPI19 set is presented and discussed. The dissociation curve of an ion–π complex is shown in Section 6.5.2. SQM and FF methods are evaluated in Section 6.5.3 and a comparison of computation times is given in Section 6.5.4.

6.5.1. Benchmark Study on IONPI19

A representative set of different DFAs including five (meta-)GGAs, seven hybrid functionals, and four double-hybrid (DH) functionals was assessed. In addition, HF and MP2 were tested. Furthermore, the recently developed efficient composite DFT-D methods B97-3c (GGA), r^2SCAN-3c (meta-GGA), and

PBEh-3c (hybrid) are evaluated in comparison. Moreover, three correction schemes for capturing long-range London dispersion interactions with DFAs were applied, the D3 correction with Becke–Johnson (BJ) or zero (0) damping, the newly developed D4 correction with three-body ATM contributions, and the nonlocal dispersion correction (VV10) in its non self-consistent implementation. First, we want to determine which combination of DFA and dispersion correction works best for the compiled IONPI19 benchmark set. The performance in terms of the mean absolute deviation (MAD) from the reference values of a subset of DFAs, for which all three dispersion corrections are available, is shown in Figure 6.3. With a mean MAD of 0.9 kcal mol^{-1} averaged over all 12 tested methods, the recently introduced

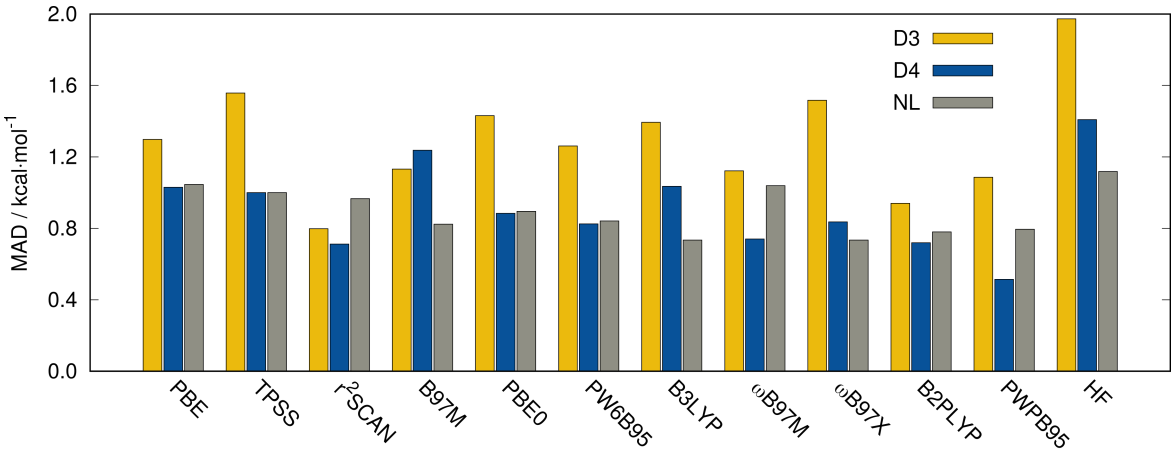


Figure 6.3.: Comparison of the performance of the D3(BJ), D4(BJ)-ATM, and NL dispersion corrections for different DFAs and HF on the IONPI19 benchmark set.

charge scaled D4 scheme outperforms its predecessor D3, which yields a mean MAD of 1.3 kcal mol^{-1}. Considering the mean interaction energy ΔE of -20.9 kcal mol^{-1}, this improvement is significant. The nonlocal dispersion correction performs on average equally well as the D4 scheme with a mean MAD of also 0.9 kcal mol^{-1}. Yet, this is mostly due to the better performance of the NL correction within the B97M-V and ωB97X-V functionals, which were developed together with the VV10 correction (see Refs. 492, 498). To further investigate the origin of the difference in performance between the D3 and D4 schemes, the role of three-body contributions is taken into account. The results are shown in Figure 6.4. Here, the ATM term which is the default in DFT-D4 was also added to the D3 correction.

Adding the three-body ATM term to the D3 scheme on average improves the mean MAD from 1.3 kcal mol^{-1} to 1.2 kcal mol^{-1}, which is still clearly off the accuracy of the D4-ATM approach with a mean MAD of 0.9 kcal mol^{-1}. As an extension to the ATM term, the many-body dispersion (MBD) approach by Tkatchenko–Scheffler has also been tested[430] in combination with the D4 scheme. Again, the MAD is shown in comparison to respective reference values. Exchanging the ATM term by the MBD approach has nearly no effect, the MADs are only slightly larger. This proves that the three-body term in D4-ATM is not the reason for the improvement over the D3 scheme for the IONPI19 but the incorporation of atomic partial charges in D4 yielding charge scaled polarizabilities. The rather negligible contribution of three-body effects is further confirmed by the previously discussed good performance of the NL correction, which also does not include ATM or MBD terms. Another important fact that can be taken

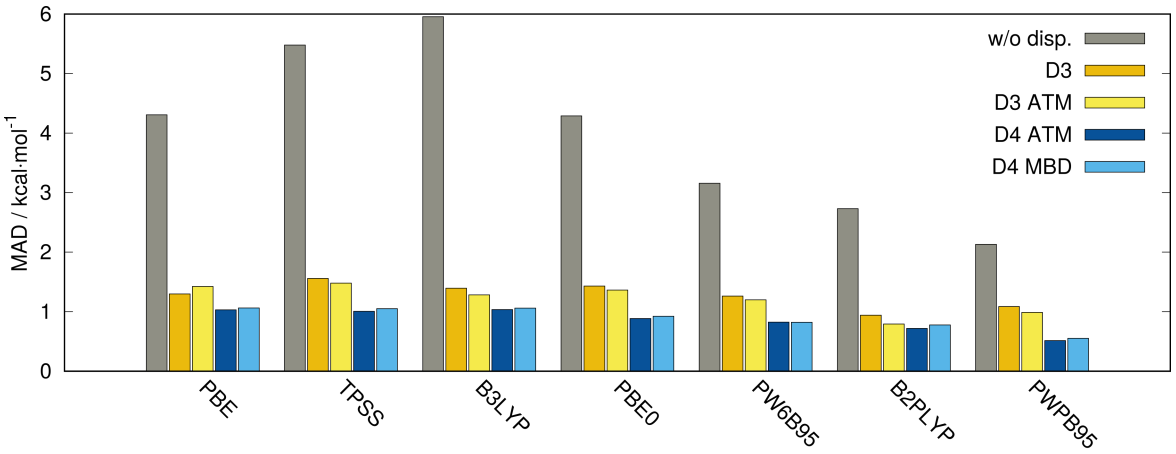


Figure 6.4.: Comparison of the performance of D3(BJ), D3(BJ)-ATM, D4(BJ)-ATM, and D4(BJ)-MBD dispersion corrections for different DFAs on the IONPI19 benchmark set. The LD-uncorrected results are shown for comparison.

from Figure 6.4 is the general influence of the dispersion correction. It is shown that LD-uncorrected DFT underestimates ion–π interactions significantly, even though electrostatic interactions dominate the overall binding. The mean MAD without an LD correction amounts to 4.0 kcal mol^{-1}. This is in line with SAPT studies, which revealed, that for polarizable systems also LD interactions contribute a crucial part of the ion–π interaction.[439] In general, LD-corrections improve also the Minnesota-type functionals (with D3(0)), although to a smaller extent, since they already capture some dispersion interactions at intermediate interatomic distances by density dependent terms and their parameterization. Yet, on this specific ion–π benchmark no improvement was achieved neither with the D3 nor the D4 scheme. Hence, the LD-uncorrected functionals M06-L, M06-2X, and MN15 will be discussed in the following.

Figure 6.5 shows the statistical data of the performance for the best combinations of all tested DFA and LD corrections. The efficient composite DFT methods, HF, and MP2 are also included for comparison. The assessed DFAs perform on average as expected according to the picture of "Jacob's Ladder" and resemble closely the results for the extensive GMTKN55 main group benchmark set.[162] The tested (meta-)GGA functionals yield a mean MAD of 1.0 kcal mol^{-1} and mean SD of 1.2 kcal mol^{-1}. With an MAD of 0.7 kcal mol^{-1} the newly developed r^2SCAN-D4 functional is the best performing DFA, closely followed by the B97M-V functional with an MAD of 0.8 kcal mol^{-1}. TPSS-D4 and PBE-D4 perform reasonably well at the limit of chemical accuracy (1.0 kcal mol^{-1}) with an MAD of 1.0 kcal mol^{-1} and 1.1 kcal mol^{-1}, respectively. The worst among all tested DFAs is M06-L with an MAD of 1.5 kcal mol^{-1}. Also, while most (meta-)GGAs tend to systematically overbind the IONPI19 set (MD < 0), M06-L is the only DFA thereof with a significant positive MD of 1.4 kcal mol^{-1}.

An improvement is obtained with hybrid DFAs. The mean MAD and SD are reduced to 0.8 kcal mol^{-1} and 1.0 kcal mol^{-1}, respectively. The MD is negative for all hybrids, indicating a small systematic error. Out of all tested hybrid DFAs, B3LYP-NL, MN15, ωB97X-V, and ωB97M-D4 perform best with an MAD of 0.7 kcal mol^{-1}, followed by PW6B95-D4 and PBE0-D4 with an MAD of 0.8 and 0.9 kcal mol^{-1},

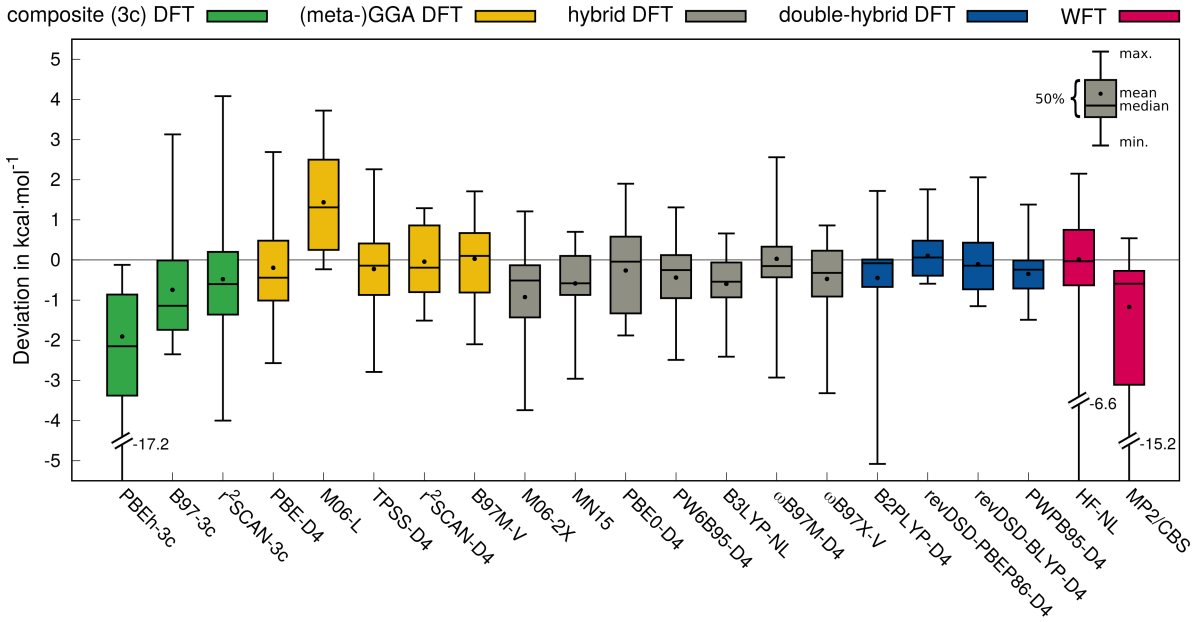


Figure 6.5.: Deviations of calculated ion–π interaction energies with different DFT and WFT methods for the IONPI19 set. The minimum deviation and maximum deviation for each data set are shown as range together with the first and third quartiles as the central box for each data set, the inter-quartile range contains 50 % of the data set. Additionally, the mean and median deviation are depicted as dots and vertical bars, respectively.

respectively. The smallest SD is obtained by B3LYP-NL (0.8 kcal mol^{-1}). The global-hybrid MN15 is the best performing Minnesota-type hybrid functional. It clearly improves upon M06-2X (54 % of Fock-exchange), which yields an MAD of 1.1 kcal mol^{-1} and also has the largest MD and SD among all tested hybrid DFAs with −0.9 kcal mol^{-1} and 1.2 kcal mol^{-1}, respectively. Overall, the improvement from (meta-)GGAs to hybrid DFAs is rather small. Hence, the SIE seems to be less severe for the IONPI19 set since the (meta-)GGAs are able to compete with the hybrid functionals. This observation is consistent with the fact that the DFA with the largest amount of Fock-exchange (M06-2X) performs worst. Tentatively it seems that at some point around 30 % to 40 % Fock-exchange contribution in the hybrid functional, the results are not improved any more.

Going to double-hybrid functionals, the mean MAD and SD are further reduced to 0.6 kcal mol^{-1} and 0.8 kcal mol^{-1}, respectively. The best performing DFAs are revDSD-PBEP86-D4 (0.4 kcal mol^{-1} MAD, 0.5 kcal mol^{-1} SD) and PWPB95-D4 (0.5 kcal mol^{-1} MAD, 0.6 kcal mol^{-1} SD), which both nearly approach the accuracy of the reference values. This generally good performance is in agreement with the conclusions of other benchmark studies.[162] revDSD-BLYP-D4 has an MAD of 0.6 kcal mol^{-1} and the DH-DFA with the largest deviation from the reference values is B2PLYP-D4 (0.7 kcal mol^{-1} MAD), whose performance is comparable to the best hybrid DFAs. It is noticeable, that B2PLYP-D4 shows the largest error range out of all DFAs discussed so far. The reason, therefore, is the overestimation of the association energy of **17** by 5.1 kcal mol^{-1} indicating an outlier of B2PLYP. For **17**, which is the largest system from the test set, a CBS(def2-TZVPP/def2-QZVPP) extrapolation and counterpoise correction were employed for MP2 to correct for the BSSE and basis set incompleteness error. To find out if this

is also necessary for double-hybrids, which include perturbative correlation energy terms, the same scheme was also applied for PWPB95-D4. The corresponding correction for MP2/CBS amounts to 1.3 kcal mol^{-1}, whereas a correction of only 0.4 kcal mol^{-1} was obtained for PWPB95-D4/CBS, which is less than the mean estimated error of the entire IONPI19 set (± 0.5 kcal mol^{-1}). Hence for the IONPI19 set, a CBS extrapolation is not necessary for DH-DFAs if they are employed together with the relatively large def2-QZVPP basis set.

The basic philosophy of "3c" composite methods is to provide a consistent description, *i.e.*, without systematic deviations to as low as possible computational cost. Hence, small but well-balanced atomic orbital (AO) basis sets are employed, while the remaining basis set errors are corrected by a geometrical counterpoise correction or in the case of B97-3c absorbed into the DFA itself via a slight reparameterization. DFT-3c methods are tested on the IONPI19 set to investigate whether the electrostatics of ion–π interactions are sufficiently described also without large amounts of polarization functions as in the def2-QZVPP basis set. PBEh-3c with its modified SVP basis yields a rather large MAD of 3.5 kcal mol^{-1} and SD of 4.4 kcal mol^{-1}. B97-3c performs already significantly better with a modified TZVP basis, yielding an MAD of 1.4 kcal mol^{-1} (SD = 1.8 kcal mol^{-1}). Best performing "3c" method is the very recently developed r^2SCAN-3c functional with an MAD and SD of 1.3 kcal mol^{-1} and 1.4 kcal mol^{-1}, respectively. Compared to, *e.g.*, the well-established TPSS-D4 meta-GGA with the large def2-QZVPP basis set, r^2SCAN-3c yields only a 0.3 kcal mol^{-1} larger MAD, whilst being one order of magnitude faster. Due to the relatively small but well balanced mTZVP basis set, the calculations need only a fraction of computation time compared to the large def2-QZVPP calculations. Hence, the performance of r^2SCAN-3c is promising for large scale computational studies of this type of chemistry. A more detailed look at computation times is given in Section 6.5.4.

For ion–π interactions, the MP2/CBS method (2.0 kcal mol^{-1} MAD) can not reach the accuracy of good DFAs, despite the application of computationally demanding CBS extrapolation schemes. As expected, the ion–π and π–π interactions are systematically overestimated, a trend that was not observed for DH-DFAs. The largest deviation from the reference is obtained for **17**, where the association energy is overbound by 15.2 kcal mol^{-1} with MP2/CBS. In general, MP2/CBS can not be recommended for the description of ion–π interactions. Here, common DFAs offer higher accuracy at significantly lower computational cost and without CBS extrapolation. Much better results are obtained by HF-NL. The MAD of 1.1 kcal mol^{-1} is comparable to good performing (meta-)GGAs. Since Hartree–Fock is SIE free, the remaining error is mainly due to the comparably poor description of electrostatic interactions.

The calculation of the functional mean deviation (MD) is shown in Figure 6.6. Here, the deviation from the reference averaged over all DFAs is given per system. The composite DFT methods are excluded due to the different basis set sizes. The functional mean is a good indication, whether certain subsets (cation, anionic, large systems) or individual systems of the IONPI19 benchmark are particularly challenging or show systematic errors. The largely inconspicuous course of the curve presented in Figure 6.6 indicates a statistically well-balanced test set, without major outliers. The average ratio of 1.3 of the MAD to the SD is a further indication of a normally distributed test set.[516] Neither the systems size nor the anions or cations induce systematic errors. Yet, it is noticeable that the largest functional MDs occur for **3** and **17** with ±1.2 kcal mol^{-1}, respectively. For many tested DFAs these two systems

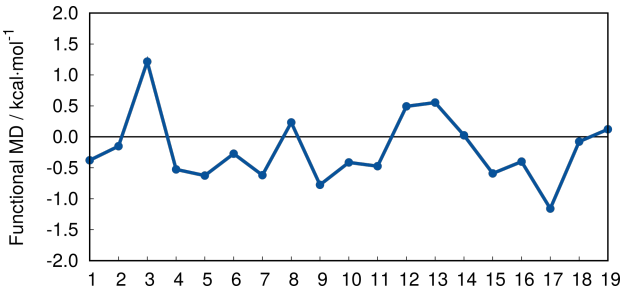


Figure 6.6.: Functional mean deviation from the reference values calculated as the average of all tested DFAs ("3c" methods excluded).

determine the error range, *e.g.*, TPSS-D4, ωB97M-D4, B3LYP-NL, and B2PLYP-D4. System **3** is the only K^+ containing complex. Here, the error seems to be attributed more to the def2-QZVPP basis set than to the DFAs.[437] For the largest and most complex system of the IONPI19 set **17**, a proportionally larger error can be expected. The estimated error of the DLPNO-CCSD(T1)/CBS reference amounts to 1.9 kcal mol^{-1} and hence, the functional MD is well below this value.

6.5.2. Dissociation of Ion–π Complexes

The results presented for equilibrium geometries (vide supra) did not suggest a major influence of the SIE on the interaction energies of the IONPI19 set. Now, we want to investigate if the SIE becomes more severe for the GGA and meta-GGA classes of DFAs when looking at the dissociation of an ion–π complex. Figure 6.7 shows the dissociation curves of C_6F_6 and the chloride anion (system **8**) computed with PBE-D4, TPSS-D4, PBE0-D4, and MP2/CBS in comparison to the W1-F12 reference for five different CMA distances within a range of 5–8 bohr. For shorter distances (5–6 bohr), all tested DFAs

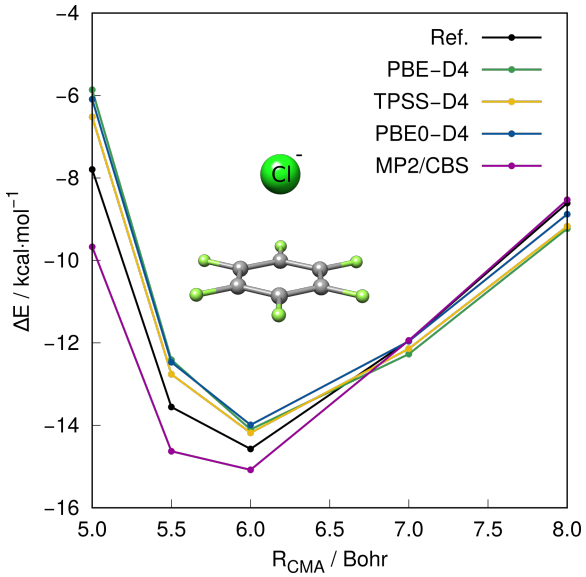


Figure 6.7.: Intermolecular potential energy curve of C_6F_6 and the chloride anion (Cl^-) obtained with PBE-D4, TPSS-D4, PBE0-D4, and MP2/CBS. All DFT calculations were performed in a def2-QZVPP basis set.

perform similarly and slightly underbind compared to the reference values, whereas MP2/CBS tends to overbind. For larger distances (6–8 bohr), all tested methods are reasonably close to the reference. At first sight, also for this system the SIE seems to be less severe since the (meta-)GGAs are able to compete with the hybrid functional (cf. PBE vs. PBE0) and the difference is marginal. To better understand this observation we conducted an EDA[517] for PBE-D4 and PBE0-D4 to investigate the effect of Fock exchange at four CMA distances taken from the dissociation curve in Figure 6.7 from 5 to 8 bohr, see Table 6.3).

For all tested CMA distances, PBE0-D4 is only slightly more accurate with an MAD of 0.7 kcal mol^{-1} than PBE-D4 (0.8 kcal mol^{-1}). The EDA interaction energy (INT) is calculated as the sum of electrostatics (EL), Pauli repulsion (REP), DFA correlation (CORR), and LD contributions. Table 6.3 lists deviations from W1-F12 reference interaction energies (denoted as $\Delta^{\text{ref}} = E_{\text{INT}}^{\text{calc}} - E_{\text{INT}}^{\text{ref}}$) for both DFAs. At CMA distances of 5, 6, and 7 bohr no significant SIE related issues occur and the energy contributions of PBE-D4 and PBE0-D4 are on the same order of magnitude. This changes, however, for the largest CMA distance of 8 bohr, where PBE-D4 results in nonphysical contributions for EL (repulsive) and REP (attractive) of 46.3 and −47.7 kcal mol^{-1}, respectively. This error is probably due to a violation of the Perdew–Parr–Levy–Balduz condition,[518–522] meaning that the total electronic energy as a function of electron number under a fixed external potential is not interpolating straight between integers. For GGA methods, this usually results in over-delocalization errors[161,523] and thus, the SIE becomes a problem for dissociating ion–π systems. Yet, despite the nonphysical contributions for EL and REP, an accurate PBE-D4 total interaction energy is obtained based on fortuitously error compensation.

Table 6.3.: Energy decomposition analyses of the $C_6F_6\cdots Cl^-$ complex for PBE-D4 and PBE0-D4. The total interaction energy (INT), the electrostatic (ES), the Pauli repulsion (REP), the short-range DFA correlation (CORR), and the LD contributions are listed. All values are given in kcal mol^{-1}.

	CMA distance of $C_6F_6\cdots Cl^-$							
	5 a_0		6 a_0		7 a_0		8 a_0	
DFA	PBE	PBE0	PBE	PBE0	PBE	PBE0	PBE	PBE0
INT	5.1	-5.6	-13.4	-13.5	-11.7	-11.5	-8.7	-8.5
EL	-34.0	-33.5	-12.2	-12.4	-6.1	-6.5	46.3	-4.4
REP	55.8	53.8	13.8	12.9	3.1	2.9	-47.7	0.5
CORR	-24.4	-23.5	-13.1	-12.2	-7.6	-6.7	-6.8	-3.9
LD‡	-3.3	-2.9	-2.5	-2.3	-1.7	-1.6	-1.1	-1.1
Δ^{ref} †	1.9	1.7	0.5	0.6	-0.3	0.0	-0.6	-0.3

† $\Delta^{\text{ref}} = E_{\text{INT}}^{\text{calc}} - E_{\text{INT}}^{\text{ref}}$. ‡ LD contribution calculated with D4.

6.5.3. Performance of SQM Methods

In recent years, SQM methods have become increasingly popular due to significant improvements in accuracy and applicability.[96] Two widely used examples are the NDDO-based PMx[86,88] methods and the more recently developed extended tight-binding methods of the GFNn-xTB[93,94,96] family. The latter

proved to be in general more robust and accurate for structure optimization and noncovalent interactions. Recently, also a partially polarizable generic FF has been introduced, termed GFN-FF,[117] which is a promising, generally applicable candidate for a very efficient description of noncovalent interactions. In this section, the performance of the introduced SQM and FF methods is tested for the IONPI19 benchmark set. The statistical data are summarized in Figure 6.8.

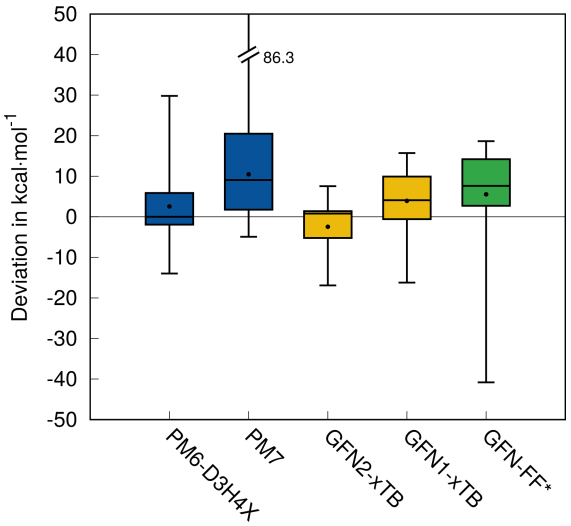


Figure 6.8.: Deviations of calculated ion–π interaction energies with SQM and FF methods for the IONPI19 set. The minimum deviation and maximum deviation for each data set are shown as range together with the first and third quartiles as the central box for each data set, the inter-quartile range contains 50 % of the data set. Additionally, the mean and median deviation are depicted as dots and vertical bars, respectively. The asterisk indicates that systems with wrong topology assignments were excluded for GFN-FF.

Out of the four tested SQM methods, GFN2-xTB is the best performer with an MAD of 4.7 kcal mol^{-1}, followed by GFN1-xTB (6.8 kcal mol^{-1} MAD). PM6-D3H4X and PM7 show larger deviations from the reference with an MAD of 7.9 and 18.8 kcal mol^{-1}, respectively. With GFN-FF, the topology assignment is initially wrong for the Li^+ containing systems **1** and **5**, where the coordination number of Li^+ is six rather than zero. Hence, these systems are excluded from the statistical evaluation indicated in Figure 6.8 by the asterisk. To circumvent wrong topology assignments, the topology file can be generated on the GFN-FF equilibrium structure. For the remaining 17 systems, an MAD of 11.9 kcal mol^{-1} is achieved. This somewhat larger deviation for the tested ion–π interactions can be attributed to the classical EEQ charge model in GFN-FF. In comparison to previous studies on mostly neutral NCI complexes,[273] the overall trend among the tested SQM and FF methods is comparable, but the absolute errors are much larger. This is mainly due to the large contribution of electrostatics to the total interaction energy of ion–π systems. Thus, it is not surprising, that GFN2-xTB performs best out of all tested SQM methods, as it contains a sophisticated multipole electrostatic model. The combination of sufficient accuracy with computational efficiency in GFN2-xTB is promising for large scale applications of ion–π interactions in biomacromolecular systems. Nevertheless, the description of electrostatic/induction interactions by SQM and FF methods can not reach the same accuracy as the tested DFAs.

6.5.4. Timing Comparison

The cost to accuracy ratio is evaluated for the best performing methods of the assessed levels of theory. Computational timings for the single-point (SP) calculations of **15** are shown in Figure 6.9 in combination with the respective MAD of the entire IONPI19 set. The wall times are given in seconds on a logarithmic scale and were calculated in parallel on ten CPU cores.

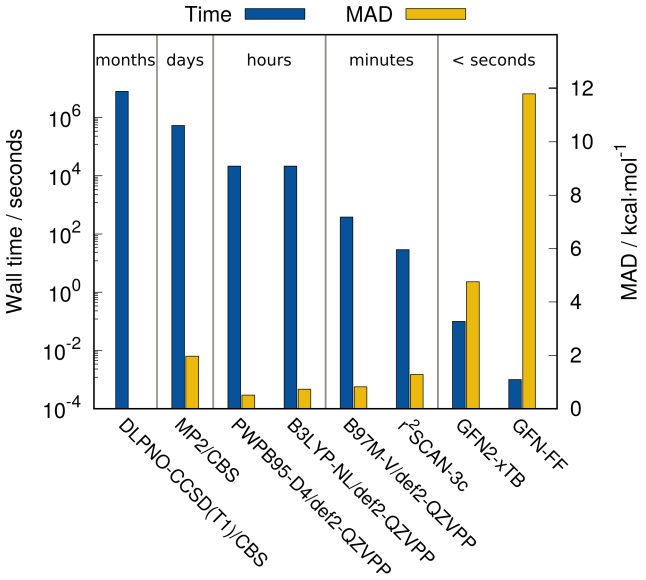


Figure 6.9.: Total wall time for the single-point energy of **15** calculated by the best performing methods of each theoretical level in parallel on ten Intel© Xeon E5-2660 v4 @ 2.00 GHz CPUs.

With 42 atoms, **15** is well suited as a representative for the IONPI19 set, where the average system size is about 32 atoms per molecule. The calculation of the reference values at the DLPNO-CCSD(T1)/*VeryTightPNO*/CBS(aug-cc-pVTZ/aug-cc-pVQZ) level of theory took about three months. RI-MP2/CBS(aug-cc-pVTZ/aug-cc-pVQZ) calculations lasted more than six days and resulted in an MAD of 1.9 kcal mol^{-1}. DFT methods show significant improvements in the total wall time compared to the MP2/CBS schemes. PWPB95-D4 and B3LYP-NL SP calculations in the large def2-QZVPP basis set converge within a few hours and show lower MADs than MP2/CBS. Here, the double-hybrid PWPB95-D4 is almost as fast as B3LYP-NL, because the MP2 part employing the RI approximation takes up only 5 % of the total wall time. Also, the PWPB95-D4 calculation requires one SCF iteration cycle less to converge than B3LYP-NL. The meta-GGA B97M-V SP calculation finishes within ten minutes and the MAD (0.8 kcal mol^{-1}) is still within chemical accuracy. r^2SCAN-3c shows almost the same deviation as B97M-V whilst being one order of magnitude faster. With MADs close to chemical accuracy, meta-GGA functionals yield in general the best cost-to-accuracy ratio. GFN2-xTB significantly reduces the computational wall time further to less than a second but the underlying approximations increase the MAD to 4.7 kcal mol^{-1}. GFN-FF as the only assessed FF method is yet two orders of magnitude faster than GFN2-xTB.

6.6. Conclusion

In this work, a comprehensive benchmark set was compiled for a wide range of ion–π interactions. This IONPI19 compilation represents a diverse set of (bio)chemically relevant molecules, also of larger size, and consists of 19 molecular structures that cover inter- as well as intramolecular interactions between anions/cations and π-conjugated systems. The IONPI19 set was used to benchmark various DFAs as well as HF, MP2, SQM, and FF methods. For all non-composite DFAs and HF, a large def2-QZVPP basis set was applied. In the context of DFT, the main focus was put on the effect of the self-interaction error and of London dispersion interactions for ion–π interactions. Second-order Møller–Plesset perturbation theory extrapolated to the complete basis set limit (MP2/CBS) and Hartree–Fock were evaluated as computationally more expensive but self-interaction error free competitors. Reference interaction and association energies were generated with high-level coupled cluster (CCSD(T)/CBS, W1-F12, W2-F12, and DLPNO-CCSD(T1)/*VeryTightPNO*/CBS) protocols.

First, the effect of different LD corrections was tested and the performance of different DFAs in combination with the D3, D4, and NL dispersion correction schemes were assessed. With a mean MAD of 0.9 kcal mol^{-1} each, the D4 and NL dispersion correction performed equally accurate. In comparison to its predecessor D3, the newly developed D4 model performed consistently better for each tested functional. The difference between these two schemes is mainly due to the inclusion of atomic partial charges in DFT-D4. The incorporation of three- and higher-body dispersion terms was found to have rather small effects. In general, the application of a dispersion correction is inevitable for the IONPI19 benchmark, as LD-uncorrected DFT underestimates ion–π interactions significantly, and the mean MAD amounts to 4.0 kcal mol^{-1}.

For the IONPI19 set the trend along Jacob's ladder functional classification scheme was mostly preserved among the tested combinations of DFAs and LD correction meaning that the average performance of (meta-)GGAs (1.0 kcal mol^{-1} MAD) was improved by hybrids (0.8 kcal mol^{-1} MAD), whereas the highest accuracy was reached by the double-hybrids (0.6 kcal mol^{-1} MAD). DH-DFAs reach in many cases an accuracy that is remarkably close to the high-level coupled cluster reference values but at up to two orders of magnitude lower computational cost than MP2/CBS. And, even more importantly, the double-hybrids are also significantly more accurate than MP2/CBS, which systematically and significantly overestimates ion–π interactions (MD = −1.9 kcal mol^{-1}, MAD = 2.0 kcal mol^{-1}). The best cost-to-accuracy ratio was obtained with the newly developed r^2SCAN-3c composite method, which yielded an accuracy close to meta-GGAs like TPSS in a much larger def2-QZVPP basis, whilst being one order of magnitude faster. It was found that the SIE has a relatively small effect on ion–π interactions. This is reflected in the fact that hybrid DFAs, which include Fock exchange to correct for the SIE, performed only slightly better than (meta-)GGAs. Energy decomposition analysis for the dissociation of ion–π complex **8** further revealed that the SIE of a GGA is rather small for the equilibrium geometry and may become significant only at larger intermolecular distances. SQM and FF methods were additionally tested for the IONPI19 set. The simpler description of electrostatic interactions in comparison to DFAs resulted in generally larger errors than obtained in previous studies on neutral systems, as electrostatic interactions are the major contribution of ion–π interactions. The best per-

forming method was GFN2-xTB, which takes into account anisotropic electronic effects by higher order multipole terms.

In conclusion, we generally recommend the use of DH-DFAs with the D4 dispersion correction in a large def2-QZVPP basis set for calculating reference interaction energies of larger (100-250 atoms) ion–π systems. DH-DFAs in combination with D4 extend the possibilities for generating reliable reference values for larger systems, which are essential for the development of low-cost methods to describe ion–π interactions in very large systems such as proteins. This conclusion only really becomes apparent, when larger systems are investigated with high-level references as in the presented IONPI19 benchmark set.

Acknowledgements

The German Science Foundation (DFG) is gratefully acknowledged for financial support through the priority program No. SPP 1807 “Control of Dispersion Interactions in Molecular Chemistry”. Sebastian Spicher thanks the “Fond der Chemischen Industrie (FCI)” for financial support.

Part IV.

Improved Vibrational Frequencies and Thermodynamics for Free Energy Calculations

To this point, geometries in gas phase and in solution, as well as noncovalent interaction energies, were studied in detail using GFN-FF and GFN*n*-xTB methods on a variety of large molecules. The important quantity to convert from energy to free energy is the thermostatistical contribution, which is yet missing. In *ab initio* or semiempirical calculations, thermodynamic quantities are most commonly obtained within the RRHO approximation (*cf.* Section 2.3.1). The required vibrational frequencies and the corresponding normal modes are calculated from second derivatives, either analytically or numerically from the first-order derivative, *i.e.*, the nuclear gradient. The polynomial increase in computation time brings DFT methods much sooner to their system size limit and hence, requires the application of simplified methods.

Part IV of this thesis goes beyond the description of geometries and gas phase energies and includes thermodynamic quantities in the description of molecules. GFN methods are applied and evaluated for the efficient computation of free energy contributions of large molecules and a new approach is presented for improved vibrational frequencies (see Figure 6.10). Solvation effects are treated implicitly and explicitly.

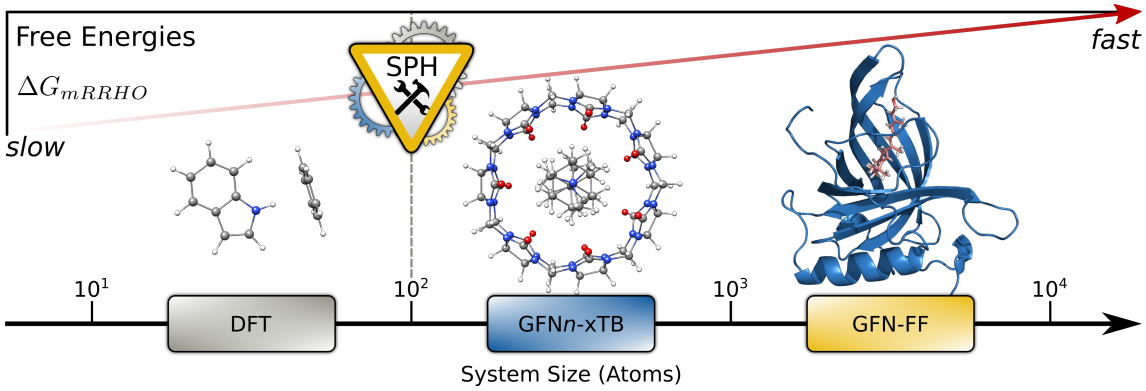


Figure 6.10.: Free energy calculations for different system sizes. For larger systems, efficient empirical methods need to be employed to derive the required thermodynamic quantities. SPH calculations help to increase the accuracy of approximated methods towards DFT accuracy.

In Chapter 7, the accuracy of GFN methods is first tested in the context of free energy contributions for standard benchmark sets of complexation reactions.[346] The deviation from DFT reference values is generally small and both methods, GFN2-xTB and GFN-FF, reach a high accuracy at a much lower computational cost. Based on this foundation, the system size is increased drastically (~5000 atoms) and association free energies are computed, *e.g.*, for the human serum retinol binding protein (RBP) encapsulating retinol (vitamin A) and rivaroxaban binding to factor Xa. For the latter example, the GFN-FF result is compared to previous QM studies.

A prerequisite for accurate thermodynamics within the RRHO approximation is usually a fully optimized input structure since the harmonic approximation is only valid in close proximity to the equilibrium geometry. For general, non-equilibrium structures, thermodynamic quantities are usually not accessible. In Chapter 8, a new method is developed for the computation of harmonic vibrational frequencies and thermodynamic contributions to the free energy within the modified RRHO approxima-

tion for general non-equilibrium geometries.[524] The method is termed single-point Hessian (SPH), in analogy to single-point energy (SPE) calculations, and can also be calculated at every position of the PES. The SPH approach is tested for supramolecular association reactions, MD simulations, reaction paths, and IR spectra. In comparison to conventional calculation on fully optimized geometries or unrelaxed structures, the SPH approach is in general superior.

In Chapter 9, the effect of explicit solvation on IR spectra and the underlying harmonic frequencies is investigated.[525,526] Here, a newly developed cluster algorithm, termed quantum cluster growth (QCG), is applied, in which explicit solvent molecules are added to energetically favorable positions around the solute molecule. The energetic screening is executed by the intermolecular force-field xTB-IFF[466] and subsequent cluster optimizations are performed by GFN2-xTB. Explicit solvation approaches account for local solute–solvent interactions, such as hydrogen bonds, which are necessary for qualitatively correct frequencies. The IR spectra calculated from QCG cluster show in general better agreement with experimental reference spectra than implicit approaches employing, *e.g.*, the COSMO model.

In general, the quality of harmonic frequencies calculated with GFN-FF and GFN*n*-xTB methods is comparable to that of *ab initio* DFT. With the GFN methods at hand, Part IV shows that the computation of association free energies are now feasible for new dimensions of molecular size, which was previously impossible at this high accuracy level.

7. Efficient Computation of Free Energy Contributions for Association Reactions of Large Molecules

Sebastian Spicher[a] and Stefan Grimme[a]

Received: 22 June 2020, Published online: 27 July 2020

Reprinted (adapted) with permission[†] from
Spicher, S.; Grimme, S. *J. Phys. Chem. Lett.* **2020**, *11*, 6606–6611.

DOI acs.jpclett.0c01930

Own manuscript contribution

- Conception of the study and choice of methods
- Performing all the calculations
- Interpretation of the results
- Writing the manuscript

[a]Mulliken Center for Theoretical Chemistry, Institute of Physical and Theoretical Chemistry, University of Bonn, Beringstraße 4, 53115 Bonn, Germany
[†]Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

Abstract Modern density functional theory (DFT) methods are capable of providing accurate association energies for supramolecular systems and even protein-ligand complexes. However, the calculation of the essential harmonic vibrational frequencies needed to correct to free energies is often computationally too demanding. In this work, the corresponding thermostatistical contributions are computed in the well established (modified) rigid-rotor-harmonic-oscillator approximation with structures and frequencies taken from low-cost quantum chemical methods, namely GFN2-xTB and PM6-D3H4. Additionally, a recently developed new general force-field (GFN-FF) is tested for this purpose. DFT reference values for 59 complexes composed of three standard noncovalent and supramolecular benchmark sets (S22, L7, S30L) are used in the evaluation. Overall the accuracy of the low-cost methods is remarkable with typical deviations of only 0.5–2 kcal mol^{-1} (5–10 %) from the DFT reference values. Especially the performance of the GFN force-field is promising considering the speedup of five and 2–3 orders of magnitude compared to DFT and GFN2-xTB, respectively. This opens new perspectives for computing thermodynamic properties of, *e.g.*, (bio-)macromolecules as shown exemplarily for the binding of retinol and rivaroxaban in protein complexes consisting of up to 4700 atoms.

7.1. Introduction

A common task of computational chemistry is to realistically model various chemical reactions and predict their outcome.[229,527,528] While those reactions are usually carried out at room temperature in solution, quantum mechanical (QM) calculations are primarily conducted for isolated molecules at zero Kelvin. To compare theory with experiment, additional corrections and computational steps are required. A common non-dynamical, super-molecule approach towards association free energies ΔG_a for the formation of a noncovalently interacting (NCI) complex at a given temperature employs Eq. 7.1

$$\Delta G_a = \Delta E_a + \Delta\delta G_{\text{solv}} + \Delta G_{\text{trvc}}\,, \tag{7.1}$$

where E_a is the total molecular gas phase energy, δG_{solv} is the corresponding free solvation energy, and G_{trvc} is the thermostatistical contribution from energy to free energy at 298 K accounting for translation, rotation, vibration and conformational degrees of freedom. The Δ implies the product/educt difference with appropriate stoichiometric factors. In principle, this scheme can be applied to any reaction but we restrict it here to noncovalently bound species, *i.e.*, typically A + B⟶ A···B (see Ref. 68 for further discussion). Out of all three terms in Eq. 7.1 it is G_{trvc} which is by far computationally the most demanding. Thus in this work, the performance of so-called low-cost methods is tested for its efficient calculation.

Employing accurate dispersion corrected DFT for the energies of single, geometry optimized structures of large supramolecular complexes, a scheme based on Eq. 7.1 was introduced in 2012.[119] Therein, G_{trvc} values were obtained in a low-frequency modified rigid-rotor-harmonic-oscillator approximation (mRRHO), also including the harmonic zero-point vibrational energy (ZPVE) and thermal/volume work corrections, but neglecting conformations for each species (*i.e.*, $G_{\text{trvc}} \approx G^{298}_{\text{mRRHO}}$). The physically sound rotor-type modification of the standard RRHO term mainly addresses the treatment of low-frequency

modes and their large (and numerically problematic) contribution to the entropy and has been successfully used as default in our group for years.

The reader should note that the first term in Eq. 7.1 is numerically often the largest, while the third one is of approximately the same order of magnitude (but different sign), and usually only the solvation contribution is smaller, at least for neutral systems. Hence, a reasonably accurate and robust, yet still fast standard treatment for $\Delta G_{\mathrm{mRRHO}}$ is mandatory for large molecules. Alternative methods for approximating ΔG_{trvc} (or ΔG_a) directly for the process are thermodynamic integration (TI)[529,530] or free energy perturbation (FEP)[531,532] mostly based on classical (force-field, FF) molecular dynamics simulations.[533,534] Apart from serious numerical problems in the DFT calculation of harmonic vibrational frequencies (HVF) due to the employed exchange-correlation integration grids,[535] the analytical (or numerical) computation of Hessian matrices by first-principles methods with reasonably large AO basis sets becomes prohibitively expensive.

The urgent need for faster methods becomes evident from Figure 7.1 showing computer timings for HVF calculations on a prototypical complex taken from our S30L benchmark set.[120] Calculations

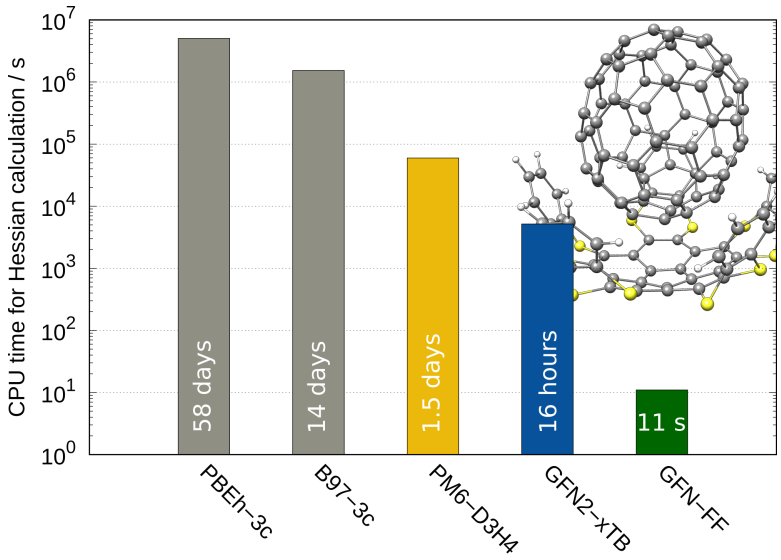


Figure 7.1.: CPU times given in seconds for the calculation of vibrational frequencies of pentakis(1,4-benzodithiino)-corannulene binding a C_{70} guest with PBEh-3c, B97-3c, PM6-D3H4, GFN2-xTB and GFN-FF methods. Computations were performed on a single Intel© Xeon E5-2660 v4 @ 2.00 GHz CPU.

were conducted with two DFT composite methods (hybrid functional PBEh-3c[231] and the GGA B97-3c[237]) and the here tested semiempirical quantum mechanical (SQM) methods GFN2-xTB,[93] PM6-D3H4 (mainly used for comparison),[86] as well as the recently presented general GFN-FF.[117] The well-established GFN methods were developed with a special focus on a good description of Geometries, Frequencies, and Noncovalent interactions (GFN) as reviewed recently in some detail.[96] Also note, that both DFT methods were specially developed with small to medium-sized (truncated) AO basis sets and hence already represent a best-practice compromise between accuracy and computational cost.

For the complex shown in Figure 7.1 consisting of 150 atoms, the HVF calculations took 58 days with

PBEh-3c and 14 days with B97-3c, respectively, which is unacceptable even when parallel computers are applied. The PM6-D3H4 SQM calculation still took almost two days, whilst GFN2-xTB finished within 16 hours. GFN-FF on the other hand took only 11 seconds for the same calculation. One of the big advantages of GFN-FF is that it can be applied out-of-the-box to all elements of the periodic table up to radon, including metal containing complexes. Hence, this approach for G_{mRRHO} computations may open completely new possibilities. Therefore, the important question arises, how accurate are the resulting free energy contributions, or in other words what is lost for the tremendous speed-up of two to five orders of magnitude? Surprisingly, this problem has rarely been investigated thoroughly in the literature so far (see *e.g.*, Ref. 536).

7.2. Results and Discussion

7.2.1. Benchmark Sets

To address this question thoroughly, G_{mRRHO} values are calculated for three standard benchmark sets, S30L, S22,[537] and L7.[538] They cover various noncovalent binding motifs, from hydrogen bonds to π–π stacking and ionic interactions. The system size varies from six atoms (water dimer in S22) to 200 atoms for the supramolecular complexes in S30L. The reference values are calculated as the arithmetic mean of the two DFT methods, PBEh-3c and B97-3c. The corresponding standard deviation is taken as an uncertainty measure of the reference data and denoted as the error bar in the plots. All thermostatistical values are computed within the mRRHO approximation at 298 K. A standard rotational cut-off value of 50 cm^{-1} is chosen in the mRRHO treatment (see Ref. 119 for details) and small imaginary frequencies below 20 cm^{-1} (mostly methyl group torsions) are set real (multiplied by $-i$). Note, that the G_{298} values by the rotational partition function also depend on the molecular geometry, meaning that in addition to the HVF, the accuracy of the optimized equilibrium (minimum energy) structure is tested as well. Further technical details are given in Appendix A7.

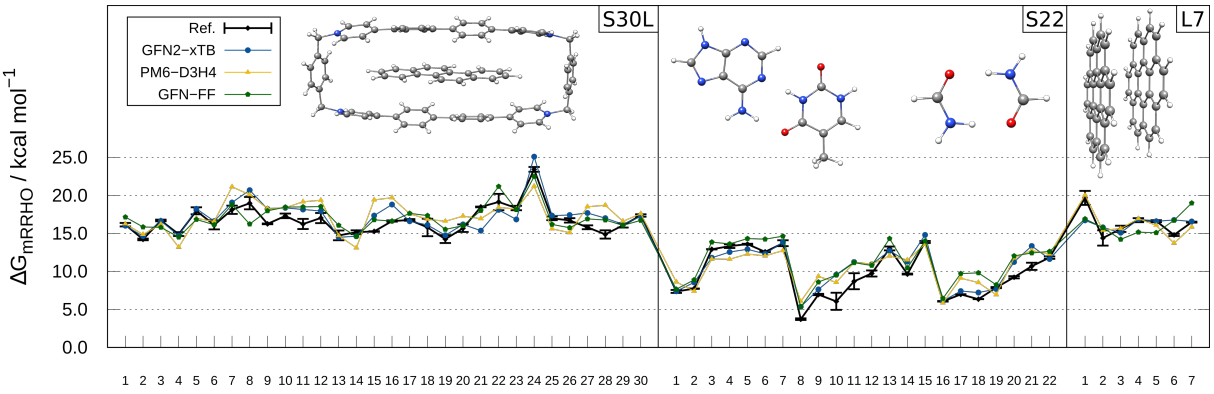


Figure 7.2.: Thermostatistical $\Delta G_{\mathrm{mRRHO}}$ values (in kcal mol^{-1}) to the association free energy of the complexes taken from the S30L, S22, and L7 benchmark sets.

The results for the thermostatistical contributions to binding for all 59 NCI complexes are shown in Figure 7.2. For each benchmark, structural examples illustrate the various binding motifs. Note again

that the ΔE_a term is not included at this point. Overall, the three tested low-cost methods provide a very good agreement with the (averaged) DFT reference values, and importantly, no outliers are observed. Their performance is evaluated statistically by the mean deviation (MD), mean absolute deviation (MAD), and root-mean-square deviation (RMSD) given in Table 7.1. Due to the averaged

Table 7.1.: Statistical evaluation (in kcal mol^{-1}) of the thermostatistical contribution $\Delta G_{\mathrm{mRRHO}}$ for the 59 complexes[a]. The *rel*MAD values are given in %.

		S30L	S22	L7	all sets
DFT	MD	±0.24	±0.12	±0.22	±0.04
	MAD	0.36	0.23	0.42	0.34
	RMSD	0.47	0.38	0.55	0.47
	*rel*MAD	2.1	2.5	2.7	2.4
GFN2-xTB	MD	-0.50	0.71	0.21	0.14
	MAD	0.98	1.03	0.95	0.99
	RMSD	1.27	1.39	1.28	1.31
	*rel*MAD	5.6	10.3	5.9	7.3
PM6-D3H4	MD	-0.70	0.67	0.10	0.02
	MAD	1.57	1.46	0.55	1.19
	RMSD	1.92	1.55	0.66	1.38
	*rel*MAD	8.8	15.0	3.5	9.1
GFN-FF	MD	-0.40	1.40	0.21	0.40
	MAD	1.06	1.40	1.86	1.44
	RMSD	1.25	1.71	1.94	1.63
	*rel*MAD	6.1	13.7	11.4	10.4

[a] The statistical measures are given with respect to the average DFT reference for GFN2-xTB, PM6-D3H4 and GFN-FF. For the DFT reference, average values for PBEh-3c and B97-3c are taken and their RMSD of 0.47 kcal mol^{-1} is taken as an uncertainty estimate.

reference values obtained from two reliable density functionals, an uncertainty estimate can be given for each value. The average of all complexes is only 0.34 kcal mol^{-1} (RMSD of 0.47 kcal mol^{-1}) but definitely not negligible. Roughly speaking this is at best what one can expect from the tested low-level methods. Vibrational anharmonicity effects not considered here are probably of the same order of magnitude.[68] For the whole set, GFN2-xTB performs best with an MAD of 0.99 kcal mol^{-1}, followed by PM6-D3H4 with an MAD of 1.19 kcal mol^{-1}. Both SQM methods yield rather accurate results across the entire test set with a maximum absolute deviation of 3.5 and 2.5 kcal mol^{-1}, respectively. GFN-FF also yields good results with an MAD of only 1.44 kcal mol^{-1}, which is excellent for a FF and comparable to the much more elaborated SQM methods. The performance of the universal force-field (UFF[114]) which is the only available competitor to GFN-FF is clearly worse (MAD=2.05 kcal mol^{-1}, *rel*MAD = 17 % and partially distorted structures, see Appendix A7). The small MD and RMSD measures of all low-cost methods indicate robustness and the absence of any systematic errors for this wide variety of binding motifs. A similar picture emerges when relative deviations are considered (*rel*MAD in Table 7.1). The reference has an uncertainty of about 2 % for $\Delta G_{\mathrm{mRRHO}}$ for all three sets. This is closely approached by GFN2-xTB, yielding only about 6 % error for S30L and L7 and an overall small *rel*MAD value of 7.3 %. For the highly important S30L benchmark, GFN-FF performs similarly well with a 6 % error on average

thus clearly outperforming PM6-D3H4. Both, GFN-FF and PM6-D3H4 yield somewhat less accurate values for the small S22 complexes such that a total *rel*MAD value of about 10 % is obtained which is, nevertheless, sufficiently small to be useful in practice.

Furthermore, we consider the somewhat simpler enthalpic contribution $\Delta H_{\mathrm{mRRHO}}$ including the ZPVE. This quantity is of course part of the already discussed $\Delta G_{\mathrm{mRRHO}}$ values but in some applications[539] it is of practical relevance in itself. The results are illustrated in Figure 7.3 and the statistical evaluation is given in Appendix A7. Similar to the case of the free energies, the reference values are well reproduced

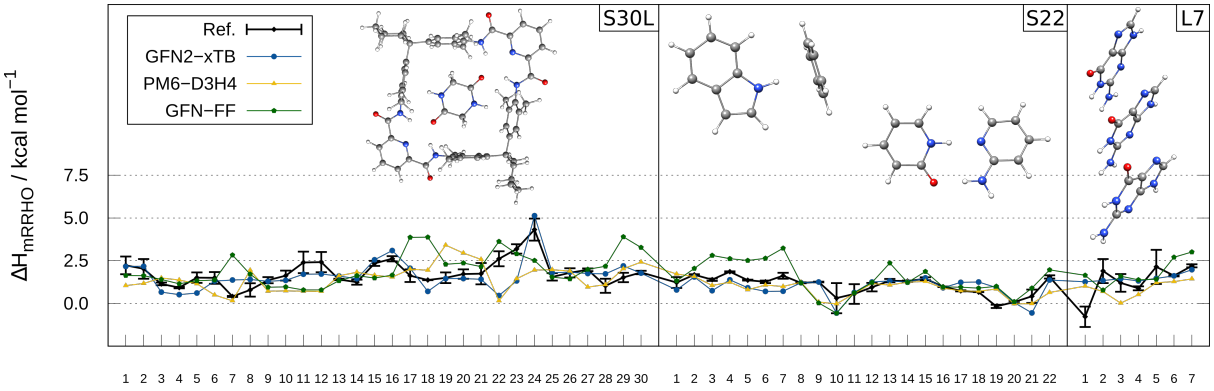


Figure 7.3.: Enthalpic contribution $\Delta H_{\mathrm{mRRHO}}$ (kcal mol^{-1}) to the association free energy of the complexes taken from the S30L, S22, and L7 benchmark set.

by the low-level methods and no significant outliers occur, confirming the above observations. Because the ΔH_{298} values are absolutely smaller compared to the ΔG_{298} ones, all deviations and statistical measures diminish. The order of their accuracy is for the low-cost methods the same as before, with GFN2-xTB being closest to the reference with an MAD of 0.49 kcal mol^{-1}, followed by PM6-D3H4 with an MAD of 0.74 kcal mol^{-1} and GFN-FF with an MAD of 0.83 kcal mol^{-1}.

7.2.2. Protein-Drug Binding

Finally, three very difficult examples are discussed which were deliberately chosen to explore the limits of the low-cost methods. Complexes involving strong hydrogen bonds are practically important and are specifically problematic for FFs because electronic many-body (polarization effects) play an important role.

In the first case study, we test the accuracy of the G_{mRRHO} values for the formation of water clusters. We investigate the association of $(H_2O)_4$ to $(H_2O)_8$ and their formation of the dodecahedron structure of $(H_2O)_{20}$. All starting geometries were taken from the WATER27 benchmark[540] and are illustrated in Figure 7.4A. Note that here and in the last two examples we additionally discuss total (free) association energies in addition to the mRRHO contributions in order to put them and their error into better perspective. The reaction to $(H_2O)_{20}$ is mainly discussed. All value are given in Table 7.2 The reference ΔG_{298} values are again obtained by PBEh-3c and B97-3c, but not averaged here to indicate their spread (1.6 kcal mol^{-1}) explicitly. Compared to the (average) reference value (28.0 kcal mol^{-1}) GFN2-xTB is spot-on. Even in this very difficult scenario, including, *e.g.*, important induction and other many-body

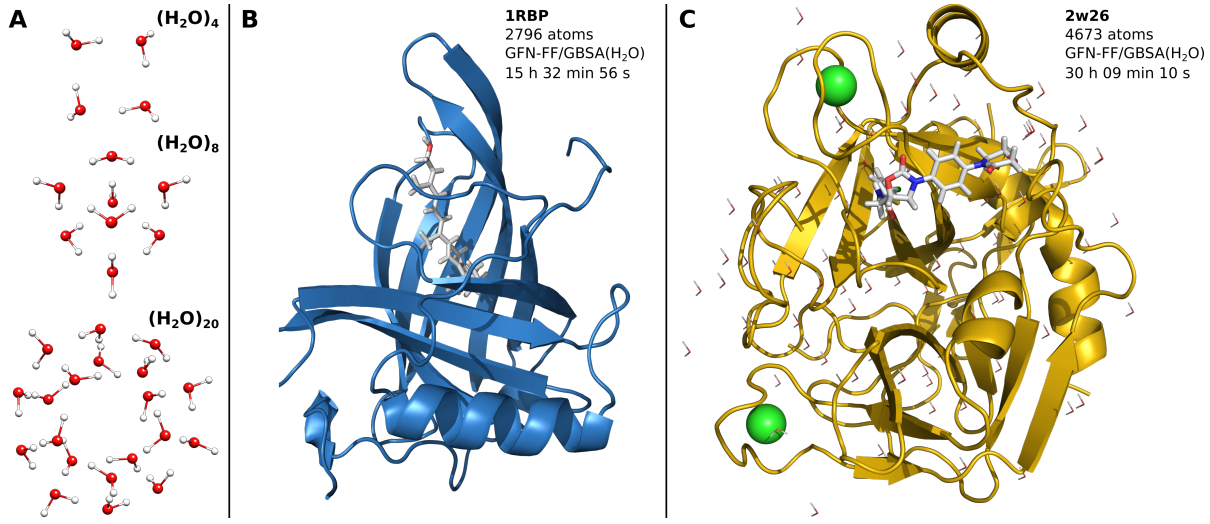


Figure 7.4.: (A) The three investigated water clusters from the WATER27 benchmark set. (B) Human serum retinol binding protein (RBP) encapsulating retinol (Vitamin A). (C) Factor Xa in complex with rivaroxaban in the presence of crystal water molecules. The PDB identifier is given in bold.

effects, GFN-FF yields a result (31.8 kcal mol^{-1}) relatively close to the DFT reference. The fact that even this error of 2–3 kcal mol^{-1} is not alarming becomes clear if the variations in the association energies ΔE_a are considered. The DFT methods differ by almost 5 kcal mol^{-1}, indicating that the accuracy limiting factor in such applications is not the use of approximate mRRHO contributions from the low-cost methods, but the highest theoretical level one can afford for the bare interaction energy. The performance of the two GFN methods (and the FF in particular) is striking for the ΔE_a values with only minor deviations from the respective DFT values. PM6-D3H4 on the other hand shows difficulties in the description of the $(H_2O)_{20}$ cluster as both, the $\Delta G_{\mathrm{mRRHO}}$ contribution and the ΔE_a value deviate clearly from the DFT values. The reaction to $(H_2O)_8$ on the other hand is reasonably described by PM6-D3H4, yet still worse than with the GFN methods regarding ΔE_a and $\Delta G_{\mathrm{mRRHO}}$ indicating higher robustness of the latter.

Table 7.2.: $\Delta G_{\mathrm{mRRHO}}$, $\Delta H_{\mathrm{mRRHO}}$ and ΔE_a values (in kcal mol^{-1}) for two water cluster formation reactions.

		PBEh-3c	B97-3c	GFN2-xTB	PM6-D3H4	GFN-FF
2 $(H_2O)_8$ + $(H_2O)_4$ → $(H_2O)_{20}$	$\Delta G_{\mathrm{mRRHO}}$	29.3	27.7	28.0	55.7	31.8
	$\Delta H_{\mathrm{mRRHO}}$	2.1	2.1	2.7	16.4	6.4
	ΔE_a	-33.4	-28.8	-24.4	-6.6	-35.0
2 $(H_2O)_4$ → $(H_2O)_8$	$\Delta G_{\mathrm{mRRHO}}$	16.6	17.6	17.1	23.8	20.6
	$\Delta H_{\mathrm{mRRHO}}$	2.3	2.2	1.8	2.4	3.1
	ΔE_a	-20.5	-19.6	-19.9	-12.9	-13.4

In the last two examples, we want to highlight the potential of low-cost methods, especially GFN-FF, in real-life biochemical applications. As a showcase the association free energy of the human serum

retinol binding protein (RBP) with Vitamin A[541] is calculated by GFN-FF employing the same static super-molecule approach. The system is illustrated in Figure 7.4B and consists of 2796 atoms and hence the calculation of $\Delta G_{\mathrm{mRRHO}}$ is only affordable at the GFN-FF level of theory. The entire computation, consisting of full, unconstrained geometry optimization starting from the PDB structure and HVF calculation took roughly 15 hours on a single Intel© Xeon E5-2660 v4 @ 2.00 GHz CPU with GFN-FF. The total association free energy is calculated according to Eq. 7.1 employing our implicit GBSA continuum solvation model[96] for water in all computational steps (geometry optimization and HVF calculation). The results are shown in Table 7.3. With GFN-FF, the association energy without the solvation contribution amounts to −30.7 kcal mol^{-1}. Adding the $\Delta\delta G_{\mathrm{solv}}$ term of 0.6 kcal mol^{-1} leads to a value of −30.1 kcal mol^{-1}. For comparison, a value of −29.6 kcal mol^{-1} is obtained with GFN2-xTB. The GFN-FF calculation of the vibrational frequencies with default technical settings yields only a single imaginary frequency of 5 cm^{-1} introduced by the finite grid size for the surface integration part in GBSA. No imaginary frequencies were obtained in gas phase computations. This technical robustness of the implementation in the *xtb* program[243] is not self-evident. Due to the mentioned numerical grid error for the exchange-correlation energy, such a calculation is technically hardly possible with DFT methods. GFN-FF yields a thermostatistical contribution $\Delta G_{\mathrm{mRRHO}}$ of 17.2 kcal mol^{-1} resulting in a predicted overall free association energy ΔG_a of −12.9 kcal mol^{-1}. This value is close to the affinities observed for typical protein-drug complexes of −5 to −15 kcal mol^{-1}.[542] Although this value is rather preliminary lacking, *e.g.*, any conformational contributions, it seems to indicate a relatively small error for all components in Eq. 7.1 with our new GFN-FF(GBSA) approach. A related protein-ligand study mainly based on Hartree–Fock (HF) and DFT calculations was performed in 2017[375] on cutouts of Factor Xa in complex with rivaroxaban.[543] With GFN-FF at hand, an unconstrained optimization of the full protein-drug complex is now possible including two calcium ions and important crystal water molecules (see Figure 7.4C). The result obtained without any empirical adjustment for a large system with complicated interactions is remarkable from a quantum chemistry perspective. With a calculated association free energy of −26.3 kcal mol^{-1} (see Table 7.3), the GFN-FF result is just about 14 kcal mol^{-1} off the experimentally measured value of −12.5 kcal mol^{-1}. The ΔE_a and $\Delta\delta G_{\mathrm{solv}}$ contributions are in good agreement with the calculated values from Ref. 375 of −50 to −70 kcal mol^{-1} for ΔE_a and 20 to 40 kcal mol^{-1} for $\Delta\delta G_{\mathrm{solv}}$, respectively. With very tight convergence thresholds for the geometry optimization, only a single imaginary frequency of 6 cm^{-1} remains, yielding a thermostatistical contribution of 22.5 kcal mol^{-1} also closely agreeing to the previous results.

Table 7.3.: Computed association free energy (in kcal mol^{-1}) for human serum retinol binding protein (RBP) with retinol (Vitamin A) and Factor Xa with rivaroxaban by GFN-FF.

	ΔE_a	$\Delta G^{298}_{\mathrm{mRRHO}}$	$\Delta\delta G_{\mathrm{solv}}(H_2O)$	ΔG_a
Retinol @ 1RBP	-30.7	17.2	0.6	-12.9
Rivaroxaban @ 2W26	-83.8	22.5	35.0	-26.3

7.3. Conclusion

In this work, we investigated the performance of low-cost SQM and FF methods for the calculation of the rigid-rotor-harmonic-oscillator contributions to the enthalpy and entropy parts of free association energies for various noncovalently bound complexes. The initial question was whether low-cost methods, whose computational efficiency is absolutely necessary for large molecules, can approach the accuracy of DFT methods for the calculation of $\Delta G_{\mathrm{mRRHO}}$. In a multi-level modeling scheme, one would like to replace the most costly DFT steps of geometry optimization and frequency calculation with corresponding SQM or even FF ones which can save up to five orders of magnitude computation time. Throughout all assessed sets, GFN2-xTB, PM6-D3H4, and GFN-FF show accurate results with MAD values < 1.4 kcal mol^{-1} and relative deviations of 5–10 % for $\Delta G_{\mathrm{mRRHO}}$. The GFN methods also perform excellently for the difficult case of water cluster formation. As an outlook for future applications, the association free energies of two protein-ligand complexes were successfully computed with GFN-FF.

In conclusion, one may reformulate the initial question and ask if one should "waste" a lot of computer time (and electrical power) for the computation of a not very system dependent property such as $\Delta G_{\mathrm{mRRHO}}$ by costly high-level DFT methods instead of using the available robust SQM or FF alternatives. Under the premise that the low-level methods can correctly describe the basic noncovalent interactions (which they mostly do), the answer is clearly "no".

Acknowledgments

This work was supported by the DFG in the framework of the "Gottfried-Wilhelm-Leibniz" prize to Stefan Grimme. Sebastian Spicher thanks the "Fond der Chemischen Industrie (FCI)" for financial support. The authors thank M. Bursch and A. Hansen for helpful discussions.

8. Single-Point Hessian Calculations for Improved Vibrational Frequencies and Rigid-Rotor-Harmonic-Oscillator Thermodynamics

Sebastian Spicher[a] and Stefan Grimme[a]

Received: 18 December 2020, Published online: 8 February 2021

Reprinted (adapted) with permission[†] from
Spicher, S.; Grimme, S. *J. Chem. Theory Comput.* **2021**, *17*, 1701−1714.

DOI acs.jctc.0c01306

Own manuscript contribution

- Co-development of the SPH procedure
- Implementation in the *xtb* code
- Performing all the calculations
- Interpretation of the results
- Writing the manuscript

[a]Mulliken Center for Theoretical Chemistry, Institute of Physical and Theoretical Chemistry, University of Bonn, Beringstraße 4, 53115 Bonn, Germany
[†]Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

Abstract The calculation of harmonic vibrational frequencies (HVF) to interpret IR spectra and to convert molecular energies into free energies is one of the essential steps in computational chemistry. A prerequisite for accurate thermostatistics so far was to optimize the molecular input structures in order to avoid imaginary frequencies, which inevitably leads to changes in the geometry if different theoretical levels are applied for geometry optimization and frequency calculations. In this work, we propose a new method termed single-point Hessian (SPH) for the computation of HVF and thermodynamic contributions to the free energy within the modified rigid-rotor-harmonic-oscillator approximation, for general non-equilibrium molecular geometries. The key ingredient is the application of a biasing potential given as Gaussian functions expressed with the root-mean-square deviation (RMSD) in Cartesian space in order to retain the initial geometry. The herein derived theory is generally applicable to quantum mechanical (QM), semiempirical QM (SQM), and force-field (FF) methods. Besides a detailed description of the underlying theory, including the important back-correction of the biased HVF, the SPH approach is tested for reaction paths, molecular dynamics snapshots of crambin, and supramolecular association free energies in comparison to high-level DFT values. Further, the effect on IR spectra is investigated for organic dimers and transition metal complexes revealing improved spectra at the low theoretical levels. On average DFT reference free energies are better reproduced by the newly developed SPH scheme than by conventional calculations on freely optimized geometries or without any relaxation.

8.1. Introduction

Harmonic vibrational frequencies (HVF) represent characteristic properties of molecular systems that are used for the calculation of infrared (IR)[544–546] and Raman spectroscopy[547,548] and, maybe even more importantly, are applied in statistical thermodynamics to derive the ro-vibrational enthalpic and entropic contributions to the Gibbs free energy G.[68,69,542] As statistical thermodynamics conveniently enable the transition from energy E to G as given in Eq. 8.1, HVF have a major contribution in the comparison of theory and experiment[21,457,549,550] and are commonly computed in quantum chemistry. Here, $G(T)$ refers to an individual molecule in a definite minimum (equilibrium) energy conformation at standard conditions in the gas phase

$$G(T) = E_{\text{gas}} + G_{\text{trv}}(T)\,, \tag{8.1}$$

where E_{gas} is the total molecular gas phase energy, and G_{trv} is the thermostatistical correction from E to G at absolute temperature T, accounting for translation, rotation, and vibration degrees of freedom (DOF), also including the zero-point vibrational energy (ZPVE) and thermal/volume work corrections. In this work, we will focus on the latter term (G_{trv}) and its computation for general non-equilibrium (R_e) structures. For the transition from gas phase to solution, the free solvation energy $\delta G_{\text{solv}}(T)$ needs to be added, which is commonly done in the framework of continuum solvation models,[551–555] but not further considered here explicitly. The indicated temperature dependence is omitted in the following.

In quantum mechanics (QM) and partially also in force-field (FF) methods, free energy contribu-

tions are often calculated within the rigid-rotor-harmonic-oscillator (RRHO) approximation[536,556–558]. Herein, vibrational frequencies and the corresponding normal modes are obtained within the harmonic-oscillator (HO) approximation by a Taylor expansion around the equilibrium geometry, which is furthermore used to compute the rotational contributions. Accordingly, until now, a prerequisite for accurate HVF and derived thermodynamics was a molecular structure, which represents a stationary point (local minimum or transition state) on the respective potential energy surface (PES).[120,559–561] Otherwise, spurious imaginary, as well as inaccurate low-lying, HVF may arise for geometries, that are not fully optimized by the applied theoretical method for the PES. The resulting loss of (real) vibrational DOFs and errors in low-lying HVF causes significant flaws for the derived G_{trv} values (see Section 8.4 for numerical examples). This is evident from Eq. 8.2, where the vibrational contribution to the entropy is expressed as[536]

$$S_{\mathrm{vib}} = R \sum_{i}^{n_{\mathrm{vib}}} \left(\frac{h\nu_i}{kT} \frac{1}{e^{h\nu_i/kT}} - ln\left(1 - e^{-h\nu_i/kT}\right) \right) . \tag{8.2}$$

Here, the sum runs over all vibrational DOFs (n_{vib}), N is Avogadro's number, h and k are Planck's and Boltzmann's constants, and ν_i are the respective HVF. Because the last term in Eq. 8.2 approaches infinity for $\nu \rightarrow 0$, small changes in the low-lying HVF are causing large changes in S_{vib} and G_{trv}.

As the HO approximation is notoriously inaccurate for the description of low-lying vibrational modes and they are furthermore heavily affected by numerical noise in the quantum chemical calculations (DFT grids in particular[535]) a modified scheme termed mRRHO from now on was already introduced in 2012[119] by one of us. It was developed with a focus on the thermodynamics of large supra-molecular complexes treated by dispersion corrected density functional theory (DFT), but is generally applicable and our group default since then. Therein, G_{trv} values are obtained in a low-frequency modified RRHO version where a physically sound and non-empirical rotor-type approximation is applied. Importantly, it yields a finite S_{vib} even for $\nu = 0$. The mRRHO treatment of low-frequency modes as internal rotations and thus, correcting for their large (and numerically problematic) contribution to the entropy, is practically done by an interpolation between rotational and HO approximations. With the help of a standard switching function and a corresponding rotor cutoff value (typically set to 50 cm^{-1}), S_{vib} is obtained mostly from a proper rotational description below this value while for large frequencies, effectively the pure vibration formula is applied. This established mRRHO procedure lays the foundation for the calculation of thermostatistical contributions for general non-equilibrium structures because it eliminates the wrong and practically relevant $\nu \rightarrow 0$ asymptotics. Note, that chemically interesting, flexible systems (*e.g.*, drugs) have typically many HVF in the 10–20 cm^{-1} regime resulting in $G_{\mathrm{v}}(298)$ values on the order of −1 to −2 kcal mol^{-1} per mode.

The need for low-frequency corrections in the HO approximation for non-equilibrium geometries is further illustrated in Figure 8.1, where a prototypical double-well potential describes some chemical reaction path. Note that Hessian matrices and harmonic frequencies can of course be computed for any point on a PES but here we specifically focus on their physical meaning and practical use. Along the transition from one minimum x_1 to the other x_2, the force constant (FC), which is proportional to $\frac{\partial^2 f(x)}{\partial x^2}$, first leaves the harmonic region. Close to the turning point (TP) of the function, FC (and ν) approaches zero and becomes negative beyond. In the standard RRHO approximation, the vibrational entropy

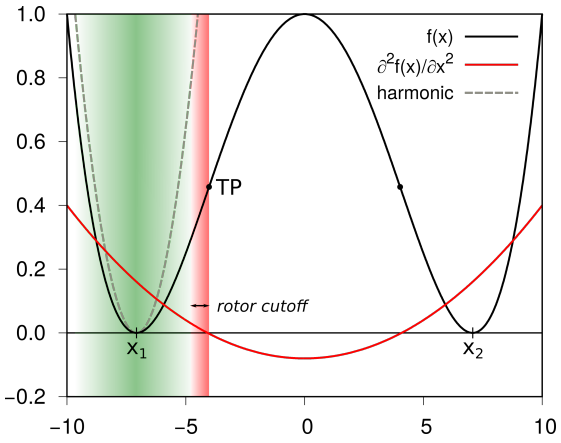


Figure 8.1.: Double well potential f(x) (black) and its second derivative (red) resembling the respective force constant (in arbitrary units). Beyond the harmonic approximation regime (grey), the force constant approaches zero and becomes negative after the turning point (TP) leading eventually to imaginary frequencies. The red-shaded area denotes the regime, where the modifications in the mRRHO scheme are most effective.

approaches infinity at the TP (see Eq. 8.2), while in mRRHO, it stays finite and is applicable up to the TP. The appearance of imaginary frequencies after the TP indicates that the (partially) unoptimized system resides in a transition state,[51,52] even though this is not necessarily the case. However, it is common practice that only real HVF are considered in Eq. 8.2,[68,562], which we follow also here in a slightly modified form. This analysis suggests that calculating the Hessian in the mRRHO framework yields meaningful G_{trv} values as long as the system is in the harmonic regime (but not necessarily at R_e). The proposed method, which we call for obvious reasons single-point Hessian (SPH) is outlined below in detail. It tries to increase this so-called "trust-region" in an absolutely automatic manner thereby minimizing the number of spurious imaginary frequencies and increasing the reliability of G_{trv}.

From another point of view, one can argue that for a multi-dimensional model PES composed of a simple sum of HO potentials, the computed frequencies are completely independent of the point where the Hessian is evaluated. Hence, in real anharmonic systems,[563,564] the dependence of the HVF on this evaluation geometry is stronger for shallow and more anharmonic modes than, *e.g.*, for stiff stretching vibrations. In fact, standard Hessian calculations on non-R_e structures often yield only many low-frequency imaginary modes, which is problematic for the calculation of S_{vib}.

Non-equilibrium structures are commonly used in quantum chemistry for energy calculations, *i.g.*, a lower theoretical level method Y is applied in the computationally demanding geometry optimization procedure, while a high theoretical level method X is taken only for a final energy calculation on the lower level structure. This is indicated by the standard $X//Y$ nomenclature. Along the same line of thought, one may wish to replace the very costly Hessian calculation at the high X level with a cheap calculation with method Y (see Ref. 346 for a recent thorough evaluation of this approach). Normally, this requires re-optimization of the $R_e(X)$ structure at the Y level as illustrated in Figure 8.2 Although a re-optimization restores the validity of the HO approximation, and hence in the ideal case no spurious

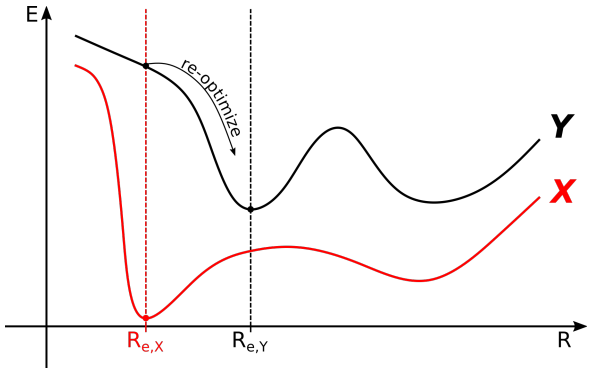


Figure 8.2.: One dimensional schematic potential energy diagram of a molecule calculated at two different levels of theory, X and Y, which differ in their equilibrium geometries. Here, X refers to high-level reference method (usually DFT) and Y represents a lower theoretical level such as SQM or FF methods.

imaginary frequencies occur, it may lead to a distortion of the geometry in an unwanted direction. In the extreme but commonly found case for larger molecules, the low-level method produces a qualitatively different structure (conformation), and in fact, this observation initiated our work. This difference in structure causes an error in the calculation of the rotational entropy,[68] which only depends on the Cartesian coordinates (and atomic masses) according to

$$S_{\text{rot}} = R\left(\frac{3}{2} + ln\left(\frac{\sqrt{\pi}}{\sigma}\left(\frac{8\pi^2 kT}{h^2}\right)^{\frac{3}{2}}\sqrt{I_1 I_2 I_3}\right)\right), \tag{8.3}$$

where the actual geometry enters directly via the moment of inertia I_k and the symmetry number σ. Calculation of geometries and free energies at different levels is thus a balancing act. Retaining the reference geometry at level X may lead to imaginary HVF at low level Y, which vanishes upon re-optimization, but changing the desired geometry and in turn introducing an error in S_{rot}. According to our experience in the field of conformational analysis of large molecules,[565] such effects can be considerable on the order of 0.5 kcal mol^{-1} for a relative conformational free energy at room temperature. Hence, a method is still needed which modifies the PES of method Y to yield the same R_e structure as method X, thereby reducing the number of imaginary modes significantly (at best to zero), but still providing the correct S_{rot} (of method Y). So far, to the best of the author's knowledge, only methods exist, which tackle the issue of partially optimized systems by considering subblocks of the Hessian matrix in the so-called partial Hessian vibrational analysis (PHVA).[566,567]

In this work, we therefore, introduce the new concept of the single-point Hessian. SPH calculations employ a biasing root-mean-square-deviation (RMSD) based constraining potential[220] in order to retain the target geometry. The systematic shift of the HVF caused by the modification of the PES due to the biasing potential is subsequently removed approximately (but rather accurately) by individual frequency scaling, giving access to accurate thermostatistical contributions for general non-equilibrium geometries with low-level methods.

As the analytical or numerical computation of HVF by *ab initio* methods[558] becomes prohibitively expensive for large molecules, semiempirical QM (SQM)[86,96,568] and force-field (FF)[201,569] methods present faster but still reasonable candidates for SPH treatments. Especially the GFN*n*-xTB methods[93,94] and the GFN force-field (GFN-FF)[117] have proven to yield good G_{mRRHO} vales in comparison to DFT calculations. They can be used for molecular sizes up to a few thousand atoms enabling the calculation of free energies for huge noncovalently interacting (NCI) complexes, and biomacromolecules.[346] The prime application of SPH calculations is thus in so-called multi-level approaches, which are common practice in computational chemistry.[570] Here, the theoretical level differs from single-point energies (SPE, cheap), over geometry optimizations (first derivative, medium effort), to HVF calculations (second derivative, expensive). Hence, the high-level (*X*) optimized reference structure represents a non-equilibrium geometry at the lower level (*Y*) at which the thermostatistical contributions are calculated. Exemplary workflows thereof are illustrated in Figure 8.3 for differently sized systems. For small systems < 50–

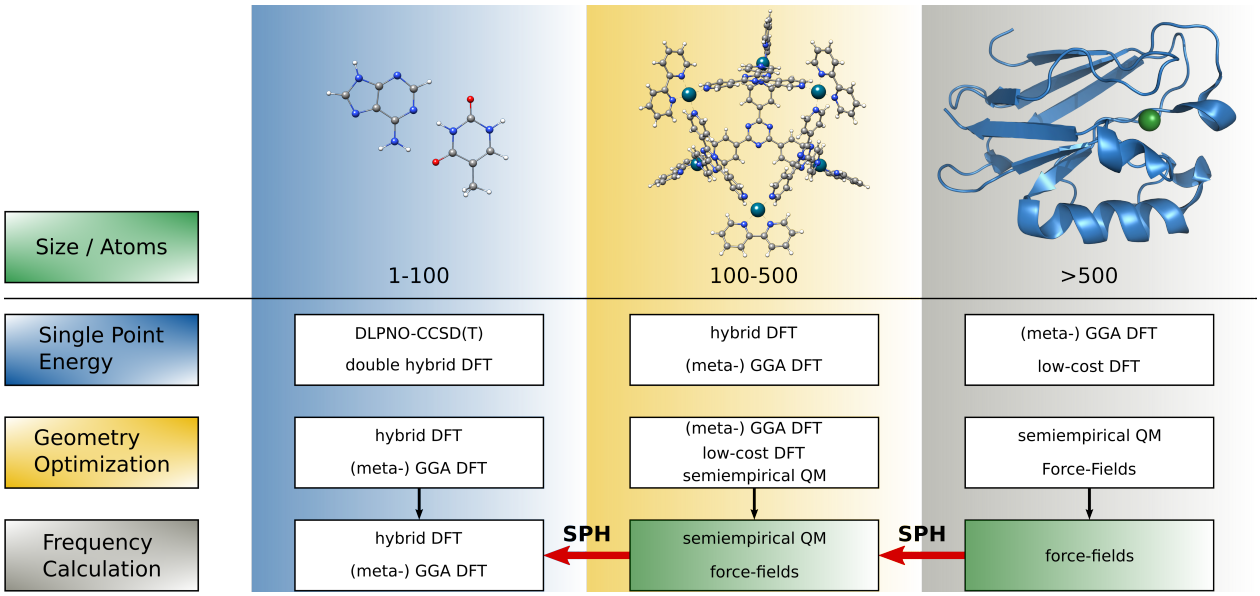


Figure 8.3.: Multi-level approach for the calculation of free energies consisting of three steps: geometry optimization, single-point energies, and the calculation of the free energy from HVF in the mRRHO framework. For small systems (< 50–100 atoms) all three steps are feasible by high level *ab initio* (mostly DFT) calculations, whereas for medium to large sized molecules SQM or FF methods have to be applied for geometries and frequencies.

100 atoms, all properties can be calculated routinely by *ab initio* QM methods.[334] Highly accurate SPE are feasible by DLPNO-CCSD(T)[147] or density functional theory (DFT) at, *e.g.*, double hybrid level.[571] Geometry optimizations, as well as HVF calculations, are still practical at the hybrid-GGA level,[166,172] while for medium to large molecules (100–500 atoms), however, the situation changes. Here, (low-cost) DFT[231,237] is maximally applicable for geometry optimizations and HVF have to be calculated at the SQM or FF level of theory. In many important areas of research, such as supramolecular chemistry,[572–574] for metal-organic frameworks,[287–290] or protein-drug binding[375,575], the relevant molecular size increases far beyond 100 atoms and multi-level approaches are often applied. For this purpose, the

SPH approach at SQM and FF low levels is presented as a new methodology, which further increases the accuracy of calculating thermostatistical contributions towards DFT values whilst retaining computational efficiency. In fact, the proposed method has already been tested internally in our group for some months with great success and is taken as a new default in our standard conformational search workflow.[60]

First, the theoretical background of the SPH method is given and its advantage for applications in computational chemistry over conventional approaches is motivated. Because of the enormous potential of the here-proposed method in combination with SQM and FF methods in many areas of chemistry, not every possible application is yet explored, and only selected illustrative examples and benchmarks are presented in this work. The performance of SPH calculations for G_{mRRHO} values is shown in comparison to the conventional standard method in which the reference structure is first fully re-optimized at a low theoretical level. Throughout this work, this approach is abbreviated by *OHESS*. Furthermore, the results of the SPH procedure for thermostatistical contributions are compared to another procedure in which the reference structure is used at the low level without any optimization (unrelaxed), which is abbreviated as *HESS*. As test systems, we show and evaluate a Diels–Alder reaction path, the molecular dynamics simulation of the small protein crambin, and IR spectra of the uracil dimer and a silver containing transition metal complex. Moreover, free association energies for standard supramolecular benchmark sets are investigated.

8.2. Theory and Implementation

8.2.1. RMSD Potential

The use of the Cartesian RMSD as a generally applicable collective variable in QM driven meta-dynamics was described by one of us in 2019 in the context of chemical space exploration[220]. Here, the total energy E_{tot} of the system is the sum of the total molecular (internal) energy E_{mol} of the applied method (*e.g.*, QM, SQM, or FF) plus the biasing root-mean-square deviation (RMSD) potential $E_{\mathrm{bias}}^{\mathrm{RMSD}}$

$$E_{\mathrm{tot}} = E_{\mathrm{mol}} + E_{\mathrm{bias}}^{\mathrm{RMSD}} \,. \tag{8.4}$$

The biasing potential was originally introduced for meta-dynamics (MTD) simulations, where a history-dependent potential fills the minima of the PES over time to overcome large reaction barriers. Such MTD simulations are successfully applied in the context of chemical space exploration by the *crest* program[60] and related applications.[273,274] Here, we pursue the opposite intention of retaining the initial geometry. More specifically, the biasing potential used is given by

$$E_{\mathrm{bias}}^{\mathrm{RMSD}} = \sum_{i=1}^{M} k_i \; exp\left(-\beta\Delta_i^2\right) \,, \tag{8.5}$$

where M is the number of reference structures associated with the pushing ($k_i > 0$) or pulling ($k_i < 0$) strength k, Δ is the collective variable, and the parameter β determines the width of the biasing

potential in the RMSD space. By limiting M to just the input structure ($M = 1$) and choosing a negative value for k_1, a restraining potential is added for any kind of molecular distortion away from the input structure. In other words, the PES is modified so that a local minimum can be created at the position of the input geometry for sufficiently negative values of k_1. Hence, in standard geometry optimizations irrespective of the actual coordinate system chosen, for large absolute k_1 values the input (reference) structure is obtained. For the collective variable, the standard RMSD in Cartesian space is chosen as a metric given by

$$\Delta = \sqrt{\frac{1}{n}\sum_{j=1}\left(r_i - r_i^{\mathrm{ref}}\right)^2}, \tag{8.6}$$

where r_i is an atomic position vector of the actual molecule, r_i^{ref} is the corresponding element in the reference structure, and n is the number of atoms. Here and in the following, the structural RMSD Δ always refers to the difference between the reference (input) structure calculated at a high-level of theory (X, usually DFT) and the lower level (Y, SQM or FF) structure. As we are interested in geometry optimizations and the Hessian, also the gradient of the RMSD as a function of the actual coordinates is required. For this purpose, the quaternion algorithm of Coutsias *et al.*[576] is employed.

8.2.2. Optimal Bias

Application of the biasing RMSD potential causes modifications of the PES and a change of the HVF. Hence, it is desirable to find the smallest possible bias, which retains the reference structure within a chosen tolerance. For SPH, a modified bisection algorithm[577] is applied to determine this bias potential. The algorithm, which is schematically shown in Figure 8.4 determines the optimal pulling strength k_{opt} that yields a structural RMSD between reference and re-optimized structure close to the chosen target value Δ_{target}.

First, an unbiased geometry optimization ($E_{\mathrm{tot}} = E_{\mathrm{mol}}$) is performed on the input structure, whereof the initial RMSD Δ_{init} between the reference structure and the fully optimized structure at a lower theoretical level is calculated. If this value is below the target ($\Delta_{\mathrm{init}} < \Delta_{\mathrm{target}}$), no additional bias is needed. Otherwise ($\Delta_{\mathrm{init}} > \Delta_{\mathrm{target}}$), the pulling strength is determined by the bisection algorithm. The required input is a continuous function, which in our case is Δ (see Eq. 8.6), an interval $[k_{\mathrm{min}}, k_{\mathrm{max}}]$, and the function value $\Delta(k_{\mathrm{bias}})$. The structural RMSD is not an analytical function of the pulling strength, but calculated from the difference of the initial structure and the optimized geometry. Hence, applying different pulling strengths within the biased geometry optimizations ($E_{\mathrm{tot}} = E_{\mathrm{mol}} + E_{\mathrm{bias}}^{\mathrm{RMSD}}$) effects the RMSD. The optimal pulling strength is then obtained iteratively. If convergence is reached, *i.e.*, $\Delta_{\mathrm{current}} - \Delta_{\mathrm{target}}$ is smaller than a chosen threshold η, k_{opt} is returned as the current pulling strength (k_{bias}). This value is applied in biased geometry optimizations yielding (almost) the reference structure for which HVF are calculated. The introduction of multiple (biased) geometry optimizations for the determination of the optimal pulling strength inevitably leads to some computational overhead before the actual Hessian calculation. Instead of one single unconstrained geometry optimization before the Hessian calculation as in the OHESS approach, SPH requires multiple iterations (limited to 10) to find k_{opt}. Yet, the number of optimization cycles decreases in each iteration due to the improved bias poten-

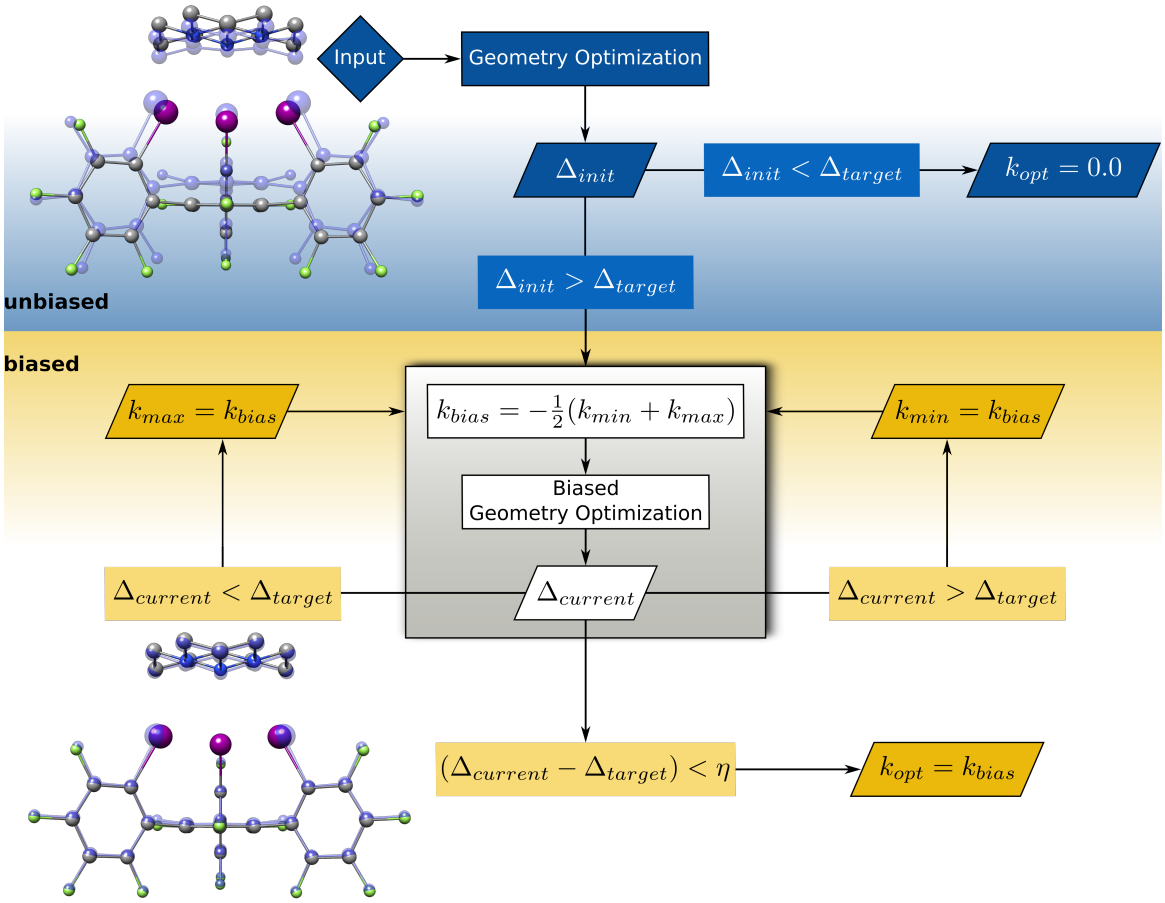


Figure 8.4.: Bisection algorithm for the determination of the optimal pulling strength. The structural RMSD is calculated between the high-level input structure and the fully unconstrained optimized structure at a lower theoretical level. If the value is above the target, the pulling strength of the bias potential is varied iteratively until biased geometry optimizations yield the desired small target RMSD.

tial. As an example, for GFN2-xTB calculations on N,N,N',N'-tetramethylpropane-1,3-diamine (TPMA) in complex with the tridentate halogen bond donor[578] (cf. Figure 8.4, 69 atoms), the SPH approach increases the optimization time by a factor of five compared to the OHESS approach. Nevertheless, for large systems, this increase is small considering the much more costly Hessian calculation. The SPH treatment is typically slower by a factor of 1.7 compared to a plain Hessian calculation and a factor of 1.5 compared to OHESS.

8.2.3. Restoring the Frequencies

Adding $E_{\text{bias}}^{\text{RMSD}}$ to the total energy is necessary to retain the reference geometry, but on the downside this modifies the HVF and hence, the latter needs to be corrected. In general, HVF are related to FC via a Taylor expansion of the PES up to second order around the equilibrium geometry. Hence, FC are the second order derivatives of the total energy with respect to nuclear displacements, which are contained

in the mass-weighted Hessian matrix

$$\mathbf{F}_{ij}^{(m)} = \frac{1}{\sqrt{m_i m_j}} \left(\frac{\partial^2 E_{\text{tot}}}{\partial R_i \partial R_j} \right), \tag{8.7}$$

where E_{tot} is the total molecular energy, $R_{i,j}$ is the displacement in Cartesian space, and $m_{i,j}$ are the respective masses of nuclei i and j. Diagonalization of the matrix $\mathbf{F}^{(m)}$ according to

$$\mathbf{F}^{(m)}\mathbf{Q} = \vec{\epsilon}\mathbf{Q} \tag{8.8}$$

yields the eigenvalues $\vec{\epsilon}$ of $\mathbf{F}^{(m)}$, and $\mathbf{Q}$ are the normal modes (eigenvectors). The corresponding HVF ν are obtained from

$$\nu_p = \frac{1}{2\pi}\sqrt{\epsilon_p}\,, \tag{8.9}$$

for each mode p. In analogy to Eq. 8.4, the total mass-weighted Hessian can further be written as

$$\mathbf{F}^{(m)} = \mathbf{F}_{\text{mol}}^{(m)} + \mathbf{F}_{\text{bias}}^{(m)}\,, \tag{8.10}$$

where $\mathbf{F}_{\text{mol}}^{(m)}$ is the Hessian matrix derived from the total energy of the applied low-level method, and $\mathbf{F}_{\text{bias}}^{(m)}$ is the second-derivative of the added RMSD biasing potential. Calculating the latter term numerically and subtracting it from $\mathbf{F^{(m)}}$ gives access to the unconstrained Hessian matrix $\mathbf{F}_{\text{mol}}^{(m)}$. The corresponding force constants $\vec{\epsilon}'_{mol}$ and $\vec{\epsilon}'_{\text{bias}}$ are calculated as expectation values (indicated by the prime) over the normal modes $\mathbf{Q}$.

$$\mathbf{Q}^{\dagger}\mathbf{F}_{\text{mol}}^{(m)}\mathbf{Q} = \vec{\epsilon}'_{\text{mol}} \tag{8.11}$$

$$\mathbf{Q}^{\dagger}\mathbf{F}_{\text{bias}}^{(m)}\mathbf{Q} = \vec{\epsilon}'_{\text{bias}} \tag{8.12}$$

To remove the frequency shift caused by the bias RMSD potential, $\vec{\epsilon}'_{\text{mol}}$ and $\vec{\epsilon}'_{\text{bias}}$ are then employed to construct a scaling factor for the HVF of every mode p given by

$$\nu_p^{\text{SPH}} = \sqrt{sgn(\epsilon'_{\text{mol},p})\frac{\epsilon'_{\text{mol},p}}{\epsilon'_{\text{mol},p} + \alpha\,\epsilon'_{\text{bias},p}}} \cdot \nu_p\,. \tag{8.13}$$

Here, ν^{SPH} are the resulting scaled frequencies of the SPH procedure, and $sgn(\epsilon'_{\text{mol}})$ is the sign function of the force constant. The functional form was chosen to obtain unbiased frequencies for $E_{\text{bias}}^{\text{RMSD}} = 0$, which is the case for small differences between the input and reference geometry. In this case, $\vec{\epsilon}'_{\text{bias}}$ is also zero and the scaling function is equal to unity. For finite $E_{\text{bias}}^{\text{RMSD}}$ and hence with $\vec{\epsilon}'_{\text{bias}} \neq 0$, the scaling term is smaller than one and removes the shift in ν_p. The great advantage of this approach is thus an individual frequency scaling, as some vibrational modes are stronger influenced by the biasing potential than others. Since $\mathbf{Q}$ are no eigenvectors of $\mathbf{F}_{\text{mol}}^{(m)}$ and $\mathbf{F}_{\text{bias}}^{(m)}$, the scaling is approximate, mainly neglecting mode couplings as well as mode changes. To correct this approximation, a single empirical parameter $\alpha = 1.27$ is applied. This parameter was determined on a small molecular test set, where reference equilibrium and (identical) input geometries were obtained at the GFN2-xTB level of theory. For different values of k_{bias}, α was varied to minimize the difference between biased and unbiased

GFN2-xTB G_{mRRHO} values over the set of molecules. The optimum value of α was found to be rather insensitive to the value of k_{bias} and to the choice of the level of theory (GFN-FF instead of GFN2-xTB). Due to the sign function, true imaginary HVF as they occur for transition states are preserved by the SPH approach.

Small imaginary HVF are generally (in HESS, OHESS, and SPH) treated by "inversion", *i.e.*, values $< i20$ cm^{-1} are set real (multiplied by $-i$) and consistently included in the thermostatistical treatment. Since molecular symmetry entering in Eq. 8.3 by σ is a discontinuous classification that creates a discontinuity in G_{mRRHO}, it is ignored (set to C_1) throughout this work. For consistent automated workflows, one may introduce continuous symmetry measures[579].

8.3. Computational Details

All quantum mechanical calculations were performed with the TURBOMOLE 7.2.1[335,336,580] (DFT), and xtb 6.3.3[515] (GFN1-xTB,[94] GFN2-xTB,[93] GFN-FF[117]) program packages. The ANCOpt optimizer was applied for GFN*n*-xTB and GFN-FF optimizations as implemented in the *xtb* program package[96] with default convergence criteria 10^{-7} E_h for energies and 10^{-5} E_h/bohr for gradients. For NCI complexes, the optimization threshold was tightened if necessary to remove all spurious imaginary HVF. The same optimization thresholds hold also true for SPH calculations by GFN*n*-xTB or GFN-FF. The default target RMSD threshold (Δ_{target}) of 0.1 Å was applied for SPH calculations. For transition states this value was eventually lowered to 0.08 Å. Computations of HVF at the DFT level were performed analytically using the *aoforce* code of TURBOMOLE, while GFN2-xTB and GFN-FF HVF were calculated numerically. G_{mRRHO} values were computed with the *thermo* module of *xtb*, applying a rotor cutoff of 50 cm^{-1} and generally refer to $T = 298$ K.

All DFT reference G_{mRRHO} values were calculated applying the PBEh-3c[231] hybrid functional or the B97-3c[237] GGA functional with default convergence criteria for energies and gradients as implemented in TURBOMOLE. For the computation of thermodynamic properties, PBEh-3c HVF are scaled by a factor of 0.95. Default effective core potentials (ECPs) ECP-28 and ECP-60[340] were used for all elements with atomic numbers larger than 36 (Kr) in order to approximately account for scalar relativistic effects in the DFT calculations. The resolution-of-identity (RI) approximation for the Coulomb integrals was generally used to speed up the DFT calculations using matching default auxiliary basis sets.[341,342] For the integration of the exchange-correlation contribution, the numerical quadrature grid *m*4 was employed. All calculations were performed on Intel© Xeon E5-2660 v4 @ 2.00 GHz machines.

The spectral similarity measure (match score) for the IR spectra calculations was obtained with a small standalone code called *newspecmatch*, which was also used in Ref. 546. For the full-width-half-maximum of the Lorentzian peaks, a value of 30 cm^{-1} was chosen. The SPH method is efficiently implemented in the freely available *xtb* program package,[96,515] but can easily be extracted and integrated into other codes for application with other DFT, SQM, or FF methods.

8.4. Results and Discussion

8.4.1. Noncovalent Interactions

For noncovalent interactions, an accurate description of $\Delta G_{\mathrm{mRRHO}}$ is notoriously difficult, as the newly formed intermolecular normal modes typically appear at low frequencies and are rather sensitive to structural changes. Yet, the effect of different R_e structures at varying theoretical levels and the resulting errors in the thermostatistical contributions are mostly neglected as it is generally assumed that these are small. Figure 8.5A shows that this can be invalid in a prototypical case, as severe structural reorganizations are not uncommon due to shallow PES if, *e.g.*, DFT geometries are re-optimized at the SQM level of theory. Free association energies for iodobenzene in cyclobis(paraquat-p-phenylene) ($CBPQT^{4+}$)[581]

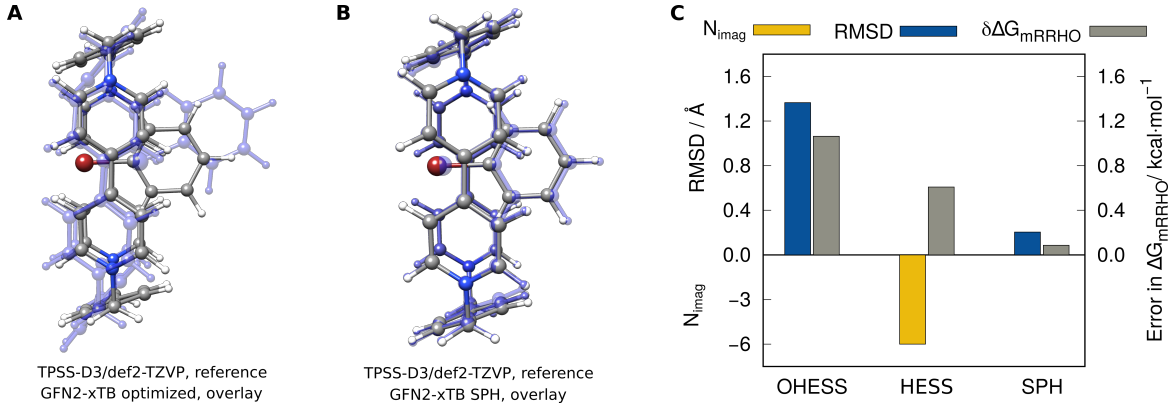


Figure 8.5.: Calculation of the thermostatistical contribution to the free energy for the inclusion of iodobenzene in $CBPQT^{4+}$. (A) TPSS-D3/def2-TZVP reference structure and GFN2-xTB optimized structure as blue overlay. (B) TPSS-D3/def2-TZVP reference and the structure obtained within the SPH approach as blue overlay. (C) Thermostatistical contribution calculated with and without re-optimization of the reference geometry, and with the SPH approach in comparison to the DFT result. The RMSD and number of imaginary modes ($> i20\ \mathrm{cm}^{-1}$) are given for the reaction.

were calculated at the GFN2-xTB level of theory starting from the reference TPSS-D3[328]/def2-TZVP[326] structure. Calculating HVF directly for the DFT structure (HESS) leads to many imaginary modes, which vanish upon re-optimization (OHESS), but lead to a large RMSD of 1.4 Å between high and low level optimized structures (Figure 8.5A,C). The number of imaginary modes ($> i20\ \mathrm{cm}^{-1}$) and the total RMSD are given as the sum for host, guest, and complex. The deviations in $\Delta G_{\mathrm{mRRHO}}$ amount to 0.7 and 1.1 kcal mol^{-1}, respectively. In comparison, SPH calculations have no imaginary modes, yield an RMSD of just 0.2 Å compared to the reference structure (Figure 8.5B,C), and the error in $\Delta G_{\mathrm{mRRHO}}$ is only 0.1 kcal mol^{-1}. One reason for the better agreement of SPH calculations with the DFT references is the improved description of S_{rot} connected with the small RMSD. OHESS calculations yield a difference for S_{rot} of 0.5 cal $\mathrm{mol}^{-1}\ \mathrm{K}^{-1}$ in comparison to the DFT value, which corresponds to a significant deviation of 0.15 kcal mol^{-1} at room temperature. SPH calculations on the other hand show a deviation of less than 0.2 cal $\mathrm{mol}^{-1}\ \mathrm{K}^{-1}$ (0.05 kcal mol^{-1} at 298 K). Hence, GFN2-xTB SPH calculations better re-

produce high-level reference DFT values for $\Delta G_{\mathrm{mRRHO}}$ compared to the standard approach of deriving frequencies from fully re-optimized geometries.

To further confirm these findings, the SPH approach was tested on three standard benchmark sets for NCIs, namely S30L.[120], S22,[162,537] and L7.[538]. Reference values for $\Delta G_{\mathrm{mRRHO}}$ were calculated with the efficient composite DFT methods B97-3c and PBEh-3c. Note that the electronic energy is not included here. GFN2-xTB free energy contributions were again calculated by the HESS, OHESS, and SPH approaches. The results for all 59 NCI complexes are shown in Figure 8.6. For each benchmark, the exemplary molecular structures shown in Figure 8.6 illustrate the various binding motifs. The results

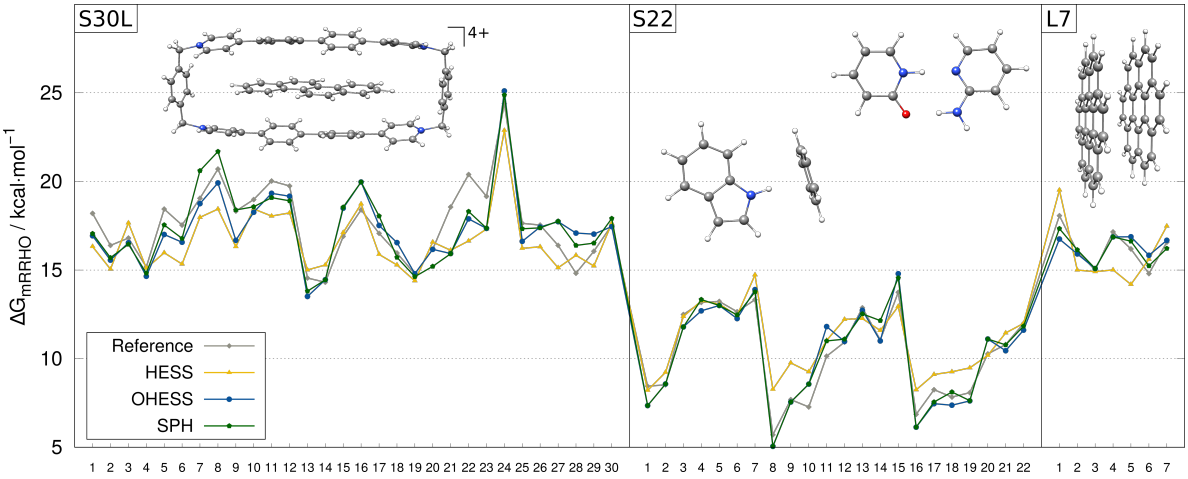


Figure 8.6.: Thermostatistical $\Delta G_{\mathrm{mRRHO}}$ values to the association free energy of the complexes taken from the S30L, S22, and L7 benchmark sets calculated directly from the Hessian, after optimization, and by the SPH approach at the GFN2-xTB level of theory.

are evaluated statistically by the mean deviation (MD), mean absolute deviation (MAD), standard deviation (SD), and relative MAD (*rel*MAD) listed in Table 8.1 (see Appendix A8 for their definitions). As can be seen in Figure 8.6, throughout the entire test set, no significant outliers occur neither for the SPH approach nor for the unconstrained OHESS calculation scheme. In contrast, the HESS scheme leads repeatedly to shifts in $\Delta G_{\mathrm{mRRHO}}$ in comparison to the reference values. Out of the three tested schemes for the calculation of $\Delta G_{\mathrm{mRRHO}}$, the newly developed SPH method performs best with an MAD of 0.59 kcal mol^{-1} and SD of 0.72 kcal mol^{-1} averaged over all three benchmarks. The relative error of $\Delta G_{\mathrm{mRRHO}}$ from the DFT reference amounts to 4.3 %. The conventional OHESS approach yields an MAD of 0.69 kcal mol^{-1} and SD of 0.88 kcal mol^{-1}, which corresponds to a relative error of 5.0 %. Calculating G_{mRRHO} without any relaxation results in an MAD of 1.12 kcal mol^{-1} and SD of 1.19 kcal mol^{-1}, which is in relative error (8.1 %) almost twice as large compared to SPH.

The reason for the differences between the three approaches becomes visible in Figure 8.7. Here, the total number of imaginary frequencies ($> i20$ cm^{-1}) and the total structural RMSD between the reference and GFN2-xTB optimized structure are shown as the sum over all 169 systems (host, guest, and complex). Calculating frequencies on unrelaxed structures leads to 156 imaginary frequencies that cause a loss of S_{vib}. Upon re-optimization (unconstrained and biased), the number of imaginary fre-

Table 8.1.: Statistical evaluation of the thermostatistical contribution $\Delta G_{\mathrm{mRRHO}}$ for the 59 complexes[a] at the GFN2-xTB level of theory in comparison to DFT reference values. The *rel*MAD values are given in %.

		S30L	S22	L7	all sets
HESS	MD	0.98	0.69	0.17	0.61
	MAD	1.27	0.89	1.21	1.12
	SD	1.19	0.93	1.47	1.19
	*rel*MAD	7.65	8.90	7.65	8.07
OHESS	MD	0.32	-0.07	-0.15	0.03
	MAD	0.96	0.57	0.54	0.69
	SD	1.17	0.72	0.74	0.88
	*rel*MAD	5.57	6.19	3.28	5.01
SPH	MD	0.14	0.03	-0.08	0.03
	MAD	0.89	0.52	0.35	0.59
	SD	1.10	0.65	0.42	0.72
	*rel*MAD	5.06	5.69	2.14	4.29

[a] The statistical measures are given with respect to the DFT reference, which is B97-3c for S30L and S22 and PBEh-3c for L7.

quencies is reduced significantly to 9 and 13 respectively, where the few remaining imaginary modes are all in the order of $i20$ to $i50$ cm^{-1}. The disadvantage of unconstrained re-optimizations is the resulting structural rearrangement. The difference in DFT and GFN2-xTB PES manifests in a total structural RMSD of 25.6 Å. In the SPH approach, this deviation is reduced drastically to 9.5 Å in total. This proves that for accurate $\Delta G_{\mathrm{mRRHO}}$ values a small RMSD in comparison to the reference structure and a minimum number of imaginary frequencies are desirable and this is only fulfilled by SPH calculations.

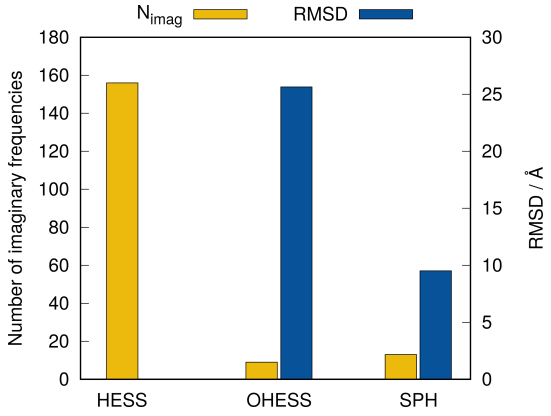


Figure 8.7.: Number of imaginary frequencies (> $i20$ cm^{-1}) and structural RMSD between the DFT and GFN2-xTB structures given as the sum over all 169 molecules (complex, host, and guest) of the S30L, S22, and L7 benchmark set.

To further investigate the contribution of the error of S_{rot} in the OHESS approach, a modified version of it was additionally tested for the L7 benchmark, where the DFT reference geometry is taken for the calculation of S_{rot}, but S_{vib} is still evaluated on the fully optimized GFN2-xTB structure. Thereby, the MAD in comparison to the DFT reference is reduced to 0.50 kcal mol^{-1} (0.54 kcal mol^{-1} for OHESS).

This shows that the effect of the choice of the structure for S_{rot} is small but significant for accurate ΔG_{mRRHO} values. However, we can also conclude that the improvements of SPH calculations over the standard OHESS approach are also due to the fact that S_{vib} is evaluated on the correct input structure. Hence, the reference geometry does not only affect S_{rot}, but also S_{vib} significantly, especially for noncovalent interactions.

8.4.2. Camphor Model Structures

To further investigate the effects of SPH calculations in covalently bonded systems, the G_{mRRHO} values of camphor were calculated for different displacements of the nuclear coordinates. With the center of mass placed at the origin in Cartesian space, the nuclear coordinates were scaled in a range from 0.90 to 1.10 to simulate a common dissociation process. The equilibrium geometry was obtained at the GFN2-xTB level of theory and the overlay of all five displacements is shown in Figure 8.8A. Re-optimization

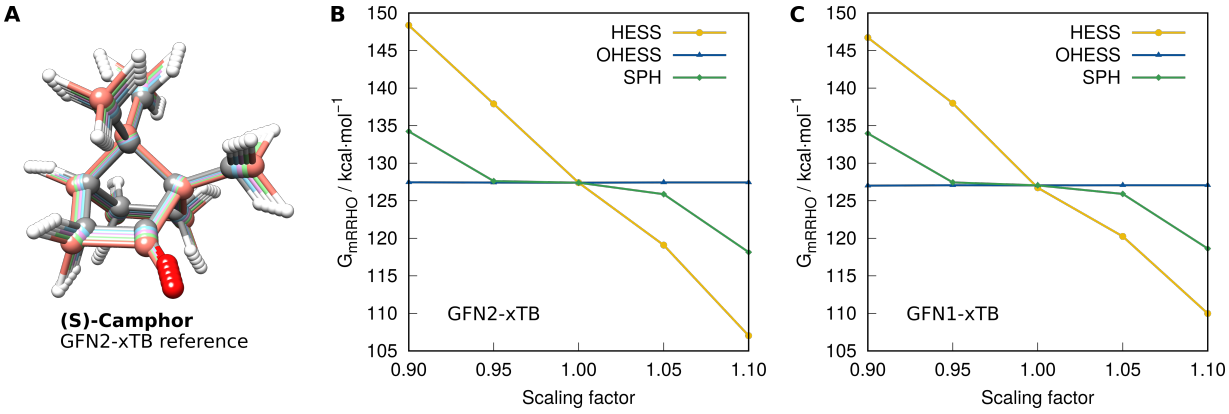


Figure 8.8.: (A) Structural overlay of different displacements of camphor. The equilibrium geometry is obtained at the GFN2-xTB level of theory. (B) G_{mRRHO} calculated by different schemes for the five structures by GFN2-xTB. (C) The corresponding G_{mRRHO} values calculated by GFN1-xTB.

of the five displaced structures by GFN2-xTB yields of course to the same minimum geometry and the corresponding (constant) G_{mRRHO} values are given by the horizontal line in Figure 8.8B. For small displacements in geometry, the change in G_{mRRHO} is also expected to be small as long as one stays well inside the harmonic trust region. This behavior is observed for SPH calculations, which show similar G_{mRRHO} values for displacements around R_e in the range 0.95–1.05 of only ±5 %. For larger displacements of ±10 %, the G_{mRRHO} values start to differ significantly from the value of the equilibrium geometry. Increased G_{mRRHO} values occur for compressed bonds, and decreased values for elongated bonds, as the entropic part increases/decreases for weaker/stronger bonds (longer/shorter bond lengths) with decreasing/increasing HVF. This qualitative behavior of the pure Hessian calculation is reproduced by the SPH approach but obviously damped, *i.e.*, values closer to that of the R_e structure are more preserved for stronger distortions. Similar results are obtained at the GFN1-xTB (see Figure 8.8C) and GFN-FF (see Appendix A8) levels. These results suggest the general use of the SPH method as a robust

default procedure for G_{mRRHO} calculations on any geometry yielding either the exact values (if the reference and low-level PES are identical) or ones that are closer to those obtained near the equilibrium structure by the high-level method. This is further exemplified with the following case studies.

8.4.3. Diels–Alder Reaction

As shown above, for a model dissociation reaction, SPH calculations represent a promising approach to compute free energies along a reaction path. Normally, paths or PES are shown in the pure energy domain, while free energies are only computed for reactant, product, and transition state (TS). For related free energy path determinations see, *e.g.*, Refs. 582–586. The reason is that unconstrained optimizations along with the reaction coordinate just yield the educt/product before/after the TS, while just conducting standard Hessian calculations on the path geometries suffer from the problem of artificial imaginary modes as discussed above. One approach to circumvent this difficulty was introduced by Miller, Handy, and Adams,[587] which uses a projection technique to remove imaginary frequencies along with an intrinsic reaction coordinate (IRC).[588]

In Figure 8.9A, the path for the Diels–Alder reaction of two cyclopentadiene molecules is shown. The exergonic reaction path was calculated by the double-ended growing string method (GSM) at the GFN2-xTB level of theory[589–591] starting from GFN2-xTB optimized reactant and product. The G_{mRRHO} values

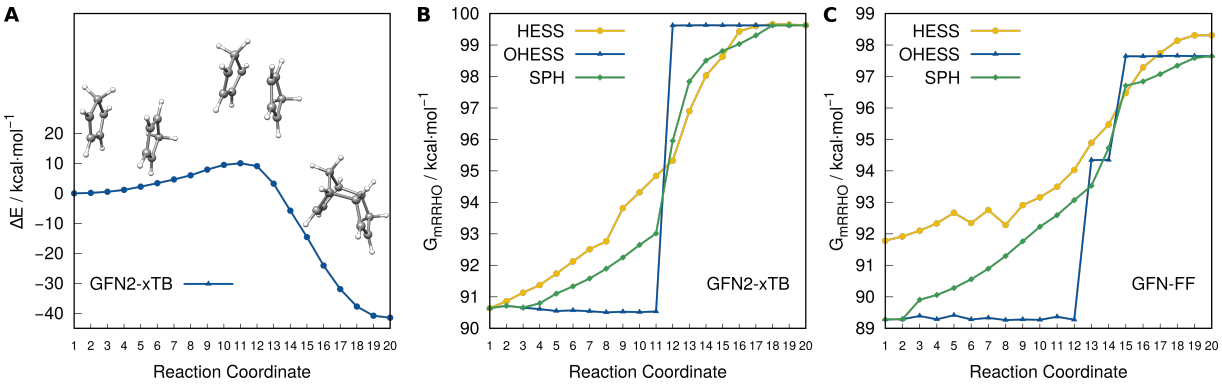


Figure 8.9.: (A) GSM reaction energy path calculated for the Diels–Alder reaction of two cyclopentadiene molecules at the GFN2-xTB level of theory. (B) G_{mRRHO} calculated with and without previous optimization and by the SPH approach along the reaction coordinate with GFN2-xTB. (C) The corresponding calculations carried out at the GFN-FF level of theory.

computed along the path at the GFN2-xTB level of theory are shown in Figure 8.9B. Optimization and subsequent Hessian calculation on the reaction path trivially leads to a step function with G_{mRRHO} values of ~90.5 kcal mol^{-1} (reactant) and ~99.5 kcal mol^{-1} (product), respectively. The increase in G_{mRRHO} is due to the loss of entropy within the [4+2] cycloaddition of cyclopentadiene by the conversion of low- to high-frequency modes. Without any relaxation, the G_{mRRHO} values almost linearly increase. Similar to the results obtained by the reaction path Hamiltonian approach of Miller *et. al.*,[587,592,593] however, only small changes for small displacements around the two minima and a larger one in the TS region (S-shaped curve) are observed for SPH calculations.

In Figure 8.9C, the corresponding thermostatistical contributions were calculated at the GFN-FF level of theory. The FF topology was recomputed for every point of the reaction path. This is a challenging test because the dissociation of chemical bonds is not easily possible with FFs and discontinuous PES may arise. In fact, GFN-FF optimizations with geometries 13 and 14 as starting points yield intermediates with one single bond between the two cyclopentadiene monomers. The constrained SPH calculations prevent these artificial products and yield the expected S-shaped curve which is in much better agreement with the one from GFN2-xTB as compared to the HESS treatment. This is a very encouraging result enabling the application of the very fast FFs even in cases for which they are originally not suited. The much higher $\Delta G_{\mathrm{mRRHO}}$ values by the HESS approach in comparison to SPH calculations, especially before the TS, are resulting from a large number of imaginary modes which are absent in the SPH treatment.

8.4.4. Transition State Example

For transition states[52,594] in general, the question arises, whether semiempirical SPH calculations are capable of reproducing the DFT imaginary frequency and transition normal mode, when starting from a DFT TS structure. Therefore, frequency calculations were performed for the TS of the *trans* addition of dihydrogen to a pincer Ir(I) complex[595,596] as an example. Figure 8.10 shows the reaction, where trace amounts of water facilitate the proton transfer reaction steps. The TS structure was obtained at the

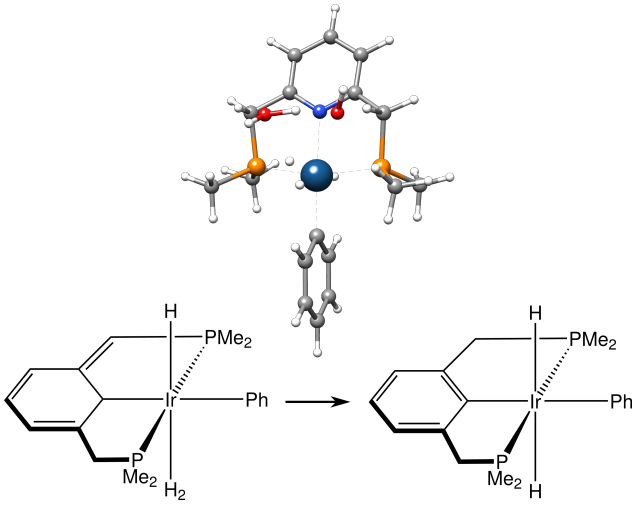


Figure 8.10.: *Trans* addition of dihydrogen to a pincer Ir(I) complex at the TPSS-D3/def2-SVP level.

TPSS-D3-def2-TZVP level of theory in Ref. 591. HVF calculations at the same level yield one imaginary frequency at $i256.1$ cm^{-1}. Starting from the DFT structure OHESS, HESS, and SPH calculations were performed at the GFN2-xTB level of theory. The frequencies of the five lowest vibrational normal modes for each calculation are listed in Table 8.2. As expected, the imaginary transition frequency vanishes upon optimization by the OHESS approach. The HESS calculation leads to more than one imaginary frequency as the DFT TS structure is no stationary point on the unconstrained GFN2-xTB PES. SPH calculations on the other hand yield only one imaginary frequency at $i183.6$ cm^{-1} and the

corresponding normal mode closely reproduces the DFT vibrational mode. Hence, SPH calculations can be used to determine if the DFT TS optimization found the correct TS without computing the costly DFT Hessian. Furthermore, the corresponding normal mode can be used to initiate an IRC at the DFT level of theory.

Table 8.2.: Vibrational frequencies[a] for the five lowest vibrational normal modes, calculated for the transition state of the Ir(I) complex by TPSS-D3/def2-SVP and GFN2-xTB. At the SQM level results are shown for OHESS, HESS, and SPH calculations.

	TPSS-D3/def2-SVP		GFN2-xTB	
Mode	OHESS	OHESS	HESS	SPH
1	*i*256.1	26.6	*i*589.6	*i*183.6
2	50.6	29.7	*i*482.3	117.8
3	60.5	36.4	*i*196.1	125.4
4	80.3	41.5	54.7	131.8
5	116.9	42.3	59.7	143.4

[a] Values are given in cm^{-1}.

8.4.5. MD of Crambin

Molecular dynamics (MD) simulations give access to thermally averaged structures of biomacromolecules.[274,597,598] The SPH approach can be used as a new and convenient tool to properly average free energy contributions over snapshots along MD trajectories. In this application, the reference structure does not necessarily refer to a structure from higher-level theoretical treatment but one that may differ from R_e additionally by the thermal effects introduced by the MD. This is shown here for the example of a GFN2-xTB/GBSA(H_2O) MD simulation on crambin (100 ps length with a 4 fs time step at an increased hydrogen mass of 4 amu). The input geometry was taken from the protein database and is illustrated in Figure 8.11A. Snapshots were taken in 1 ps steps after an equilibration phase of 50 ps. The calculated free energies are shown in Figure 8.11B (GFN2-xTB) and 8.11C (GFN-FF).

In general, the computation of G_{mRRHO} without any relaxation is not recommended for larger, flexible systems. In this case, typically more than 50 imaginary HVF (up to $i600\,cm^{-1}$) are observed for the snapshots. The resulting loss of S_{vib} leads to about 10–15 kcal mol^{-1} larger G_{mRRHO} values at 298 K compared to computations on fully relaxed structures for both GFN methods. For GFN2-xTB, OHESS yields a mean G_{mRRHO} value of 3006 kcal mol^{-1} with a standard deviation (SD) of 1.6 kcal mol^{-1}. The corresponding SPH results are very similar with a mean of 3001 kcal mol^{-1} and SD of 2.0 kcal mol^{-1}. The increase in entropy (decrease in G_{mRRHO}) from optimized to thermally equilibrated structures as obtained by the SPH calculations seems to be realistic and is intuitively understandable if many higher-energy conformational states are thermally populated. GFN-FF G_{mRRHO} calculations on the respective GFN2-xTB MD snapshots are shown in Figure 8.11C. Here, re-optimization results in GFN-FF minimum snapshot structures, which strongly differ from the thermally equilibrated GFN2-xTB input structures, and G_{mRRHO} values are far off the GFN2-xTB result. The G_{mRRHO} values calculated without relaxation show large jumps along the trajectory manifested in a large SD of 4.8 kcal mol^{-1}. In contrast, SPH

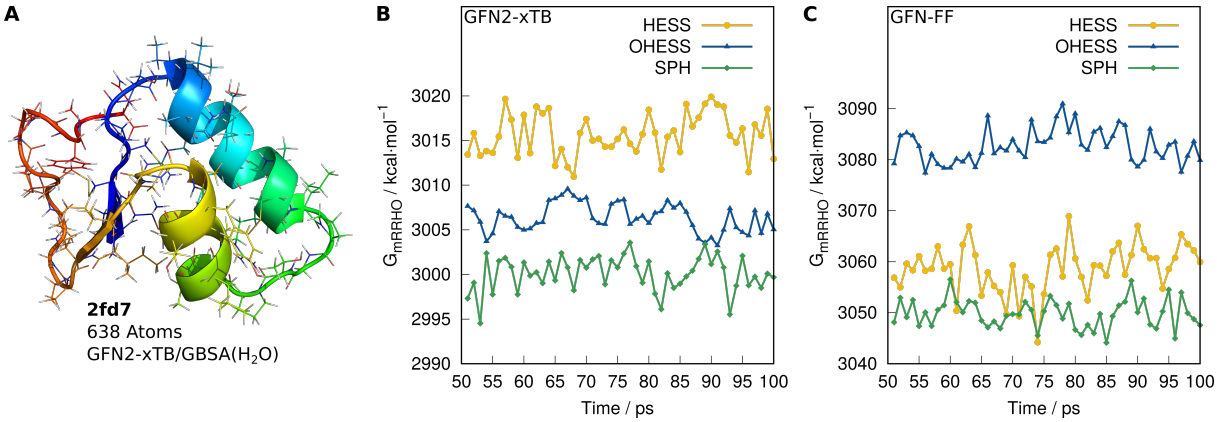


Figure 8.11.: (A) Molecular structure of crambin, the PDB identifier is given in bold. (B) G_{mRRHO} values calculated with and without relaxation and by the SPH approach at the GFN2-xTB level of theory for 50 MD snapshots. (C) The corresponding calculations computed at the GFN-FF level of theory on the same GFN2-xTB snapshots.

calculations closely approach the GFN2-xTB result with a mean G_{mRRHO} value of 3049 kcal mol^{-1} and SD of 2.8 kcal mol^{-1}. This suggests a general procedure for obtaining averaged G_{mRRHO} values for flexible systems based on low-level SPH calculations on higher-level MD trajectories. Replacing the Hessian calculation at the GFN2-xTB level of theory by a GFN-FF calculation additionally leads to significant savings in computation time by about two orders of magnitude.

8.4.6. IR Spectra

The performance of the semiempirical GFN tight-binding and force-field methods for the simulation of IR spectra in direct comparison with experimental gas phase data was recently investigated in Ref. 546. As IR spectra are quite sensitive to structural changes,[599] the SPH approach seems to be promising to further improve the accuracy of SQM and FF methods by constraining to the higher-level structure. An example is given in Figure 8.12, where the IR spectrum of the uracil dimer is calculated at the GFN-FF level of theory in comparison to the B97-3c reference spectrum. In contrast to the DFT result, GFN-FF favors a hydrogen bonded dimer which is lower in energy by ~9 kcal mol^{-1} compared to the π-stacked conformation. The B97-3c reference and the fully relaxed GFN-FF structures are shown in Figure 8.12A. The effect of re-optimization on the GFN-FF IR spectrum is significant, as the N-H stretching intensity is artificially amplified by the hydrogen bonding which is absent in the DFT structure, resulting in a spectral match score between DFT and GFN-FF of just 0.26 (unity means perfect agreement of the two spectra while zero corresponds to absolutely no agreement). The GFN-FF IR spectrum without relaxation is shown in Figure 8.12B. Here, the structure is naturally identical with the DFT structure, but two significant imaginary modes arise for this non-R_e structure and the N-H stretching consists of three peaks instead of two as in the DFT spectrum. Overall, this leads to a slightly better match score of 0.32. In contrast, GFN-FF SPH calculations maintain the input geometry as shown in Figure 8.12C, whilst eliminating all imaginary frequencies. This IR spectrum matches the DFT computed in-

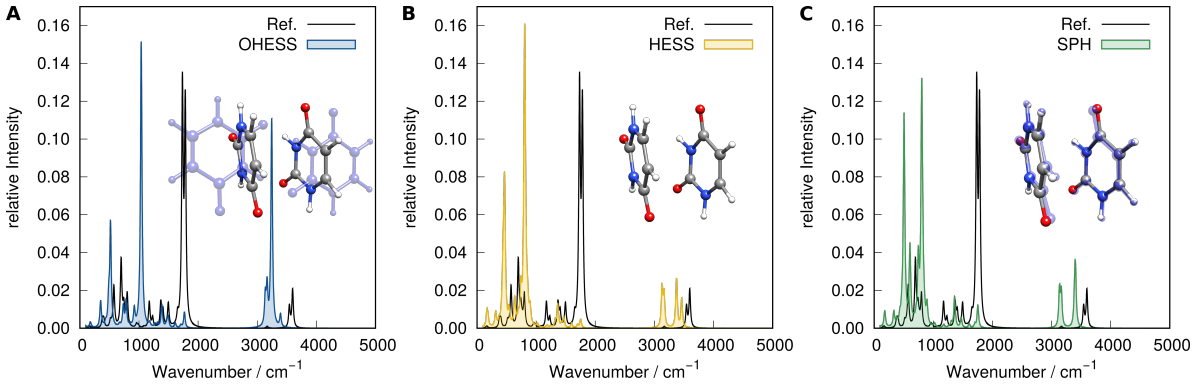


Figure 8.12.: IR spectra for the uracil dimer calculated by GFN-FF in comparison to the reference B97-3c spectrum. (A) IR spectrum calculated after re-optimization. (B) Without relaxation on the B97-3c structure. (C) SPH calculation. Structural overlays of the DFT structure and the GFN-FF re-optimized structures are shown as blue overlays.

tensity of the N-H stretching mode better and the corresponding doublet splitting is also qualitatively reproduced, which improves the match score to a value of 0.38. Hence, for large structural changes upon re-optimization, SPH calculations represent a clear improvement over the conventional OHESS approach for the simulation of IR spectra in comparison to the DFT reference.

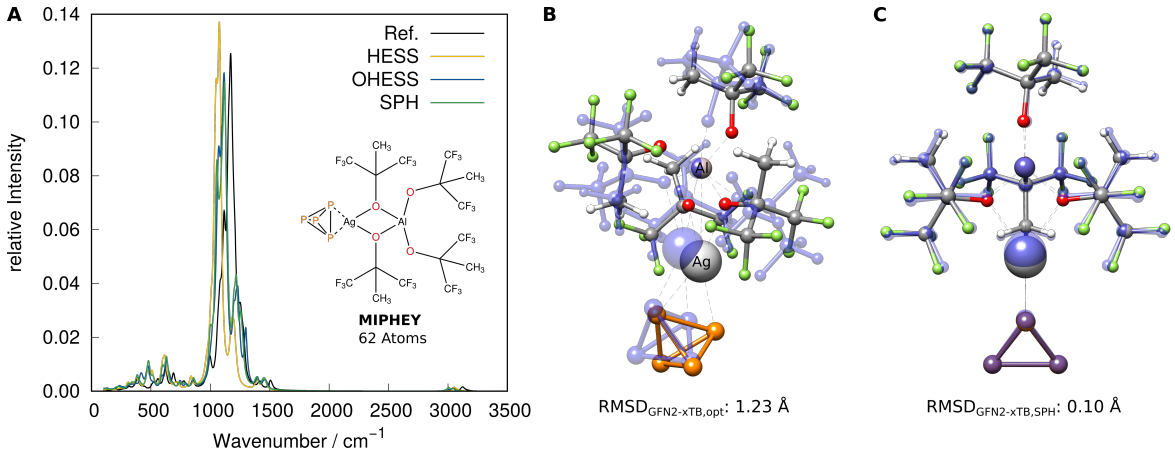


Figure 8.13.: (A) IR spectrum of a silver containing transition metal complex (CSD ID in bold) calculated by GFN2-xTB in comparison to the reference B97-3c spectrum. The GFN2-xTB spectra obtained without relaxation, after unconstrained re-optimization, and with the SPH approach are given. (B) Structural overlay of the DFT structure and the fully relaxed GFN2-xTB structure in blue. (C) Structural overlay of the DFT structure and the SPH structure in blue.

But even for less significant differences between high- and low-level PES, SPH calculations lead to some improvement over the conventional approach. Figure 8.13A shows the IR spectrum of a silver-P_4 transition metal complex with a weakly coordinating aluminum anion $(P_4)AgAl[OC(CH_3)(CF_3)_2]_4$.[221,600] The GFN2-xTB optimized structure is shown in Figure 8.13B with a structural RMSD of 1.23 Å in com-

parison to the DFT structure while by construction, the SPH structure has the default RMSD of just 0.10 Å and is shown in Figure 8.13C. Without relaxation, a match score of 0.67 is obtained, whereas unconstrained re-optimization yields a match score of 0.84. The best agreement with the reference IR spectrum is obtained by SPH calculations with a match score of 0.86. Even though smaller than in the previous example, an improvement could yet be achieved by the SPH approach. As pointed out in Ref. 546, the decisive factor for high match scores with low-level methods is the intensity, rather than the HVF. Since SPH calculations improve mainly the latter, only slightly increased match scores can be expected if the structural differences are small.

8.5. Conclusion

In this work, we introduced the new concept of single-point Hessian (SPH) calculations for thermostatistical free energy contributions for general non-equilibrium geometries within the modified RRHO approximation and the computation of IR spectra. The HVF obtained from the SPH method improve the accuracy of SQM or FF computed G_{mRRHO} values in comparison to high-level DFT reference values if a corresponding DFT structure is used as input. Multi-level approaches for the calculation of free energies are thus the prime field of application. The key ingredient is the application of a biasing potential given as Gaussian functions expressed with the structural RMSD in Cartesian space to retain the input geometry. Effects on the HVF caused by the applied constraining potential are approximately removed by an empirical scaling function, which employs only one global parameter derived from normal mode expectation values in order to recover the unconstrained frequencies. The whole procedure is fully automatic and general. Moreover, it requires no further user input other than reference coordinates.

The performance of the SPH method was tested for G_{mRRHO} values in comparison to calculations with and without structure relaxation at low theoretical level. High-level DFT input structures and G_{mRRHO} values computed at the same DFT level served as reference. On a versatile test set composed of the S30L, S22, and L7 noncovalent interaction benchmarks, GFN2-xTB SPH calculations reduced the mean absolute deviation of $\Delta G_{\mathrm{mRRHO}}$ from 0.69 kcal mol^{-1} with the conventional OHESS approach to 0.59 kcal mol^{-1} in comparison to DFT reference values. The two main reasons for the improvement are the reduced structural RMSD between the DFT input and SQM optimized structure (better rotational entropy) and the much smaller number of imaginary frequencies. We can only speculate on a secondary general "shaping" effect of the biasing potential on the low-level PES and concomitantly improving the HVF.

For two model reaction paths (nuclear displacements around R_e for (S)-camphor and the Diels–Alder reaction of two cyclopentadiene molecules), SPH calculations showed the correct behavior of small deviations in G_{mRRHO} for small changes around the R_e geometry and an increase for larger distortions. At the GFN-FF level of theory, SPH calculations additionally solved the problem of artificial intermediates near the transition state of the Diels–Alder reaction.

In the context of MD simulations, it was shown that SPH calculations give access to G_{mRRHO} values for thermally equilibrated MD snapshots in a robust and automated way. The computational savings are significant (two orders of magnitude) if GFN-FF SPH are computed on GFN2-xTB snapshots instead of

always taking GFN2-xTB. Improvements were also observed for the simulation of IR spectra. For the π-stacked uracil dimer, which converts to a hydrogen bonded conformation upon GFN-FF re-optimization as required by the standard OHESS approach, SPH calculation significantly improved the match score with the corresponding DFT IR spectrum by almost 50 % compared to the relaxed calculation. Even for the challenging case of predicting the IR spectrum of a transition metal complex, SPH calculations slightly improved the match score with the DFT spectrum in comparison to the standard approach of unconstrained re-optimization.

In general, calculations of G_{mRRHO} values from HVF without relaxation should be avoided as the results are often totally artificial due to the frequent occurrence of significant imaginary modes. The reason that sometimes reasonable values were calculated is attributable to the modified RRHO approach, which partly corrects for inaccurate low-lying frequencies. In fact, RRHO calculations on non-equilibrium structures are generally useless yielding totally unreasonable results. Although this seems to be common sense, many computational chemistry textbooks do not provide a detailed explanation of this issue and we hope to shed light on this.

In comparison to the conventional approach of re-optimizing the input structure at lower, *e.g.*, SQM level prior to the Hessian calculation, the SPH approach achieved improvements in almost every test case. The computational overhead required for the automatic determination of the optimum bias strength is moderate, *i.e.*, the SPH treatment is typically slower by a factor of 1.5 compared to a re-optimized Hessian calculation. Thus, SPH calculations represent a step towards higher accuracy in routine computations of free energies or IR spectra for large systems with efficient SQM or FF methods. The implementation in the *xtb* program package is open source and a transfer to other codes and QM/FF methods is easily achievable.

Acknowledgments

This work was supported by the DFG in the framework of the “Gottfried-Wilhelm-Leibniz” prize to Stefan Grimme. Sebastian Spicher thanks the “Fond der Chemischen Industrie (FCI)” for financial support. The authors thank F. Bohle, M. Bursch, A. Hansen, and P. Pracht for helpful discussions.

9. Fast and Accurate Quantum Chemical Modeling of Infrared Spectra of Condensed-Phase Systems

Sergey A. Katsyuba,[a] Sebastian Spicher,[b] Tatiana P. Gerasimova,[a] and Stefan Grimme[b]

Received: 29 June 2020, Published online: 7 July 2020

Reprinted (adapted) with permission[†] from
Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. *J. Phys. Chem. B* **2020**, *124*, 6664–6670.

DOI 10.1021/acs.jpcb.0c05857

Own manuscript contribution

- Conception of the study
- Development of the QCG algorithm
- Generating the clusters
- Writing parts of the manuscript

[a] Arbuzov Institute of Organic and Physical Chemistry, FRC Kazan Scientific Centre of RAS, Arbuzov St. 8, 420088 Kazan, Russia
[b] Mulliken Center for Theoretical Chemistry, Institute of Physical and Theoretical Chemistry, University of Bonn, Beringstraße 4, 53115 Bonn, Germany
[†] Permission requests to reuse material from this chapter should be directed to the American Chemical Society.

Abstract An efficient approach for an accurate quantum mechanical (QM) modeling of infrared (IR) spectra of condensed-phase systems is described. An ensemble of energetically low-lying cluster structures of a solute molecule surrounded by an explicit shell of solvent molecules is efficiently generated at the semiempirical tight-binding QM level and then reoptimized at the density functional theory level of theory. The IR spectrum of the solvated molecule is obtained as a thermodynamic average of harmonically computed QM spectra for all significantly populated cluster structures. The accuracy of such simulations in comparison to experimental data for some organic compounds and their solutions is shown to be the same or even better than the corresponding QM computations of the gas phase IR spectrum for the isolated molecule.

9.1. Introduction

The infrared (IR) absorption spectrum is a unique characteristic of a chemical substance, which is one of the main reasons for the wide applicability of IR spectroscopy in chemistry and related areas of science. An IR spectrum contains information not only about the electronic and spatial structure of the molecule but also about its intermolecular interactions. Modern spectroscopic techniques allow IR experiments to detect the response of the molecular system to subtle changes in the surroundings. This richness of information is accompanied by increased complexity in the interpretation of the experiment, which is often impossible without computational analysis/modeling of IR spectra. The latter task is difficult and truly nonstandard for condensed-phase systems representing an overwhelming majority of samples experimentalists deal with.

Quantum chemical computations, widely used for the simulation and interpretation of gas phase IR spectra, are also applicable to molecular crystals, provided that single-crystal X-ray data on their structure are available.[601,602] In the more common case of liquids, solutions, and amorphous substances, where information on the supramolecular or dynamical structure of condensed matter is limited, periodic *ab initio* molecular dynamics (AIMD)[603,604] represent a rather rigorous method for IR spectra computations. However, this approach involves a huge amount of computer resources while the assignment of IR bands to nuclear motions (spectrum interpretation) is not straightforward and requires additional efforts.[605] Thus, less-demanding approaches combining a quantum mechanical (QM) description of the target molecule responsible for the spectral signal and classical molecular mechanical treatment of the environment, that is, mostly the solvent, come into prominence.[606–608] However, the accuracy of the theoretical spectra strongly depends on the quality of the applied force fields,[608] which are furthermore not parameterized for many, more uncommon elements, (*e.g.*, neat organometallic compounds). Further simplification of the theoretical treatment of the environmental effects, for example, by implicit polarizable continuum models, produces unsatisfactory results in particular for the computed IR intensities.[609] Furthermore, typical continuum models are available only for common solvents and hence many pure liquids cannot be treated adequately.

Here, we propose a new, rather general approach, in which the system under consideration is represented by an ensemble of molecular clusters. Hereby, the semiempirical GFN2-xTB[93] QM method is applied to generate structures of a solute surrounded by explicit solvent molecules forming solva-

tion shells. By taking the solute molecule also as a solvent, neat liquids or amorphous solids can be described, and the extension to arbitrary mixtures is straightforward. All available GFN methods[96] are consistently parameterized for the entire periodic table up to radon ($Z \leq 86$). Full-structure optimization of the GFN2-xTB-generated clusters at the low-cost B97-3c[237] level of density functional theory (DFT) provides equilibrium structures suitable for standard vibrational frequency calculations. The B97-3c method (internally employing a reasonably large modified TZVP atomic orbital (AO) basis set) has been recently shown[610] to perform very well for IR spectra computations in the usual harmonic approximation (which is also applied in this work). The IR spectrum of the system is then calculated as a thermodynamic average of the individual QM spectra computed for each complete cluster at the DFT level. This approximation was tested by comparison of calculated IR spectra of clusters, comprising an organic molecule ("solute") surrounded by N solvent molecules (CCl_4 or CS_2) with the experimental spectra of the corresponding dilute solutions in these solvents featuring a moderately transparent mid-IR spectral range. N was chosen so that the first solvation shell of the solute molecule was formed. Typical values of $N \approx 10–20$ were taken for the test molecules described in Appendix A9, though in some cases, we intentionally increased N to estimate the possible influence of a further increased cluster size (*vide infra*). Neat liquids or amorphous solids were modeled in a similar way. In order to minimize finite cluster size effects in the latter case, the IR spectrum of the central "solute" molecule was separated from the spectra of the surrounding molecules and compared to the corresponding experimental spectra of the neat compound. To achieve such separation, atomic masses for all species in the cluster, except the "solute", were increased to 400 amu, and thus, their spectra were shifted to the low-frequency region.

9.2. Methodology

9.2.1. Cluster Generation

The clusters are generated with a new algorithm termed quantum cluster growth (QCG).[611] The QCG algorithm combines a force-field based docking algorithm (xTB-IFF[466]) with the fast and robust semiempirical QM method GFN2-xTB[93] to add a given number of solvent molecules iterative at energetically favorable positions around a given solute. Molecular dynamics (MD) simulations of the thereby generated cluster followed by full geometry optimizations of each snapshot generate a large ensemble of energetically low lying clusters. Thereof, the most populated representatives serve as the starting point for the calculation of IR spectra in solution. QCG consists of the following steps:

1 One solvent molecule is added to the solute with xTB-IFF. At this point, all structures are treated rigid, *i.e.*, only intermolecular degrees of freedom are optimized.

2 The complex is fully optimized on the GFN2-xTB level of theory.

3 Steps 1–2 are repeated until a sufficient (or selected) number of solvent molecules are added.

4 MD simulations are performed for 50 ps at 298 K on the generated cluster under gas phase conditions.

5 Every snapshot (typically about 500 in total) is fully optimized on the GFN2-xTB level of theory and energetically sorted.

6 The final cluster ensemble consists of the lowest clusters found within a 6 kcal mol^{-1} energy window.

A relative difference of 3 kcal mol^{-1} between different clusters in the ensemble already leads to a negligible Boltzmann population of less than 1 % at room temperature. Only clusters that are populated by more than 10 % are taken into account for the IR spectra calculations.

9.2.2. Computational Details

All calculations were carried out using the Turbomole 7.2 program package.[612] Following full geometry optimizations at each level of theory, harmonic vibrational frequencies and IR intensities were calculated analytically for the DFT methods employed in this study, corresponding to the B97-3c[237] generalized gradient approximation (GGA) functional with the polarized valence-triple-zeta basis set mTZVP.[326] Note, B97-3c applies the D3 London dispersion correction in the Becke–Johnson sampling scheme,[252,253] but additionally a short-range bond length correction potential as well as a specific adjustment of the electronic parameters in the B97 Taylor expansion. We have shown very recently,[610] that B97-3c, being much faster than the well-recognized hybrid functional B3LYP, offers similarly good quantitative performance in comparison to experimental data for relative IR intensities and fundamental frequencies ($\nu \leq 2200\ cm^{-1}$) for isolated molecules. Infrared intensities were computed in the double harmonic approximation, ignoring cubic and higher force constants and omitting second and higher order dipole moment derivatives. To minimize the influence of this neglect of the anharmonicity effects on a comparison of the computed and experimental (anharmonic) intensities, the bands of the most anharmonic CH stretching modes were excluded. Moreover, only experimental IR bands reliably assigned to fundamental vibrations were chosen for the comparison. The harmonic frequencies obtained using the B97-3c computations were compared with their experimental counterparts directly without any frequency scaling. The correlation between experimental and theoretical frequencies was based on a manual normal mode assignment of individual bands. The shown spectra represent the experimentally measured or computed absolute intensities plotted against experimental or non-scaled computed frequencies, respectively, with a Lorentzian broadening (fwhm = 30 cm^{-1}).

9.3. Results and Discussion

Cluster generation is performed with the QCG algorithm combining a force field-based rigid-structure docking algorithm with a subsequent fast and robust semiempirical QM GFN2-xTB treatment to add a given number of solvents molecules at energetically favorable positions around a solute. MD simulations followed by full geometry optimizations of each snapshot generate a large ensemble of structures. The final ensemble consists of the lowest energy clusters found within a 6 kcal mol^{-1} energy window. The number of clusters usually varies from 10 to 20 depending on the system. Clusters with a population of more than 10 % are included for the IR spectra (an example for individual, weighted spectra is

given in Appendix A9). Standard deviations (SD) of the averaged vibrational frequencies are typically less than 5 cm^{-1} while for the intensities, SD values usually did not exceed 10 % of the corresponding absolute intensities. In the majority of cases, the deviations of frequencies and intensities computed for the energetically lowest-lying cluster from the thermodynamically averaged values over all the clusters did not exceed the SD.

Molecules used to test our approach are chosen based on two principles: (1) availability of reliably interpreted IR spectra in gas and condensed state with quantitatively measured frequencies and intensities; (2) conformational rigidity of compounds to avoid difficulties with averaging simulated spectra of different conformers. Conformationally flexible species can be treated similarly by an initial QM-based conformational search using a continuum model as discussed recently,[60] followed by subsequent cluster generation for individual conformers and proper averaging of their spectra. The molecules cyclohexane (C_6H_{12}), cyclohexane-d_{12} (C_6D_{12}), 1,4-dioxane ($C_4H_8O_2$), 1,4-dioxane-d_8 ($C_4D_8O_2$), acetone (CH_3COCH_3), acetonitrile (CH_3CN), tetrahydropyrane ($C_5H_{10}O$), pyridine (C_5H_5N), benzene (C_6H_6), benzene-d_6 (C_6D_6), methanol (CH_3OH), and methanol-d_4 (CD_3OD) were considered. These molecules, except methanol, do not form hydrogen bonds in neat liquids/solids or CCl_4/CS_2 solutions. Nevertheless, their IR intensities change significantly when going from the gas to a condensed state. For example, the solution of nonpolar cyclohexane-d_{12} in an innocent solvent like CCl_4 results in an almost two-fold increase of absolute IR intensities (A) of some bands (Table 9.1). Many attempts have

Table 9.1.: Experimental IR Intensities (A, km·mol^{-1})[613] of cyclohexane-d_{12} in the gas phase and in CCl_4/CS_2 solution and the corresponding B97-3c computed values. Values are given in cm^{-1}.

Integration range	1240–1130	1130–1045	1045–950	950–880	770–600	450–350
A^{gas}(exp.)	6.4	6.5	5.9	3.5	3.5	0.4
A^{gas}(comp.)	9.4	7.8	6.4	4.0	4.4	0.4
$A^{sol.}$(exp.)	4.6	8.7	5.4	4.6	5.6	0.8
$A^{sol.}$(comp.)	6.1 (0.3)[a]	9.9 (0.5)[a]	4.7 (0.4)[a]	4.5 (0.1)[a]	8.8 (2.0)[b], 8.3 (1.0)[c]	0.7 (0.1)[a]
$A^{sol.}/A^{gas}$(exp.)	0.7	1.4	0.9	1.3	1.6	2.0
$A^{sol.}/A^{gas}$(comp.)	0.7 (0.04)[a]	1.3 (0.1)[a]	0.7 (0.1)[a]	1.1 (0.04)[a]	2.0 (0.5)[b], 1.9 (0.2)[c]	1.8 (0.2)[a]
$A^{sol.}/A^{gas}$(COSMO)	1.2	1.2	1.2	1.4	1.5	2.0

[a] for 1:19 C_6D_{12}/CCl_4 ratio.
[b] for 1:20 C_6D_{12}/CS_2 ratio.
[c] for 1:30 C_6D_{12}/CS_2 ratio.

been made to explain the ratio of band intensities in solution and gas phases assuming the absence of specific intermolecular interactions. In particular, according to the Polo–Wilson equation[614]

$$\frac{A^{sol.}}{A^{gas}} = \frac{\left(n^2+2\right)^2}{9n}, \tag{9.1}$$

where n is the refractive index of the medium, the intensity enhancement should be 1.3 for CCl_4. Solvent effects can be modeled by a QM description of a solute including a classical polarizable continuum within the framework of the COSMO model.[215] This approach predicts an intensification of each IR band by factors from 1.2 to 2.0 while experimental $A^{sol.}/A^{gas}$ ratios for cyclohexane-d_{12} range from 0.7

to 2.0 (Table 9.1). The values of A(cluster)/A(isolated molecule) computed with our cluster model also varies from 0.7 to 2.0 and reproduce all qualitative trends (increase or decrease of A) found experimentally for each pair of compared IR bands (Table 9.1).

The excellent quality of the cluster modeling approach for C_6D_{12} is demonstrated in Figure 9.1. The main redistribution of relative intensities of the IR bands found in experimental spectra in passing from gas to nonpolar solution is reproduced by the simulations. In contrast, the COSMO model yields qualitatively wrong condensed state-induced changes for relative intensities of IR bands in the spectral region of 1200–900 cm^{-1}.

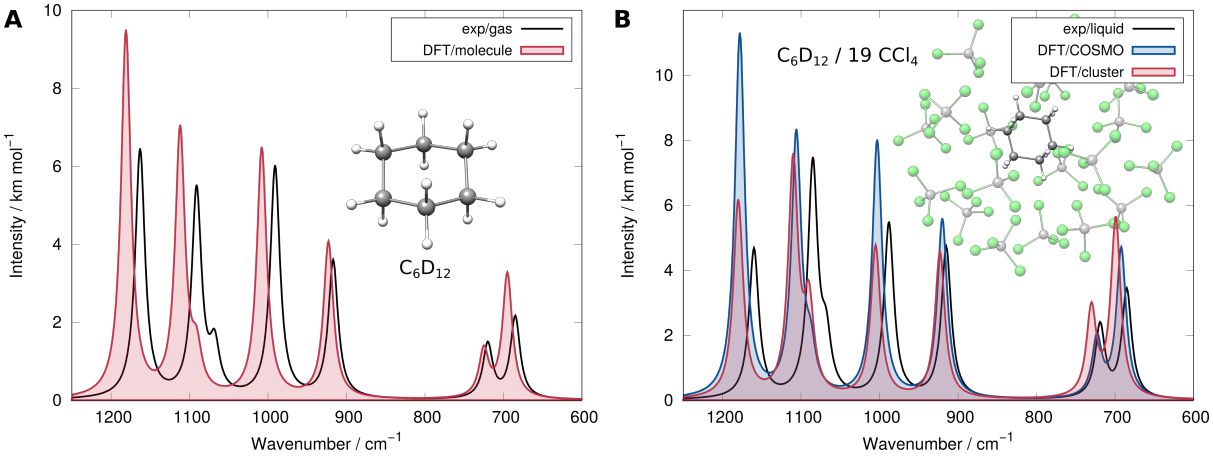


Figure 9.1.: (A) Experiment for gas (black) and B97-3c simulations for the isolated molecule (red). (B) Experiment for CCl_4/CS_2 solutions (black) and the corresponding cluster (red) comprising one C_6D_{12} and 19 CCl_4 molecules (or 20 CS_2 molecules for 730–680 cm^{-1} range) or COSMO modeling (blue).

The same is true for the comparative analysis of the IR intensities of polar acetone in the gas phase and in an amorphous solid sample (Figure 9.2, Table 9.2). Gas/amorphous-solid intensity ratios collected in Table 9.2 demonstrate that the observed failure is not specific for the COSMO model but is similarly observed for other widely used continuum methods like PCM[207–209] or SMD[210]. Thus, the problem seems to be unrelated to the different treatments (or entire neglect) of cavitation/dispersion contribution to solvation but instead is pointing to deficiencies in the electrostatic description.

The results obtained for other test molecules are collected in Tables A9.1–A9.3 (Appendix A9) and illustrate a similarly good performance of our approach. The general agreement between absolute IR intensities computed for the whole series of cluster models and their experimental condensed-phase counterparts is practically the same as in the case of the isolated gas phase molecules (Figure 9.3). Even slightly better results are obtained for relative IR intensities (Figure A9.2).

Our cluster model also nicely reproduces dramatic differences in both the intensities and vibrational frequencies of liquid methanol (CH_3OH) and methanol-d_4 (CD_3OD) compared to their vapors,[616,617] which are caused by formation of strong hydrogen bonds in the liquid state (Figure 9.4 and Table A9.3). The huge intensification and pronounced red shift of the hydroxyl group stretching bands (ν_{OH}/ν_{OD}) are reproduced already by the smallest clusters, modeling the first solvation shell of the

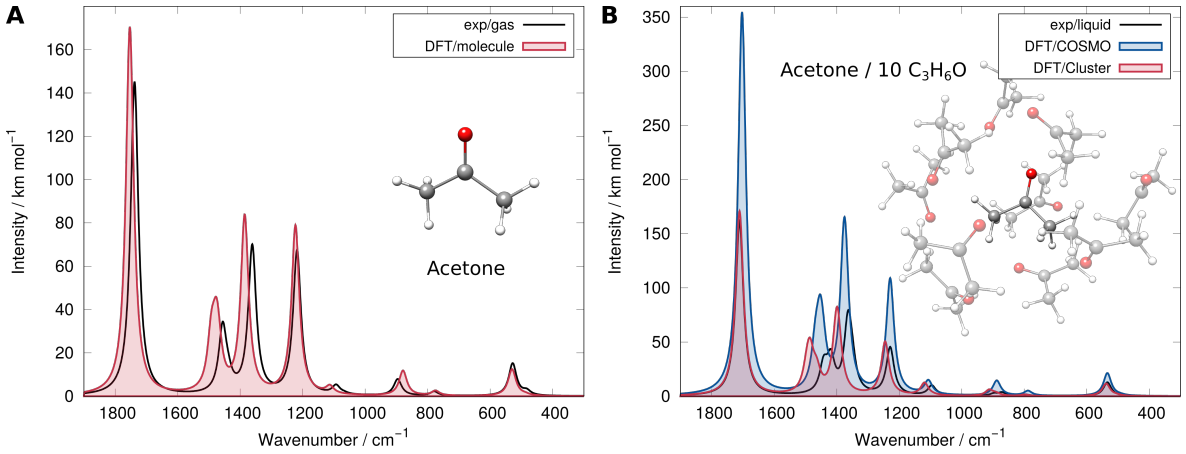


Figure 9.2.: IR spectra of acetone in the spectral region of 1800–400 cm^{-1}. (A) Experiment for gas (black) and amorphous solid at 100 K (red). (B) B97-3c simulations for the isolated molecule (black), cluster comprising 11 acetone molecules (red), and COSMO model (blue).

Table 9.2.: Experimental IR intensities (A, km·mol^{-1}) of acetone in gas[613] and amorphous low-temperature solid[615] and the corresponding B97-3c Computed Values[a]. Values are given in cm^{-1}.

Frequency[613]	1738	1456–1438	1363–1360	1218	1093–1067	896	779	528–483
A^{gas}(exp.)	145 (5)	35	69	66.5 (1.5)	4.5 (0.2)	7.6 (0.3)	1.9 (0.4)	17
A^{gas}(comp.)	170.2	56.4	81.9	77.9	3.4	11.5	2.2	12.8
Integration range[615]	1740–1680	1500–1380	1380-1335	1250–1210	1100–1080	940–840	820–760	545–524
$A^{sol.}$(exp.)	161	55	84	44	9.5	5.0	1.0	12.8
$A^{sol.}$(comp.)[b]	170 (9)	67 (6)	82 (4)	49 (3)	12 (2)	7 (1)	1.4 (0.4)	12.2 (0.6)
$A^{sol.}/A^{gas}$(exp.)	1.1	1.6	1.2	0.7	2.1	0.7	0.5	0.8
$A^{sol.}/A^{gas}$(comp)	1.0	1.2	1.0	0.6	3.4	0.6	0.6	0.9
$A^{sol.}/A^{gas}$(COSMO)	2.1	1.9	2.0	1.4	3.7	1.2	2.0	1.7
$A^{sol.}/A^{gas}$(CPCM)	1.8	1.7	1.7	1.4	3.1	1.3	1.6	1.6
$A^{sol.}/A^{gas}$(IEFPCM)	1.8	1.7	1.6	1.4	3.0	1.3	1.6	1.6
$A^{sol.}/A^{gas}$(SMD)	1.8	1.9	2.0	1.4	4.5	1.5	1.6	1.7

[a] In parentheses: SD for experimental measurements or cluster computations.
[b] For cluster comprising 11 acetone molecules.

central methanol molecule. In particular, the red shifts of ν_{OH}/ν_{OD} computed in this simplest model, which comprises 20 methanol molecules ($\Delta\nu_{OH/OD}$= 292/210 cm^{-1}), are very close to the corresponding experimental values (318/224 cm^{-1}). The same is true for the computed absolute intensity $A_{\nu_{OH}}/A_{\nu_{OD}}$: 601/319 km·mol^{-1} versus experimental 585/310 km·mol^{-1}. Increase of the cluster size to 30 molecules results in an increase of the computed red shifts $\Delta\nu_{OH/OD}$ to 483/349 cm^{-1} caused by cooperative strengthening of hydrogen bonds. Further increase of the number of H-bonded methanol molecules in the cluster practically does not influence the $\Delta\nu_{OH/OD}$ values.

The dependence of the $A_{\nu_{OH}}/A_{\nu_{OD}}$ value on the cluster size is more pronounced: 631/334 and 790/407 km·mol^{-1} for 30 and 40 methanol molecules, respectively. The computed $A_{\nu_{OH}}/A_{\nu_{OD}}$ and ν_{OH}/ν_{OD} val-

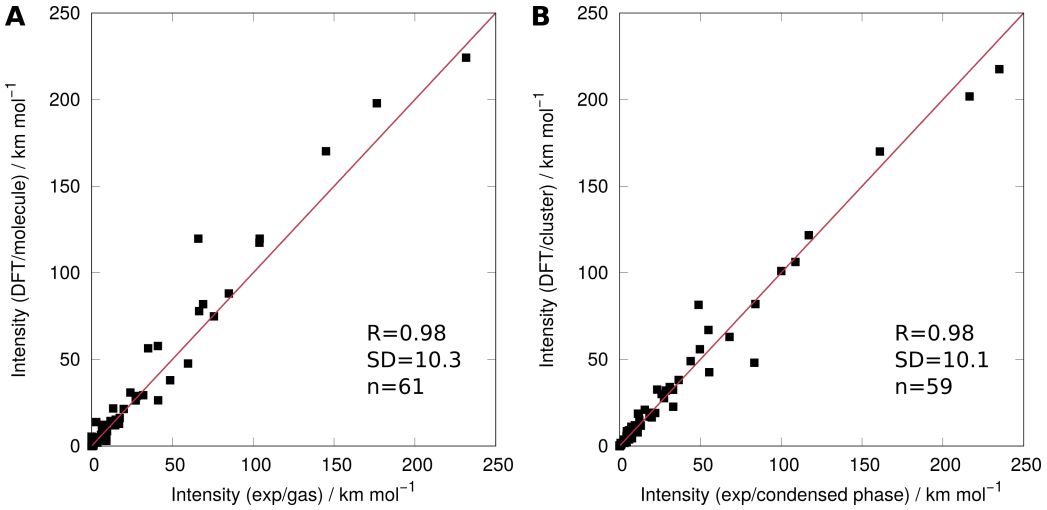


Figure 9.3.: Correlation plot of absolute IR intensities (km·mol^{-1}) calculated for the isolated molecules of cyclohexane, cyclohexane-d_{12}, 1,4-dioxane, 1,4-dioxane-d_8, acetone, tetrahydropyrane, pyridine, benzene, and benzene-d_6 (A) and corresponding cluster models (B) vs experimental gas phase (A) and condensed-phase (B) intensities, respectively. *R*, correlation coefficient; SD, standard deviation; and *n*, total number of data included in the analysis.

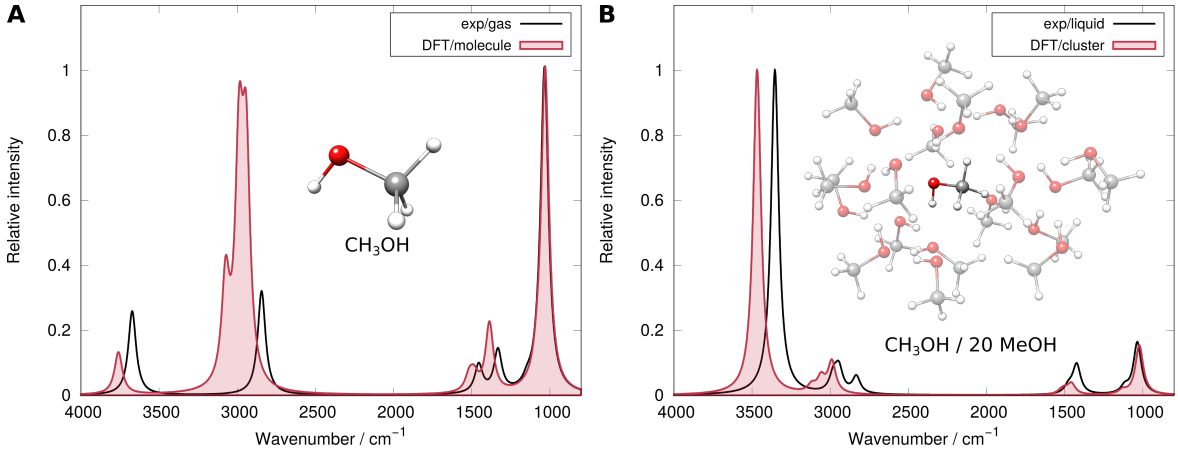


Figure 9.4.: IR spectra of CH_3OH in the spectral region of 4000–800 cm^{-1}. (A) Experiment for gas (black) and B97-3c simulations for isolated molecule (red). (B) Experiment for liquid (black) and the corresponding cluster model (red) comprising 20 methanol molecules.

ues are also affected by a rather strong coupling of vibrations of neighboring methanol molecules (see Appendix A9 for details). Nevertheless, both the coupling and the cooperative strengthening of the H-bonds in the larger clusters produce only minor re-distributions of the relative IR intensities and moderate shifts of the bands in the simulated spectra, still allowing a reliable correlation of the computed and experimental bands (Figure A9.1). The same is true for the possible influence of anharmonicity effects hampering direct comparison of the computed (harmonic) and experimental (anharmonic) IR spectra. According to recent experimental[618] and quantum chemical[619] data, anharmonicity effects

on the $\Delta\nu_{OH/OD}$ and $A_{\nu_{OH}}/A_{\nu_{OD}}$ values are similar for the isolated methanol molecule and its various H-bonded clusters and does not exceed 20 %.

To rationalize possible main reasons for the condensed state induced IR intensity variations in the absence of H-bonding, which are well described by our cluster models (*vide supra*), we discuss the following factors: (1) geometry changes; (2) intermolecular charge transfer; (3) charge redistribution within the solute molecule; and (4) changes of normal modes of the solute molecule induced by the solvent. Geometry effects turned out to be rather small. For example, CC and CH bond lengths in cyclohexane/CCl_4 clusters shorten by less than 0.003 Å relative to the gas phase, and bond angles change by $\leq$ 0.1–0.2°. Charge transfer between solute/solvent molecules in the clusters is rather small, for example, $\leq$ 0.01 electrons for relatively weakly interacting systems like 1,4-dioxane or cyclohexane in CCl_4. Changes of Mulliken atomic charges, as a qualitative measure of electron density redistribution, within a solute molecule in the cluster compared to the gas are somewhat larger with about 0.05 electrons at most (up to 0.07 in methanol clusters). Normal modes remain qualitatively the same irrespective of the state, but some atomic amplitudes vary within ca. ±20 % in passing from the isolated molecule to the same molecule in a cluster. An example illustrating the above effects of atomic charge and normal mode changes on IR intensity variations can be found in Appendix A9, Table A9.1. Clearly, factors (3) and (4) can simultaneously influence the molecular dipole moment derivative determining the IR intensity of the vibration. Tentatively, we can state that factor (4) is most relevant, at least for relative intensities. For example, Table A9.4 in Appendix A9 demonstrates that more pronounced cluster-induced "deformation" of the normal modes of cyclohexane-d_{12} is accompanied by stronger relative intensity change.

In order to assess more generally the accuracy of the cluster approach for condensed state vibrational frequencies, we included, in addition to methanol and methanol-d_4, frequencies of acetone,[613,615] which also undergo rather strong shifts upon the "vapor-amorphous solid" transition but for reasons different from the H-bond formation. To minimize the influence of the neglected anharmonicity effects, the bands of the most anharmonic CH/CD and OH/OD stretching modes were excluded. The agreement between the frequencies computed for cluster models of these compounds and their experimental condensed-phase counterparts is the same as in the case of the isolated gas phase molecules (Figure A9.3).

It should be noted that the influence of the condensed state on relative IR intensities can also be dramatic even in the absence of H-bonds, which is demonstrated by the example of acetonitrile.[620,621] Again, a minimal-size cluster model qualitatively reproduces the very strong intensification of the ν_{CN} band at ca. 2300 cm^{-1} relative to the CH3-bending bands at ca. 1500 cm^{-1} (Figure 9.5). The performance of our approach is rather independent of the cluster size (11, 15, or 21 acetonitrile molecules, as shown in Figure 9.4).

Fast convergence of the quality of computed IR spectra with the cluster size is found for all studied systems and this is further demonstrated in Appendix A9, for the example of modeling liquid benzene. This enables a huge reduction of the computational costs because clusters comprising only the first solvation shell of a solute molecule produce IR spectra of sufficient accuracy (Tables 9.1, 9.2, A9.1–A9.3). The wall time required for such computations carried out in parallel on 14 processors varies from about 2 hours (*e.g.*, for clusters comprising 11 acetone molecules, Figure 9.2) to about 1 day (*e.g.*, for 16 C_6H_{12}).

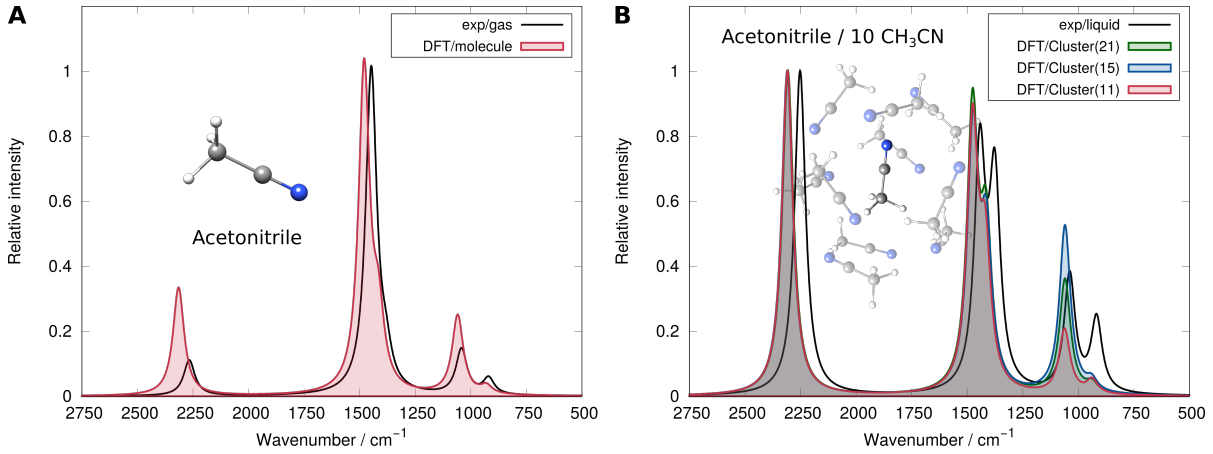


Figure 9.5.: IR spectra of acetonitrile in the spectral region of 2500–750 cm^{-1}. (A) Experiment[620] for gas (black) and B97-3c simulations for the isolated molecule (red). (B) Experiment[621] for liquid (black), and the corresponding cluster model comprising 11 (red), 15 (blue), and 21 (green) acetonitrile molecules.

The entire procedure from cluster generation to the final spectrum took from one to four days of wall computation time, which is suitable even for application in routine spectrochemical practice. Further reduction of the required computational resources can be achieved in a simplified workflow, where the IR spectrum is calculated only for the energetically most-stable cluster. The difference between this and the thermodynamically averaged spectrum is mostly statistically insignificant. This is true even for the case of methanol clusters, where minor variations of mutual positions of methanol molecules influence the H-bonding pattern and produce differences of OH stretching frequencies of ca. 130 cm^{-1} accompanied with changes of IR intensities by ca. 200 km·mol^{-1} (see Appendix A9).

9.4. Conclusion

We presented an efficient approach to an accurate QM modeling of IR spectra of condensed-phase systems. A fully automatic cluster generation algorithm is applied to construct an ensemble of explicit solvated molecular clusters that are reoptimized at a reasonable but efficient DFT level of theory. The latter was represented by the B97-3c composite scheme but in principle could be replaced by any reasonable quantum chemical method. For example, the B3LYP-D3/def2-TZVP level may yield a slightly more accurate description, of course at a much higher computational cost (factor of 5–10). The IR spectrum of the system, calculated as a Boltzmann average of the DFT spectra, matches the corresponding condensed-phase IR experiment better, or at least no worse, than the theoretical spectrum of the isolated molecule compared to the gas phase IR spectrum. In order to simulate IR spectra in the condensed phase with similar accuracy as for individual molecules in vacuum, it was sufficient to take into account mainly the first solvation shell of the solute. This holds even for strongly hydrogen-bonded systems like methanol and suggests that the vibrational modes and dipole moment derivatives are mainly influenced locally by their immediate surroundings. We hope that the outcome of our study will help

researchers to simplify the modeling of common condensed-phase systems and furthermore represents a practical tool for computational chemists for various analytical purposes.

Acknowledgments

Travel grant from scholarship programs of the German Academic Exchange Service (DAAD, personal Ref. number 91577762) is gratefully acknowledged (S.A.K.).

Part V.

Final Summary and Conclusion

Science evolves rapidly and new areas of research continue to emerge. This is reflected in the constantly rising number of articles published in major scientific journals. As an example, the number of published articles in the German general chemistry flagship journal *Angewandte Chemie* has almost doubled within five years from 2009 to 2014.[622] Chemical compounds that are state-of-the-art in energy storage, drug delivery, biocatalysis, and supramolecular chemistry always increase in molecular size, complexity, and the variety of elements. This progress continues to pose new challenges to the field of computational chemistry, which provides theoretical models to predict or interpret experimental results. Here, a valid and often chosen approach is to design a separate specially adapted method for each chemical problem. Consequently, the number of theoretical models must be equally large and versatile as the whole of chemistry. From a purist's point of view, a general approach is more desirable, where the underlying physics describes a broad range of molecules with equal accuracy. Following this philosophy, the presented doctoral thesis was devoted to the development and application of efficient computational methods for describing large molecules with an elemental composition across the periodic table. This thesis aimed to develop methods for the improvement of free energy calculations and all underlying contributions (*cf.* Section 2.3) or to make it possible for large molecules (> 1000 atoms) in the first place.

The starting point was given by the recently introduced GFN*n*-xTB methods developed by Grimme and co-workers, whose conception allows the fully quantum mechanical treatment of a wide range of organic as well as inorganic systems (*cf.* Section 2.1.6). GFN*n*-xTB methods feature a full periodic table parameterization and are thus applicable to a broad range of molecules. Nevertheless, the QM nature limits the molecular size to roughly 1000 atoms for geometries, harmonic frequencies, and noncovalent interaction energies. To extend the treatable molecule size scale even further, a generic force-field was developed in the course of this thesis (Part II, Chapter 3), building upon the existing GFN framework. In relation to the anticipated target properties, this FF method was termed GFN-FF.[117] Formally, GFN-FF can be regarded as a zeroth-order-only TB scheme, where the QM terms in form of extended-Hückel-type theory (EHT) were replaced by classical bond, angle, and torsion terms. An exponentially damped $1/R$ term for the nuclear repulsion, a variation of the D4 dispersion correction, and a sophisticated electrostatic interaction term based on the EEQ model completed the total GFN-FF energy expression (*cf.* Section 3.2). In close accordance with the GFN*n*-xTB methods, mostly global and element-specific parameters for the entire periodic table up to radon ($Z \leq 86$) were included in GFN-FF. With this generic FF method at hand, new possibilities in the description of large and complex molecules opened up. Chapter 3 started with a showcase of five large metal-organic frameworks, where none of them was fully optimized before by any other method, due to either the sheer size or the complicated electronic structure. Here, the GFN-FF optimized structures showed excellent agreement with the experimental crystal structure cutouts. A remarkable accuracy for a FF method was also achieved for the small to medium sized transition metal complex structures of the TMG145 benchmark set,[221] where the structural RMSD of GFN-FF was close to the RMSD of GFN2-xTB. In the context of biomacromolecules, geometry optimizations were performed for the metallo-protein hemoglobin (~10.000 atoms, RMSD = 1.02 Å) and for an entire benchmark set of 70 organic peptide and protein structures.[224] For the latter set, GFN-FF showed about the same or even better accuracy as the special-purpose methods OPLS2005 and AMBER*.

The performance for noncovalent interaction energies was tested on standard benchmark sets, such as the GMTKN55[162] and the S30L.[120] On average, GFN-FF yielded an accuracy for conformation and association energies that was comparable to GFN1-xTB, only slightly worse than GFN2-xTB, and much better than other SQM methods such as PM6-D3H4X and PM7. Even though GFN-FF was mainly fitted to *ab initio* values of small and medium sized molecules, the transferability to large (bio-)metal-organic compounds proved to be outstanding. This is mainly due to the sophisticated and physically motivated interaction terms, which distinguishes GFN-FF from other force-fields. GFN-FF provided throughout Chapter 3 high universality paired with almost QM accuracy at high computational speed and hence, a new quality standard was set for general FFs.

Part III of this thesis presented the application of GFN-FF along with the other GFN*n*-xTB methods to interdisciplinary problems ranging from material science to structural biology. In Chapter 4, the fuel storage, greenhouse gas capture, and drug delivery in porous (metal-)organic materials were modeled by GFN methods. Therefore, optimal binding sites of H_2, N_2, CO_2, CH_3OH, and C_6H_6 in various metal-organic frameworks and porous organic cages were determined by the fully automated combination scheme of GFN-FF and CREST.[273] Re-optimization of the structure ensembles at the B97-3c level of theory delivered a versatile test set of 117 different gas–cage combinations. It was demonstrated that GFN methods, in general, reproduce the DFT reference values (PBE0-D4/def2-TZVP) remarkably well and that GFN-FF reaches a comparable accuracy as the GFN*n*-xTB methods, whilst being two orders of magnitude faster. The study on gas adsorption in MOFs and POCs showed that GFN methods can be applied reliably for the generation and energetic sorting of gas–cage association complexes, which was impossible before for the given size of the investigated systems. The presented approach is already used by other groups in the context of ethylene purification.[623]

Combining CREST and GFN-FF as the underlying level of theory emerged as a powerful tool for the automatic conformational space exploration. After successfully determining the gas binding sites in porous materials, this scheme was further applied to the structure elucidation of biochemical systems. Therefore in Chapter 5, a new methodology (CREST/MD) was developed for the computation of distance distributions in spin-labeled mutants of azurin and T4 lysozyme (*cf.* Chapter 5).[274] EPR distance measurements of the corresponding spin–spin distances served as a reliable reference. The CREST/MD approach employed the default iMTD-GC algorithm at the GFN-FF level for the entire protein to determine the conformations of the MTSL label. This is to date the largest application of the *crest* program for more than 2500 atoms in total. On the energetically lowest conformations subsequent GFN-FF/GBSA(H_2O) MD simulations were performed. The overall distance distribution was then calculated as the Boltzmann weighted sum of distributions obtained from the MD trajectories of each evaluated conformer. For the tested systems, CREST/MD showed superior accuracy and the deviation from the experiment was on average less than 2 Å, which was within the experimental error. This clearly outperformed other computational approaches such as MMM and mtsslWizard. The novel CREST/MD approach provided valuable insights into the structures and dynamics of biomacromolecules and also captured temperature effects. This opens up new possibilities for the structure elucidation of biochemical systems in the future. A common binding motif in biological systems is the interaction of charged amino groups with aromatic side chains.[276] Thus, the so-called ion–π interactions, have an important

contribution to the structure and function of proteins.[275] In Chapter 6, the IONPI19 benchmark set was introduced to apply and evaluate the most advanced computational methods for the difficult ion–π interactions.[184] As references, (local) coupled cluster values were provided. It was shown that the functional performance trend along Jacob's ladder is generally obeyed and double-hybrid DFAs reproduce the CC reference values with the closest accuracy. Dispersion uncorrected DFT underestimated ion–π interactions significantly. The description of ion–π interactions with GFN methods was quite sufficient, but the error was larger than for neutral systems. This study on charged systems including ion–π interactions revealed two important findings. On the one hand, it showed that modern DFAs are well suited to describe complex molecules (up to 133 atoms) with almost CC accuracy at a much lower computational cost. On the other hand, the IONPI19 set showed that GFN methods are in general backward compatible and hence, the presented molecules with an average size of 31 atoms are equally well described as the much large systems in the previous Chapters 3–5.

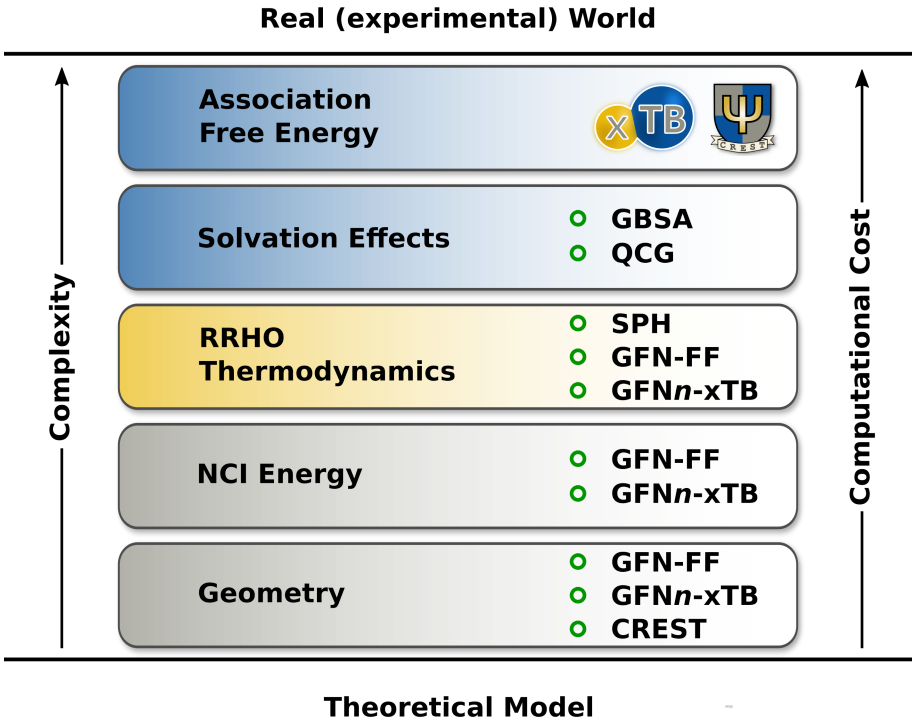


Figure 9.6.: Summarized findings and achievements of this thesis. With CREST, the GFN-FF and GFN*n*-xTB methods, SPH calculations and the GBSA solvation model, robust and efficient methods were evaluated for calculating free association energies of large molecules and all contributing terms.

Summarizing Parts II and III of this thesis, it was demonstrated that GFN-FF manages the balancing act between complex inorganic frameworks, large biochemical proteins, and small charged systems. This versatility is unique among all existing force-fields. At the beginning of this thesis, the calculation of association free energies for large molecules was set as an overall target. As an interim result, it can already be recorded that the first two principal components to ΔG, namely accurate geometries and gas

phase association energies, can now be computed with sufficient accuracy by the GFN-FF and GFN*n*-xTB methods for previously inaccessible molecule sizes (see Figure 9.6). The next consequent step towards free energy computations was the incorporation of statistical thermodynamic effects. Therefore, Part IV of this thesis evaluated and improved the quality of vibrational frequencies and thermodynamics computed with GFN methods in the modified RRHO approximation.

Due to the naming, GFN methods suggest to yield accurate harmonic frequencies and thermostatistical contributions derived therefrom. Nevertheless, this was not extensively tested in Refs. 93–95,117. In Chapter 7, the thermostatistical contribution within the mRRHO approximation was first computed by the GFN methods for a test set composed of the S22,[537] L7,[538] and S30L[120] benchmark set. For harmonic frequencies and thermodynamic quantities, the GFN methods were able to convince with small deviations from the DFT reference (B97-3c, PBEh-3c), whilst being three (GFN2-xTB) to five (GFN-FF) orders of magnitude faster. Building on this observation, the molecular size was increased to study association free energies of protein-drug binding. For rivaroxaban binding to factor Xa[543] (~5000 atoms), GFN-FF/GBSA(H_2O) yielded a binding free energy that was much closer to the reference than previous QM studies.[375] This result is highly encouraging for future studies of protein-drug binding by GFN-FF, having in mind that the calculations took on average just one day on a standard desktop computer with 4 CPUs. So far, the calculation of accurate frequencies and thermodynamics within the RRHO approximation always required a full geometry optimization to eliminate spurious imaginary frequencies. Hence, for general non-equilibrium geometries, the harmonic approximation becomes invalid and thermodynamic quantities are not accessible. This was changed in Chapter 8, where the concept of single-point Hessians was introduced. For the very first time, this novel approach enabled the calculation of G_{mRRHO} values for distorted structures, reactions paths, transition states, and MD simulation snapshots. Starting from DFT input structures, SPH calculations at the SQM (or FF) level showed to be superior to conventional calculations on fully re-optimized geometries or unrelaxed structures. Thus, SPH calculations represent the next step towards higher accuracy in routine calculations of free energies or IR spectra for large systems using efficient SQM or FF methods.

Solvent effects in the context of IR spectra were additionally investigated in Chapter 9. The strict separation of vibration and solvation effects according to Eq. 2.74 is only an approximation in solution and the vibrational frequencies of a solute molecule depend on the solvent environment. So far, solvent effects to energies, geometries, and frequencies were included only implicitly. In Chapter 9, the new quantum cluster growth approach towards explicit solvation was applied. The QCG algorithm combined an FF-based genetic docking algorithm with GFN2-xTB structure optimizations to generate explicitly solvated clusters of neat liquids. The explicit treatment of local solute–solvent hydrogen bonds improved the quality of the calculated IR spectra significantly in comparison to implicit solvent models such as COSMO or SDM. This study showed that for the high accuracy of liquid phase IR spectra, the inclusion of explicit solvent molecules is inevitable for polar solvents.

Summarizing Part IV of this thesis, also the remaining two contributions to ΔG, *i.e.*, thermodynamic quantities and solvation effects, were extensively tested and significantly improved (see Figure 9.6). Just like in the previous parts, the game-changing components were the GFN-FF and GFN*n*-xTB methods, which enabled the efficient calculation of thermodynamic properties. SPH calculation yielded further

improvements in thermodynamic quantities, not only in terms of accuracy but also in terms of applicability to general non-equilibrium geometries. Overall, it was demonstrated extensively that the GFN-FF and GFN*n*-xTB methods combine robustness, efficiency, and accuracy for the calculation of ΔG and all its contributing terms. This unique combination provides access to large molecules, versatile in their elemental composition, that could not previously be described with such high accuracy. Yet, there is still much to be done in the research field of simplified methods and numerous challenges remain.

In total, the GFN methods consist of three SQM and one FF method. Thus, the next consequent step is their mutual combination within so-called QM/MM techniques,[624] which were rewarded with the Nobel prize in 2013.[625,626] FF methods, and GFN-FF is no exception, are unable to describe the changes in the electronic structure of a system within a chemical reaction, *i.e.*, bond-breaking, bond-forming, and charge transfer.[624] Here, proper treatment requires the application of QM methods. In large systems, often only small parts are included in the actual chemical reaction and hence, just the reaction center must be treated at the QM level (GFN2-xTB), whereas the rest of the molecule is properly described by a force-field (GFN-FF). In fact, an ONIOM scheme[627] was already implemented in the *xtb* code, but it has not yet been tested extensively. This will be part of future work. Furthermore, the application of GFN methods in this work was limited mainly to the description of molecular structures. While periodic implementations of the GFN*n*-xTB methods already exist,[628] a corresponding periodic version of GFN-FF is yet missing. For further studies on materials and biological systems, at some point, periodic boundary conditions must be implemented. This will also attract many users of other free available molecular simulation software packages such as *GROMACS*[629] that feature PBCs. GFN-FF was developed from the beginning as a generic FF. Even though the general accuracy is outstanding for a force-field, certain fields of application remain where an even higher accuracy closer to (S)QM methods is desirable. Especially for conformational energies, which are in general small numbers, it was recently shown that DFT re-ranking is necessary to calculate properties, such as partition coefficients, circular dichroism (CD), or NMR spectra, with chemical accuracy.[59] In order not to change the original parameterization of GFN-FF, neural networks represent an appealing approach to construct an additive potential for special purpose applications.[630] This approach benefits from the correct inherent physics of GFN-FF and the machine learning potential only has to learn the difference to the DFT result. This is by far much simpler than constructing an ML potential from scratch. ML corrections can further be applied to improve the description of thermochemistry, protonations, and tautomerizations. For the description of molecules in solution, with GBSA and QCG, two different approaches were applied to capture solvent effects. Recently, the ALPB model[188] was implemented in the *xtb* code, as the successor of the GBSA model. For future studies, the implicit ALPB solvation model will also be applied and evaluated, since it promises to be more accurate in the description of polar solvents. The QCG algorithm represents an explicit approach not only for IR spectra computations but also for the calculation of solvation free energies. The latter quantity will be the subject of future studies. The QCG algorithm may also be used in the context of microsolvation.[631]

At the end of this thesis, the introductory scientific question and the formulated objectives are conclusively addressed. First, the development of accurate polarizable force-fields was named as one of the remaining holy grails for computational chemistry.[2] With GFN-FF, a generic, partially polarizable FF

was presented in the course of this work, which is currently one of the most sophisticated FFs for the description of molecular systems, combining unique universality and accuracy. Nevertheless, thinking of the infinitely large chemical space, plenty of systems remain, where the description with GFN-FF is insufficient or just wrong. To correct this, promising approaches including periodic boundary conditions, explicit solvation, and machine learning corrections already exist and will be applied in the near future. From today's perspective, GFN-FF already fulfills to a large extent all the specified requirements in Ref. 2 and comes closest to the desired accuracy of polarizable FFs. Yet, through the continuous progress in theoretical chemistry, a method that is state-of-the-art today can be improved even further tomorrow. The second objective of this work was the calculation of association free energies and all contributing terms (see Figure 9.6) for large molecules that were inaccessible before. Free energies enable the direct comparison between theory and experiment and are the starting point for the derivation of important molecular properties. The calculation of ΔG for every conceivable chemical compound is a difficult task and will continue to occupy theoretical chemistry in the future. Nevertheless, the methods used in this thesis (GFN-FF, GFN*n*-xTB, CREST, SPH, GBSA, QCG) represent a step forward in the accurate description of large molecules, ranging from organometallic to biochemical systems, and thus, further reduce the gap between theoretical simulations and experimental findings.

In conclusion, the results presented in this thesis are pioneering for prospective computational studies in numerous fields of chemistry. The developed theoretical models continue to approach chemical accuracy for an ever-increasing amount of compounds. This may help to predict ligand-protein binding energies, design materials and devices, and better understand the functionality of biomacromolecules. Already today, computer-aided methods are commonly applied in the support of experimentation. Theoretical models help to verify experimental findings, and computational pre-screening replaces to a certain extend laborious and costly synthetic approaches. The joint collaboration of theory and experiment can increase economic and ecological sustainability in future chemical research.

Bibliography

[1] Thiel, W. *Angew. Chem. Int. Ed.* **2011**, *50*, 9216–9217.

[2] Houk, K.; Liu, F. *Acc. Chem. Res.* **2017**, *50*, 539–543.

[3] Grimme, S.; Schreiner, P. R. *Angew. Chem. Int. Ed.* **2018**, *57*, 4170–4176.

[4] Cavalleri, M. *Int. J. Quantum Chem.* **2013**, *113*, 1–1.

[5] McArdle, S.; Endo, S.; Aspuru-Guzik, A.; Benjamin, S. C.; Yuan, X. *Rev. Mod. Phys.* **2020**, *92*, 015003.

[6] Krylov, A.; Windus, T. L.; Barnes, T.; Marin-Rimoldi, E.; Nash, J. A.; Pritchard, B.; Smith, D. G.; Altarawy, D.; Saxe, P.; Clementi, C.; Crawford, T. D.; Harrison, R. J.; Jha, S.; Pande, V. S.; Head-Gordon, T. *J. Chem. Phys.* **2018**, *149*, 180901.

[7] Breslow, R.; Tirrell, M. V.; Barton, J. K.; Barteau, M. A.; Bertozzi, C. R.; Brown, R. A.; Gast, A. P. et al. *Beyond the Molecular Frontier, Challenges for Chemistry and Chemical Engineering*; National Academies Press: Washington (DC), **2003**.

[8] *Challenges and Advances in Computational Chemistry and Physics*; Springer, **2006–2021**, Vol. 1–28.

[9] Young, D. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; John Wiley & Sons, Inc.: New York (NY), **2001**.

[10] Linder, M. *Comput. Struct. Biotechnol. J.* **2012**, *2*, e201209009.

[11] Lin, X.; Li, X.; Lin, X. *Molecules* **2020**, *25*, 1375.

[12] Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. *Pharmacol. Rev.* **2014**, *66*, 334–395.

[13] Hafner, J.; Wolverton, C.; Ceder, G. *MRS Bulletin* **2006**, *31*, 659–665.

[14] Rogl, P.; Podloucky, R.; Wolf, W. *J. Phase Equilibria Diffus.* **2014**, *35*, 221–222.

[15] Jensen, F. *Introduction to Computational Chemistry*; Wiley, **2007**, Vol. 2.

[16] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover Publications, **1996**.

[17] Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; J. Wiley: New York, **2000**.

[18] Schaefer, H. F.; *Methods of electronic structure theory*; Springer Science & Business Media, **2013**.

[19] Peat, T. S.; Dolezal, O.; Newman, J.; Mobley, D.; Deadman, J. J. *J. Comput. Aided Mol. Des.* **2014**, *28*, 347–362.

[20] Skillman, A. G. *J. Comput. Aided Mol. Des.* **2012**, *26*, 473–474.

[21] Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K. *J. Comput. Aided Mol. Des.* **2014**, *28*, 305–317.

[22] Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1139–1149.

[23] MacMillan, D. W. *Nature* **2008**, *455*, 304–308.

[24] Enders, D.; Niemeier, O.; Henseler, A. *Chem. Rev.* **2007**, *107*, 5606–5655.

[25] Dalko, P. I.; Moisan, L. *Angew. Chem. Int. Ed.* **2004**, *43*, 5138–5175.

[26] Crabtree, R. H. *J. Organomet. Chem.* **2004**, *689*, 4083–4091.

[27] Sinko, W.; Lindert, S.; McCammon, J. A. *Chem. Biol. Drug Des.* **2013**, *81*, 41–49.

[28] Verma, J.; Khedkar, V. M.; Coutinho, E. C. *Curr. Top. Med. Chem.* **2010**, *10*, 95–115.

[29] Hartenfeller, M.; Schneider, G. *Chemoinformatics and computational chemical biology* **2010**, 299–323.

[30] Noble, M. E.; Endicott, J. A.; Johnson, L. N. *Science* **2004**, *303*, 1800–1805.

[31] Marom, R.; Amalraj, S. F.; Leifer, N.; Jacob, D.; Aurbach, D. *J. Mater. Chem.* **2011**, *21*, 9938–9954.

[32] Yersin, H. *Highly efficient OLEDs with phosphorescent materials*; Wiley Online Library, **2008**.

[33] Sagara, Y.; Mutai, T.; Yoshikawa, I.; Araki, K. *J. Am. Chem. Soc.* **2007**, *129*, 1520–1521.

[34] Li, H.; Wang, K.; Sun, Y.; Lollar, C. T.; Li, J.; Zhou, H.-C. *Mater. Today* **2018**, *21*, 108–121.

[35] Mason, J. A.; Veenstra, M.; Long, J. R. *Chem. Sci.* **2014**, *5*, 32–51.

[36] Farrusseng, D. *Metal-organic frameworks: applications from catalysis to gas storage*; John Wiley & Sons, **2011**.

[37] Morris, R. E.; Wheatley, P. S. *Angew. Chem. Int. Ed.* **2008**, *47*, 4966–4981.

[38] Sheldon, R. A.; Woodley, J. M. *Chem. Rev.* **2018**, *118*, 801–838.

[39] Fotouhi, A.; Auger, D. J.; Propp, K.; Longo, S.; Wild, M. *Renew. Sustain. Energy Rev.* **2016**, *56*, 1008–1021.

[40] Ishida, T.; Haruta, M. *Angew. Chem. Int. Ed.* **2007**, *46*, 7154–7156.

[41] Collins, T. *Science* **2001**, *291*, 48–49.

[42] Hillisch, A.; Heinrich, N.; Wild, H. *Chem. Med. Chem.* **2015**, *10*, 1958–1962.

[43] Boyd, D. B. *Rev. Comput. Chem.* **2007**, *23*, 401.

[44] Shakhnovich, E. *Chem. Rev.* **2006**, *106*, 1559–1588.

[45] Alberty, R. A. *Thermodynamics of biochemical reactions*; John Wiley & Sons, **2005**.

[46] Rowan, S. J.; Cantrill, S. J.; Cousins, G. R.; Sanders, J. K.; Stoddart, J. F. *Angew. Chem. In. Ed.* **2002**, *41*, 898–952.

[47] Gibbs, J. W. *Trans. Conn. Acad. Arts Sci.* **1873**, 382–404.

[48] Atkins, P.; Jones, L. *Chemical principles: The quest for insight*; Macmillan, **2007**.

[49] Georg, I.; Teichmann, J.; Bursch, M.; Tillmann, J.; Endeward, B.; Bolte, M.; Lerner, H.-W.; Grimme, S.; Wagner, M. *J. Am. Chem. Soc.* **2018**, *140*, 9696–9708.

[50] Stepen, A. J.; Bursch, M.; Grimme, S.; Stephan, D. W.; Paradies, J. *Angew. Chem. Int. Ed.* **2018**, *57*, 15253–15256.

[51] Laidler, K. J.; King, M. C. *J. Phys. Chem.* **1983**, *87*, 2657–2664.

[52] Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. *J. Phys. Chem.* **1996**, *100*, 12771–12800.

[53] Makov, G.; Payne, M. *Phys. Rev. B* **1995**, *51*, 4014.

[54] Sauer, J. *Chem. Rev.* **1989**, *89*, 199–255.

[55] Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. *J. Cheminformatics* **2012**, *4*, 17.

[56] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. *Acta Cryst. B* **2016**, *72*, 171–179.

[57] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

[58] Schlegel, H. B. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 790–809.

[59] Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

[60] Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

[61] Pracht, P.; Grimme, S. *Chem. Sci.* **2021**, *12*, 6551–6568.

[62] Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198–210.

[63] Zhu, X.; Lopes, P. E. M.; MacKerell, A. D. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 167–185.

[64] Brooks, B. R.; Brooks III, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

[65] Banks, J. L. et al. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

[66] Born, M.; Oppenheimer, R. *Ann. Phys.* **1927**, *389*, 457–484.

[67] Schrödinger, E. *Phys. Rev.* **1926**, *28*, 1049.

[68] Jensen, J. H. *Phys. Chem. Chem. Phys.* **2015**, *17*, 12441–12451.

[69] Zhou, H.-X.; Gilson, M. K. *Chem. Rev* **2009**, *109*, 4092–4107.

[70] Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.

[71] Tomasi, J. *Theor. Chem. Acc.* **2004**, *112*, 184–203.

[72] Mark, P.; Nilsson, L. *J. Phys. Chem. A* **2001**, *105*, 9954–9960.

[73] Dirac, P. A. M.; Fowler, R. H. *Proc. R. Soc. Lond.* **1929**, *123*, 714–733.

[74] Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1988**, *145*, 514–522.

[75] Waldrop, M. M. *Nature News* **2016**, *530*, 144.

[76] Lundstrom, M. *Science* **2003**, *299*, 210–211.

[77] Yang, K. *Clin. Transl. Sci.* **2020**, *13*, 842–844.

[78] Hartree, D. R. *Math. Proc. Cambridge* **1928**, *24*, 89–110.

[79] Fock, V. *Z. Physik* **1930**, *61*, 126–148.

[80] Slater, J. C. *Phys. Rev.* **1929**, *34*, 1293.

[81] Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

[82] Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

[83] Burke, K. *J. Chem. Phys.* **2012**, *136*, 150901.

[84] Kohn, W. *Rev. Mod. Phys.* **1999**, *71*, 1253.

[85] Furukawa, S.; Horike, N.; Kondo, M.; Hijikata, Y.; Carne-Sanchez, A.; Larpent, P.; Louvain, N.; Diring, S.; Sato, H.; Matsuda, R. *Inorg. Chem.* **2016**, *55*, 10843–10846.

[86] Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141–151.

[87] Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.

[88] Stewart, J. J. P. *J. Mol. Model.* **2013**, *19*, 1–32.

[89] Elstner, M.; Seifert, G. *Philos. Trans. R. Soc. A* **2014**, *372*, 20120483.

[90] Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947–12957.

[91] Seifert, G.; Porezag, D.; Frauenheim, T. *Int. J. Quantum Chem.* **1996**, *58*, 185–192.

[92] Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

[93] Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

[94] Grimme, S.; Bannwarth, C.; Shushkov, P. *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

[95] Pracht, P.; Caldeweyher, E.; Ehlert, E.; Grimme, S. ChemRxiv, **2019**, DOI: 10.26434/chemrxiv.8326202.v1.

[96] Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020**, e01493.

[97] Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. *Chem. Rev.* **2016**, *116*, 5301–5337.

[98] Thiel, W. *WIREs Comput. Mol. Sci.* **2014**, *4*, 145–157.

[99] Nar, H.; Huber, R.; Messerschmidt, A.; Filippou, A. C.; Barth, M.; Jaquinod, M.; van de Kamp, M.; Canters, G. W. *Eur. J. Biochem.* **1992**, *205*, 1123–1129.

[100] Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–242.

[101] Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell Jr, A. D. *Chem. Rev.* **2016**, *116*, 4983–5013.

[102] Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

[103] Yu, A.; Pak, A. J.; He, P.; Monje-Galvan, V.; Casalino, L.; Gaieb, Z.; Dommer, A. C.; Amaro, R. E.; Voth, G. A. *Biophys. J.* **2021**, *120*, 1097–1104.

[104] Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027–1044.

[105] Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694–701.

[106] Tropsha, A. *Mol. Inform.* **2010**, *29*, 476–488.

[107] Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R. *J. Med. Chem.* **2014**, *57*, 4977–5010.

[108] Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. *Nat. Commun.* **2019**, *10*, 1–8.

[109] Fariselli, P.; Pazos, F.; Valencia, A.; Casadio, R. *Eur. J. Biochem.* **2002**, *269*, 1356–1361.

[110] Goh, G. B.; Hodas, N. O.; Vishnu, A. *J. Comput. Chem.* **2017**, *38*, 1291–1307.

[111] Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. *Sci. Adv.* **2017**, *3*, e1603015.

[112] Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Sorensen, D. *LAPACK Users' Guide, 3rd ed.*; Society for Industrial and Applied Mathematics: Philadelphia, PA, **1999**.

[113] Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

[114] Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

[115] Mayo, S. L.; Olafson, B. D.; Goddard, W. A. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

[116] Shi, S.; Yan, L.; Yang, Y.; Fisher-Shaulsky, J.; Thacher, T. *J. Comput. Chem.* **2003**, *24*, 1059–1076.

[117] Spicher, S.; Grimme, S. *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.

[118] Hall, P. *J. Chem. Edu.* **1995**, *72*, A71.

[119] Grimme, S. *Chem. Eur. J.* **2012**, *18*, 9955–9964.

[120] Sure, R.; Grimme, S. *J. Chem. Theory Comput.* **2015**, *11*, 3785–3801.

[121] Bohm, D *Quantum Theory*; Dover Publications: New York, **1989**.

[122] Mohr, P. J.; Taylor, B. N.; Newell, D. B. *J. Phys. Chem. Ref. Data* **2008**, *80*, 633–1284.

[123] Dirac, P. A. M. *Math. Proc. Cambridge* **1939**, *35*, 416–418.

[124] Ekeland, I. *J. Math. Anal. and Appl.* **1974**, *47*, 324–353.

[125] Pauli, W. *Phys. Rev.* **1940**, *58*, 716.

[126] Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69.

[127] Hall, G. *Proc. R. Soc. Lond.* **1951**, *205*, 541–552.

[128] Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

[129] Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265–3269.

[130] Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[131] Peterson, K. A. *J. Chem. Phys.* **2003**, *119*, 11099–11112.

[132] Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2003**, *119*, 11113–11123.

[133] Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

[134] Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

[135] Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753–12762.

[136] Hill, J. G. *Int. J. Quantum Chem.* **2013**, *113*, 21–34.

[137] Gutowski, M.; Van Lenthe, J.; Verbeek, J.; Van Duijneveldt, F.; Chałasinski, G. *Chem. Phys. Lett.* **1986**, *124*, 370–375.

[138] Xantheas, S. S. *J. Chem. Phys.* **1996**, *104*, 8821–8824.

[139] Wigner, E. *Phys. Rev.* **1934**, *46*, 1002.

[140] Löwdin, P.-O. *Phys. Rev.* **1955**, *97*, 1509.

[141] Sherrill, C. D.; Schaefer III, H. F. *Adv. Quant. Chem.* **1999**, *34*, 143–269.

[142] Szalay, P. G.; Muller, T.; Gidofalvi, G.; Lischka, H.; Shepard, R. *Chem. Rev.* **2012**, *112*, 108–181.

[143] Bartlett, R. J. *Annu. Rev. Phys. Chem.* **1981**, *32*, 359–401.

[144] Bartlett, R. J.; Musiał, M. *Rev. Mod. Phys.* **2007**, *79*, 291.

[145] Knowles, P. J.; Handy, N. C. *Chem. Phys. Lett.* **1984**, *111*, 315–321.

[146] Bauschlicher Jr, C. W.; Langhoff, S. R.; Taylor, P. R.; Partrigge, H. *Chem. Phys. Lett.* **1986**, *126*, 436–440.

[147] Riplinger, C.; Neese, F. *J. Chem. Phys.* **2013**, *138*, 034106.

[148] Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. *J. Chem. Phys.* **2016**, *144*, 024109.

[149] Pavošević, F.; Peng, C.; Pinski, P.; Riplinger, C.; Neese, F.; Valeev, E. F. *J. Chem. Phys.* **2017**, *146*, 174108.

[150] Kruse, H.; Mladek, A.; Gkionis, K.; Hansen, A.; Grimme, S.; Sponer, J. *J. Chem. Theory Comput.* **2015**, *11*, 4972–4991.

[151] Hirschfelder, J. O.; Brown, W. B.; Epstein, S. T. **1964**, *1*, 255–374.

[152] Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

[153] Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1988**, *153*, 503–506.

[154] Engel, E.; Dreizler, R. M. *Density Functional Theory*; Springer, **2011**.

[155] Thomas, L. H. *Math. Proc. Camb. Philos. Soc.* **2008**, *23*, 542–548.

[156] Fermi, E. *Z. Phys.* **1928**, *48*, 73–79.

[157] Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press, **1989**.

[158] Koch, W.; Holthausen, M. C. *A Chemists Guide to Density Functional Theory*; Wiley, **2001**.

[159] Perdew, J. P.; Schmidt, K. *AIP Conf. Proc.* **2001**, *577*, 1–20.

[160] Bao, J. L.; Gagliardi, L.; Truhlar, D. G. *J. Phys. Chem. Lett.* **2018**, *9*, 2353–2358.

[161] Zhang, Y.; Yang, W. *J. Chem. Phys.* **1998**, *109*, 2604–2608.

[162] Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

[163] Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. *J. Chem. Phys.* **2005**, *123*, 062201.

[164] Dirac, P. A. **1930**, *26*, 376–385.

[165] Slater, J. C. *Phys. Rev.* **1951**, *81*, 385.

[166] Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[167] Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

[168] Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

[169] Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439–10452.

[170] Sun, J.; Ruzsinszky, A.; Perdew, J. P. *Phys. Rev. Lett.* **2015**, *115*, 036402.

[171] Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

[172] Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

[173] Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A. *J. Chem. Theory Comput.* **2010**, *6*, 3688–3703.

[174] Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.

[175] Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 4786–4791.

[176] Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. *Chem. Rev.* **2016**, *116*, 5301–5337.

[177] Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

[178] Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

[179] Hourahine, B. et al. *J. Chem. Phys.* **2020**, *152*, 124101.

[180] Jones, J. E. *Proc. R. Soc. A* **1924**, *106*, 463–477.

[181] Buckingham, R. A. *Proc. R. Soc. A* **1938**, *168*, 264–283.

[182] Hoffmann, R. *J. Chem. Phys.* **1963**, *39*, 1397–1412.

[183] Imamura, A. *J. Chem. Phys.* **1970**, *52*, 3168–3175.

[184] Spicher, S.; Caldeweyher, E.; Hansen, A.; Grimme, S. *Phys. Chem. Chem. Phys.* **2021**, *23*, 11635–11648.

[185] Ohno, K. *Theor. Chim. Act.* **1964**, *2*, 219.

[186] Klopman, G. *J. Am. Chem. Soc.* **1964**, *86*, 4450.

[187] Nishimoto, K.; Mataga, N. *Z. Phys. Chem.* **1957**, *12*, 335–338.

[188] Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. *J. Chem. Theory Comput.* **2021**, 10.1021/acs.jctc.1c00471.

[189] Clark Still, W.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

[190] Hehre, W. J.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1969**, *51*, 2657–2664.

[191] Koester, A. M.; Leboeuf, M.; Salahub, D. R. In *Molecular Electrostatic Potentials: Concepts and Applications* (Eds. Murray, J. S., Sen, K.); Elsevier, **1996**, Vol. 3, pp. 105–142.

[192] Sokalski, W.; Shibata, M.; Rein, R.; Ornstein, R. *J. Comput. Chem.* **1992**, *13*, 883–887.

[193] Sokalski, W. A.; Sawaryn, A. *J. Mol. Struct.* **1992**, *256*, 91–112.

[194] Mortier, W. J.; Ghosh, S. K.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315–4320.

[195] Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. *Phys. Rev. B* **2015**, *92*, 045131.

[196] Morse, P. M. *Phys. Rev.* **1929**, *34*, 57.

[197] Fujimoto, K.; Payal, R. S.; Hattori, T.; Shinoda, W.; Nakagaki, M.; Sakaki, S.; Okazaki, S. *J. Comput. Chem.* **2019**, *40*, 2571–2576.

[198] Tschumper, G. S. *Rev. Comput. Chem.* **2009**, *26*, 39.

[199] London, F. *Zeitschrift für Physik* **1930**, *63*, 245–279.

[200] Eisenschitz, R.; London, F. *Zeitschrift für Physik* **1930**, *60*, 491–527.

[201] Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.

[202] Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127–8134.

[203] Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.

[204] Cox, S.; Williams, D. *J. Comput. Chem.* **1981**, *2*, 304–323.

[205] Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519.

[206] Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–1069.

[207] Mennucci, B. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 386–404.

[208] Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117–129.

[209] Miertus, S.; Tomasi, J. *Chem. Phys.* **1982**, *65*, 239–245.

[210] Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

[211] Sigalov, G.; Fenley, A.; Onufriev, A. *J. Chem. Phys.* **2006**, *124*, 124902.

[212] Onufriev, A. V.; Case, D. A. *Annu. Rev. Biophys.* **2019**, *48*, 275–296.

[213] Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.

[214] Im, W.; Lee, M. S.; Brooks III, C. L. *J. Comput. Chem.* **2003**, *24*, 1691–1702.

[215] Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, *0*, 799–805.

[216] Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

[217] Eckert, F.; Klamt, A. *AIChE Journal* **2002**, *48*, 369–385.

[218] Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

[219] Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D. *J. Chem. Theory Comput.* **2017**, *13*, 1034–1043.

[220] Grimme, S. *J. Chem. Theory. Comput.* **2019**, *15*, 2847–2862.

[221] Bursch, M.; Neugebauer, H.; Grimme, S. *Angew. Chem. Int. Ed.* **2019**, *58*, 11078–11087.

[222] McDonald, D. Q.; Still, W. C. *Tetrahedron Lett.* **1992**, *33*, 7743–7746.

[223] Ferguson, D. M.; Kollman, P. A. *J. Comput. Chem.* **1991**, *12*, 620–626.

[224] Schmitz, S.; Seibert, J.; Ostermeir, K.; Hansen, A.; Göller, A. H.; Grimme, S. *J. Phys. Chem. B* **2020**, *124*, 3636–3646.

[225] Skorupskii, G.; Trump, B. A.; Kasel, T. W.; Brown, C. M.; Hendon, C. H.; Dincă, M. *Nat. Chem.* **2020**, *12*, 131–136.

[226] Bryson, J. W.; Betz, S. F.; Lu, H. S.; Suich, D. J.; Zhou, H. X.; O'Neil, K. T.; DeGrado, W. F. *Science* **1995**, *270*, 935–941.

[227] Ruijter, E.; Scheffelaar, R.; Orru, R. V. *Angew. Chem. Int. Ed.* **2011**, *50*, 6234–6246.

[228] Vriend, G. *J. Mol. Graph.* **1990**, *8*, 52–56.

[229] Gund, P.; Andose, J. D.; Rhodes, J. B.; Smith, G. M. *Science* **1980**, *208*, 1425–1431.

[230] Poliakoff, M.; Fitzpatrick, J. M.; Farren, T. R.; Anastas, P. T. *Science* **2002**, *297*, 807–810.

[231] Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. *J. Chem. Phys.* **2015**, *143*, 054107.

[232] Yilmazer, N. D.; Korth, M. *Comput. Struc. Biotechnol. J.* **2015**, *13*, 169–175.

[233] Bursch, M.; Hansen, A.; Grimme, S. *Inorg. Chem.* **2017**, *56*, 12485–12491.

[234] Momany, F. A.; Rone, R. *J. Comput. Chem.* **1992**, *13*, 888–900.

[235] MacKerell Jr., A. D.; Banavali, N.; Foloppe, N. *Biopolymers* **2000**, *56*, 257–265.

[236] Bureekaew, S.; Amirjalayer, S.; Tafipolsky, M.; Spickermann, C.; Roy, T. K.; Schmid, R. *Phys. Status Solidi B* **2013**, *250*, 1128–1141.

[237] Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. *J. Chem. Phys.* **2018**, *148*, 064104.

[238] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[239] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.

[240] Smith, J. S.; Isayev, O.; Roitberg, A. E. *Chem. Sci.* **2017**, *8*, 3192–3203.

[241] Behler, J. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930–17955.

[242] Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. *J. Chem. Phys.* **2019**, *150*, 154122.

[243] *Semiempirical Extended Tight-Binding Program Package xtb*, Version 6.3.0., **2020**, https://github.com/grimme-lab/xtb.

[244] Grimme, S.; Hansen, A. *Angew. Chem. Int. Ed.* **2015**, *54*, 12308–12313.

[245] Grimme, S. *J. Chem. Theory Comput.* **2014**, *10*, 4497–4514.

[246] Pyykkö, P.; Atsumi, M. *Chem. Eur. J.* **2009**, *15*, 186–197.

[247] Axilrod, B. M.; Teller, E. *J. Chem. Phys.* **1943**, *11*, 299–300.

[248] Muto, Y. *Proc. Phys. Math. Soc. Jpn.* **1943**, *17*, 629–631.

[249] Caldeweyher, E.; Bannwarth, C.; Grimme, S. *J. Chem. Phys.* **2017**, *147*, 034112.

[250] Floyd, R. W. *Commun. ACM* **1962**, *5*, 345.

[251] Mantina, M.; Valero, R.; Cramer, C. J.; Truhlar, D. G. *CRC Handbook of Chemistry and Physics* **2013**, *94*.

[252] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*.

[253] Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

[254] Ghosh, D. C.; Islam, N. *Int. J. Quant. Chem.* **2010**, *110*, 1206–1213.

[255] Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2005**, *123*, 024101.

[256] Bleiholder, C.; Werz, D. B.; Köppel, H.; Gleiter, R. *J. Am. Chem. Soc.* **2006**, *128*, 2666–2674.

[257] Bleiholder, C.; Gleiter, R.; Werz, D. B.; Köppel, H. *Inorg. Chem.* **2007**, *46*, 2249–2260.

[258] Férey, G. *Chem. Soc. Rev.* **2008**, *37*, 191–214.

[259] Addicoat, M. A.; Vankova, N.; Akter, I. F.; Heine, T. *J. Chem. Theory. Comput.* **2014**, *10*, 880–891.

[260] Fujita, D.; Ueda, Y.; Sato, S.; Mizuno, N.; Kumasaka, T.; Fujita, M. *Nature* **2016**, *540*, 563–566.

[261] Gong, Y.; Zhang, Y.; Qin, C.; Sun, C.; Wang, X.; Su, Z. *Angew. Chem. Int. Ed.* **2019**, *58*, 780–784.

[262] Herm, Z. R.; Wiers, B. M.; Mason, J. A.; van Baten, J. M.; Hudson, M. R.; Zajdel, P.; Brown, C. M.; Masciocchi, N.; Krishna, R.; Long, J. R. *Science* **2013**, *340*, 960–964.

[263] Rizzuto, F. J.; Nitschke, J. R. *Nat. Chem.* **2017**, *9*, 903–908.

[264] Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C. *Nature* **2020**, *582*, 289–293.

[265] Abdullin, D.; Matsuoka, H.; Yulikov, M.; Fleck, N.; Klein, C.; Spicher, S.; Hagelueken, G.; Grimme, S.; Lützen, A.; Schiemann, O. *Chem. Eur. J.* **2019**, *25*, 8820–8828.

[266] Hagelueken, G.; Ward, R.; Naismith, J. H.; Schiemann, O. *Appl. Magn. Reson.* **2012**, *42*, 377–391.

[267] Bogetti, X.; Ghosh, S.; Gamble Jarvi, A.; Wang, J.; Saxena, S. *The Journal of Physical Chemistry B* **2020**, *124*, 2788–2797.

[268] Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

[269] Boyd, P. G.; Moosavi, S. M.; Witman, M.; Smit, B. *J. Phys. Chem. Lett.* **2017**, *8*, 357–363.

[270] MacKerell Jr, A. D.; Banavali, N. K. *J. Comput. Chem.* **2000**, *21*, 105–120.

[271] Li, Y.; Yang, R. T. *Langmuir* **2007**, *23*, 12937–12944.

[272] Xu, D.; Zhang, Y. *Proteins* **2012**, *80*, 1715–1735.

[273] Spicher, S.; Bursch, M.; Grimme, S. *J. Phys. Chem. C* **2020**, *124*, 27529–27541.

[274] Spicher, S.; Abdullin, D.; Grimme, S.; Schiemann, O. *Phys. Chem. Chem. Phys.* **2020**, *22*, 24282–24290.

[275] Lund-Katz, S.; Phillips, M. C.; Mishra, V. K.; Segrest, J. P.; Anantharamaiah, G. *Biochemistry* **1995**, *34*, 9219–9226.

[276] Burley, S.; Petsko, G. *FEBS Lett.* **1986**, *203*, 139–143.

[277] Chen, Z.; Li, P.; Anderson, R.; Wang, X.; Zhang, X.; Robison, L.; Redfern, L. R.; Moribe, S.; Islamoglu, T.; Gómez-Gualdrón, D. A.; Yildirim, T.; Stoddart, J. F.; Farha, O. K. *Science* **2020**, *368*, 297–303.

[278] Cohen, R. L.; Wernick, J. *Science* **1981**, *214*, 1081–1087.

[279] Mercer, J. H. *Nature* **1978**, *271*, 321–325.

[280] Jacobson, M. Z.; Colella, W.; Golden, D. *Science* **2005**, *308*, 1901–1905.

[281] Liu, H.; Song, C.; Zhang, L.; Zhang, J.; Wang, H.; Wilkinson, D. P. *J. Power Sources* **2006**, *155*, 95–110.

[282] Hamnett, A. *Catal. Today* **1997**, *38*, 445–457.

[283] Kamarudin, S. K.; Achmad, F.; Daud, W. R. W. *Int. J. Hydrogen Energ.* **2009**, *34*, 6902–6916.

[284] Post, W. M.; Pastor, J.; Zinke, P. J.; Stangenberger, A. G. *Nature* **1985**, *317*, 613–616.

[285] Dhakshinamoorthy, A.; Garcia, H. *Chem. Soc. Rev.* **2014**, *43*, 5750–5765.

[286] Milton, R. D.; Cai, R.; Abdellaoui, S.; Leech, D.; De Lacey, A. L.; Pita, M.; Minteer, S. D. *Angew. Chem. Int. Ed.* **2017**, *56*, 2680–2683.

[287] Li, J.-R.; Sculley, J.; Zhou, H.-C. *Chem. Rev.* **2012**, *112*, 869–932.

[288] Horcajada, P.; Gref, R.; Baati, T.; Allan, P. K.; Maurin, G.; Couvreur, P.; Ferey, G.; Morris, R. E.; Serre, C. *Chem. Rev.* **2012**, *112*, 1232–1268.

[289] Kreno, L. E.; Leong, K.; Farha, O. K.; Allendorf, M.; Van Duyne, R. P.; Hupp, J. T. *Chem. Rev.* **2012**, *112*, 1105–1125.

[290] Liu, J.; Chen, L.; Cui, H.; Zhang, J.; Zhang, L.; Su, C.-Y. *Chem. Soc. Rev.* **2014**, *43*, 6011–6061.

[291] Sumida, K.; Rogow, D. L.; Mason, J. A.; McDonald, T. M.; Bloch, E. D.; Herm, Z. R.; Bae, T.-H.; Long, J. R. *Chem. Rev.* **2012**, *112*, 724–781.

[292] Ma, S.; Zhou, H.-C. *Chem, Comm,* **2010**, *46*, 44–53.

[293] Farha, O. K.; Yazaydın, A. Ö.; Eryazici, I.; Malliakas, C. D.; Hauser, B. G.; Kanatzidis, M. G.; Nguyen, S. T.; Snurr, R. Q.; Hupp, J. T. *Nat. Chem.* **2010**, *2*, 944–948.

[294] Kitagawa, S. *Angew. Chem. Int. Ed.* **2015**, *54*, 10686–10687.

[295] Yang, Q.; Zhong, C.; Chen, J.-F. *J. Phys. Chem. C* **2008**, *112*, 1562–1569.

[296] Gascon, J.; Corma, A.; Kapteijn, F.; Llabres i Xamena, F. X. *Acs Catal.* **2014**, *4*, 361–378.

[297] Horcajada, P.; Chalati, T.; Serre, C.; Gillet, B.; Sebrie, C.; Baati, T.; Eubank, J. F.; Heurtaux, D.; Clayette, P.; Kreuz, C. *Nat. Mater.* **2010**, *9*, 172–178.

[298] Vardhan, H.; Yusubov, M.; Verpoort, F. *Coord. Chem. Rev.* **2016**, *306*, 171–194.

[299] Perry Iv, J. J.; Perman, J. A.; Zaworotko, M. J. *Chem. Soc. Rev.* **2009**, *38*, 1400–1417.

[300] Tranchemontagne, D. J.; Ni, Z.; O'Keeffe, M.; Yaghi, O. M. *Angew. Chem. Int. Ed.* **2008**, *47*, 5136–5147.

[301] Mastalerz, M. *Acc. Chem. Res.* **2018**, *51*, 2411–2422.

[302] Tozawa, T.; Jones, J. T.; Swamy, S. I.; Jiang, S.; Adams, D. J.; Shakespeare, S.; Clowes, R.; Bradshaw, D.; Hasell, T.; Chong, S. Y. *Nat. Mater.* **2009**, *8*, 973–978.

[303] Briggs, M. E.; Cooper, A. I. *Chem. Mater.* **2017**, *29*, 149–157.

[304] Evans, J. D.; Huang, D. M.; Hill, M. R.; Sumby, C. J.; Sholl, D. S.; Thornton, A. W.; Doonan, C. J. *J. Phys. Chem. C* **2015**, *119*, 7746–7754.

[305] Hasell, T.; Cooper, A. I. *Nat. Rev. Mater.* **2016**, *1*, 1–14.

[306] Chen, L.; Reiss, P. S.; Chong, S. Y.; Holden, D.; Jelfs, K. E.; Hasell, T.; Little, M. A.; Kewley, A.; Briggs, M. E.; Stephenson, A. *Nat. Mater.* **2014**, *13*, 954–960.

[307] Liang, J.; Nuhnen, A.; Millan, S.; Breitzke, H.; Gvilava, V.; Buntkowsky, G.; Janiak, C. *Angew. Chem. Int. Ed.* **2020**, *59*, 6068–6073.

[308] Yang, Q.; Liu, D.; Zhong, C.; Li, J.-R. *Chem. Rev.* **2013**, *113*, 8261–8323.

[309] Dzubak, A. L.; Lin, L.-C.; Kim, J.; Swisher, J. A.; Poloni, R.; Maximoff, S. N.; Smit, B.; Gagliardi, L. *Nat. Chem.* **2012**, *4*, 810–816.

[310] He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. *J. Phys. Chem. Lett.* **2018**, *9*, 4562–4569.

[311] Haldoupis, E.; Borycz, J.; Shi, H.; Vogiatzis, K. D.; Bai, P.; Queen, W. L.; Gagliardi, L.; Siepmann, J. I. *J. Phys. Chem. C* **2015**, *119*, 16058–16071.

[312] Kharissova, O. V.; Kharisov, B. I.; González, L. T. *J. Mater. Res.* **2020**, *35*, 1424–1438.

[313] Mancuso, J. L.; Mroz, A. M.; Le, K. N.; Hendon, C. H. *Chem. Rev.* **2020**,

[314] Buda, C.; Dunietz, B. D. *J. Phys. Chem. B* **2006**, *110*, 10479–10484.

[315] S. Brahmkshatriya, P.; Dobeš, P.; Fanfrlík, J.; Řezáč, J.; Paruch, K.; Bronowska, A.; Lepšík, M.; Hobza, P. *Curr. Comput.-Aid. Drug.* **2013**, *9*, 118–129.

[316] Tafipolsky, M.; Schmid, R. *J. Phys. Chem. B* **2009**, *113*, 1341–1352.

[317] Mercado, R.; Vlaisavljevich, B.; Lin, L.-C.; Lee, K.; Lee, Y.; Mason, J. A.; Xiao, D. J.; Gonzalez, M. I.; Kapelewski, M. T.; Neaton, J. B. *J. Phys. Chem. C* **2016**, *120*, 12590–12604.

[318] Zang, J.; Nair, S.; Sholl, D. S. *J. Phys. Chem. C* **2013**, *117*, 7519–7525.

[319] Dürholt, J. P.; Fraux, G.; Coudert, F.-X.; Schmid, R. *J. Chem. Theory Comput.* **2019**, *15*, 2420–2432.

[320] Greathouse, J. A.; Allendorf, M. D. *J. Am. Chem. Soc.* **2006**, *128*, 10678–10679.

[321] Addicoat, M. A.; Vankova, N.; Akter, I. F.; Heine, T. *J. Chem. Theory Comput.* **2014**, *10*, 880–891.

[322] Vanduyfhuys, L.; Vandenbrande, S.; Verstraelen, T.; Schmid, R.; Waroquier, M.; Van Speybroeck, V. *J. Comput. Chem.* **2015**, *36*, 1015–1027.

[323] Siwaipram, S.; Bopp, P. A.; Soetens, J.-C.; Schmid, R.; Bureekaew, S. *J. Mol. Liq.* **2019**, *285*, 526–534.

[324] Yıldız, Z.; Uzun, H. *Micropor. Mesopor. Mat.* **2015**, *208*, 50–54.

[325] Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

[326] Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

[327] Mardirossian, N.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904–9924.

[328] Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

[329] Li, H.; Eddaoudi, M.; O'Keeffe, M.; Yaghi, O. M. *Nature* **1999**, *402*, 276–279.

[330] Takezawa, H.; Shitozawa, K.; Fujita, M. *Nat. Chem.* **2020**, 1–5.

[331] Alexandre, P.-E.; Zhang, W.-S.; Rominger, F.; Elbert, S. M.; Schröder, R. R.; Mastalerz, M. *Angew. Chem. Int. Ed.* **2020**, *59*, 19675–19679.

[332] Zhang, G.; Presly, O.; White, F.; Oppel, I. M.; Mastalerz, M. *Angew. Chem. Int. Ed.* **2014**, *53*, 1516–1520.

[333] Kunde, T.; Nieland, E.; Schröder, H. V.; Schalley, C. A.; Schmidt, B. M. *Chem. Comm.* **2020**, *56*, 4761–4764.

[334] Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. *Chem. Rev.* **2016**, *116*, 5105–5154.

[335] Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 91–100.

[336] Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

[337] *Semiempirical Extended Tight-Binding Program Package xtb*, Version 6.3.2., **2020**, https://github.com/grimme-lab/xtb.

[338] MOPAC2016, Stewart Computational Chemistry, **2016**; available from http://OpenMOPAC.net.

[339] Bursch, M.; Caldeweyher, E.; Hansen, A.; Neugebauer, H.; Ehlert, S.; Grimme, S. *Acc. Chem. Res.* **2019**, *52*, 258–266.

[340] Andrae, D.; Häußermann, U.; Dolg, M.; Stoll, H.; Preuß, H. *Theor. Chim. Acta* **1990**, *77*, 123–141.

[341] Eichkorn, K.; Treutler, O.; Oehm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 652–660.

[342] Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.

[343] Rappe, A. K.; Goddard III, W. A. *J. Phys. Chem.* **1991**, *95*, 3358–3363.

[344] Daniel, C. R.; Rodrigues, N. M.; da Costa Jr, N. B.; Freire, R. O. *J. Phys. Chem. C* **2015**, *119*, 23398–23406.

[345] Coutsias, E. A.; Seok, C.; Dill, K. A. *J. Comput. Chem.* **2004**, *25*, 1849–1857.

[346] Spicher, S.; Grimme, S. *J. Phys. Chem. Lett.* **2020**, *11*, 6606–6611.

[347] Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

[348] Vydrov, O. A.; Van Voorhis, T. *J. Chem. Phys.* **2010**, *133*, 244103.

[349] Pianwanit, A.; Kritayakornupong, C.; Vongachariya, A.; Selphusit, N.; Ploymeerusmee, T.; Remsungnen, T.; Nuntasri, D.; Fritzsche, S.; Hannongbua, S. *Chem. Phys.* **2008**, *349*, 77–82.

[350] Vilar, R. *Angew. Chem. Int. Ed.* **2003**, *42*, 1460–1477.

[351] Custelcean, R.; Bonnesen, P. V.; Duncan, N. C.; Zhang, X.; Watson, L. A.; Van Berkel, G.; Parson, W. B.; Hay, B. P. *J. Am. Chem. Soc.* **2012**, *134*, 8525–8534.

[352] Rausch, M.; Dyson, P. J.; Nowak-Sliwinska, P. *Adv. Ther.* **2019**, *2*, 1900042.

[353] Weiss, A.; Berndsen, R. H.; Dubois, M.; Müller, C.; Schibli, R.; Griffioen, A. W.; Dyson, P. J.; Nowak-Sliwinska, P. *Chem. Sci.* **2014**, *5*, 4742–4748.

[354] Kuhlman, B.; Bradley, P. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 681–697.

[355] Shi, Y. *Cell* **2014**, *159*, 995–1014.

[356] Dubochet, J.; Lepault, J.; Freeman, R.; Berriman, J.; Homo, J.-C. *J. Microsc.* **1982**, *128*, 219–237.

[357] Adrian, M.; Dubochet, J.; Lepault, J.; McDowall, A. W. *Nature* **1984**, *308*, 32–36.

[358] Frank, J. *Ultramicroscopy* **1975**, *1*, 159–162.

[359] Förster, T. *Ann. Phys.* **1948**, *437*, 55–75.

[360] Clapp, A. R.; Medintz, I. L.; Mattoussi, H. *Chem. Phys. Chem.* **2006**, *7*, 47–57.

[361] Ghisaidoobe, A. B.; Chung, S. J. *Int. J. Mol. Sci.* **2014**, *15*, 22518–22538.

[362] Kainosho, M.; Torizawa, T.; Iwashita, Y.; Terauchi, T.; Ono, A. M.; Güntert, P. *Nature* **2006**, *440*, 52–57.

[363] Castellani, F.; Van Rossum, B.; Diehl, A.; Schubert, M.; Rehbein, K.; Oschkinat, H. *Nature* **2002**, *420*, 99–102.

[364] Goldfarb, D.; Stoll, S. *EPR spectroscopy: fundamentals and methods*; John Wiley & Sons, **2018**.

[365] Tsvetkov, Y. D.; Bowman, M. K.; Grishin, Y. A. *Pulsed Electron–Electron Double Resonance*; Springer, **2019**.

[366] Schiemann, O.; Prisner, T. F. *Q. Rev. Biophys.* **2007**, *40*, 1.

[367] Abdullin, D.; Schiemann, O. *ChemPlusChem* **2020**, *85*, 353–372.

[368] Elsässer, C.; Brecht, M.; Bittl, R. *J. Am. Chem. Soc.* **2002**, *124*, 12606–12611.

[369] Denysenkov, V.; Prisner, T.; Stubbe, J.; Bennati, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13386–13390.

[370] Hubbell, W. L.; Cafiso, D. S.; Altenbach, C. *Nature Struct. Biol.* **2000**, *7*, 735–739.

[371] Timmel, C. R.; Harmer, J. R. *Structural Information from Spin-Labels and Intrinsic Paramagnetic Centres in the Biosciences*; Springer, **2014**, Vol. 125.

[372] Berliner, L. J.; Grunwald, J.; Hankovszky, H. O.; Hideg, K. *Anal. Biochem.* **1982**, *119*, 450–455.

[373] Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589–1615.

[374] Levitt, M. *Nat. Struct. Biol.* **2001**, *8*, 392–393.

[375] Ehrlich, S.; Göller, A. H.; Grimme, S. *Chem. Phys. Chem.* **2017**, *18*, 898–905.

[376] Cui, Q. *J. Chem. Phys.* **2016**, *145*, 140901.

[377] Stewart, J. J. *J. Mol. Model.* **2009**, *15*, 765–805.

[378] Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.

[379] Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. *Annu. Rev. Biophys.* **2001**, *30*, 211–243.

[380] Mackerell Jr, A. D.; Feig, M.; Brooks III, C. L. *J. Comput Chem.* **2004**, *25*, 1400–1415.

[381] Sale, K.; Song, L.; Liu, Y.-S.; Perozo, E.; Fajer, P. *J. Am. Chem. Soc.* **2005**, *127*, 9334–9335.

[382] Hagelueken, G.; Abdullin, D.; Ward, R.; Schiemann, O. *Mol. Phys.* **2013**, *111*, 2757–2766.

[383] Hagelueken, G.; Abdullin, D.; Schiemann, O. *Method Enzymol.* **2015**, *563*, 595–622.

[384] Hatmal, M. M.; Li, Y.; Hegde, B. G.; Hegde, P. B.; Jao, C. C.; Langen, R.; Haworth, I. S. *Biopolymers* **2012**, *97*, 35–44.

[385] Hirst, S. J.; Alexander, N.; Mchaourab, H. S.; Meiler, J. *J. Struct. Biol.* **2011**, *173*, 506–514.

[386] Polyhach, Y.; Bordignon, E.; Jeschke, G. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2356–2366.

[387] Jeschke, G. *Prog. Nucl. Magn. Reson. Spectrosc.* **2013**, *72*, 42–60.

[388] Islam, S. M.; Roux, B. *J. Phys. Chem. B* **2015**, *119*, 3901–3911.

[389] Robinson, B.; Slutsky, L.; Auteri, F. *J. Chem. Phys.* **1992**, *96*, 2609–2616.

[390] Steinhoff, H.-J.; Hubbell, W. L. *Biophys. J.* **1996**, *71*, 2201–2212.

[391] Budil, D. E.; Sale, K. L.; Khairy, K. A.; Fajer, P. G. *J. Phys. Chem. A* **2006**, *110*, 3703–3713.

[392] Tombolato, F.; Ferrarini, A.; Freed, J. H. *J. Phys. Chem. B* **2006**, *110*, 26248–26259.

[393] DeSensi, S. C.; Rangel, D. P.; Beth, A. H.; Lybrand, T. P.; Hustedt, E. J. *Biophys. J.* **2008**, *94*, 3798–3809.

[394] Oganesyan, V. S. *Phys. Chem. Chem. Phys.* **2011**, *13*, 4724–4737.

[395] Sezer, D.; Freed, J. H.; Roux, B. . *Phys. Chem. B* **2008**, *112*, 5755–5767.

[396] Sezer, D.; Freed, J. H.; Roux, B. *J. Am. Chem. Soc.* **2009**, *131*, 2597–2605.

[397] Brehm, M.; Kirchner, B. *J. Chem. Inf. Model.* **2011**, *51*, 2007–2023.

[398] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

[399] Schrödinger Release 2020-2: Maestro, Schrödinger, LLC, New York, NY, **2020**.

[400] Kauzmann, W. *Chem. Rev.* **1948**, *43*, 219–256.

[401] Angell, C. A.; Sare, J. M.; Sare, E. J. *J. Phys. Chem.* **1978**, *82*, 2622–2629.

[402] Florin, N.; Schiemann, O.; Hagelueken, G. *BMC Struct. Biol.* **2014**, *14*, 1–10.

[403] Abdullin, D.; Florin, N.; Hagelueken, G.; Schiemann, O. *Angew. Chem.* **2015**, *127*, 1847–1851.

[404] Abdullin, D.; Hagelueken, G.; Schiemann, O. *Phys. Chem. Chem. Phys.* **2016**, *18*, 10428–10437.

[405] Kazmier, K.; Alexander, N. S.; Meiler, J.; Mchaourab, H. S. *J. Struct. Biol.* **2011**, *173*, 549–557.

[406] Borbat, P. P.; Mchaourab, H. S.; Freed, J. H. *J. Am. Chem. Soc.* **2002**, *124*, 5304–5314.

[407] Schiemann, O.; Piton, N.; Mu, Y.; Stock, G.; Engels, J. W.; Prisner, T. F. *J. Am. Chem. Soc.* **2004**, *126*, 5722–5729.

[408] Müller-Dethlefs, K.; Hobza, P. *Chem. Rev.* **2000**, *100*, 143–168.

[409] Rodgers, M.; Armentrout, P. *Chem. Rev.* **2016**, *116*, 5642–5687.

[410] Gutierrez, O.; Aubé, J.; Tantillo, D. J. *J. Org. Chem.* **2012**, *77*, 640–647.

[411] Kennedy, C. R.; Lin, S.; Jacobsen, E. N. *Angew. Chem., Int. Ed.* **2016**, *55*, 12596–12624.

[412] Goldstein, R.; Cheng, J.; Stec, B.; Roberts, M. F. *Biochemistry* **2012**, *51*, 2579–2587.

[413] Estarellas, C.; Frontera, A.; Quiñonero, D.; Deyà, P. M. *Angew. Chem., Int. Ed.* **2011**, *50*, 415–418.

[414] Faraldos, J. A.; Antonczak, A. K.; González, V.; Fullerton, R.; Tippmann, E. M.; Allemann, R. K. *J. Am. Chem. Soc.* **2011**, *133*, 13906–13909.

[415] Raines, D. E.; Gioia, F.; Claycomb, R. J.; Stevens, R. J. *J. Pharmacol. Exp. Ther.* **2004**, *311*, 14–21.

[416] Tantry, S.; Ding, F.-X.; Dumont, M.; Becker, J. M.; Naider, F. *Biochemistry* **2010**, *49*, 5007–5015.

[417] Wu, D.; Hu, Q.; Yan, Z.; Chen, W.; Yan, C.; Huang, X.; Zhang, J.; Yang, P.; Deng, H.; Wang, J. *Nature* **2012**, *484*, 214–219.

[418] Kapoor, K.; Duff, M. R.; Upadhyay, A.; Bucci, J. C.; Saxton, A. M.; Hinde, R. J.; Howell, E. E.; Baudry, J. *Biochemistry* **2016**, *55*, 6056–6069.

[419] Chen, C.-C.; Hsu, W.; Kao, T.-C.; Horng, J.-C. *Biochemistry* **2011**, *50*, 2381–2383.

[420] Chen, C.-C.; Hsu, W.; Hwang, K.-C.; Hwu, J. R.; Lin, C.-C.; Horng, J.-C. *Arch. Biochem. Biophys.* **2011**, *508*, 46–53.

[421] Mardirossian, N.; Head-Gordon, M. *Mol. Phys.* **2017**, *115*, 2315–2372.

[422] Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2006**, *125*, 201102.

[423] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Science* **2008**, *321*, 792–794.

[424] Becke, A. D. *J. Chem. Phys.* **2014**, *140*, 18A301.

[425] Rudberg, E. *J. Phys. Condens. Matter* **2012**, *24*, 072202.

[426] Ruiz, E.; Salahub, D. R.; Vela, A. *J. Am. Chem. Soc.* **1995**, *117*, 1141–1142.

[427] Lundberg, M.; Siegbahn, P. E. *J. Chem. Phys.* **2005**, *122*, 224103.

[428] Lonsdale, D. R.; Goerigk, L. *Phys. Chem. Chem. Phys.* **2020**, *22*, 15805–15830.

[429] Kristyán, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175–180.

[430] Tkatchenko, A.; DiStasio Jr, R. A.; Car, R.; Scheffler, M. *Phys. Rev. Lett.* **2012**, *108*, 236402.

[431] Hermann, J.; DiStasio Jr, R. A.; Tkatchenko, A. *Chem. Rev.* **2017**, *117*, 4714–4758.

[432] Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. *J. Phys. Org. Chem.* **2009**, *22*, 1127–1135.

[433] Grimme, S. *WIREs Comput. Mol. Sci.* **2011**, *1*, 211–228.

[434] Lee, K.; Murray, É. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. B* **2010**, *82*, 081101.

[435] Szalewicz, K. *WIREs Comput. Mol. Sci.* **2012**, *2*, 254–272.

[436] Jansen, G. *WIREs Comput. Mol. Sci.* **2014**, *4*, 127–144.

[437] Lao, K. U.; Schäffer, R.; Jansen, G.; Herbert, J. M. *J. Chem. Theory Comput.* **2015**, *11*, 2473–2486.

[438] Wheeler, S. E. *Acc. Chem. Res.* **2013**, *46*, 1029–1038.

[439] Kim, D.; Tarakeshwar, P.; Kim, K. S. *J. Phys. Chem. A* **2004**, *108*, 1250–1258.

[440] Grimme, S.; Hujo, W.; Kirchner, B. *Phys. Chem. Chem. Phys.* **2012**, *14*, 4875–4883.

[441] Perlt, E.; Ray, P.; Hansen, A.; Malberg, F.; Grimme, S.; Kirchner, B. *J. Chem. Phys.* **2018**, *148*, 193835.

[442] Abbas, H. *J. Biol. Phys.* **2017**, *43*, 105–111.

[443] Ferretti, A.; d'Ischia, M.; Prampolini, G. *J. Phys. Chem. A* **2020**, *124*, 3445–3459.

[444] Řezáč, J.; Greenwell, C.; Beran, G. J. *J. Chem. Theory Comput.* **2018**, *14*, 4711–4721.

[445] Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. *J. Chem. Phys.* **2004**, *120*, 647–659.

[446] Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.

[447] Janowski, T.; Pulay, P. *J. Am. Chem. Soc.* **2012**, *134*, 17520–17525.

[448] Cybulski, S. M.; Lytle, M. L. *J. Chem. Phys.* **2007**, *127*, 141102.

[449] Heßelmann, A. *J. Chem. Phys.* **2008**, *128*, 144112.

[450] Estarellas, C.; Lucas, X.; Frontera, A.; Quiñonero, D.; Deyà, P. M. *Chem. Phys. Lett.* **2010**, *489*, 254–258.

[451] Goldey, M.; Head-Gordon, M. *J. Phys. Chem. Lett.* **2012**, *3*, 3592–3598.

[452] Giese, M.; Albrecht, M.; Rissanen, K. *Chem. Rev.* **2015**, *115*, 8867–8895.

[453] Yamada, S. *Chem. Rev.* **2018**, *118*, 11353–11432.

[454] Karton, A.; Martin, J. M. *J. Chem. Phys.* **2012**, *136*, 124114.

[455] Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

[456] Minoux, H.; Chipot, C. *J. Am. Chem. Soc.* **1999**, *121*, 10366–10372.

[457] Sure, R.; Antony, J.; Grimme, S. *J. Phys. Chem. B.* **2014**, *118*, 3431–3440.

[458] Kumar, K.; Woo, S. M.; Siu, T.; Cortopassi, W. A.; Duarte, F.; Paton, R. S. *Chem. Sci.* **2018**, *9*, 2655–2665.

[459] Turupcu, A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2020**, *16*, 7184–7194.

[460] Liu, H.; Fu, H.; Shao, X.; Cai, W.; Chipot, C. *J. Chem. Theory Comput.* **2020**, *16*, 6397–6407.

[461] Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 154108.

[462] Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 124108.

[463] Van Gisbergen, S.; Snijders, J.; Baerends, E. *J. Chem. Phys.* **1995**, *103*, 9347–9354.

[464] Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. *Physical Review B* **2015**, *92*, 045131.

[465] Caldeweyher, E.; Mewes, J.-M.; Ehlert, S.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 8499–8512.

[466] Grimme, S.; Bannwarth, C.; Caldeweyher, E.; Pisarek, J.; Hansen, A. *J. Chem. Phys.* **2017**, *147*, 161708.

[467] Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.

[468] Johnson, E. R.; DiLabio, G. A. *Chem. Phys. Lett.* **2006**, *419*, 333–339.

[469] Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2006**, *124*, 174104.

[470] Otero-de-la Roza, A.; Johnson, E. R. *J. Chem. Phys.* **2013**, *138*, 054103.

[471] DiStasio Jr, R. A.; Gobre, V. V.; Tkatchenko, A. *J. Phys.: Condens. Matter* **2014**, *26*, 213202.

[472] Berland, K.; Cooper, V. R.; Lee, K.; Schröder, E.; Thonhauser, T.; Hyldgaard, P.; Lundqvist, B. I. *Rep. Prog. Phys.* **2015**, *78*, 066501.

[473] Hujo, W.; Grimme, S. *J. Chem. Theory Comput.* **2013**, *9*, 308–315.

[474] Najibi, A.; Goerigk, L. *J. Chem. Theory Comput.* **2018**, *14*, 5725–5738.

[475] Ernst, B. G.; Lao, K. U.; Sullivan, A. G.; DiStasio Jr, R. A. *J. Phys. Chem. A* **2020**, *124*, 4128–4140.

[476] Schottel, B. L.; Chifotides, H. T.; Dunbar, K. R. *Chem. Soc. Rev.* **2008**, *37*, 68–83.

[477] Wang, D.-X.; Wang, M.-X. *J. Am. Chem. Soc.* **2013**, *135*, 892–897.

[478] Parker, A. J. *Chem. Rev.* **1969**, *69*, 1–32.

[479] McCurdy, A.; Jimenez, L.; Stauffer, D. A.; Dougherty, D. A. *J. Am. Chem. Soc.* **1992**, *114*, 10314–10321.

[480] Dougherty, D. A.; Stauffer, D. A. *Science* **1990**, *250*, 1558–1560.

[481] Aoki, K.; Murayama, K.; Nishiyama, H. *J. Chem. Soc., Chem. Commun.* **1995**, 2221–2222.

[482] Yamada, S.; Yamamoto, N.; Takamori, E. *Org. Lett.* **2015**, *17*, 4862–4865.

[483] Yamada, S.; Yamamoto, N.; Takamori, E. *J. Org. Chem.* **2016**, *81*, 11819–11830.

[484] Goerigk, L.; Mehta, N. *Aust. J. Chem.* **2019**, *72*, 563–573.

[485] Zhao, Y.; Truhlar, D. G. *J. Chem. Phys* **2006**, *125*, 194101.

[486] Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[487] *Generally Applicable Atomic-Charge Dependent London Dispersion Correction dftd4*, **2021**, https://github.com/grimme-lab/dftd4.

[488] ORCA – an ab initio, density functional and semiempirical program package, V. 4.2.1, F. Neese, MPI für Kohlenforschung, Mülheim a. d. Ruhr (Germany), **2019**.

[489] Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. *J. Chem. Phys.* **2021**, *154*, 064103.

[490] Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. *J. Phys. Chem. Lett.* **2020**, *11*, 8208–8215.

[491] Ehlert, S.; Huniar, U.; Ning, J.; Furness, J. W.; Sun, J.; Kaplan, A. D.; Perdew, J. P.; Brandenburg, J. G. *J. Chem. Phys.* **2021**, *154*, 061101.

[492] Mardirossian, N.; Head-Gordon, M. *J. Chem. Phys.* **2015**, *142*, 074111.

[493] Haoyu, S. Y.; He, X.; Li, S. L.; Truhlar, D. G. *Chem. Sci.* **2016**, *7*, 5032–5051.

[494] Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

[495] Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.

[496] Becke, A. D. *J. Chem. Phys* **1993**, *98*, 5648–5652.

[497] Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

[498] Mardirossian, N.; Head-Gordon, M. *J. Chem. Phys.* **2016**, *144*, 214110.

[499] Santra, G.; Sylvetsky, N.; Martin, J. M. *J. Phys. Chem. A* **2019**, *123*, 5129–5143.

[500] Goerigk, L.; Grimme, S. *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

[501] TURBOMOLE V7.5.1 **2020**, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

[502] Neese, F. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1327.

[503] Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. *J. Chem. Phys.* **2018**, *148*, 011101.

[504] Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

[505] Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

[506] Neese, F.; Valeev, E. F. *J. Chem. Theory Comput.* **2011**, *7*, 33–43.

[507] Stoychev, G. L.; Auer, A. A.; Neese, F. *J. Chem. Theory Comput.* **2017**, *13*, 554–562.

[508] Lee, T. J.; Taylor, P. R. *Int. J. Quantum Chem.* **1989**, *36*, 199–207.

[509] Gersdorf, P.; John, W.; Perdew, J. P.; Ziesche, P. *Int. J. Quantum Chem.* **1997**, *61*, 935–941.

[510] Fogueri, U. R.; Kozuch, S.; Karton, A.; Martin, J. M. *Theor. Chem. Acc.* **2013**, *132*, 1–9.

[511] Werner, H. J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. *WIREs Comput. Mol. Sci.* **2012**, *2*, 242–253.

[512] *Molpro V15.1 2015, a package of ab initio programs*, **2015**, http://www.www.molpro.net.

[513] Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

[514] *Conformer-Rotamer Ensemble Sampling Tool based on the xtb Semiempirical Extended Tight-Binding Program Package crest*, **2021**, https://github.com/grimme-lab/crest.

[515] *Semiempirical Extended Tight-Binding Program Package xtb*, Version 6.3.3., **2020**, https://github.com/grimme-lab/xtb.

[516] Geary, R. C. *Biometrika* **1935**, *27*, 310–332.

[517] Su, P.; Li, H. *J. Chem. Phys.* **2009**, *131*, 014102.

[518] Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz Jr, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691.

[519] Zhang, Y.; Yang, W. *Theor. Chem. Acc.* **2000**, 346–348.

[520] Yang, W.; Zhang, Y.; Ayers, P. W. *Phys. Rev. Lett.* **2000**, *84*, 5172.

[521] Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Staroverov, V. N.; Tao, J. *Phys. Rev. A* **2007**, *76*, 040501.

[522] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. *Phys. Rev. B* **2008**, *77*, 115123.

[523] Johnson, E. R.; Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 204112.

[524] Spicher, S.; Grimme, S. *J. Chem. Theory Comput.* **2021**, *17*, 1701–1714.

[525] Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. *J. Phys. Chem. B* **2020**, *124*, 6664–6670.

[526] Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. *J. Chem. Phys.* **2021**, 10.1063/5.0057024.

[527] Houk, K. N.; Paddon-Row, M. N.; Rondan, N. G.; Wu, Y.-D.; Brown, F. K.; Spellmeyer, D. C.; Metz, J. T.; Li, Y.; Loncharich, R. J. *Science* **1986**, *231*, 1108–1117.

[528] Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. *Mol. Sys. Des. Eng.* **2018**, *3*, 442–452.

[529] Kirkwood, J. G. *J. Phys. Chem.* **1935**, *3*, 300–313.

[530] Henriksen, N. M.; Gilson, M. K. *J. Chem. Theory Comput.* **2017**, *13*, 4253–4269.

[531] Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[532] Aqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.

[533] Mobley, D. L.; Gilson, M. K. *Annu. Rev. Biophys.* **2017**, *46*, 531–558.

[534] Hansen, N.; Van Gunsteren, W. F. *J. Chem. Theory Comput.* **2014**, *10*, 2632–2647.

[535] Bootsma, S., Andrea N.; Wheeler *ChemRxiv* **2019**, https://doi.org/10.26434/chemrxiv.8864204.v5.

[536] Scott, A. P.; Radom, L. *J. Phys. Chem.* **1996**, *100*, 16502–16513.

[537] Jurečka, P.; Šponer, J.; Černỳ, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

[538] Sedlak, R.; Janowski, T.; Pitonak, M.; Rezac, J.; Pulay, P.; Hobza, P. *J. Chem. Theory Comput.* **2013**, *9*, 3364–3374.

[539] Del Bene, J. E.; Mettee, H. D.; Frisch, M. J.; Luke, B. T.; Pople, J. A. *J. Phys. Chem.* **1983**, *87*, 3279–3282.

[540] Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C.; Goddard III, W. A. *J. Chem. Theory Comput.* **2009**, *5*, 1016–1026.

[541] Cowan, S. W.; Newcomer, M. E.; Jones, T. A. *Proteins* **1990**, *8*, 44–61.

[542] Rizzi, A.; Jensen, T.; Slochower, D. R.; Aldeghi, M.; Gapsys, V.; Ntekoumes, D.; Bosisio, S.; Papadourakis, M.; Henriksen, N. M.; De Groot, B. L. *J. Comput. Aided Mol. Des.* **2020**, 1–33.

[543] Roehrig, S.; Straub, A.; Pohlmann, J.; Lampe, T.; Pernerstorfer, J.; Schlemmer, K.-H.; Reinemer, P.; Perzborn, E. *J. Med. Chem.* **2005**, *48*, 5900–5908.

[544] Schrader, B. *Infrared and Raman spectroscopy: methods and applications*; John Wiley & Sons, **2008**.

[545] McCarthy, M.; Lee, K. L. K. *J. Phys. Chem. A* **2020**, *124*, 3002–3017.

[546] Pracht, P.; Grant, D. F.; Grimme, S. *J. Chem. Theory Comput.* **2020**, *16*, 7044–7060.

[547] Peticolas, W. L.; Rush III, T. *J. Comput. Chem.* **1995**, *16*, 1261–1270.

[548] Neugebauer, J.; Reiher, M.; Kind, C.; Hess, B. A. *J. Comput. Chem.* **2002**, *23*, 895–910.

[549] Kollman, P. *Chem. Rev.* **1993**, *93*, 2395–2417.

[550] Decherchi, S.; Cavalli, A. *Chem. Rev.* **2020**, *120*, 12788–12833.

[551] Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

[552] Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

[553] Klamt, A.; Schüürmann, G. *J. Chem. Soc.* **1993**, 799–805.

[554] Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

[555] Eckert, F.; Klamt, A. *AIChE J.* **2002**, *48*, 369–385.

[556] Pople, J.; Schlegel, H.; Krishnan, R.; DeFrees, D.; Binkley, J.; Frisch, M.; Whiteside, R.; Hout, R.; Hehre, W. *Int. J. Quantum Chem.* **1981**, *20*, 269–278.

[557] Fogarasi, G.; Pulay, P. *Annu. Rev. Phys. Chem.* **1984**, *35*, 191–213.

[558] Stratmann, R. E.; Burant, J. C.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1997**, *106*, 10175–10183.

[559] Temelso, B.; Archer, K. A.; Shields, G. C. *J. Phys. Chem. A* **2011**, *115*, 12034–12046.

[560] Antony, J.; Sure, R.; Grimme, S. *Chem. Commun.* **2015**, *51*, 1764–1774.

[561] Volbach, L.; Struch, N.; Bohle, F.; Topić, F.; Schnakenburg, G.; Schneider, A.; Rissanen, K.; Grimme, S.; Lützen, A. *Chem. Eur. J.* **2020**, *26*, 3335–3347.

[562] Canneaux, S.; Bohr, F.; Henon, E. *J. Comput. Chem.* **2014**, *35*, 82–93.

[563] Panek, P. T.; Jacob, C. R. *J. Chem. Phys.* **2016**, *144*, 164111.

[564] Panek, P. T.; Jacob, C. R. *J. Phys. Chem. Lett.* **2016**, *7*, 3084–3090.

[565] Grimme, S.; Bannwarth, C.; Dohm, S.; Hansen, A.; Pisarek, J.; Pracht, P.; Seibert, J.; Neese, F. *Angew. Chem. Int. Ed.* **2017**, *56*, 14763–14769.

[566] Ghysels, A.; Van Neck, D.; Van Speybroeck, V.; Verstraelen, T.; Waroquier, M. *J. Chem. Phys.* **2007**, *126*, 224102.

[567] Li, H.; Jensen, J. H. *Theor. Chem. Acc.* **2002**, *107*, 211–219.

[568] Stewart, J. J. *J. Comput. Chem.* **1989**, *10*, 221–264.

[569] Nevins, N.; Allinger, N. L. *J. Comput. Chem.* **1996**, *17*, 730–746.

[570] Neugebauer, H.; Bohle, F.; Bursch, M.; Hansen, A.; Grimme, S. *J. Phys. Chem. A* **2020**, *124*, 7166–7176.

[571] Goerigk, L.; Grimme, S. *WIREs Comput. Mol. Sci.* **2014**, *4*, 576–600.

[572] Lehn, J.-M. *Science* **1993**, *260*, 1762–1764.

[573] Lehn, J.-M. *Angew. Chem. Int. Ed.* **1988**, *27*, 89–112.

[574] Steed, J. W.; Atwood, J. L. *Supramolecular chemistry*; John Wiley & Sons, **2013**.

[575] Gilson, M. K.; Zhou, H.-X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*.

[576] Coutsias, E. A.; Seok, C.; Dill, K. A. *J. Comput. Chem.* **2004**, *25*, 1849–1857.

[577] Richard, B.; Douglas, J. *The bisection algorithm*; PWS Publishers Boston, **1985**.

[578] Jungbauer, S. H.; Bulfield, D.; Kniep, F.; Lehmann, C. W.; Herdtweck, E.; Huber, S. M. *J. Am. Chem. Soc.* **2014**, *136*, 16740–16743.

[579] Zabrodsky, H.; Peleg, S.; Avnir, D. *J. Am. Chem. Soc.* **1992**, *114*, 7843–7851.

[580] TURBOMOLE V7.2.1 **2017**, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

[581] Dron, P. I.; Fourmentin, S.; Cazier, F.; Landy, D.; Surpateanu, G. *Supramol. Chem.* **2008**, *20*, 473–477.

[582] Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800–2806.

[583] Gonzalez, C.; Schlegel, H. B. *J. Chem. Phys.* **1989**, *90*, 2154–2161.

[584] Hratchian, H. P.; Schlegel, H. B. *J. Chem. Phys.* **2004**, *120*, 9918–9924.

[585] Yang, S.-Y.; Fleurat-Lessard, P.; Hristov, I.; Ziegler, T. *J. Phys. Chem. A* **2004**, *108*, 9461–9468.

[586] Yang, S.-Y.; Hristov, I.; Fleurat-Lessard, P.; Ziegler, T. *J. Phys. Chem. A* **2005**, *109*, 197–204.

[587] Miller, W. H.; Handy, N. C.; Adams, J. E. *J. Chem. Phys.* **1980**, *72*, 99–112.

[588] Fukui, K. *Acc. Chem. Res.* **1981**, *14*, 363–368.

[589] Zimmerman, P. M. *J. Chem. Phys.* **2013**, *138*, 184102.

[590] Zimmerman, P. *J. Chem. Theory Comput.* **2013**, *9*, 3043–3050.

[591] Dohm, S.; Bursch, M.; Hansen, A.; Grimme, S. *J. Chem. Theory Comput.* **2020**, *16*, 2002–2012.

[592] Page, M.; McIver Jr, J. W. *J. Chem. Phys.* **1988**, *88*, 922–935.

[593] Hu, H.; Kobrak, M. N.; Xu, C.; Hammes-Schiffer, S. *J. Phys. Chem. A* **2000**, *104*, 8058–8066.

[594] Schlegel, H. B. *Adv. Chem. Phys.* **1987**, *67*, 249.

[595] Iron, M. A.; Ben-Ari, E.; Cohen, R.; Milstein, D. *Dalton Trans.* **2009**, 9433–9439.

[596] Iron, M. A.; Janes, T. *J. Phys. Chem. A* **2019**, *123*, 3761–3781.

[597] Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

[598] Karplus, M.; Petsko, G. A. *Nature* **1990**, *347*, 631–639.

[599] Hunt, N. T. *Chem. Soc. Rev.* **2009**, *38*, 1837–1848.

[600] Krossing, I.; van Wüllen, L. *Chem. Eur. J.* **2002**, *8*, 700–711.

[601] Katsyuba, S. A.; Vener, M. V.; Zvereva, E. E.; Fei, Z.; Scopelliti, R.; Brandenburg, J. G.; Siankevich, S.; Dyson, P. J. *J. Phys. Chem. Lett.* **2015**, *6*, 4431–4436.

[602] Katsyuba, S. A.; Vener, M. V.; Zvereva, E. E.; Fei, Z.; Scopelliti, R.; Laurenczy, G.; Yan, N.; Paunescu, E.; Dyson, P. J. *J. Phys. Chem. B* **2013**, *117*, 9094–9105.

[603] Marx, D.; Hutter, J. *Ab initio molecular dynamics: basic theory and advanced methods*; Cambridge University Press, **2009**.

[604] Thomas, M.; Brehm, M.; Fligg, R.; Vöhringer, P.; Kirchner, B. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6608–6622.

[605] Massarczyk, M.; Rudack, T.; Schlitter, J.; Kuhne, J.; Kötting, C.; Gerwert, K. *J. Phys. Chem. B* **2017**, *121*, 3483–3492.

[606] Mennucci, B. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6583–6594.

[607] Ghidinelli, S.; Abbate, S.; Boiadjiev, S. E.; Lightner, D. A.; Longhi, G. *J. Phys. Chem. B* **2018**, *122*, 12351–12362.

[608] Giovannini, T.; Del Frate, G.; Lafiosca, P.; Cappelli, C. *Phys. Chem. Chem. Phys.* **2018**, *20*, 9181–9197.

[609] Katsyuba, S. A.; Zvereva, E. E.; Burganov, T. I. *J. Phys. Chem. A* **2013**, *117*, 6664–6670.

[610] Katsyuba, S. A.; Zvereva, E. E.; Grimme, S. *J. Phys. Chem. A* **2019**, *123*, 3802–3808.

[611] *Quantum Cluster Growth: An explicit Approach towords Solvation Free Energies QCG*, **2021**, https://github.com/grimme-lab/crest.

[612] TURBOMOLE V7.2 **2017**, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

[613] Hogan, T. R.; Steele, D. *J. Mol. Struct.* **1986**, *141*, 315–323.

[614] Polo, S. R.; Wilson, M. K. *J. Chem. Phys.* **1955**, *23*, 2376–2377.

[615] Rogers, J.; Rub, B.; Goldman, S.; Person, W. *J. Phys. Chem.* **1981**, *85*, 3727–3729.

[616] Serrallach, A.; Meyer, R.; Günthard, H. H. *J. Mol. Spectrosc.* **1974**, *52*, 94–129.

[617] Bertie, J. E.; Zhang, S. L. *J. Mol. Struct.* **1997**, *413*, 333–363.

[618] Kollipost, F.; Papendorf, K.; Lee, Y.-F.; Lee, Y.-P.; Suhm, M. A. *Phys. Chem. Chem. Phys.* **2014**, *16*, 15948–15956.

[619] Behera, B.; Chakraborty, S. *Vib. Spectrosc.* **2020**, *106*, 102981.

[620] Koga, Y.; Kondo, S.; Saeki, S.; Person, W. B. *J. Phys. Chem.* **1984**, *88*, 3152–3157.

[621] Kosower, E. M.; Markovich, G.; Borz, G. *Chem. Phys. Chem.* **2007**, *8*, 2513–2519.

[622] Gölitz, P. *Angew. Chem. Int. Ed.* **2015**, *54*, 4–6.

[623] Su, K.; Wang, W.; Du, S.; Ji, C.; Yuan, D. *Nat. Commun.* **2021**, *12*, 3703.

[624] Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185–199.

[625] Karplus, M. *Angew. Chem. Int. Ed.* **2014**, *53*, 9992–10005.

[626] Warshel, A. *Angew. Chem. Int. Ed.* **2014**, *53*, 10020–10031.

[627] Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct.* **1999**, *461*, 1–21.

[628] Kühne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F. *J. Chem. Phys.* **2020**, *152*, 194103.

[629] Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

[630] Behler, J. *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050.

[631] Sure, R.; El Mahdali, M.; Plajer, A.; Deglmann, P. *J. Comput. Aided Mol. Des.* **2021**, *35*, 473–492.

[632] Harris, J. *Phys. Rev. A* **1984**, *29*, 1648.

[633] Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.

[634] Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

[635] Eshuis, H.; Bates, J. E.; Furche, F. *Theor. Chem. Acc.* **2012**, *131*, 1–18.

[636] Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095–9102.

[637] Dohm, S.; Hansen, A.; Steinmetz, M.; Grimme, S.; Checinski, M. P. *J. Chem. Theory. Comput.* **2018**, *14*, 2596–2608.

[638] Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

[639] Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.

[640] Caldeweyher, E.; Brandenburg, J. G. *J. Phys. Condens. Matter* **2018**, *30*, 213001.

[641] Sure, R.; Grimme, S. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

[642] Cutini, M.; Civalleri, B.; Corno, M.; Orlando, R.; Brandenburg, J. G.; Maschio, L.; Ugliengo, P. *J. Chem. Theory Comput.* **2016**, *12*, 3340–3352.

[643] Kruse, H.; Grimme, S. *J. Chem. Phys.* **2012**, *136*, 154101.

[644] Mermin, N. D. *Phys. Rev.* **1965**, *137*, A1441.

[645] Berendsen, H. J.; Postma, J. v.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[646] Van Gunsteren, W. F.; Berendsen, H. J. *Mol. Simul.* **1988**, *1*, 173–185.

[647] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

[648] Levenberg, K. *Q. Appl. Math.* **1944**, *2*, 164–168.

[649] Marquardt, D. *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431–441.

[650] Brandenburg, J. G.; Grimme, S. *Top. Curr. Chem.* **2014**, *345*, 1–23.

[651] TURBOMOLE V7.3.1 **2018**, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

[652] ORCA – an ab initio, density functional and semiempirical program package, V. 4.1.1, F. Neese, MPI für Kohlenforschung, Mülheim a. d. Ruhr (Germany), **2018**.

[653] Neese, F. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 73–78.

[654] Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, J. M. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20905–20925.

[655] Li, Y.-P.; Bell, A. T.; Head-Gordon, M. *J. Chem. Theory Comput* **2016**, *12*, 2861–2870.

[656] *MtsslWizard Server Version*, **2020**, http://www.mtsslsuite.isb.ukbonn.de.

[657] *The PyMOL Molecular Graphics System*, **2020**, Version 2.2.0, Schrödinger, LLC.

[658] Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[659] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

[660] Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

[661] Andersson, M. P.; Uvdal, P. *J. Phys. Chem. A* **2005**, *109*, 2937–2941.

Part VI.

Appendix

A1. Abbreviations

AES	Anisotropic electrostatics
AO	Atomic orbital
ATM	Axilrod–Teller–Muto
AXC	Anisotropic exchange
BJ	Becke–Johnson
BO	Bond order
BSIE	Basis set incompleteness error
BSSE	Basis set superposition error
CAMM	Cumulative atomic multipole moments
CBS	Complete basis set
CC	Coupled cluster
CI	Configuration interaction
CN	Coordination number
COSMO	Conductor-like screening model
COSMO-RS	Conductor-like screening model for real solvents
CP	Counterpoise
CPCM	Conductor-like Polarizable Continuum Model
CPU	Central processing unit
CREST	Conformer-rotamer ensemble sampling tool
CSD	Cambridge Structural Database
CT	Charge transfer
DFA	Density functional approximation
DFG	Deutsche Forschungsgemeinschaft
DFT	Density functional theory
DFTB	Density functional tight binding
DLPNO	Domain based local pair natural orbital
DNP	Dynamic nuclear polarization
DOF	Degrees of freedom
ECP	Effective core potential
EDA	Energy decomposition analysis
EEQ	Electronegativity equilibrium
EHT	Extended Hückel theory
EN	Electronegativity
EPR	Electron paramagnetic resonance
FC	Force constant
FCI	Full configuration interaction
FEP	Free energy perturbation

FF	Force-field
FRET	Förster resonance energy transfer
GB	Generalized Born
gCP	Geometrical counterpoise
GGA	Generalized gradient approximation
GSM	Growing string method
GTO	Gaussian type orbital
HB	Hydrogen bond
HF	Hartree–Fock
HOMO	Highest occupied molecular orbital
HVF	Harmonic vibrational frequencies
IES	Isotropic electrostatics
IR	Infrared
IXC	Isotropic exchange
KS-DFT	Kohn–Sham density functional theory
LCAO	Linear combination of atomic orbitals
LD	London dispersion
LMO	Localized molecular orbital
LP	Lone-pair
LUMO	Lowest unoccupied molecular orbital
MAD	Mean absolute deviation
MBD	Many-body dispersion
MD	Molecular dynamics
MEP	Molecular electrostatic potential
ML	Machine learning
MO	Molecular orbital
MOF	Metal organic framework
MOP	Metal organic polyhedron
MP	Møller–Plesset
MTD	Metadynamics
MTSL	Methanethiosulfonate spin-label
NCI	Noncovalent interactions
NDDO	Neglect of diatomic differential overlap
NMR	Nuclear magnetic resonance
PCM	Polarizable Continuum Model
PELDOR	Pulsed electron–electron double resonance
PES	Potential energy surface
PHVA	Partial Hessian vibrational analysis
POC	Porous organic cage
PRE	Paramagnetic relaxation enhancement
QM	Quantum mechanical
QSAR	Quantitative structure-activity relationships
RI	Resolution-of-identity
RMSD	Root-mean-square deviation
RPA	Random phase approximation

RRHO	Rigid-rotor-harmonic-oscillator
SAPT	Symmetry adapted perturbation theory
SCF	Self-consistent field
SD	Standard deviation
SDSL	Site-directed spin-labeling
SE	Schrödinger equation
SIE	Self-interaction error
SPE	Single-point energy
SPH	Single-point Hessian
SQM	Semiempirical quantum mechanical
SRB	Short-ranged basis set
STO	Slater type orbital
TI	Thermodynamic integration
TM	Transition metal
TS	Transition state
TST	Transition state theory
UEG	Uniform electron gas
WFT	Wave function theory
ZPVE	Zero-point vibrational energy

A2. Supporting Information to Chapter 2

Appendix A2 contains:

- Kohn–Sham Density Functional Theory
- Extended Tight-Binding Methods
- Molecular Dynamics and Structure Ensembles

A2.1. Kohn–Sham Density Functional Theory

A2.1.1. Density Functional Approximations

Local (Spin) Density Approximation

The local density approximation (LDA) is derived from the uniform electron gas (UEG), which assumes only small variations in the electron density of the system.[154] The more general form accouting for individual α and β densities is termed local spin density approximation (LSDA). Both directly depend on the local density $\rho(\mathbf{r})$ and the exchange functional is based on the formulas derived by Dirac (Eq. A2.1)[164,165]

$$E_{\mathrm{X}}^{\mathrm{LSDA}}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}} \int \left(\rho_\alpha(\mathbf{r})^{\frac{4}{3}} + \rho_\beta(\mathbf{r})^{\frac{4}{3}}\right) d\mathbf{r}. \tag{A2.1}$$

The correlation part is discussed in Refs. 167,168. Due to the UEG foundation, systems with a homogenous electron density like metals are described reasonably well by LSDA approaches. For systems with inhomogeneous electron densities, as found in most molecules, the accuracy is rather low.

Generalized Gradient Approximation

In the general gradient approximation (GGA), E_{XC} of LSDAs is modified by an enhancement factor $F_{\mathrm{XC}}^{\mathrm{GGA}}[\rho, \nabla\rho]$.[154] This enhancement factor depends on the local electron density and its gradient (Eq. A2.2). It is further required that the Fermi and Coulomb holes integrate to the correct values of -1 and 0, respectively.[15]

$$E_{\mathrm{XC}}^{\mathrm{GGA}} = \int \epsilon_{\mathrm{XC}}^{LSDA}[\rho(\mathbf{r})] F_{\mathrm{XC}}^{\mathrm{GGA}}[\rho(\mathbf{r}), \nabla\rho(\mathbf{r})] d\mathbf{r} \tag{A2.2}$$

GGAs are typically referred to as semi-local DFAs, because $\nabla\rho(\mathbf{r})$ introduces information about the gradient of the electron density. Semi-local DFAs suffer from the self-interaction error (SIE) resulting in an over-delocalization of the electron density.

Meta-Generalized Gradient Approximation

Meta-generalized gradient approximations (meta-GGA) include the second order derivatives of the electron density in the enhancement factor $F_{\mathrm{XC}}^{\text{meta-GGA}}[\rho, \nabla\rho]$.[154] As the Laplacian $\nabla^2\rho$ is known to show numerical instability, it is often replaced by the kinetic energy density $\tau(\mathbf{r})$ (Eq. A2.3)[169].

$$\tau(\mathbf{r}) = \frac{1}{2} \sum_{\sigma=a,b} \sum_{i}^{N_\sigma} |\nabla\phi_i(\mathbf{r})|^2 \tag{A2.3}$$

Here, the density is divided into different spin components, and N_σ is the total number of electrons with spin $\sigma \in \{\alpha, \beta\}$. In contrast to the significant improvement of GGA over LSDA functionals, the meta-GGA approach improves the accuracy of GGAs only slightly. The computational costs formally scale as $O(M^4)$, due to the semi-local two-electron integral evaluation, but can be accelerated to $O(M^3)$ by appropriate approximations such as the resolution-of-identity (RI).[341,342]

Hybrid DFT

Semi-local DFAs are improved by adding non-local information about the occupied Kohn-Sham orbitals. Therefore, the exchange functional is modified by a fraction a_{X} of exact exchange from HF, called Fock exchange $E_{\mathrm{X}}^{\mathrm{HF}}$. By invoking the adiabatic connection[632] and assuming a linear dependence[172] between the Fock and (meta-)GGA exchange as the limits of integration, an empirical expression is obtained

$$E_{\mathrm{XC}}^{\text{hybrid}} = E_{\mathrm{C}}^{\text{(meta-)GGA}} + (1 - a_{\mathrm{X}}) E_{\mathrm{X}}^{\text{(meta-)GGA}} + a_{\mathrm{X}} E_{\mathrm{X}}^{\mathrm{HF}}. \tag{A2.4}$$

The Fock exchange reduces the SIE in hybrid DFAs. As a result, most hybrids outperform the semi-local DFAs in terms of accuracy for the calculation of many chemical properties[162,424] This accuracy gain comes at the cost of an increased formal scaling $\left(O(M^4)\right)$, since the Fock exchange is evaluated non-locally by integrating $\rho(\mathbf{r}, \mathbf{r}')$ over two spatial parts $\mathbf{r}$ and $\mathbf{r}'$.[154]

In global hybrid functionals, as well as meta-GGAs, the exchange potential is not described correctly in the asymptotic limit. A correction for this is established in range-separated hybrid DFAs. Therein, the exchange part is divided into a short-ranged and a long-ranged part (Eq. A2.5)[633,634]

$$\frac{1}{r_{12}} = \underbrace{\frac{1 - a'_X - a''_X \mathrm{erf}(\mu\, r_{12})}{r_{12}}}_{\text{short-range}} + \underbrace{\frac{a'_X + a''_X \mathrm{erf}(\mu\, r_{12})}{r_{12}}}_{\text{long-range}}. \tag{A2.5}$$

Here, a modified (semi-)local DFA is employed for the short-range part, while the long-range part is exactly evaluated by a regular and modified Fock exchange term scaled by a'_X and a''_X, respectively. The error function is employed to interpolate smoothly between both parts. μ is a parameter controlling the steepness of the interpolation.

Double-hybrid DFT

By construction, mean-field methods typically lack a sufficient description of the electron correlation, and DFT is no different. The double-hybrid functionals introduce information about the virtual Kohn-Sham orbitals to the correlation part in terms of modified second order Møller–Plesset perturbation theory,[152] or the random phase approximation (RPA)[635] (Eq. A2.6).[174,175,636] Most variants apply an MP2-like treatment, which leads to the general expression

$$E_{\mathrm{XC}}^{\text{double-hybrid}} = (1 - a_{\mathrm{X}})E_{\mathrm{X}}^{\text{(meta-)GGA}} + a_{\mathrm{X}}E_{\mathrm{X}}^{\mathrm{HF}} + (1 - b_{\mathrm{C}})E_{\mathrm{C}}^{\text{(meta-)GGA}} + b_{\mathrm{C}}E_{\mathrm{C}}^{\mathrm{PT2}}, \tag{A2.6}$$

where a_{X} and b_{C} are typically treated as a fitting parameters. Due to the MP2 correlation energy, typical double-hybrid functionals scale as $O(M^5)$ with the number of AOs. This limits the applicability to large systems. Nevertheless, double-hybrid functionals are computationally more efficient than coupled cluster methods, whilst almost reaching their accuracy.[162,184,637]

A2.1.2. London Dispersion Corrected DFT

Electronic structure methods based on mean-field approximations like HF and (semi-)local Kohn–Sham DFT do not describe long-range electronic correlation effects, and hence they cannot account for so-called London dispersion (LD) interactions.[334,638] As a result, DFAs (and HF) yield wrong asymptotic interaction energies that decay exponentially at short and medium range, instead of $-1/R^6$.[339] Dispersion interactions can be empirically defined as the attractive part of a van der Waals (vdW) interaction between non-bonded atoms or molecules. More precisely, LD interactions result from relatively long-ranged electron correlation effects in many-electron systems and involve coupled local components that can be explained in a stationary, time-independent electronic state picture. London and Eisenschitz derived the famous asymptotic formula for the dispersion energy between two atoms A and B at large distance R in 1930 (Eq. A2.7) from perturbation theory[199,200]

$$E_{\mathrm{disp}}^{AB} \approx -\frac{C_{6,\mathrm{approx}}^{AB}}{R^{-6}}, \tag{A2.7}$$

with the pair-specific C_6 dispersion coefficient. In a more general form, the pair-wise dispersion energy $E_{\mathrm{disp}}(R_{AB})$ is given in A2.8

$$E_{\mathrm{disp}}^{AB} = \sum_{n=6,8,10,\ldots} E_{\mathrm{disp}}^{(n)} = -\sum_{AB} \sum_{n=6,8,10,\ldots} \frac{C_n^{AB}}{R_{AB}^{(n)}} f_{damp}^{(n)}. \tag{A2.8}$$

The dispersion coefficients C_n^{AB} of n^{th} order are chosen to be generally positive and the damping function $f_{damp}^{(n)}$ is introduced to avoid artificial overbinding in typical covalent bonding regimes. The most prominent and efficient LD correction schemes is the DFT-D approach of Grimme and co-workers, which is simply additive to the electronic mean-field energy.[334] In the most used DFT-D3 variant,[252,334] the pair-wise dispersion correction term is obtained form precalculated dispersion coefficient C_n^{AB}, the DFA dependent parameter s_n, and the distance dependent damping function $f_{damp}^{(n)}(R)$. The resulting

term (Eq. A2.9)

$$E_{\text{disp}}^{\text{D3},AB} = \sum_{n=6,8} E_{\text{disp}}^{(n)} = -\sum_{AB}\sum_{n=6,8} s_n \frac{C_n^{AB}}{R_{AB}^{(n)}} f_{damp}^{(n)}(R) \tag{A2.9}$$

depends on the molecular structure and the atomic coordination numbers (CN) that is used to calculate the dispersion coefficients $C_n^{AB}(CN_A, CN_B)$. As damping functions $f_{damp}^{(n)}(R)$ two variants can be applied, the originally applied zero-damping (Eq. A2.10) or the more frequently used Becke–Johnson damping[253,255,639] (Eq. A2.11), termed as D3(0) and D3(BJ), respectively

$$f_{damp,0}^{(n)}(R) = \frac{1}{1 + 6(R/(s_{r,n}R_0))^{a_n}} \tag{A2.10}$$

$$f_{damp,BJ}^{(n)}(R) = \frac{R^n}{R^n + (a_1 R_0 + a_2)^n}\,. \tag{A2.11}$$

For these damping functions $R_0 = \sqrt{C_8^{AB}/C_6^{AB}}$, and $s_{r,n}$, a_1, a_2 are functional specific parameters. Recently, the DFT-D4 model has been introduced as the successor of DFT-D3.[242,339] In DFT-D4, atomic charge dependent reference polarizabilities $\alpha(i\omega)$ and modified, electronegativity dependent CNs are used to calculate the dispersion coefficient C_n^{AB}. The usage of DFT in combination with an LD correction is standard nowadays. For a more detailed evaluation of different LD correction schemes, see Section 6.

A2.1.3. Composite DFT Methods

Composite DFT methods introduce (empirical) corrections to enhance the efficiency of DFT methods.[640] The basic philosophy of the so-called "3c" composite methods is to provide a consistent description, *i.e.*, without systematic deviations at the lowest possible computational cost.[231,237,489,641,642] Hence, small but well-balanced AO basis sets are employed, while the remaining basis set errors are corrected by a geometrical counterpoise correction (gCP) or absorbed into the DFA itself *via* the parameterization.[643] Further, short-ranged basis set (SRB) errors, mainly influencing covalent bond lengths in most (meta-) GGA DFAs, can be addressed by a specific correction. For some of the 3c methods, even direct modifications in the exchange-correlation part are made. All 3c composite methods apply either the D3 or D4 London dispersion correction. 3c methods reach in general a high accuracy at comparably small computational costs and are intensively used in the course of this thesis. An overview of the most important 3c composite methods is given in Table A2.1.

Table A2.1.: Overview of 3c composite methods used in this thesis and their specific components and modifications.

	PBEh-3c[231]	B97-3c[237]	r^{2}SCAN-3c[489]
AO basis set	def2-mSV(P)	mTZVP	mTZVPP
# parameters in F_X	2	4	0
# parameters in F_C	1	6	0
Fock exchange / %	42	0	0
Dispersion correction	D3	D3	D4
SRB correction	no	yes	no
gCP correction	yes	no	yes

A2.2. Extended Tight-Binding Methods

The following equations are taken from Ref. 96. The general GFN*n*-xTB energy terms are:

$$E_{\text{GFN1-xTB}} = E_{\text{rep}} + E_{\text{disp}}^{\text{D3}} + E_{\text{XB}} + E_{\text{EHT}} + E_{\gamma} + E_{\Gamma} + G_{\text{Fermi}}\,, \tag{A2.12a}$$

$$E_{\text{GFN2-xTB}} = E_{\text{rep}} + E_{\text{EHT}} + E_{\text{disp}}^{\text{D4'}} + E_{\gamma} + E_{\text{AES}} + E_{\text{AXC}} + E_{\Gamma} + G_{\text{Fermi}}\,, \tag{A2.12b}$$

$$E_{\text{GFN0-xTB}} = E_{\text{rep}} + E_{\text{disp}}^{\text{D4}} + E_{\text{SRB}} + E_{\text{EEQ}} + E_{\text{EHT}} + G_{\text{Fermi}}\,. \tag{A2.12c}$$

All information that is ommited here for brevity can be found in the original literature.[93–96]

A2.2.1. Common GFN*n*-xTB Ingredients

The GFN*n*-xTB wave functions are formulated in terms of partially polarized, (mostly) minimal valence basis sets, consisting of spherical Gaussian-type atomic orbital (GTO) basis functions. Contracted GTOs ϕ_μ are constructed from primitive GTOs χ_z^μ, where $d_{z\mu}$ are the corresponding contraction coefficients. Each contracted GTO ϕ_μ then approximates a Slater-type orbital (STO) ϕ_μ^{STO}[190]

$$\phi_\mu^{\text{STO}}(\mathbf{r}) \approx \phi_\mu(\mathbf{r}) = \sum_z^{N_{\text{prim}}^\mu} d_{z\mu}\chi_z^\mu(\mathbf{r})\,. \tag{A2.13}$$

The molecular orbitals ψ_j are expanded as a linear combination in this atomic orbital (AO) basis

$$\psi_j(\mathbf{r}) = \sum_\mu^{N_{\text{AO}}} C_{\mu j}\phi_\mu(\mathbf{r})\,. \tag{A2.14}$$

Deriving the respective energy expressions for the orbital coefficients yields a Roothaan–Hall-type generalized eigenvalue equation (*cf.* Section 2.1.3). GFN*n*-xTB methods employ a finite electronic temperature ($T_{\text{el}} = 300\,\text{K}$),[644] which introduces fractional orbital occupations and static electron correlation effects with formally a single-reference treatment. The xTB energy expressions (Eq. A2.12) are

augmented with an electronic entropy term

$$G_{\text{Fermi}} = k_{\text{B}} T_{\text{el}} \sum_{\sigma=\alpha,\beta} \sum_{i} \left[n_{i\sigma} \ln\left(n_{i\sigma}\right) + \left(1 - n_{i\sigma}\right) \ln\left(1 - n_{i\sigma}\right) \right] . \tag{A2.15}$$

T_{el} refers to the electronic temperature, k_{B} the Boltzmann constant, and $n_{i\sigma}$ is the fractional occupation number of the spin-MO $\psi_{i\sigma}$, which is given by the Fermi-distribution

$$n_{i\sigma} = \frac{1}{\exp[(\epsilon_i - \epsilon_{\text{F}}^{\sigma})/(k_{\text{B}} T_{\text{el}})] + 1} . \tag{A2.16}$$

Here, ϵ_i is the energy of the orbital ψ_i and $\epsilon_{\text{F}}^{\sigma}$ is the Fermi level within the respective spin orbital space (α or β). The xTB variants are formally spin-restricted and thus, the fractional occupation can be used to introduce spin dependencies and open-shell systems. The repulsion energy in all GFN-type methods is given as an atom-pairwise expression (Eq. A2.17)

$$E_{\text{rep}} = \frac{1}{2} \sum_{A,B} \frac{Z_A^{\text{eff}} Z_B^{\text{eff}}}{R_{AB}} e^{-\sqrt{\alpha_A \alpha_B} (R_{AB})^{k_f}} . \tag{A2.17}$$

Here, Z^{eff} are element-specific fitting parameter that correspond to effective nuclear charges, which are screened by the core reference density $\rho_0^{A,\text{core}}$. $k_f = \frac{3}{2}$ is a global parameter, while the α exponents are element-specific parameters. The extended Hückel term is calculated in all GFNn-xTB schemes as

$$E_{\text{EHT}} = \sum_{\mu\nu} P_{\mu\nu} H_{\nu\mu}^{\text{EHT}} , \tag{A2.18}$$

where the Hamiltonian elements are constructed *via*

$$H_{\mu\nu}^{\text{EHT}} = \frac{1}{2} K_{AB}^{ll'} S_{\mu\nu} \left(H_{\mu\mu} + H_{\nu\nu}\right) X(EN_A, EN_B) \, \Pi(R_{AB}, l, l') \, Y(\zeta_l^A, \zeta_{l'}^B) . \tag{A2.19}$$

Herein A/B are atoms, the indices μ/ν indicate AOs, l/l' indicate shells, and $\mu \in l(A)$ and $\nu \in l'(B)$. $K_{AB}^{ll'}$ is a shell-specific scaling constant, $X(EN_A, EN_B)$ is a electronegativity dependent function, and $Y(\zeta_l^A, \zeta_{l'}^B)$ is a shell-exponent dependent term that only occurs in GFN2-xTB. The distance-dependent polynomial scaling function $\Pi(R_{AB}, l, l')$ is

$$\Pi(R_{AB}, l, l') = \left(1 + k_{A,l}^{poly} \left(\frac{R_{AB}}{R_{cov,AB}}\right)^{\frac{1}{2}}\right) \left(1 + k_{B,l'}^{poly} \left(\frac{R_{AB}}{R_{cov,AB}}\right)^{\frac{1}{2}}\right) , \tag{A2.20}$$

where $R_{cov,AB}$ are the the summed covalent radii and $k_{A,l}^{poly}$ the fitted parameters.

A2.2.2. Energies in GFN1-xTB

In GFN1-xTB, the EHT Hamiltonian components $H_{\mu\mu}$ and $X(EN_A, EN_B)$ are given by

$$H_{\mu\mu}^{\text{GFN1}} = h_A^l \left(1 + k_{CN,l} CN_A\right) \; \forall \, \mu \in l \in A , \tag{A2.21}$$

with the element-specific parameter h_A^l, the global angular momentum-specific parameter $k_{CN,l}$, and

$$X(EN_A, EN_B) = \left(1 + k_{EN}\Delta EN_{AB}^2\right) , \tag{A2.22}$$

where k_{EN} is a global parameter. CN_A is the coordination number of atom A, which is calculated in analogy to the D3 coordination number.[252,253] The second-order exchange-correlation and isotropic electrostatic energy contributions are described by equation A2.23

$$E_\gamma^{\text{GFN1}} = \frac{1}{2}\sum_{A,B}^{N_{\text{atoms}}}\sum_{l\in A}\sum_{l'\in B} q_l q_{l'}\gamma_{AB,ll'} , \tag{A2.23}$$

with q_l representing the Mulliken[128] partial shell charges. $\gamma_{AB,ll'}$ is a Mataga–Nishimoto–Ohno–Klopman damping function for the Coulomb interaction (Eq. A2.24), damped by the term $\eta_{AB,ll'}$ (Eq. A2.25)

$$\gamma_{AB,ll'} = \frac{1}{\sqrt{R_{AB}^2 + \eta_{AB,ll'}^{-2}}} . \tag{A2.24}$$

In this short-range damping function, η is an element-specific atomic hardness parameter and κ a shell-dependent scaling parameter

$$\eta_{AB,ll'} = 2\left(\frac{1}{\eta_A(1+\kappa_A^l)} + \frac{1}{\eta_B(1+\kappa_B^{l'})}\right)^{-1} . \tag{A2.25}$$

At third order the GFN1-xTB energy is given by

$$E_\Gamma^{\text{GFN1}} = \frac{1}{3}\sum_A (q_A)^3\Gamma_A , \tag{A2.26}$$

which is an on-site third order electrostatic and exchange-correlation correction term. Γ_A is an element-specific parameter. What remains in GFN1-xTB are the D3 dispersion energy (*cf.* Eq. A2.9) and a halogen bonds (XB) correction

$$E_{\text{XB}}^{\text{GFN1}} = \sum_{AB}^{N_{XB}} f_{damp}^{AXB} k_X \left[\left(\frac{k_{XR}R_{cov,AX}}{R_{AX}}\right)^{12} - k_{X2}\left(\frac{k_{XR}R_{cov,AX}}{R_{AX}}\right)^{6}\right]\left[\left(\frac{k_{XR}R_{cov,AX}}{R_{AX}}\right)^{12} - 1\right]^{-1} . \tag{A2.27}$$

Here, k_{XR} and k_{X2} are global parameters, k_X is the halogen-specific parameter, and f_{damp}^{AXB} is the angular three-body damping function.

A2.2.3. Energies in GFN2-xTB

In GFN2-xTB, the electronegativitiy dependent scaling function $X(EN_A, EN_B)$ has the same form as in GFN1-xTB (Eq. A2.22). The Hamiltonian elements are reformulated as

$$H_{\mu\mu}^{\text{GFN2}} = h_A^l - \delta h_{CN'_A}^l CN'_A , \tag{A2.28}$$

where h_A^l and $\delta h^l_{CN'_A}$ are shell and element specific parameters. CN'_A is a modified[93] version of the D3 coordination number. GFN2-xTB additionally introduces the AO exponent ζ dependent term as a further scaling function of the Hamiltonian elements

$$Y(\zeta_A, \zeta_B) = \left(\frac{2\sqrt{\zeta_l^A \zeta_{l'}^B}}{\zeta_l^A + \zeta_{l'}^B} \right)^{\frac{1}{2}} . \tag{A2.29}$$

The isotropic second order electrostatic energy E_γ is given in Eq. A2.23. The short-range damping is modified according to[93]

$$\eta_{AB,l\,l'} = \frac{1}{2}\left[\eta_A\left(1+\kappa_A^l\right) + \eta_B\left(1+\kappa_B^{l'}\right)\right] , \tag{A2.30}$$

where 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 $l = 0$). GFN2-xTB is the first TB scheme that employs at second order anisotropic electrostatic (AES) and exchange-correlation (AXC) terms according to

$$\begin{aligned} E_{\text{AES}} =& E_{q\mu} + E_{q\Theta} + E_{\mu\mu} && \text{(A2.31)} \\ =& \frac{1}{2}\sum_{A,B}\{f_3(R_{AB})\left[q_A(\boldsymbol{\mu}_B^T\mathbf{R}_{BA}) + q_B(\boldsymbol{\mu}_A^T\mathbf{R}_{AB})\right] \\ &+ f_5(R_{AB})[q_A\mathbf{R}_{AB}^T\boldsymbol{\Theta}_B\mathbf{R}_{AB} + q_B\mathbf{R}_{AB}^T\boldsymbol{\Theta}_A\mathbf{R}_{AB} \\ &- 3(\boldsymbol{\mu}_A^T\mathbf{R}_{AB})(\boldsymbol{\mu}_B^T\mathbf{R}_{AB}) + (\boldsymbol{\mu}_A^T\boldsymbol{\mu}_B)R_{AB}^2]\} && \text{(A2.32)} \\ E_{\text{AXC}} =& \sum_A \left(f_{XC}^{\mu_A}|\boldsymbol{\mu}_A|^2 + f_{XC}^{\Theta_A}||\boldsymbol{\Theta}_A||^2\right) . && \text{(A2.33)} \end{aligned}$$

The cumulative atomic multipole moments (CAMM) describe the anisotropic interaction between the atomic charge q_A, the dipole moment μ_A, and the quadrupole moment Θ_A. $f_n(R_{AB})$ are distance dependent damping functions with global parameters. $f_{XC}^{\mu_A}$ and $f_{XC}^{\Theta_A}$ are element-specific parameters. The dispersion energy at second order is described by a self-consistent version of the D4 model[242,249] according to

$$\begin{aligned} E_{\text{disp}}^{\text{D4'}} =& -\sum_{A>B}\sum_{n=6,8} s_n \frac{C_n^{AB}(q_A, CN_A, q_B, CN_B)}{R_{AB}^n} f_{damp,BJ}^{(n)}(R_{AB}) \\ &- s_9 \sum_{A>B>C} \frac{(3\cos(\theta_{ABC})\cos(\theta_{BCA})\cos(\theta_{CAB}) + 1)\, C_9^{ABC}(CN_{cov}^A, CN_{cov}^B, CN_{cov}^C)}{(R_{AB}R_{AC}R_{BC})^3} \\ &\times f_{damp,0}^{(9)}(R_{AB}, R_{AC}, R_{BC}) . \end{aligned} \tag{A2.34}$$

This model depends on the atomic charges taken from the Mulliken population and employs both, Becke–Johnson $f_{damp,BJ}^{(n)}$, and zero damping $f_{damp,0}^{(9)}$ functions. The third order IES/IXC term in GFN2-xTB is closely related to Eq. A2.26. In addition, it is shell-dependent and introduces a global shell-

specific parameter K_l^Γ

$$E_\Gamma^{\mathrm{GFN2}} = \frac{1}{3}\sum_A \sum_{l\in A} (q_l)^3 \, K_l^\Gamma \, \Gamma_A \,. \tag{A2.35}$$

In GFN2-xTB, an empirical XB correction is obsolete due to the sophisticated AES/AXC description.

A2.2.4. Energies in GFN0-xTB

GFN0-xTB truncates the TB energy expansion already at first order and is thus the most empirical GFN*n*-xTB scheme. To account for the isotropic electrostatic interactions, which appear at second order, a classical charge equilibrium model (EEQ) is employed (*cf.* Chapter 3 in Part II). With this, a charge dependence is built into the EHT Hamiltonian elements

$$H_{\mu\mu}^{\mathrm{GFN1}} = h_A^l - \delta h_{mCN_A}^l \, mCN_A - \delta h_{q_A}^l \, q_A - \Gamma_{q_a}^l q_A^2 \,. \tag{A2.36}$$

Here, h_A^l, $\delta h_{mCN_A}^l$, $\delta h_{q_A}^l$ and $\Gamma_{q_A}^l$ are element-specific parameters, where the latter two are related to the chemical hardness. mCN_A is a modified coordination number.[95] The GFN0-xTB EHT Hamiltonian is further modified by the electronegativity dependent scaling function $X(EN_A, EN_B)$, which is shell dependent according to

$$X^{ll'}(EN_A, EN_B) = 1 + k_{EN}^{ll'} \Delta EN_{AB}^2 + k_{EN}^{ll'} b_{EN} \Delta EN_{AB}^4 \,, \tag{A2.37}$$

where $k_{EN}^{ll'}$ is a shell-specific parameter and b_{EN} a global parameter. The electronegativity is additionally applied within the short-range basis (SRB) correction[237,641]

$$E_{\mathrm{SRB}} = k_{srb} \sum_{A,B} \exp\left[-\eta_{srb}\left(1 + g_{scal}\Delta EN_{AB}^2\right)\left(R_{AB} - R_{AB}^{srb}\right)\right] \,, \tag{A2.38}$$

with the global fit parameters k_{srb}, η_{srb}, and g_{scal}. The covalent bond radii R_{srb}^{AB} are modified by the electronegativities.[95] In GFN0-xTB, the dispersion energy is described by the standard D4 model.[242,249] The D4 model and the GFN0 Hamiltonian require atomic charges, which are not obtained from Mulliken populations but from the EEQ model. This model yields the zeroth order electrostatic energy according to

$$E_{\mathrm{EEQ}} = \sum_A \left[\chi_A \, q_A + \frac{1}{2}\left(J_{AA} + \frac{2}{\sqrt{\pi}}\gamma_{AA}\right)\right] + \frac{1}{2}\sum_{A,B} q_A \, q_B \frac{\mathrm{erf}(\gamma_{AB} R_{AB})}{R_{AB}} \,, \tag{A2.39}$$

with, J_{AA}, $\chi_A(EN_A, \kappa_A, mCN_A)$ as fitted element-specific parameters. γ_{AB} is related to the inverse root mean square of the atomic radii.[95]

A2.3. Molecular Dynamics and Structure Ensembles

A convenient approach to sample the conformational space of a flexible molecule is given by molecular dynamics (MD) simulations. MD simulations analyze the physical movements of atoms and molecules and give a view of the dynamic "evolution" of the system. Given the KS-DFT, GFN*n*-xTB, or GFN-FF expressions for E_{tot}, the negative gradient with respect to the nuclear positions, *i.e.*, the forces may be computed. In Born-Oppenheimer MD simulations, the molecular trajectories are determined by numerically solving Newton's equations of motion for a system of interacting particles (Eq. A2.40)[645,646]

$$M_I \frac{d^2\vec{R}_I}{dt^2} = \underbrace{-\vec{\nabla}_I E_{\text{tot}}}_{\vec{F}_I} . \tag{A2.40}$$

Here, M_I is the mass of nucleus I, $\vec{R}_I$ its position, and $\vec{F}_I$ the force acting on it. The right-hand side of Eq. A2.40 is obtained from a molecular energy calculation. The nuclear positions at a new point in time $t + \Delta t$ are computed numerically, using a finite time step Δt, which is governed by the fastest nuclear motions. The nuclear position and velocity v_I are computed, *e.g.*, by the leap-frog algorithm given in Eq. A2.41[646]

$$\vec{R}_I(t + \Delta t) = \vec{R}_I(t) + \vec{v}_I\left(t + \frac{\Delta t}{2}\right)\Delta t \tag{A2.41a}$$

$$\vec{v}_I\left(t + \frac{\Delta t}{2}\right) = \vec{v}_I\left(t - \frac{\Delta t}{2}\right) + \frac{\vec{F}_I}{M_I}\Delta t . \tag{A2.41b}$$

In Eqs. A2.41a and A2.41b, the position and velocity updates are out of phase by half a time step. This interleaving evaluation of positions and velocities gains numerical stability concerning the conservation of energy.

With standard MD simulations it is often difficult to overcome larger energetic barriers and hence, only small parts of the chemical space are within the reach of MD simulations. To improve upon that, the concept of meta-dynamics (MTD) was introduced,[220] in which a biasing potential is added to globally explore the conformer space. The bias potential is as a sum of Gaussian functions expressed with the root-mean-square deviation (RMSD) in Cartesian space as a metric for the collective variables (*cf.* Eq. 8.4, Chapter 8)

$$E_{\text{bias}}^{\text{RMSD}} = \sum_{i=1}^{n} k_i e^{\left(-\alpha\Delta_i^2\right)} . \tag{A2.42}$$

Here, n is the number of reference structures associated with the pushing (or pulling) strength k, Δ is the structural RMSD, and the parameter α determines the width of the biasing potential. Adding $E_{\text{bias}}^{\text{RMSD}}$ to the molecular energy within a MTD simulation leads to a chemically realistic exploration of the PES by allowing otherwise impossible barrier crossings. The operating principle of MTD simulations is illustrated in Figure A2.1A.

The *crest* program represents an efficient scheme for the *in silico* sampling for parts of the molecular low-energy chemical space by semiempirical TB or FF methods, combined with a meta-dynamics

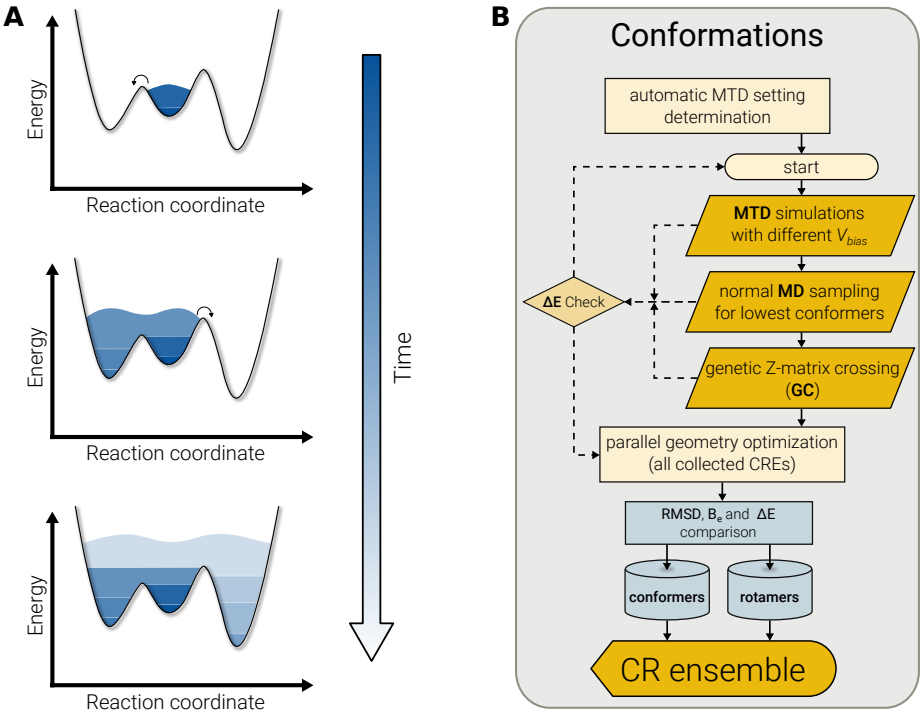


Figure A2.1.: (A) Schematic one dimensional PES that is "filled" by several biaspotentials over time, which allows larger barrier heights to be overcome. (B) Simpliefied workflow of the CREST algorithm. Figure adapted from Ref. 60

driven search algorithm. A simplified scheme is depicted in Figure A2.1B. Here the key aspect is the combination of MTD and MD simulations followed by full geometry optimizations of the respective trajectories. Based on the RMSD, the rotation constant B_e, and the difference in energy ΔE, the resulting structures are identified as conformers or rotamers, and duplicates are sorted out. The CREST algorithm is applied several times throughout this thesis and highly important for the generation of the starting structures. For further information, see the original publication Ref. 60.

A3. Supporting Information to Chapter 3

Appendix A3 contains:

- General remarks
- Supplementary text
- Materials and methods

General remarks

The Supplementary Material of the original publication is available:
Spicher, S.; Grimme, S. "Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems", *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.

The free *xtb* source code[243] can be obtained online from:
https://github.com/grimme-lab/xtb.

Supplementary text

Use of additional fragment charge information

In GFN-FF the computed atomic charges from the EEQ model may be improved by constraints if additional information about the charge distribution in the system is known. There are three different ways to incorporate this information. If the system consists of more than one NCI fragment, the charges per fragment can be written by the user into a specific file (named *.CHRG*) and will be constrained accordingly in the EEQ model, thus preventing artificial charge transfer between the NCI fragments. If a GFN-xTB calculation is performed in advance, the written file "charges" is read by the program, and the corresponding QM charges are used to constrain the values on the molecular fragments. The last option is useful, especially for biological systems. If a "pdb" file is provided with known charges on specific residues, the *xtb* program reads this information, determines the overall charge of the system automatically, and applies this charge constrain per residue, again preventing (inter-residue) charge transfer.

Potential energy curves

GFN-FF is a dissociative force-field allowing cleavage of all covalent bonds that are assigned by the initial topological procedure. Forming new covalent bonds after the setup stage (i.e., a reactive force-field) is not possible. Figure A3.1 shows the computed potential energy curves for five diatomic molecules and includes reference dissociation energies taken from Ref. 647. The asymptotic correct value is indicated by dashed lines in the same color as the corresponding GFN-FF result. For a broad range of bonds with different polarity and bond orders, GFN-FF is able to describe the dissociation qualitatively correct. No attempt in the parameter fitting procedure was made to improve the description of atomization energies, i.e., this test can be seen as a difficult cross-check of the covalent part of the GFN-FF potential.

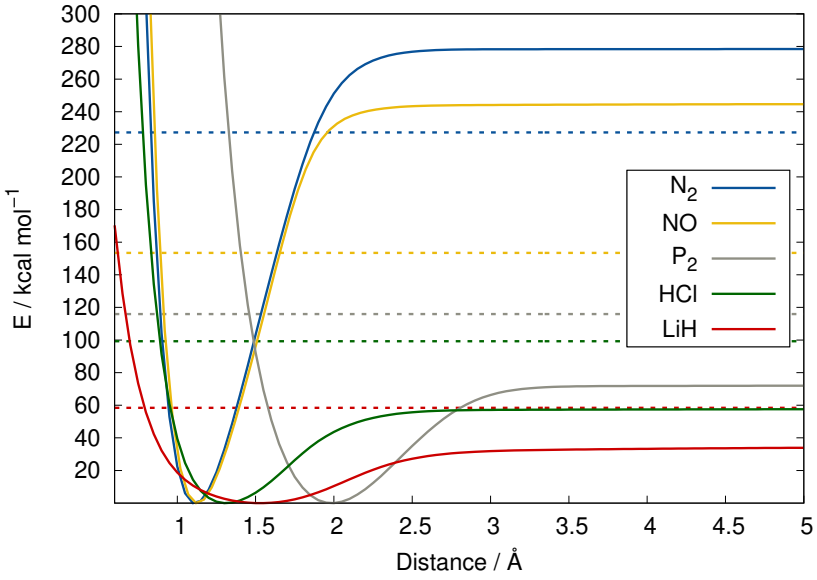


Figure A3.1.: Potential energy curves for diatomic molecules. Computed GFN-FF potential energy curves and reference dissociation energies (asymptotic value indicated by the dashed line) for five diatomic molecules.

Parameterization strategy

The fitting procedure is technically done with the Levenberg–Marquardt algorithm.[648,649] Applying this method, global and the element-specific parameters are determined by minimizing the root-mean-square deviation (RMSD) between reference (PBEh-3c[231] and B97-3c[237]) and GFN-FF computed data. The reference data consist of equilibrium and distorted geometries for force matching, harmonic frequencies, and NCI energies of subsets from the GMTKN55 database. The reference structure set is an extended and refined version of the GFN2-xTB training set, owing to the fact that larger systems can be treated by the force-field. Special attention is paid towards the correct description of transition-metal (TM) complexes by the inclusion of highly coordinated TM systems. Furthermore, the general fitting strategy is identical, meaning the global parameters have been determined along with the element spe-

cific parameters for the elements H, C, N, and O first and the parameters for all other elements were fitted while keeping the existing (HCNO)-parameters fixed. The element specific parameters for the lanthanoids are linearly interpolated between the elements La and Lu by means of the nuclear charge Z. Due to the number of parameters, they will not be listed in the appendix. However, the *xtb* code is open source and parameters can be looked up in the code provided on our github website.[243]

Trouble-shooting

The implementation of GFN-FF was tested to the best of our ability for very many different chemical systems. However, incorrect assignments of bond topology leading to bad results may still occur in particular for unusual element combinations. Therefore we refer to our github website,[243] where a detailed manual of how to use GFN-FF in *xtb* can be found. This website also offers the possibility to open up issues and ask questions directly. Since it is impossible for us to test every possible molecular system, in some cases GFN-FF might fail. If such a case is observed we are pleased about user feedback and we will try to correct it. Updates of the program in this respect as well as other revisions will always be announced on this website. This procedure worked already well informally during the development of the program/method thanks to many beta-users worldwide.

Materials and Methods

In this section, all the necessary information is given to reproduce the data in this work. The computational methods and software packages used are listed as well as all the data necessary to construct the graphics. For the given statistics, the following abbreviations are used throughout: mean deviation (MD), mean absolute deviation (MAD), standard deviation (SD), and root mean square deviation (RMSD).

Computational details

Detailed comparisons between GFN1-xTB and GFN2-xTB with other semiempirical methods have been conducted in the respective original method publications.[93,94] Therefore we will compare the performance of GFN-FF to its predecessors and results for other semiempirical calculations, such as DFTB3-D3(BJ),[92,650] PM6-D3H4X,[315] and PM7.[88] All GFN calculations have been performed with our standalone program termed *xtb* (version 6.3.0). PMx calculations have been performed with MOPAC2016 (version 18.151).[338] DFT and coupled cluster calculations have been conducted with the TURBOMOLE.7.3.1[651] and ORCA.4.1.1[652,653] program packages. Computational data for each benchmark set was taken from the respective publication. For the geometry optimizations of metal-organic frameworks and proteins in Section 3.1 and 3.3, an implicit GBSA solvation model is used throughout. For all optimizations a convergence thresholds of $E_{\text{conv}} = 5 \cdot 10^{-6} E_h$ are employed. For the optimizations of all protein structures in Section 3.3 this value is changed to $E_{\text{conv}} = 5 \cdot 10^{-5} E_h$. MD simulations with GFN-FF were carried out for 1 ns at 298 K employing the implicit GBSA(H_2O) solvation model. A time step of 4 fs (at an increased hydrogen mass of 4 amu) and equilibration phase of 200 ps was chosen.

Benchmark data

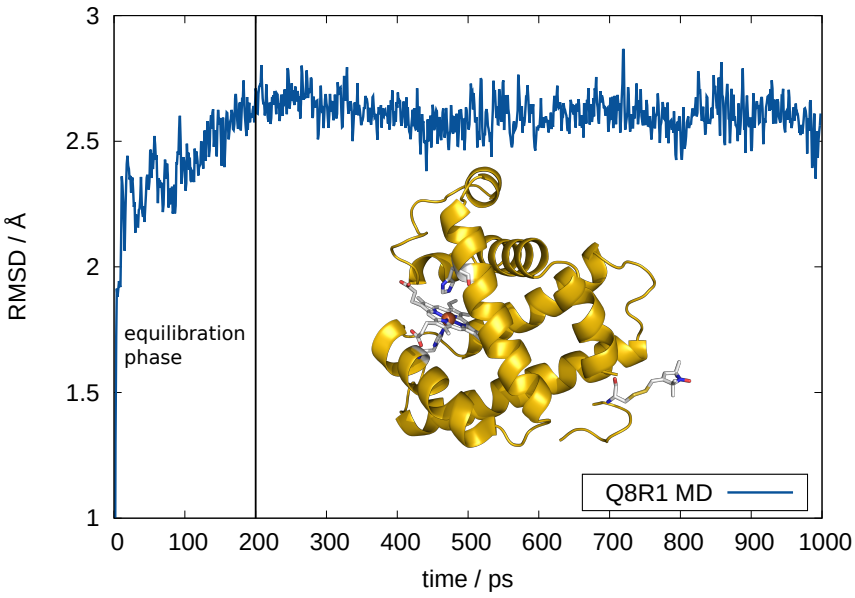


Figure A3.2.: RMSD of Q8R1 as a function of time. The first 200 ps are chosen as the equilibration phase and not considered for further structural evaluation.

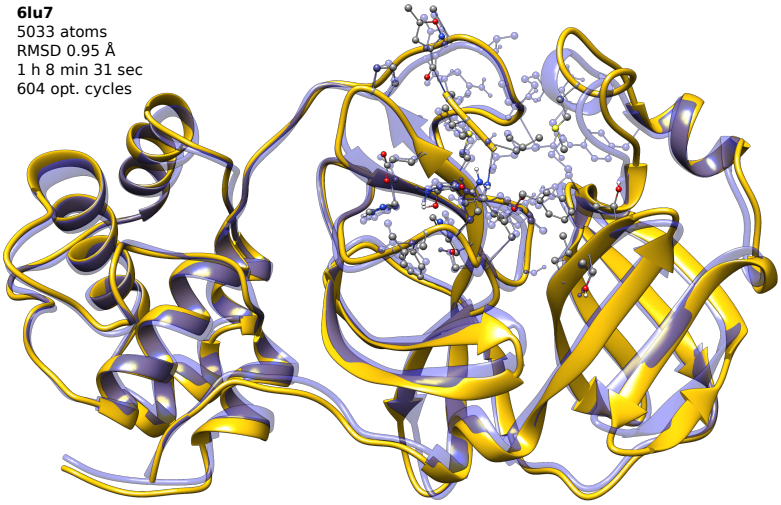


Figure A3.3.: Geometry optimization of the COVID-19 main protease in complex with an inhibitor N3. Within 1 hour and 622 optimization cycles on four Intel© Xeon CPUs, a stationary point on the PES is found. The RMSD between the experimental and the GFN-FF structure is 0.95 Å. A GBSA(H_2O) solvation model was used. The structure is taken from Ref. 264.

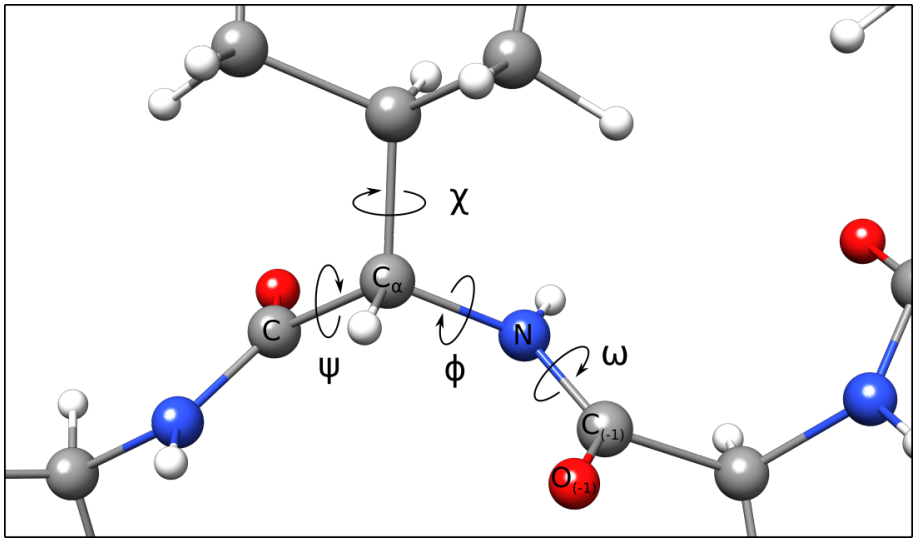


Figure A3.4.: Depiction of the four dihedral angles. The angles ϕ, ψ and ω discussed in this work are backbone angles, whereas χ is the first side chain angle.

Table A3.1.: Organic protein benchmark set. Average deviations of four types of dihedral angles (in degree) for the test set of 70 protein structures as well as average C_α and heavy atom (ha) RMSD (in Å) in each case with respect to the crystal structure.

	angles / °				RMSD / Å	
	$\Delta\phi$	$\Delta\psi$	$\Delta\xi$	$\Delta\omega$	C_α	ha
UFF	24.2	22.3	11.7	23.2	1.82	2.22
GFN2-xTB	14.2	14.7	10.1	5.4	0.59	0.87
AMBER*	19.8	15.3	12.1	5.0	0.69	0.97
OPLS2005	12.4	12.3	10.7	4.6	0.58	0.84
GFN-FF	12.4	11.4	9.4	7.5	0.60	0.81

Table A3.2.: TMG145 benchmark set. Statistical evaluation of the performance in terms of bond lengths and angles of the TMG145 benchmark set. Bond lengths in pm and angles in degree.

	bonds(M-A) / pm			angles(A-M-B) / °		
	MD	MAD	SD	MD	MAD	SD
GFN2-xTB	2.1	8.3	15.1	-0.50	3.9	6.6
UFF	2.6	14.6	21.1	-0.26	8.4	12.0
GFN-FF	-2.4	9.7	14.7	-0.03	5.7	8.7

Table A3.3.: MAD of the noncovalent interaction energies computed with GFN-FF and other semiempirical QM methods. Subsets taken from the GMTKN55 benchmark set. The values are given in kcal mol^{-1}.

benchmark	GFN0-xTB	GFN1-xTB	GFN2-xTB	PM7	GFN-FF
S22	1.630	1.330	0.758	0.768	0.89
S66	1.298	1.080	0.730	0.755	0.76
RG18	0.442	0.324	0.112	0.644	0.39
ADIM6	0.508	1.007	1.152	0.202	0.61
CARBHB12	2.316	0.670	1.084	1.883	1.81
HAL59	1.769	1.345	1.276	4.106	1.88
HEAVY28	0.908	0.658	0.608	2.948	0.77
PNICO23	2.000	2.331	1.104	5.673	1.98

Table A3.4.: MAD of the conformational energies computed with GFN-FF and other semiempirical QM methods. Subsets taken from the GMTKN55 benchmark set. The values are given in kcal mol^{-1}.

benchmark	GFN0-xTB	GFN1-xTB	GFN2-xTB	PM6-D3H4X	GFN-FF
ACONF	0.80	0.66	0.19	0.46	0.13
Amino20x4	1.00	1.11	0.95	1.47	1.52
BUT14DIOL	0.81	0.95	1.25	1.49	1.12
ICONF	2.33	2.63	1.63	3.13	3.02
MCONF	1.66	1.44	1.72	1.42	0.48
PCONF21	1.79	2.17	1.76	2.58	1.60
SCONF	1.69	2.50	1.64	5.32	1.41
UPU23	1.34	1.24	2.90	2.38	2.95

Table A3.5.: Association energies of the S30L benchmark set. Values are computed with GFN-FF and semiempirical methods for 30 large noncovalent complexes containing only main group elements (S30L). The values are given in kcal mol^{-1}.

System	GFN-FF	GFN2-xTB	PM6-D3H4X	ref.
1	-28.41	-25.93	-28.83	-29.04
2	-20.81	-18.02	-18.63	-20.78
3	-20.07	-22.50	-22.41	-23.54
4	-10.47	-21.72	-19.18	-20.27
5	-28.78	-33.88	-33.96	-28.99
6	-25.05	-25.57	-20.83	-25.50
7	-36.64	-42.20	-30.95	-35.06
8	-39.38	-48.70	-35.57	-36.79
9	-33.84	-34.83	-27.66	-28.38
10	-35.60	-35.86	-29.12	-29.78
11	-41.72	-41.74	-38.69	-32.95
12	-41.78	-42.21	-38.45	-33.92
13	-26.86	-22.30	-29.32	-30.83
14	-31.36	-25.64	-29.45	-31.33
15	-18.01	-24.10	-21.86	-17.39
16	-25.72	-25.85	-30.03	-25.12
17	-29.12	-26.78	-39.73	-33.38
18	-22.88	-20.38	-29.11	-23.31
19	-17.16	-13.05	-19.32	-17.47
20	-19.64	-15.23	-23.48	-19.25
21	-33.02	-22.14	-31.35	-24.21
22	-40.62	-36.57	-44.06	-42.63
23	-57.76	-60.72	-61.72	-61.32
24 1	-160.1	-136.59	-162.49	-135.5
25	-26.75	-28.08	-25.95	-25.96
26	-26.82	-28.21	-25.89	-25.77
27	-87.95	-83.16	-104.03	-82.18
28	-84.20	-79.49	-101.22	-80.11
29	-42.98	-50.95	-59.62	-53.54
30	-43.28	-50.52	-56.39	-49.28

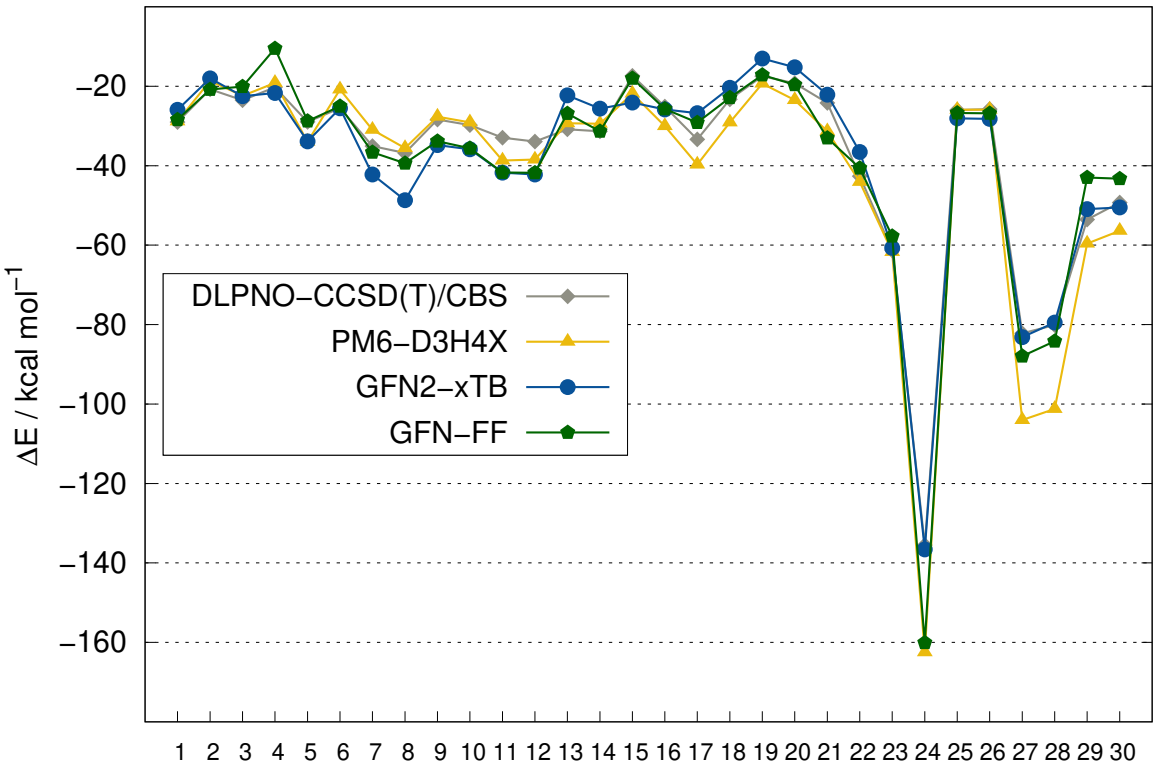


Figure A3.5.: Performance on the S30L benchmark set. Association energies for all systems in the S30L set computed with GFN-FF, GFN2-xTB and PM6-D3H4X. The reference energies are obtained at the DLPNO-CCSD(T)/CBS level.

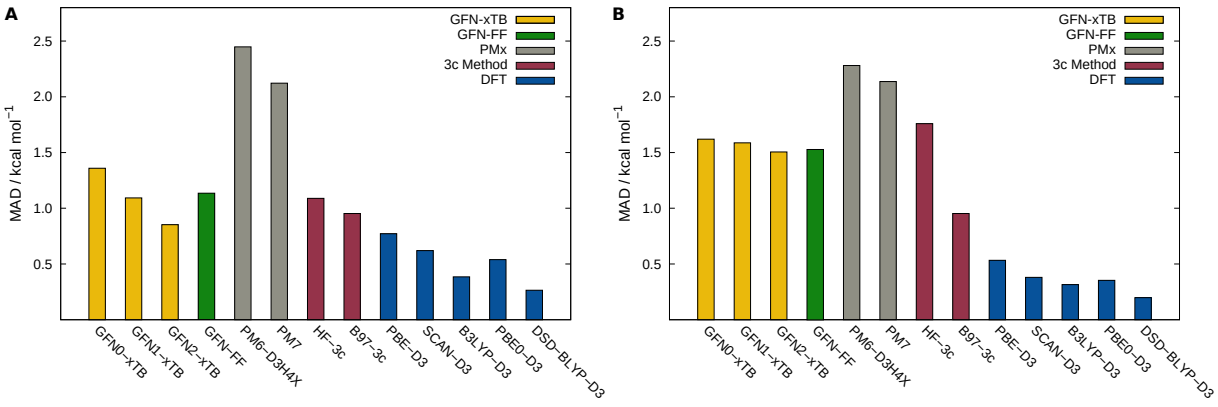


Figure A3.6.: Noncovalent interaction (A) and conformational (B) energies from the subsets of the GMTKN55 benchmark set. The MAD is given in kcal mol^{-1}.

Table A3.6.: MAD of the S30L benchmark set and the noncovalent and conformational energies for the GMTKN55 benchmark set. Values are computed with GFN-FF and other QM methods and are given in kcal mol^{-1}.

	S30L	GMTKN55	
Method		noncovalent interactions	conformational energies
GFN0-xTB	–	1.36	1.62
GFN1-xTB	6.08	1.09	1.59
GFN2-xTB	4.05	0.85	1.51
GFN-FF	4.15	1.13	1.53
PM6-D3H4X	5.15	2.45	2.28
PM7	16.4	2.12	2.14
HF-3c	5.20	1.09	1.76
PBE-D3	3.10	0.77	0.53
SCAN-D3	2.00	0.62	0.38
B97-3c	3.20	0.95	0.47
B3LYP-D3	5.88	0.39	0.32
PBE0-D3	–	0.54	0.35
DSD-BLYP-D3	–	0.27	0.20

The description of water

The performance of GFN-FF for small to medium sized clusters is shown in Figure A3.7. In part A, the dissociation curve of the water dimer is shown and compared to accurate CCSD(T)/CBS reference values.[654] Energies are shown relative to the equilibrium structure. GFN-FF shows the energetic minimum at the same inter-monomer distance as the reference, indicating the correct length of the hydrogen bond. For the water dimer, the interaction energy is semi-quantitatively correct but slightly overestimated at larger distances with GFN-FF. In part B dissociation energies of small to medium sized neutral water clusters are shown, taken from the WATER27 benchmark set.[540] The larger clusters are slightly underbound by GFN-FF but the error is in all cases <10% of D_e which is difficult to achieve even with sophisticated DFT methods. The GFN2-xTB method shown for comparison performs exceptionally well for this set. The errors of the UFF competitor method (not shown) exceed hundreds of kcal/mol for this benchmark set.

For a water cluster consisting of 1451 molecules an MD simulation at 298 K was performed for 5 ps with an equilibration time of the same length. An average density of 1.13 ± 0.003 g cm^{-1} is obtained. The deviation of 0.13 g cm^{-1} from the well-known experimental value corresponds to an average underestimation of the intermolecular distances by about 4 %. However, this is not seen for the water dimer as depicted in Figure A3.7, where the inter-monomer distance corresponds well with the reference.

Thermostatistical corrections

GFN-FF computed zero-point vibrational energies (ZPVE) and total molecular free energies at 298 K (G_{298}) are compared to corresponding values at the low-cost B97-3c DFT theoretical level. Figure A3.8 shows a comparison of ZPVE values at B97-3c and GFN2-xTB values for a set of 39 medium sized

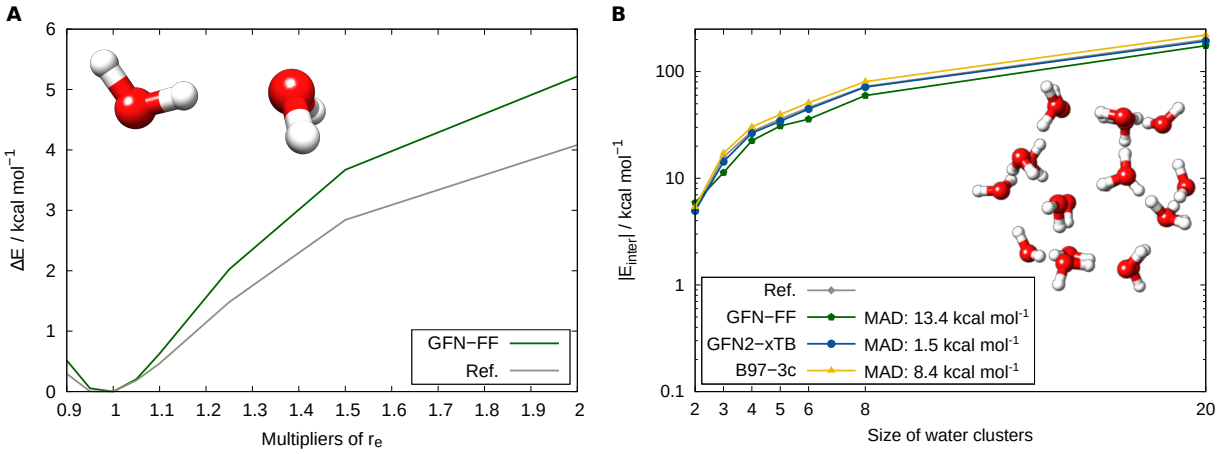


Figure A3.7.: Performance of GFN-FF for the description of water clusters. (A) Dissociation curve of the water dimer at different inter-monomer separations, where the equilibrium distance r_e is multiplied by the scaling factor shown (taken from revised S66x8 set, see Ref. 654). (B) Performance of different methods on the neutral complexes of the WATER27 benchmark set. Given are the cluster binding energies on a logarithmic scale and the corresponding MAD values compared to the CCSD(T)/CBS reference data shown as solid grey line.

organic molecules taken from a benchmark study of Li *et al.*[655]

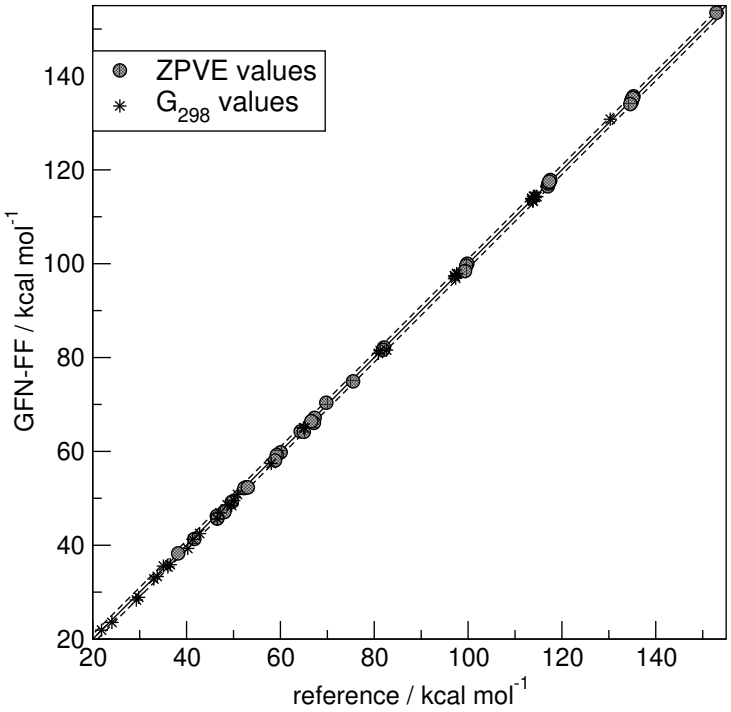


Figure A3.8.: Comparison of GFN-FF thermostatistical data with corresponding B97-3c DFT reference values. The set consists of 39 organic molecules ranging from ethane (smallest) to n-octane (largest). The solid line shows the one-to-one correspondence and the dashed ones indicate a common error range for chemical accuracy, i.e., ±1 kcal mol^{-1}.

The data have been used to determine as usual a frequency scale factor for GFN-FF. An optimum value of 1.03 was found and should be used as default in corresponding vibrational or thermostatistical applications. The scaled frequencies are listed in the output of the *xtb* program. As can be clearly seen from the graph, there is a very good reproduction of the DFT reference thermostatistical properties by GFN-FF. The MAD for the ZPVE data is only 0.37 kcal mol^{-1} (MD=0.20 kcal mol^{-1}). In actual applications where normally differences of the values for reactants and products are taken, the effective error is maybe even smaller because of cancellation. The performance for the free energies is similar with a MAD of only 0.42 kcal mol^{-1} (MD=0.22 kcal mol^{-1}).

Timings

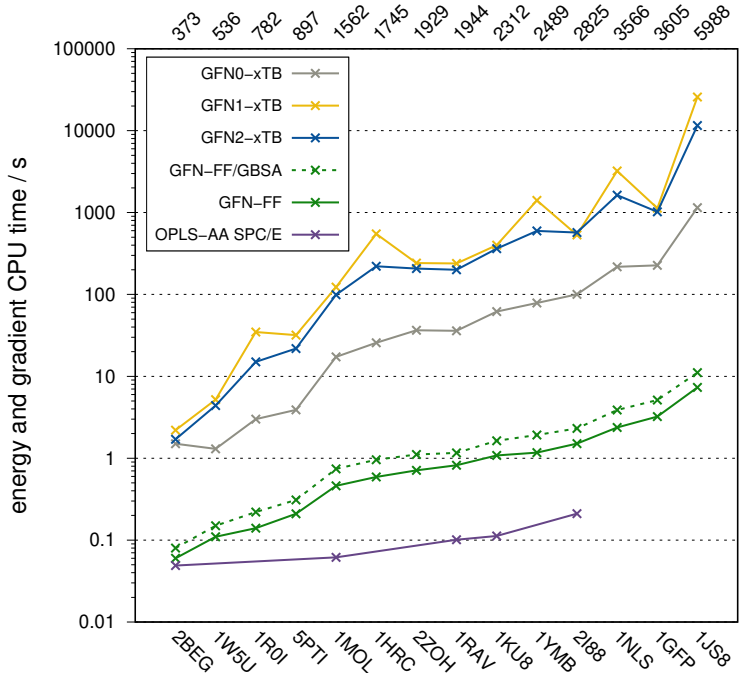


Figure A3.9.: Timings for energy and gradient computation of different GFN methods. CPU times (given in seconds) for single-point energy/gradient calculations of 14 proteins. GBSA(H_2O) had to be used for GFN1-xTB and GFN2-xTB in order to achieve SCC convergence. This is not necessary for GFN0-xTB and GFN-FF. Thus, for GFN-FF the timings including the GBSA model are shown separately. Additionally, the timings for OPLS-AA SPC/E with counter ions under periodic boundary conditions are shown for those proteins, where parameters are available. The PDB identifiers are given on the bottom x-axis, the corresponding number of atoms is given on top. Computations were performed on a single core of a quad-core desktop machine with 4.20 GHz Intel© i7-7700K CPUs.

A4. Supporting Information to Chapter 4

Appendix A4 contains:

- General remarks
- Statistical error measures
- Computational results

General remarks

The *xtb*[337] program package is available free of charge for academic use. Collected atomic Cartesian coordinates (XYZ format) of all input and optimized structures are available in addition to the Supporting Information:
Spicher, S.; Bursch, M.; Grimme, S. "Efficient Calculation of Small Molecule Binding in Metal-Organic Frameworks and Porous Organic Cages", *J. Phys. Chem. C* **2020**, *124*, 27529–27541.

Statistical error measures

Statistical measure for a set $x_1, \cdots, x_n$ of data points with references $r_1, \cdots, r_n$ are:

- Average:

$$\overline{x} = \frac{1}{n} \sum_{i}^{n} x_i \tag{A4.1}$$

- Mean deviation (MD):

$$MD = \frac{1}{n} \sum_{i}^{n} (x_i - r_i) \tag{A4.2}$$

- Mean absolute deviation (MAD):

$$MAD = \frac{1}{n} \sum_{i}^{n} |x_i - r_i| \tag{A4.3}$$

- relative MAD (*rel*MAD):

$$relMAD = \frac{1}{n} \sum_{i}^{n} \frac{|x_i - r_i|}{r_i} \tag{A4.4}$$

- Standard deviation (SD):

$$SD = \sqrt{\frac{\sum_i^n |(x_i - r_i) - MD|^2}{n-1}} \tag{A4.5}$$

- Root-mean-square deviation (RSMD):

$$RMSD = \sqrt{\frac{\sum_i^n |x_i - r_i|^2}{n}}, \tag{A4.6}$$

- *Pearson* correlation coefficient (r_p):

$$r_p = \frac{\sum_{i=1}^n (x_i - \overline{x}_i)(r_i - \overline{r}_i)}{\sqrt{\sum_{i=1}^n (x_i - \overline{x}_i)^2 \sum_{i=1}^n (r_i - \overline{r}_i)^2}}, \tag{A4.7}$$

- Empirical *Spearman* rank correlation coefficient (r_s):

$$r_s = \frac{\sum_{i=1}^n (R(x_i) - \overline{R}_x)(R(y_i) - \overline{R}_y)}{\sqrt{\sum_{i=1}^n (R(x_i) - \overline{R}_x)^2} \sqrt{\sum_{i=1}^n (R(y_i) - \overline{R}_y)^2}}, \tag{A4.8}$$

- Binding energy ΔE_{bind}:

$$\Delta E_{\text{bind}} = E(Complex) - E(Host) - E(Guest). \tag{A4.9}$$

Computational results

Table A4.1.: MAD_{mean}, $r_{p,mean}$, and $r_{s,mean}$ between PBE0-D4/def2-TZVP (reference) and all other tested methods, namely PBEh-3c, B97-3c, TPSS-/def2-TZVP, PBE-D4/def2-TZVP, GFN2-xTB, GFN1-xTB, GFN-FF, PM6-D3H4, PM7, and UFF respectively.

	MAD_{mean} kcal mol^{-1}	$r_{p,mean}$	$r_{s,mean}$
PBEh-3c	0.73	0.98	0.98
B97-3c	0.20	0.99	0.99
TPSS-D4	0.19	0.99	0.99
PBE-D4	0.40	0.99	0.99
GFN2-xTB	1.07	0.95	0.94
GFN1-xTB	1.39	0.96	0.97
GFN-FF	1.40	0.94	0.95
PM6-D3H4	6.06	0.84	0.86
PM7	4.91	0.86	0.91
UFF	12.4	0.02	0.09

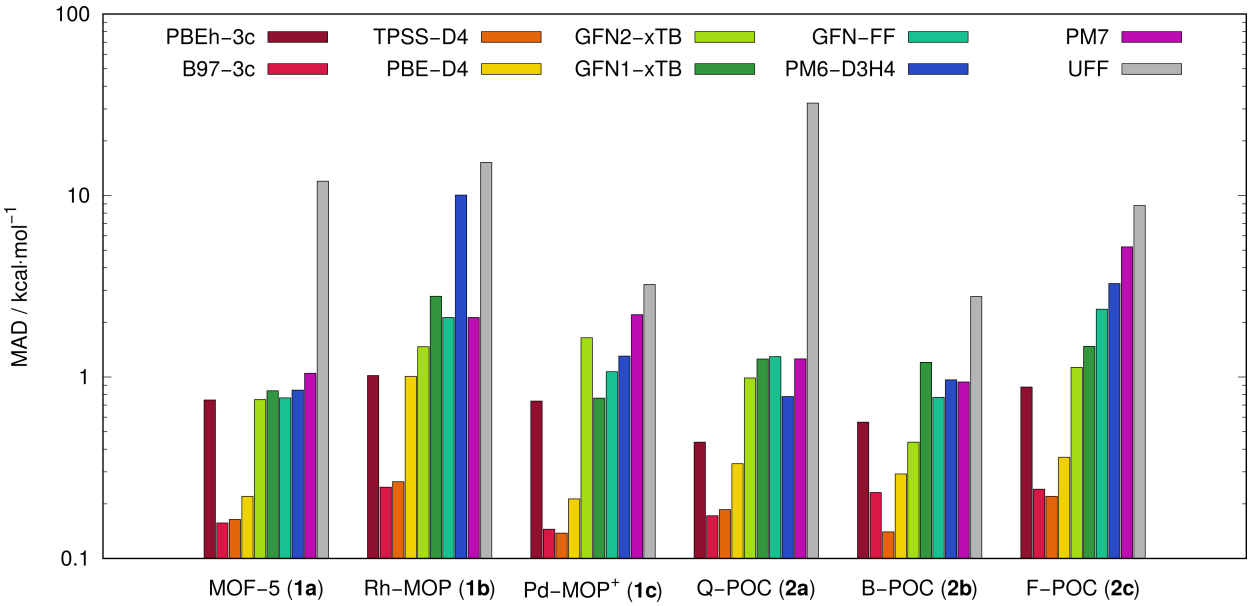


Figure A4.1.: MAD in kcal mol^{-1} for binding energies of H_2, N_2, CO_2, CH_3OH, and C_6H_6 in 6 different porous cages (**1a**-**2c**) computed by ten different methods in comparison to PBE-D4/def2-TZVP reference values. For PBE-D4 and TPSS-D4 a def2-TZVP basis set is employed as well.

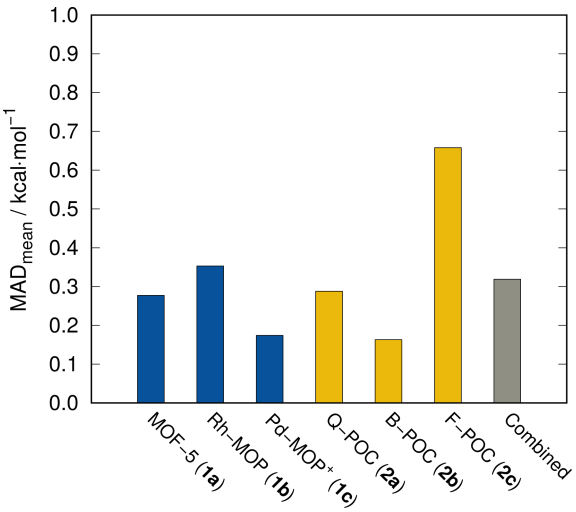


Figure A4.2.: MAD$_{mean}$ in kcal mol^{-1} for binding energies of H_2, N_2, CO_2, CH_3OH, and C_6H_6 in 6 different porous cages (**1a**-**2c**) computed by ωB97X-V/def2-TZVP in comparison to PBE-D4/def2-TZVP reference values.

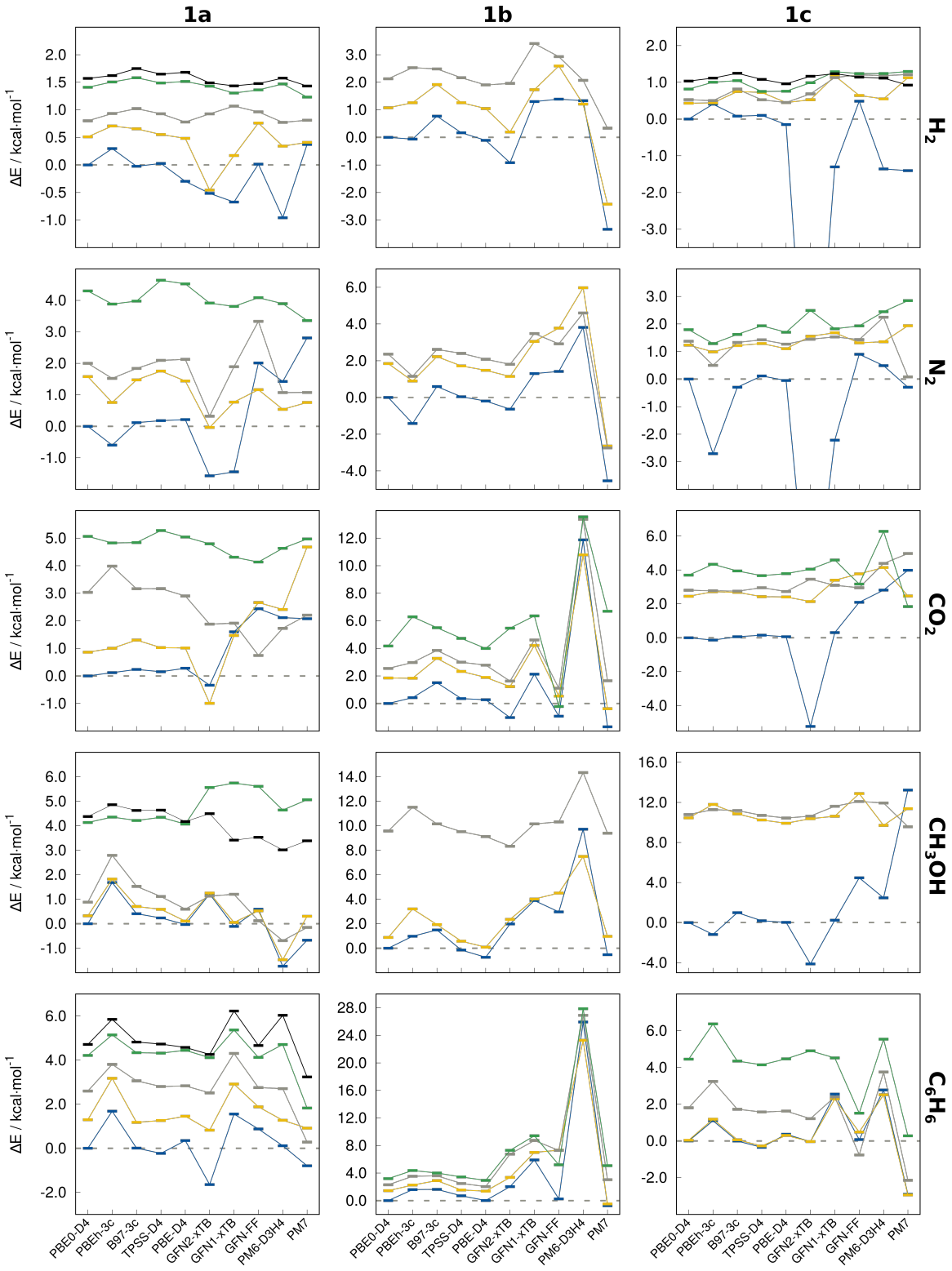


Figure A4.3.: Binding energies ΔE for H_2, N_2, CO_2, CH_3OH, and C_6H_6 adsorbed on metal-organic frmaworks **1a**-**1c** visualized by energy level diagrams. Energies are given relative to the lowest conformation at the PBE0-D4/def2-TZVP level of theory. For TPSS-D4 and PBE-D4, the same basis set is employed.

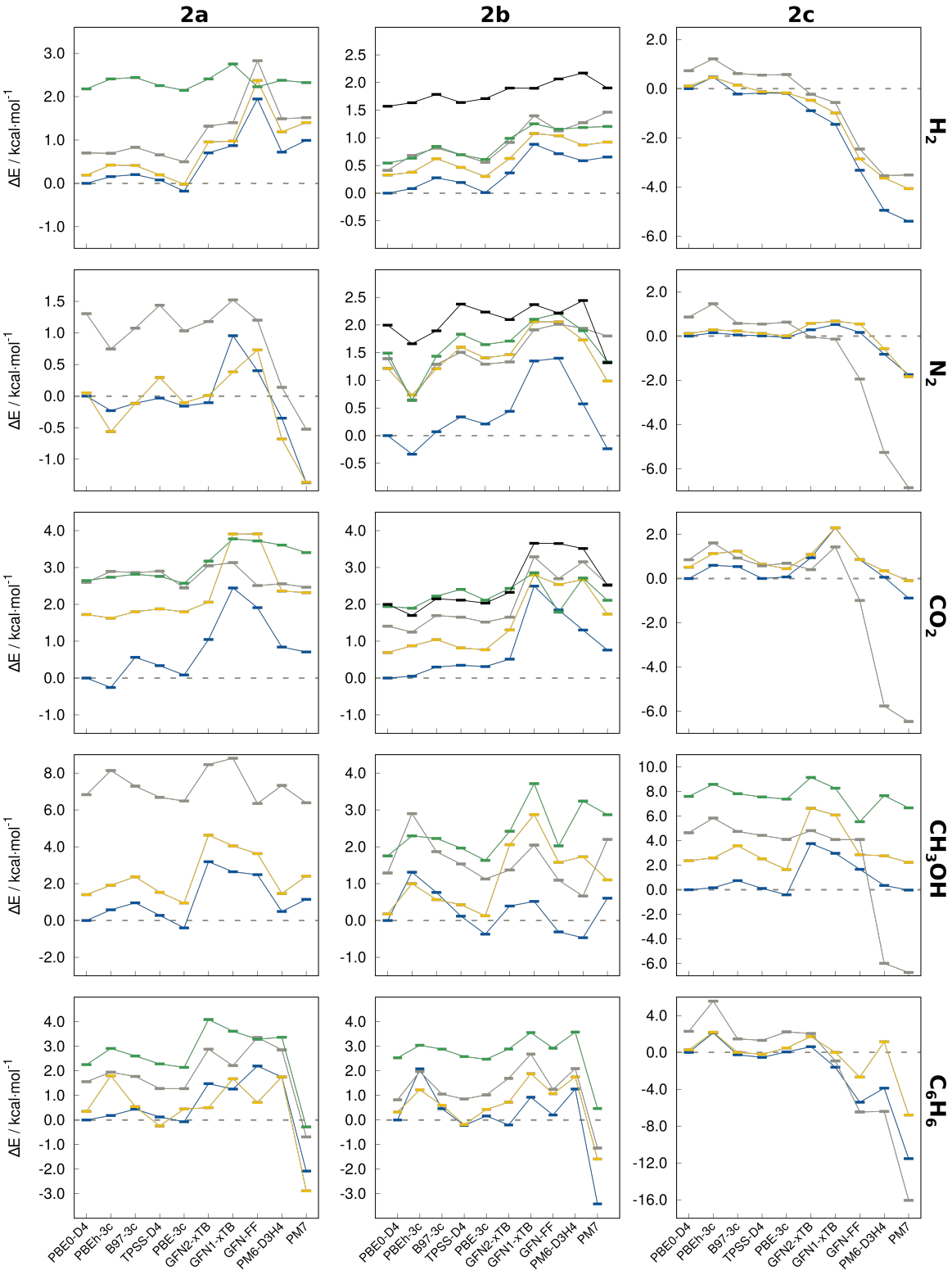


Figure A4.4.: Binding energies ΔE for H_2, N_2, CO_2, CH_3OH, and C_6H_6 adsorbed on porous organic cages **2a**-**2c** visualized by energy level diagrams. Energies are given relative to the lowest conformation at the PBE0-D4/def2-TZVP level of theory. For TPSS-D4 and PBE-D4, the same basis set is employed.

Table A4.2.: MAD_{mean}, $r_{p,mean}$, and $r_{s,mean}$ for the binding energies of all listed methods calculated on the GFN2-xTB and GFN-FF optimized structures in comparison to the respective binding energies an the reference structure (B97-3c).

	MAD_{mean} kcal mol^{-1}	$r_{p,mean}$	$r_{s,mean}$
PBE0-D4//GFN2-xTB	1.588	0.906	0.881
B97-3c//GFN2-xTB	1.623	0.904	0.888
GFN2-xTB//GFN2-xTB	0.330	0.985	0.974
GFN1-xTB//GFN2-xTB	0.551	0.988	0.964
GFN-FF//GFN2-xTB	0.606	0.974	0.929
PBE0-D4//GFN-FF	1.134	0.924	0.899
B97-3c//GFN-FF	1.236	0.899	0.901
GFN2-xTB//GFN-FF	0.481	0.966	0.976
GFN1-xTB//GFN-FF	0.525	0.956	0.955
GFN-FF//GFN-FF	0.580	0.974	0.967

Table A4.3.: Binding energies for the interaction of H_2, N_2, CO_2, CH_3OH, and C_6H_6 with MOF-5 (**1a**) calculated with the listed methods on B97-3c optimized structures. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

ΔE / kcal mol^{-1}		PBE0	PBEh-3c	B97-3c	TPSS	PBE	GFN2-xTB	GFN1-xTB	GFN-FF	PM6-D3H4	PM7	UFF
H_2	1	-1.52	-1.22	-1.54	-1.49	-1.81	-2.03	-2.19	-1.50	-2.48	-1.15	4.31
	2	-1.01	-0.81	-0.86	-0.97	-1.03	-1.97	-1.35	-0.76	-1.18	-1.11	-85.22
	3	-0.72	-0.58	-0.50	-0.59	-0.74	-0.59	-0.45	-0.55	-0.74	-0.71	3.06
	4	-0.11	-0.01	0.06	-0.03	-0.00	-0.09	-0.21	-0.16	-0.05	-0.29	3.43
	5	0.05	0.11	0.23	0.13	0.16	-0.03	-0.08	-0.04	0.06	-0.08	2.33
N_2	1	-3.93	-4.53	-3.81	-3.75	-3.72	-5.51	-5.38	-1.92	-2.51	-1.12	-11.94
	2	-2.35	-3.18	-2.46	-2.18	-2.49	-3.97	-3.16	-2.77	-3.39	-3.18	-11.30
	3	-1.93	-2.41	-2.10	-1.84	-1.81	-3.61	-2.04	-0.60	-2.86	-2.86	-12.13
	4	0.37	-0.05	0.05	0.71	0.59	-0.02	-0.13	0.15	-0.04	-0.57	-12.43
CO_2	1	-6.11	-5.99	-5.88	-5.96	-5.83	-6.45	-4.51	-3.67	-4.00	-4.04	-23.47
	2	-5.25	-5.11	-4.81	-5.09	-5.10	-7.11	-4.64	-3.45	-3.70	-1.43	-18.42
	3	-3.08	-2.13	-2.95	-2.94	-3.21	-4.23	-4.20	-5.37	-4.39	-3.90	-20.29
	4	-1.05	-1.28	-1.27	-0.84	-1.07	-1.31	-1.80	-1.98	-1.48	-1.14	-21.39
CH_3OH	1	-9.17	-7.49	-8.76	-8.93	-9.20	-7.95	-9.28	-8.58	-10.90	-9.84	-9.77
	2	-8.84	-7.35	-8.46	-8.58	-9.06	-7.91	-9.12	-8.64	-10.64	-8.86	-8.27
	3	-8.29	-6.37	-7.64	-8.06	-8.57	-8.04	-7.96	-9.05	-9.86	-9.32	3.31
	4	-5.03	-4.81	-4.95	-4.82	-5.10	-3.60	-3.42	-3.56	-4.53	-4.11	-6.71
	5	-4.79	-4.31	-4.54	-4.53	-5.00	-4.67	-5.76	-5.64	-6.15	-5.78	-10.30
C_6H_6	1	-10.43	-8.74	-10.42	-10.66	-10.08	-12.07	-8.87	-9.55	-10.31	-11.22	2.57
	2	-9.13	-7.25	-9.25	-9.17	-8.97	-9.61	-7.51	-8.55	-9.16	-9.51	2.17
	3	-7.83	-6.62	-7.37	-7.63	-7.59	-7.91	-6.12	-7.67	-7.72	-10.14	-1.60
	4	-6.21	-5.28	-6.09	-6.11	-5.98	-6.31	-5.06	-6.30	-5.72	-8.60	5.90
	5	-5.72	-4.58	-5.61	-5.70	-5.85	-6.17	-4.20	-5.76	-4.39	-7.19	3.67

Table A4.4.: Binding energies for the interaction of H_2, N_2, CO_2, CH_3OH, and C_6H_6 with Rh-MOP (**1b**) calculated with the listed methods on B97-3c optimized structures. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

ΔE / kcal mol^{-1}		PBE0	PBEh-3c	B97-3c	TPSS	PBE	GFN2-xTB	GFN1-xTB	GFN-FF	PM6-D3H4	PM7	UFF
H_2	1	-6.22	-6.28	-5.45	-6.05	-6.33	-7.14	-4.92	-4.83	-4.89	-9.55	-2.54
	2	-5.14	-4.96	-4.30	-4.96	-5.17	-6.02	-4.49	-3.63	-5.01	-8.64	-3.05
	3	-4.09	-3.69	-3.74	-4.06	-4.32	-4.26	-2.81	-3.29	-4.15	-5.88	-0.41
N_2	1	-8.64	-10.05	-8.05	-8.60	-8.84	-9.27	-7.34	-7.22	-4.83	-13.18	-11.18
	2	-6.79	-7.75	-6.42	-6.92	-7.16	-7.50	-5.59	-4.86	-2.67	-11.28	-10.27
	3	-6.28	-7.49	-6.03	-6.24	-6.56	-6.83	-5.16	-5.72	-4.04	-11.38	-4.49
CO_2	1	-14.73	-14.31	-13.22	-14.38	-14.46	-15.76	-12.61	-15.66	-2.86	-16.43	-23.85
	2	-12.88	-12.90	-11.46	-12.40	-12.84	-13.51	-10.52	-14.20	-3.95	-15.12	-9.47
	3	-12.19	-11.76	-10.88	-11.74	-11.95	-13.11	-10.12	-13.63	-1.37	-13.08	-9.42
	4	-10.56	-8.44	-9.23	-10.01	-10.74	-9.27	-8.38	-14.95	-1.18	-8.04	-5.13
CH_3OH	1	-19.26	-18.28	-17.77	-19.41	-20.00	-17.28	-15.36	-16.30	-9.54	-19.79	17.94
	2	-18.38	-16.05	-17.32	-18.69	-19.17	-16.90	-15.24	-14.76	-11.76	-18.29	80.98
	3	-9.69	-7.74	-9.11	-9.74	-10.14	-10.93	-9.11	-8.94	-4.92	-9.87	-14.41
C_6H_6	1	-22.09	-20.50	-20.46	-21.39	-22.08	-20.06	-16.19	-21.84	3.84	-22.83	-2.08
	2	-20.65	-19.83	-19.18	-20.55	-20.72	-18.72	-15.09	-14.77	1.20	-22.58	-8.47
	3	-19.80	-18.55	-18.49	-19.59	-20.06	-15.33	-13.36	-14.80	4.81	-19.07	9.74
	4	-18.90	-17.71	-18.07	-18.65	-19.16	-14.79	-12.68	-16.89	5.77	-17.01	-1.27

Table A4.5.: Binding energies for the interaction of H_2, N_2, CO_2, CH_3OH, and C_6H_6 with Pd-MOP$^+$ (**1c**) calculated with the listed methods on B97-3c optimized structures. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

ΔE / kcal mol^{-1}		PBE0	PBEh-3c	B97-3c	TPSS	PBE	GFN2-xTB	GFN1-xTB	GFN-FF	PM6-D3H4	PM7	UFF
H_2	1	-2.16	-1.76	-2.09	-2.06	-2.32	-11.97	-3.47	-1.68	-3.52	-3.57	-1.91
	2	-1.73	-1.72	-1.41	-1.44	-1.72	-1.64	-0.99	-1.52	-1.62	-1.04	-3.48
	3	-1.64	-1.66	-1.35	-1.64	-1.71	-1.48	-1.03	-0.96	-0.98	-0.95	-1.20
	4	-1.35	-1.16	-1.12	-1.41	-1.41	-1.18	-0.88	-0.93	-0.93	-0.87	-1.09
	5	-1.13	-1.05	-0.92	-1.08	-1.20	-1.00	-0.93	-1.02	-1.05	-1.24	-2.40
N_2	1	-4.47	-7.18	-4.77	-4.36	-4.53	-13.87	-6.69	-3.57	-3.98	-4.77	-0.51
	2	-3.24	-3.48	-3.26	-3.18	-3.37	-2.92	-2.79	-3.16	-3.12	-2.54	-4.26
	3	-3.10	-3.97	-3.14	-3.05	-3.21	-3.03	-2.94	-3.04	-2.23	-4.39	-4.47
	4	-2.68	-3.19	-2.85	-2.54	-2.77	-1.98	-2.65	-2.54	-2.03	-1.63	-5.87
CO_2	1	-7.34	-7.48	-7.28	-7.19	-7.28	-12.57	-7.04	-5.25	-4.53	-3.36	-3.96
	2	-4.89	-4.63	-4.66	-4.92	-4.94	-5.20	-3.94	-3.57	-3.20	-4.88	-4.26
	3	-4.54	-4.57	-4.60	-4.38	-4.61	-3.89	-4.25	-4.39	-2.96	-2.37	-12.48
	4	-3.65	-3.01	-3.40	-3.68	-3.56	-3.29	-2.76	-4.17	-1.07	-5.50	-10.11
CH_3OH	1	-17.47	-18.65	-16.48	-17.28	-17.45	-21.60	-17.23	-12.99	-15.00	-4.24	-4.85
	2	-7.02	-5.68	-6.61	-7.23	-7.55	-7.10	-6.85	-4.59	-7.76	-6.10	1.79
	3	-6.69	-6.19	-6.29	-6.76	-7.03	-6.84	-5.86	-5.38	-5.53	-7.90	-7.75
C_6H_6	1	-11.36	-10.26	-11.36	-11.71	-10.99	-11.39	-8.81	-11.28	-8.58	-14.26	-10.76
	2	-11.32	-10.17	-11.29	-11.63	-11.05	-11.39	-9.07	-10.88	-8.84	-14.30	-10.86
	3	-9.55	-8.11	-9.63	-9.79	-9.74	-10.14	-8.96	-12.12	-7.61	-13.50	-13.45
	4	-6.90	-4.99	-7.01	-7.21	-6.89	-6.45	-6.84	-9.84	-5.82	-11.08	-12.19

Table A4.6.: Binding energies for the interaction of H_2, N_2, CO_2, CH_3OH, and C_6H_6 with Q-POC (**2a**) calculated with the listed methods on B97-3c optimized structures. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

ΔE / kcal mol^{-1}		PBE0	PBEh-3c	B97-3c	TPSS	PBE	GFN2-xTB	GFN1-xTB	GFN-FF	PM6-D3H4	PM7	UFF
H_2	1	-3.67	-3.52	-3.47	-3.60	-3.85	-2.97	-2.80	-1.73	-2.95	-2.68	59.15
	2	-3.48	-3.25	-3.26	-3.48	-3.69	-2.72	-2.70	-1.30	-2.48	-2.27	55.37
	3	-2.97	-2.98	-2.84	-3.02	-3.18	-2.36	-2.27	-0.85	-2.18	-2.16	66.24
	4	-1.50	-1.27	-1.23	-1.42	-1.53	-1.27	-0.92	-1.45	-1.30	-1.35	-11.84
N_2	1	-3.57	-3.80	-3.69	-3.60	-3.73	-3.68	-2.62	-3.17	-3.92	-4.95	-7.45
	2	-3.52	-4.14	-3.69	-3.28	-3.68	-3.56	-3.19	-2.84	-4.25	-4.94	19.37
	3	-2.27	-2.83	-2.50	-2.13	-2.54	-2.39	-2.05	-2.37	-3.44	-4.10	-8.33
CO_2	1	-7.68	-7.93	-7.11	-7.34	-7.59	-6.63	-5.23	-5.77	-6.83	-6.97	25.63
	2	-5.95	-6.05	-5.88	-5.80	-5.88	-5.62	-3.77	-3.76	-5.31	-5.36	20.50
	3	-5.08	-4.78	-4.82	-4.78	-5.23	-4.63	-4.54	-5.17	-5.12	-5.21	21.60
	4	-5.03	-4.94	-4.86	-4.92	-5.10	-4.50	-3.90	-3.95	-4.07	-4.27	29.12
CH_3OH	1	-13.93	-13.35	-12.97	-13.65	-14.33	-10.74	-11.29	-11.44	-13.44	-12.78	30.79
	2	-12.53	-12.01	-11.56	-12.40	-12.98	-9.29	-9.88	-10.30	-12.47	-11.52	-7.78
	3	-7.10	-5.79	-6.64	-7.24	-7.45	-5.46	-5.12	-7.58	-6.61	-7.54	3.80
C_6H_6	1	-10.64	-10.46	-10.20	-10.52	-10.71	-9.16	-9.38	-8.45	-8.89	-12.72	65.80
	2	-10.29	-8.84	-10.10	-10.88	-10.20	-10.14	-8.97	-9.93	-8.89	-13.53	0.67
	3	-9.09	-8.70	-8.87	-9.36	-9.37	-7.76	-8.43	-7.29	-7.79	-11.33	49.40
	4	-8.39	-7.73	-8.04	-8.36	-8.50	-6.56	-7.03	-7.36	-7.27	-10.93	12.51

Table A4.7.: Binding energies for the interaction of H_2, N_2, CO_2, CH_3OH, and C_6H_6 with B-POC (**2b**) calculated with the listed methods on B97-3c optimized structures. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

ΔE / kcal mol^{-1}		PBE0	PBEh-3c	B97-3c	TPSS	PBE	GFN2-xTB	GFN1-xTB	GFN-FF	PM6-D3H4	PM7	UFF
H_2	1	-2.26	-2.18	-1.99	-2.07	-2.25	-1.90	-1.38	-1.55	-1.68	-1.61	-4.81
	2	-1.94	-1.89	-1.64	-1.80	-1.96	-1.64	-1.19	-1.23	-1.40	-1.34	-3.60
	3	-1.85	-1.58	-1.45	-1.57	-1.71	-1.34	-0.87	-1.14	-0.99	-0.80	-4.88
	4	-1.72	-1.63	-1.42	-1.57	-1.65	-1.27	-1.01	-1.11	-1.08	-1.06	-4.39
	5	-0.69	-0.63	-0.48	-0.62	-0.55	-0.36	-0.37	-0.20	-0.09	-0.36	-1.78
N_2	1	-4.15	-4.49	-4.08	-3.81	-3.94	-3.71	-2.80	-2.75	-3.58	-4.39	-6.34
	2	-2.93	-3.42	-2.94	-2.55	-2.75	-2.69	-2.09	-2.09	-2.42	-3.16	-5.10
	3	-2.76	-3.50	-2.86	-2.65	-2.86	-2.82	-2.24	-2.13	-2.21	-2.35	-3.53
	4	-2.66	-3.51	-2.72	-2.32	-2.50	-2.44	-2.05	-1.94	-2.25	-2.82	-7.99
	5	-2.15	-2.49	-2.26	-1.77	-1.92	-2.05	-1.78	-1.93	-1.71	-2.83	-6.05
CO_2	1	-5.72	-5.67	-5.43	-5.38	-5.41	-5.21	-3.23	-3.88	-4.42	-4.96	-10.07
	2	-5.03	-4.85	-4.68	-4.91	-4.96	-4.42	-2.93	-3.18	-3.05	-3.99	-5.99
	3	-4.32	-4.48	-4.03	-4.07	-4.20	-4.07	-2.44	-3.03	-2.57	-3.18	-5.75
	4	-3.78	-3.83	-3.50	-3.32	-3.61	-3.29	-2.87	-3.94	-3.01	-3.62	-10.93
	5	-3.72	-4.03	-3.58	-3.61	-3.69	-3.40	-2.07	-2.07	-2.21	-3.21	-4.44
CH_3OH	1	-7.30	-5.99	-6.54	-7.19	-7.67	-6.91	-6.78	-7.61	-7.77	-6.70	1.15
	2	-7.13	-6.30	-6.74	-6.87	-7.18	-5.25	-4.43	-5.72	-5.57	-6.20	-7.68
	3	-6.01	-4.40	-5.44	-5.77	-6.18	-5.93	-5.26	-6.21	-6.64	-5.10	-0.22
	4	-5.55	-5.01	-5.07	-5.34	-5.67	-4.88	-3.59	-5.28	-4.06	-4.43	-3.24
C_6H_6	1	-8.46	-6.38	-8.00	-8.69	-8.30	-8.67	-7.54	-8.26	-7.20	-11.88	-11.87
	2	-8.14	-7.24	-7.86	-8.65	-8.03	-7.74	-6.58	-7.40	-6.71	-10.06	-10.51
	3	-7.64	-6.49	-7.42	-7.61	-7.44	-6.77	-5.79	-7.22	-6.37	-9.60	-7.96
	4	-5.93	-5.42	-5.58	-5.88	-5.99	-5.58	-4.91	-5.54	-4.89	-8.00	-5.24

Table A4.8.: Binding energies for the interaction of H_2, N_2, CO_2, CH_3OH, and C_6H_6 with F-POC (**2c**) calculated with the listed methods on B97-3c optimized structures. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

ΔE / kcal mol^{-1}		PBE0	PBEh-3c	B97-3c	TPSS	PBE	GFN2-xTB	GFN1-xTB	GFN-FF	PM6-D3H4	PM7	UFF
H_2	1	-0.66	-0.19	-0.88	-0.84	-0.85	-1.56	-2.11	-3.98	-5.61	-6.06	-4.47
	2	-0.56	-0.20	-0.52	-0.79	-0.83	-1.13	-1.65	-3.53	-4.30	-4.73	-5.93
	3	0.07	0.56	-0.04	-0.10	-0.09	-0.89	-1.23	-3.12	-4.20	-4.17	-1.86
N_2	1	-4.00	-3.85	-3.95	-3.99	-4.06	-3.72	-3.48	-3.84	-4.82	-5.75	-5.44
	2	-3.88	-3.72	-3.77	-3.88	-4.00	-3.42	-3.32	-3.45	-4.57	-5.83	-4.23
	3	-3.14	-2.54	-3.42	-3.45	-3.37	-4.05	-4.14	-5.94	-9.26	-10.86	-0.20
CO_2	1	-6.99	-6.40	-6.45	-6.99	-6.92	-6.05	-4.70	-6.14	-6.93	-7.88	-5.89
	2	-6.48	-5.87	-5.76	-6.35	-6.55	-5.90	-4.70	-6.14	-6.65	-7.09	-5.68
	3	-6.14	-5.39	-6.07	-6.41	-6.30	-6.59	-5.56	-7.99	-12.76	-13.46	-2.02
CH_3OH	1	-14.38	-14.21	-13.63	-14.27	-14.78	-10.62	-11.41	-12.71	-14.02	-14.40	17.36
	2	-12.01	-11.78	-10.79	-11.85	-12.73	-7.75	-8.29	-11.51	-11.61	-12.14	16.68
	3	-9.73	-8.54	-9.63	-9.95	-10.27	-9.56	-10.29	-10.29	-20.36	-21.11	21.20
	4	-6.77	-5.79	-6.56	-6.82	-6.99	-5.23	-6.11	-8.83	-6.71	-7.71	-11.11
C_6H_6	1	-14.47	-12.35	-14.73	-15.01	-14.42	-13.86	-16.08	-19.88	-18.35	-26.00	-20.67
	2	-14.19	-12.28	-14.47	-14.67	-13.98	-12.75	-14.48	-17.15	-13.32	-21.26	-26.26
	3	-12.18	-8.90	-13.00	-13.16	-12.24	-12.42	-15.39	-20.93	-20.88	-30.52	-17.06

Table A4.9.: Binding energies for B-POC (2b) and H_2, N_2, CO_2, CH_3OH, and C_6H_6 calculated with the listed methods based on GFN2-xTB and GFN-FF optimized structures. For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

ΔE / kcal mol^{-1}		PBE0// GFN2-xTB	B97-3c-3c// GFN2-xTB	GFN2-xTB// GFN2-xTB	GFN1-xTB// GFN2-xTB	GFN-FF// GFN2-xTB	PBE0// GFN-FF	B97-3c-3c// GFN-FF	GFN2-xTB// GFN-FF	GFN1-xTB// GFN-FF	GFN-FF// GFN-FF
H_2	1	-1.60	-1.22	-1.35	-0.71	-1.15	-0.86	-0.46	-0.31	-0.04	-1.25
	2	-1.27	-0.76	-1.12	-0.39	-0.71	-1.41	-1.07	-1.09	-0.77	-0.88
	3	-0.73	-0.50	-1.18	-0.37	-0.71	-0.30	0.23	0.68	0.87	-0.77
	4	0.28	0.41	0.33	0.34	0.33	-0.51	-0.48	-0.47	-0.51	0.16
	5	-0.87	-0.68	-1.22	-0.46	-0.81	-1.05	-0.57	-0.27	-0.14	-1.05
N_2	1	-2.26	-2.17	-3.94	-2.19	-2.16	-4.19	-4.25	-4.40	-3.31	-2.77
	2	-0.50	-0.26	-2.22	-1.85	-2.33	-0.43	-0.07	-1.85	-2.22	-2.62
	3	2.15	1.88	-3.70	-1.03	-0.00	-2.37	-2.47	-3.12	-2.49	-1.94
	4	1.23	1.24	-3.27	-1.11	-0.72	-2.83	-3.00	-3.48	-2.59	-2.05
	5	0.33	0.14	-2.69	-1.29	-1.09	-2.68	-2.93	-2.98	-2.47	-1.87
CO_2	1	-3.23	-2.48	-3.21	-2.62	-4.67	-1.92	-0.58	-3.33	-2.28	-5.58
	2	-1.67	-1.51	-3.64	-2.61	-3.68	-2.13	-1.97	-4.41	-3.35	-3.96
	3	-1.49	-1.37	-3.63	-2.45	-3.53	-1.85	-1.68	-4.30	-3.12	-3.91
	4	-2.43	-2.28	-2.99	-1.19	-1.40	-1.38	-1.19	-3.80	-2.72	-3.68
	5	-4.41	-4.03	-4.86	-2.27	-3.07	-4.65	-4.35	-5.05	-2.82	-3.58
CH_3OH	1	-6.39	-5.39	-6.80	-6.36	-8.42	-5.54	-4.44	-6.89	-6.85	-9.02
	2	-5.14	-4.46	-5.70	-4.83	-6.51	-3.89	-3.40	-5.04	-4.46	-7.82
	3	-6.43	-6.15	-4.84	-4.03	-5.23	-7.31	-6.92	-5.41	-4.68	-5.80
	4	-5.07	-4.58	-4.76	-3.07	-4.81	-5.69	-5.33	-5.08	-4.05	-5.17
C_6H_6	1	-7.33	-6.78	-8.72	-7.21	-8.35	-7.32	-6.93	-9.11	-7.80	-8.87
	2	-6.79	-6.56	-6.34	-5.10	-7.45	-6.37	-6.31	-6.55	-5.39	-8.04
	3	-6.52	-6.40	-7.67	-6.09	-7.00	-7.43	-7.30	-8.13	-6.84	-7.35
	4	-5.21	-4.47	-5.55	-5.07	-5.97	-4.42	-3.58	-5.49	-4.71	-6.56

Table A4.10.: Binding energy ΔE for RAPTA-C in Q-POC (**2a**) and B-POC (**2b**) for all methods listed as well as the mean MAD in comparison to PBE0-D4/def2-TZVP (reference). For all density functionals, beside the composite methods, a def2-TZVP basis set is employed.

ΔE / kcal mol^{-1}	RAPTA-C@**2a**	RAPTA-C@**2b**	MAD_{mean}
PBE0-D4	-37.57	-17.11	–
PBEh-3c	-39.30	-13.87	2.48
B97-3c	-36.10	-16.89	0.84
TPSS-D4	-36.32	-17.62	0.88
PBE-D4	-36.72	-16.68	0.64
PM6-D3H4	-19.18	-3.01	16.2
PM7	-48.71	-25.02	9.53
GFN2-xTB	-31.15	-17.14	3.22
GFN1-xTB	-34.13	-18.02	2.18
GFN-FF	-35.62	-21.95	3.40
UFF	16.55	-25.49	31.3

A5. Supporting Information to Chapter 5

Appendix A5 contains:

- General remarks
- Computational details
- Statistical error measures
- Computational results

General remarks

The *xtb*[337] program package is available free of charge for academic use. Collected atomic Cartesian coordinates (XYZ format) of all input structures are available in addition to the Supporting Information: Spicher, S.; Abdullin, D.; Grimme, S.; Schiemann, O "Modeling of spin-spin distance distributions for nitroxide labeled biomacromolecules", *Phys. Chem. Chem. Phys.* **2020**, *22*, 24282–24290.

Computational details

Input structures were generated with *mtsslWizard*.[266] Structures were converted from pdb to xyz format with the *maestro*[399] software package. All CREST/MD computations were performed with the *crest*[60] and *xtb*[337] stand alone programs. The default convergence criteria (10^{-7} E_h for energies and 10^{-5} E_h/bohr for gradients) were used throughout. MMM calculations were performed with version *2020.2* employing the R1A-298K-UFF-216-r1-CASD rotamer library for the R1 side chain. MtsslWizard calculations were performed with the server version[656] with clashes settings tight. For the azurin mutants, the distribution of the Cu(II) spin density was taken into account with each respective method. Trajectory evaluation was performed with the program *travis*.[397] Structure visualization was done in *pymol*.[657] MD simulations with GFN-FF were carried out for 1 ns at the respective freezing temperature of the solvents, employing the implicit GBSA(H_2O) solvation model. A time step of 2 fs at an increased hydrogen mass of 4 amu and equilibration phase of 200 ps was chosen.
For the mutants of azurin (1952 atoms), 100 ps of the MD simulation took on average 7.3 hours on 4 Intel© Xeon E5-2660 v4 @ 2.00 GHz CPUs. On the same machine, the 100 ps took on average 16.4 hours for the mutants of T4L (2683 atoms), again on 4 CPUs.

Statistical error measures

In this work, the following statistical measures were used. p is an arbitrary property.

$$\delta p = p_{calc.} - p_{ref.} \tag{A5.1}$$

The error measures are defined by:

- Mean absolute deviation (MAD):

$$MAD = \frac{1}{N}\sum_{i}^{N} |\delta p_i| \tag{A5.2}$$

- Standard deviation (SD):

$$SD = \sqrt{\frac{\sum_i^N |\delta p_i - MD|^2}{N-1}} \tag{A5.3}$$

- relative MAD (relMAD):

$$relMAD = \frac{1}{N}\sum_{i}^{N} \frac{|\delta p_i|}{p_{ref}}. \tag{A5.4}$$

For the calculation of the distance distributions fur azurin, the distribution of spin densities is taken into account. ρ_{Cu} = 0.35, ρ_S = 0.60 and for nitroxide, the center of the N-O bond is taken (ρ_N = ρ_O = 0.50). The spin density weighted distances are calculated according to:

$$\frac{1}{r^3} = \frac{1}{\rho_{Cu}+\rho_S}\left(\rho_{Cu}\frac{1}{r^3_{NO-Cu}} + \rho_S \frac{1}{r^3_{NO-S}}\right), \tag{A5.5}$$

$$r = \left(\frac{0.95}{\frac{0.35}{r^3_{NO-Cu}} + \frac{0.60}{r^3_{NO-S}}}\right)^{\frac{1}{3}}. \tag{A5.6}$$

The most probable distance r_p is the distance with the highest intensity/probability within the distance distribution. The mean distance $\bar{r}$ is calculated as follows,

$$\bar{r} = \frac{\sum_i r\, P(r_i)}{\sum_r P(r_i)} \tag{A5.7}$$

where $I(r)$ is the respective probability associated with each distance.

Computational results

Molecular dynamics settings

MD simulations with GFN-FF were carried out for 1 ns at the freezing point of the respective solvent employing the implicit GBSA(H_2O) solvation model. A time step of 2 fs (at an increased hydrogen mass of 4 amu) and equilibration phase of 200 ps was chosen. In the following, the effects of different MD simulation lengths and temperatures are investigated (see Figure A5.1A). Regarding the simulation

time, the limiting factor is the computation time. GFN-FF is a physically motivated partially polarizable force-field with many sophisticated energy terms that lead to a scaling that is roughly a factor of 10 slower than specialized protein FFs. For 1 ns an overall wall time of four to seven days (dependent on the system size) is reached on 4 Intel© Xeon E5-2660 v4 @ 2.00 GHz CPUs.

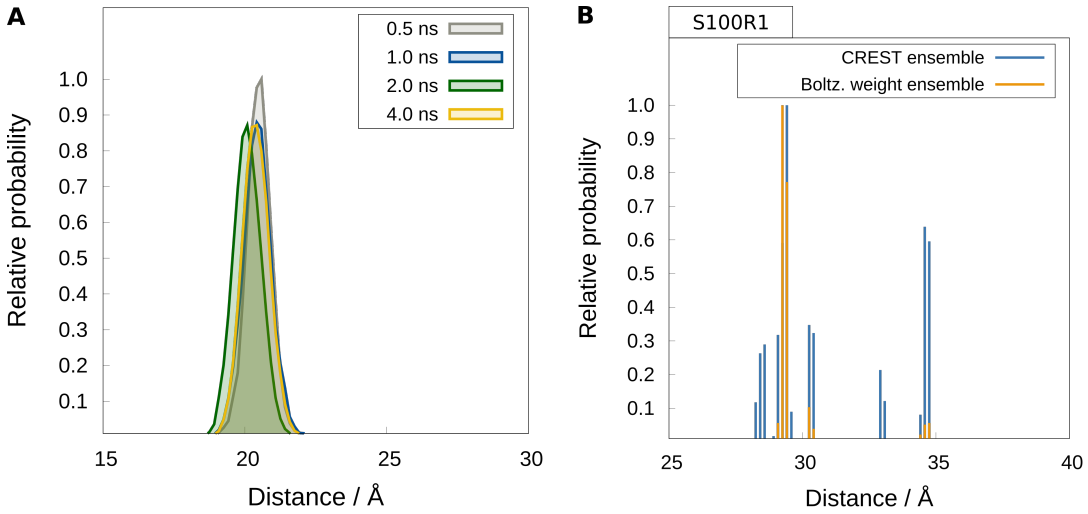


Figure A5.1.: (A) Effect of different MD settings on the radial distribution function of azurin mutant T21R1. For longer MD simulation times the intensities decrease while the distribution width becomes slightly larger. The mean distances remain roughly the same. (B) Conformer ensemble for azurin mutant S100R1. In blue, the sum of all distances obtained from the entire CREST ensemble. In yellow, the Boltzmann weighted sum of all distances.

CREST conformations

From the constrained CREST calculations for the R1 side chain, conformer ensembles with sizes from 20 (for azurin mutants) to 200 (for T4L mutants) were obtained. For each conformer, the distance between the respective spin center was calculated. The distance distribution was obtained as the sum of all the distances from the conformers as illustrated for S100R1 in Figure A5.1B. The Boltzmann population of each conformer determines the intensity. Relative probabilities were calculated by dividing the intensity of each conformer by the maximum value of the energetically lowest conformer. If multiple conformers existed that showed a very similar distance between the spin center (<1 Å), they were considered as one R1 conformer cluster. For S100R1, 34 conformers existed, which lead to three R1 conformer clusters as shown in Figure A5.1B. The division into conformer clusters was done manually. Since only one R1 conformer cluster had a relative probability of more than 0.5, one GFN-FF MD simulation was carried out for the conformer with the highest Boltzmann population within the respective R1 conformer cluster. In blue, the conformer ensemble without Boltzmann weighting is shown. Here, a critical aspect of the applied procedure should be addressed. All conformational energies were obtained at the GFN-FF level of theory for a constrained system. Even though this was tested in Ref. 117 for small to medium sized systems, the extrapolation to large biomolecules can not be assumed.

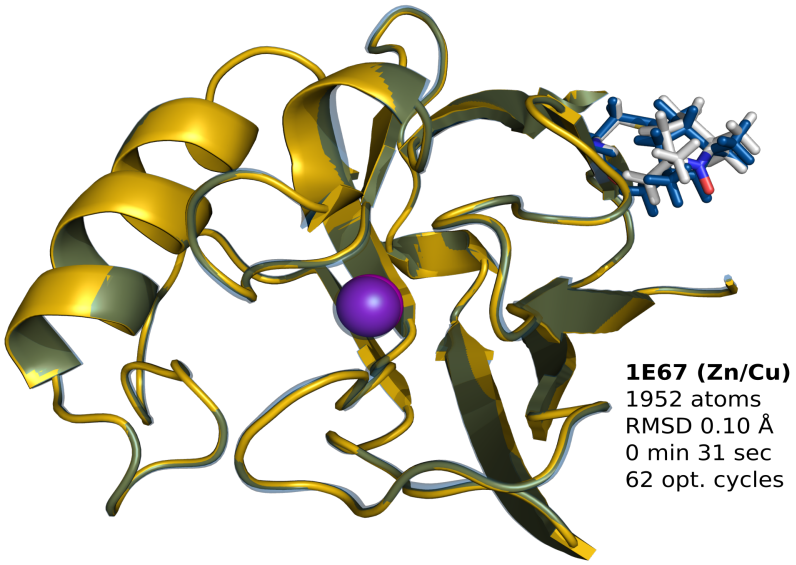


Figure A5.2.: RMSD minimized structure overlay between the optimized GFN-FF geometries for zinc-azurin (yellow) and the copper-azurin analogue (transparent blue). The CSD identifier are given as well as heavy atom RMSD values, total computation wall-times, and the required number of geometry optimization cycles.

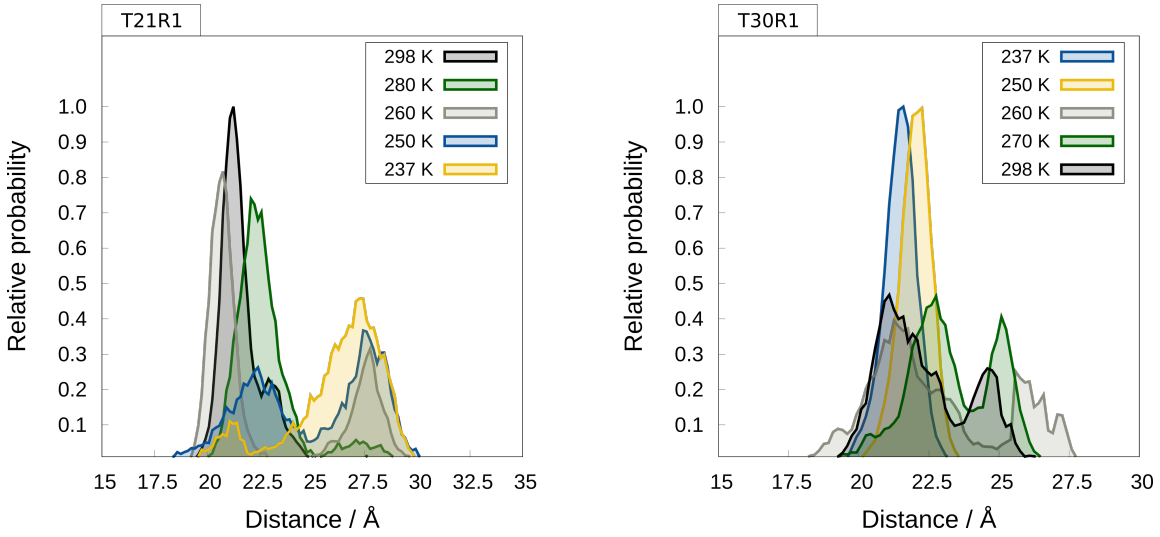


Figure A5.3.: Effect of different MD temperatures on the radial distribution function of azurin mutant T21R1 and T30R1. For T21R1 the starting conformation is taken from the MTSL conformation of larger distance. For T30R1 the shorter conformation was the starting point of the MD. More information is in the manuscript.

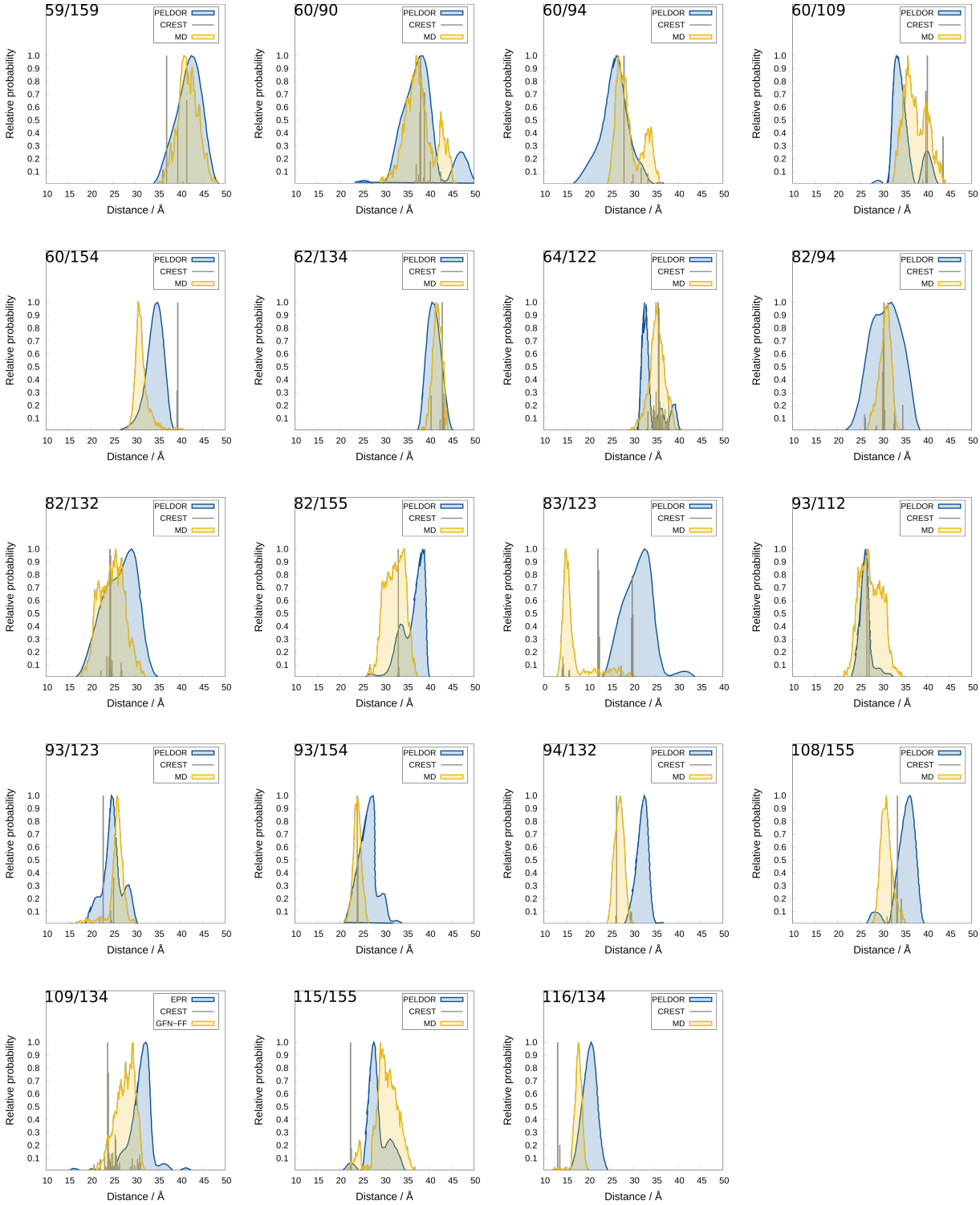


Figure A5.4.: Spin label distance distribution functions for the 19 T4L mutants. The results from the CREST conformations without any MD averaging are shown in gray, the distribution obtained by the GFN-FF MD simulation in yellow and experimental EPR data are shown in blue.

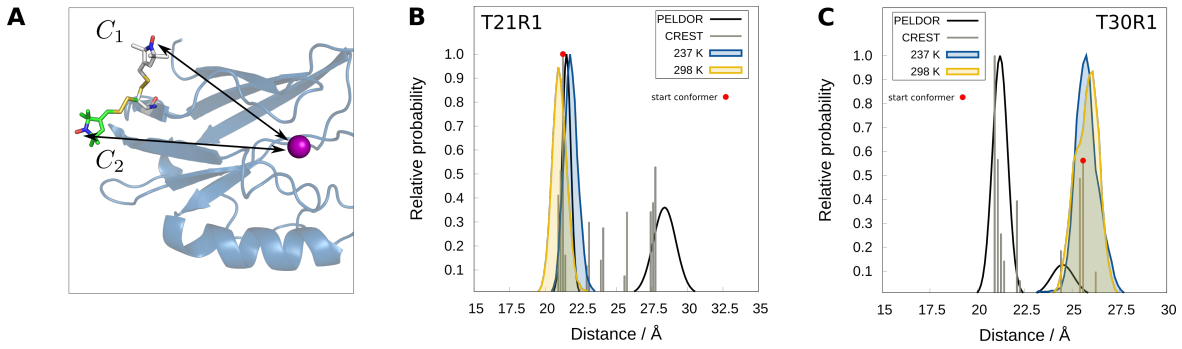


Figure A5.5.: Investigation of temperature effects in the MD simulation. In **A**, two conformations with shorter (C_1) and longer (C_2) inter-spin distances of T21R1 are illustrated. Computations are performed at the freezing point of the solvent (237 K) and at room temperature (298 K), respectively, for the azurin mutants T21R1 (**B**) and T30R1 (**C**). The starting conformation for the MD simulation is indicated by a red dot.

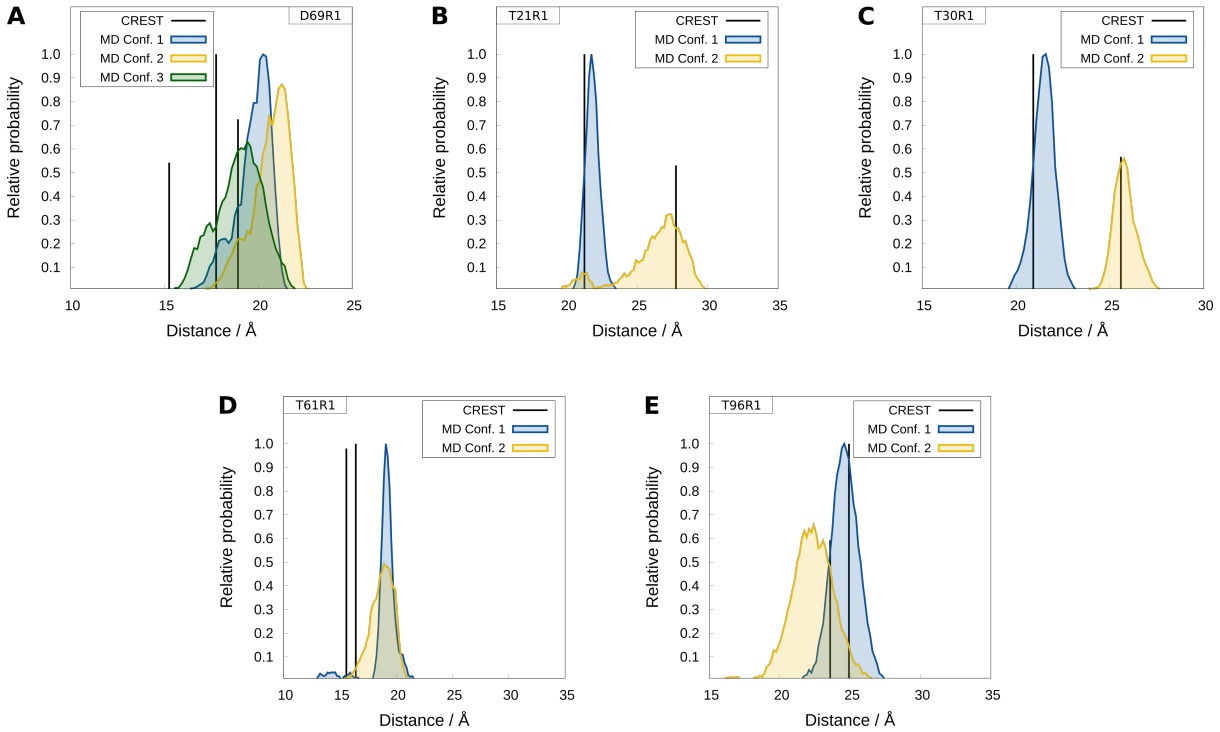


Figure A5.6.: Contributions of the individual conformers to the radial distribution. For D69R1 MD simulations are performed for three conformers. T21R1, T30R1, and T62R1 show two conformations with a relative probability larger than 0.5. The MTSL conformers are named according to their relative probability, e.g Conf. 1 for the highest probability. Only conformers with a relative probability larger than 0.5 are shown.

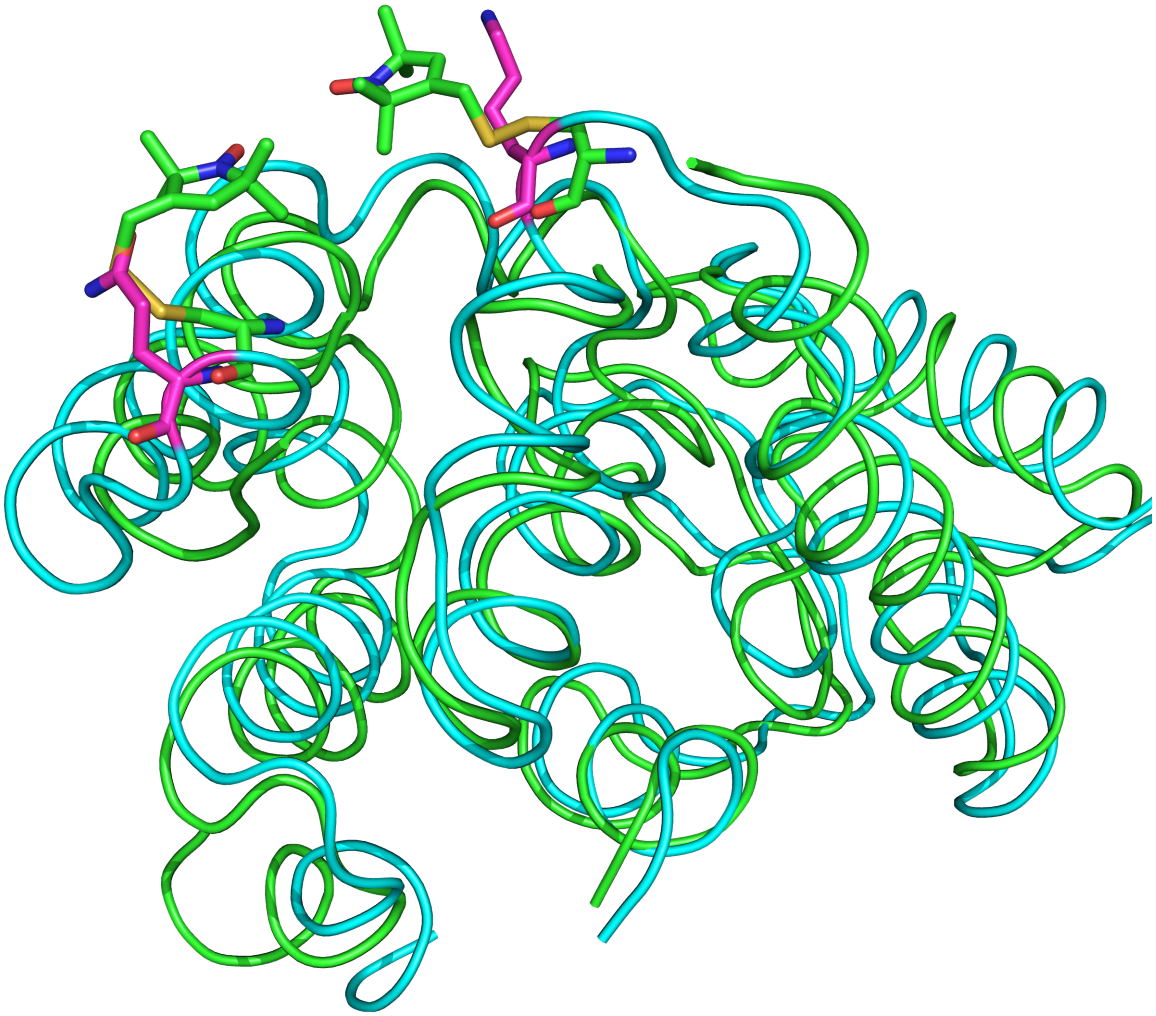


Figure A5.7.: Overlay of an MD snapshots (green) with the initial X-ray structure (blue) for mutant 83/123 revealing non-covalent interactions between the two spin-labels. Even though the tertiary structure is similar on a global view, significant differences in the positions of individual loops and helices are visible. These conformational changes cause the short mean distance of only 5.8 Å.

A6. Supporting Information to Chapter 6

Appendix A5 contains:

- General remarks
- Statistical measures
- Detailed results

General remarks

The *xtb*[515] and *crest*[60] program packages are available free of charge for academic use. Collected atomic Cartesian coordinates (XYZ format) of all benchmark structures are available as additional supplementary material (geometries.zip):
Spicher, S.; Caldeweyher, E.; Hansen, A.; Grimme, S. "Benchmarking London dispersion corrected density functional theory for noncovalent ion–π interactions", *Phys. Chem. Chem. Phys.* **2021**, *23*, 11635–11648.

The nomenclature of the benchmark set is the following. For systems **1**–**17** the complex is termed AB and A/B are the respective fragments of which B is the ion. For systems **18** and **19**, A and B are the respective conformers. The CHRG files contain the molecular charge.

Statistical measures

For statistical analysis of a set $\{x_1, \ldots, x_n\}$ of data points with references $\{r_1, \ldots, r_n\}$ we use the following measures:

- Average : $\overline{x} = \frac{1}{n} \sum_i x_i$
- Mean deviation (MD): $MD = \frac{1}{n} \sum_i (\mathrm{x}_i - \mathrm{r}_i)$
- Mean absolute deviation (MAD): $MAD = \frac{1}{n} \sum_i |\mathrm{x}_i - \mathrm{r}_i|$
- Standard deviation (SD) : $\mathrm{SD} = \sqrt{\frac{\sum_i (x_i - \overline{x})^2}{n}}$
- relative MAD (relMAD): $relMAD = \frac{1}{n} \sum_i \frac{|\mathrm{x}_i - \mathrm{r}_i|}{\mathrm{r}_i}$

Detailed results

Results of the tested methods

For all DFAs and HF, a large def2-QZVPP basis set was employed. The composite 3c methods have their own adjusted basis set. For MP2, a CBS extrapolation was performed. For all systems including alkali metals (**1**–**6**), a counterpoise (CP) correction was additionally applied.

Table A6.1.: Interaction energies in kcal mol^{-1} for the IONPI19 benchmark calculated with the listed WFT methods and different LD corrections.

System	Ref.	HF			MP2/CBS
		D3	D4	NL	–
1	-39.09	-42.30	-37.36	-41.64	-39.36
2	-25.63	-27.00	-27.00	-26.99	-25.71
3	-19.90	-14.78	-16.85	-19.79	-20.49
4	-14.81	-14.85	-14.51	-13.35	-15.31
5	-25.65	-34.19	-23.13	-25.94	-25.11
6	-19.74	-25.18	-18.62	-19.67	-19.84
7	-21.51	-22.40	-20.21	-19.36	-24.92
8	-14.57	-13.82	-15.56	-14.60	-15.08
9	-10.41	-10.23	-12.72	-11.04	-10.89
10	-1.93	0.33	-2.40	-0.83	-2.81
11	-5.70	-4.21	-6.86	-5.36	-6.48
12	-18.56	-17.38	-19.18	-17.61	-19.07
13	-33.66	-32.41	-34.69	-32.91	-35.59
14	-45.03	-44.38	-46.92	-45.45	-48.49
15	-29.43	-28.84	-30.62	-30.33	-32.54
16	-26.27	-25.25	-27.80	-25.72	-29.83
17	-37.17	-39.57	-40.86	-43.73	-52.40
18	-5.01	-5.66	-4.89	-5.59	-6.22
19	-2.42	-2.87	-2.81	-2.86	-2.66

Table A6.2.: Interaction energies in kcal mol^{-1} for the IONPI19 benchmark calculated with the listed double hybrid functionals and different LD corrections.

System	Ref.	PWPB95			revDSD-BLYP	revDSD-PBEP86	B2PLYP		
		D3	D4	NL	D4	D4	D3	D4	NL
1	-39.09	-41.86	-40.56	-41.49	-38.39	-38.53	-40.30	-39.09	-39.96
2	-25.63	-27.47	-26.31	-26.78	-25.37	-25.15	-26.33	-25.63	-25.80
3	-19.90	-20.34	-18.52	-19.66	-17.84	-18.14	-19.15	-18.18	-19.06
4	-14.81	-16.80	-15.75	-16.03	-14.95	-14.75	-15.83	-15.15	-15.11
5	-25.65	-28.67	-27.14	-27.82	-24.97	-25.03	-27.58	-25.74	-26.32
6	-19.74	-21.66	-20.00	-20.67	-19.23	-19.25	-21.24	-19.76	-20.10
7	-21.51	-22.55	-22.22	-22.18	-21.41	-21.49	-22.64	-21.99	-21.69
8	-14.57	-14.30	-14.51	-14.58	-14.86	-14.49	-14.39	-14.65	-14.70
9	-10.41	-10.61	-11.02	-10.87	-11.29	-11.00	-10.77	-11.18	-11.05
10	-1.93	-1.74	-2.14	-1.99	-2.57	-2.36	-2.04	-2.55	-2.36
11	-5.70	-5.55	-5.94	-5.85	-6.37	-6.09	-5.88	-6.37	-6.25
12	-18.56	-18.00	-18.47	-18.44	-18.8	-18.36	-17.88	-18.17	-18.35
13	-33.66	-33.05	-33.74	-33.85	-34.39	-33.72	-33.03	-33.33	-33.84
14	-45.03	-45.29	-45.86	-46.41	-46.08	-45.21	-44.78	-45.10	-45.88
15	-29.43	-28.41	-29.39	-29.70	-30.58	-29.99	-30.08	-30.92	-31.56
16	-26.27	-25.90	-26.28	-26.27	-27.3	-26.71	-27.11	-27.66	-27.42
17	-37.17	-40.62	-37.69	-40.31	-37.19	-37.09	-42.53	-42.25	-42.06
18	-5.01	-5.49	-5.12	-5.48	-4.58	-4.75	-5.28	-5.00	-5.16
19	-2.42	-2.48	-2.45	-2.49	-2.31	-2.36	-2.31	-2.31	-2.30

Table A6.3.: Interaction energies in kcal mol^{-1} for the IONPI19 benchmark calculated with the listed range-separated hybrid functionals and different LD corrections as well as two Minnesota-type hybrid functionals.

System	Ref.	ωB97M			ωB97X			M06-2x	MN15
		D3	D4	NL	D3	D4	NL	–	–
1	-39.09	-40.71	-39.52	-41.33	-37.22	-37.30	-41.21	-42.20	-39.88
2	-25.63	-25.73	-25.57	-26.72	-23.89	-24.42	-26.55	-27.23	-25.53
3	-19.90	-17.51	-17.34	-19.78	-17.71	-18.63	-19.58	-20.30	-19.20
4	-14.81	-14.66	-14.56	-15.38	-13.37	-13.28	-15.06	-15.87	-14.71
5	-25.65	-28.11	-25.89	-27.24	-23.91	-23.47	-26.75	-28.06	-26.23
6	-19.74	-20.83	-19.43	-20.84	-18.62	-18.47	-20.46	-21.04	-19.50
7	-21.51	-22.38	-21.23	-21.62	-20.46	-19.79	-20.65	-22.02	-21.85
8	-14.57	-13.76	-14.24	-14.79	-13.34	-14.29	-14.34	-15.07	-15.38
9	-10.41	-10.55	-11.35	-11.55	-10.24	-11.30	-11.11	-11.45	-11.75
10	-1.93	-1.42	-2.33	-2.39	-1.37	-2.34	-2.02	-1.96	-2.64
11	-5.70	-5.34	-6.21	-6.46	-5.13	-6.19	-6.02	-6.15	-6.61
12	-18.56	-17.13	-17.66	-18.40	-16.30	-17.78	-17.95	-19.07	-18.77
13	-33.66	-31.70	-32.30	-34.00	-30.32	-33.09	-33.22	-35.09	-34.52
14	-45.03	-43.16	-43.78	-46.47	-41.69	-44.88	-45.35	-48.77	-47.99
15	-29.43	-28.87	-30.07	-31.01	-29.08	-30.31	-30.34	-29.38	-29.22
16	-26.27	-25.69	-26.61	-27.14	-25.16	-26.48	-26.24	-26.40	-26.47
17	-37.17	-40.72	-40.10	-42.27	-41.18	-37.05	-40.49	-35.96	-38.89
18	-5.01	-5.70	-5.19	-5.67	-5.64	-5.11	-5.48	-5.53	-5.88
19	-2.42	-2.59	-2.57	-2.62	-2.52	-2.38	-2.63	-2.47	-2.60

Table A6.4.: Interaction energies in kcal mol^{-1} for the IONPI19 benchmark calculated with the listed hybrid functionals and different LD corrections.

System	Ref.	B3LYP			PW6B95			PBE0		
		D3	D4	NL	D3	D4	NL	D3	D4	NL
1	-39.09	-40.45	-38.18	-40.25	-42.53	-41.38	-42.34	-42.07	-40.97	-42.14
2	-25.63	-26.57	-25.98	-26.17	-27.86	-27.34	-27.54	-27.07	-26.85	-27.21
3	-19.90	-18.47	-17.55	-19.24	-20.06	-18.84	-20.00	-19.34	-18.95	-20.25
4	-14.81	-16.18	-15.67	-15.35	-17.03	-16.62	-16.58	-16.31	-16.15	-16.25
5	-25.65	-28.97	-25.40	-26.79	-29.61	-28.14	-28.79	-29.16	-27.40	-28.22
6	-19.74	-22.19	-19.88	-20.46	-21.91	-20.69	-21.16	-21.88	-20.80	-21.3
7	-21.51	-23.55	-22.34	-21.80	-22.61	-22.36	-22.18	-23.48	-22.84	-23.02
8	-14.57	-14.03	-14.75	-14.78	-14.22	-14.65	-14.70	-13.52	-13.99	-14.37
9	-10.41	-10.59	-11.63	-11.34	-10.60	-11.19	-11.05	-10.18	-10.81	-11.01
10	-1.93	-1.67	-2.85	-2.40	-1.62	-2.18	-2.01	-1.22	-1.93	-1.95
11	-5.70	-5.56	-6.72	-6.42	-5.44	-6.01	-5.92	-5.04	-5.74	-5.93
12	-18.56	-17.00	-17.90	-18.22	-17.65	-18.44	-18.41	-17.09	-17.66	-18.36
13	-33.66	-31.48	-32.60	-33.54	-32.40	-33.61	-33.76	-31.37	-32.15	-33.64
14	-45.03	-42.78	-44.08	-45.59	-44.21	-45.50	-46.12	-42.14	-43.13	-45.33
15	-29.43	-29.09	-30.65	-31.52	-27.79	-28.82	-29.38	-28.27	-29.37	-30.53
16	-26.27	-26.57	-27.83	-27.06	-25.34	-25.91	-25.89	-25.74	-26.63	-26.86
17	-37.17	-42.48	-41.96	-39.58	-38.81	-35.86	-37.74	-38.92	-38.66	-38.44
18	-5.01	-5.33	-4.80	-5.07	-5.32	-4.90	-5.28	-5.35	-5.03	-5.35
19	-2.42	-2.23	-2.23	-2.23	-2.36	-2.35	-2.38	-2.41	-2.42	-2.45

Table A6.5.: Interaction energies in kcal mol^{-1} for the IONPI19 benchmark calculated with the listed meta GGA functionals and different LD corrections.

System	Ref.	B97M			r^2SCAN			TPSS			M06-L
		D3	D4	NL	D3	D4	NL	D3	D4	NL	–
1	-39.09	-39.63	-35.58	-37.42	-38.98	-38.65	-39.05	-40.50	-38.82	-40.69	-35.37
2	-25.63	-27.03	-24.02	-24.30	-26.15	-26.04	-26.35	-25.93	-25.77	-26.22	-22.63
3	-19.90	-19.35	-16.57	-18.19	-18.89	-18.72	-19.40	-18.08	-17.64	-19.49	-16.78
4	-14.81	-17.36	-14.52	-14.35	-15-98	-15.88	-16.19	-16.21	-16.17	-16.20	-13.58
5	-25.65	-29.09	-24.02	-25.26	-26.20	-25.63	-26.00	-29.03	-26.45	-27.75	-24.10
6	-19.74	-23.54	-18.93	-19.64	-21.11	-20.72	-21.15	-21.60	-20.14	-20.91	-18.43
7	-21.51	-22.73	-22.14	-21.57	-23.49	-23.03	-23.08	-24.81	-24.27	-24.38	-21.26
8	-14.57	-13.41	-14.14	-14.02	-13.51	-13.61	-13.92	-13.33	-14.16	-14.54	-12.52
9	-10.41	-10.91	-11.99	-11.54	-10.77	-10.91	-11.13	-10.27	-11.38	-11.46	-10.22
10	-1.93	-1.95	-3.15	-2.67	-2.02	-2.20	-2.32	-1.60	-2.77	-2.63	-1.77
11	-5.70	-5.74	-6.89	-6.51	-5.80	-5.97	-6.17	-5.28	-6.46	-6.54	-5.29
12	-18.56	-17.31	-18.06	-17.88	-17.57	-17.70	-18.15	-16.87	-17.96	-18.72	-16.19
13	-33.66	-31.95	-32.93	-32.99	-32.27	-32.41	-33.33	-30.77	-32.33	-34.00	-30.35
14	-45.03	-43.35	-44.43	-44.94	-43.54	-43.74	-45.03	-41.11	-43.09	-45.66	-43.61
15	-29.43	-28.89	-30.85	-31.25	-30.16	-30.57	-31.01	-27.63	-29.35	-30.71	-29.66
16	-26.27	-26.09	-27.78	-27.11	-26.28	-26.46	-26.58	-25.57	-27.14	-27.18	-25.70
17	-37.17	-37.60	-38.95	-39.27	-38.85	-37.97	-36.52	-39.64	-39.96	-39.50	-34.67
18	-5.01	-4.98	-4.73	-5.07	-5.23	-5.04	-5.01	-5.17	-4.82	-5.28	-5.17
19	-2.42	-1.96	-1.97	-1.99	-2.08	-2.07	-2.06	-2.05	-2.08	-2.14	-1.86

Table A6.6.: Interaction energies in kcal mol^{-1} for the IONPI19 benchmark calculated with the listed GGA functional and different LD corrections.

System	Ref.	PBE		
		D3	D4	NL
1	-39.09	-40.43	-39.25	-40.60
2	-25.63	-26.06	-26.06	-26.39
3	-19.90	-18.32	-18.20	-19.58
4	-14.81	-16.38	-16.47	-16.51
5	-25.65	-28.71	-26.83	-27.77
6	-19.74	-21.62	-20.64	-21.24
7	-21.51	-24.54	-24.08	-24.23
8	-14.57	-13.44	-14.09	-14.51
9	-10.41	-10.57	-11.42	-11.61
10	-1.93	-2.08	-3.00	-3.03
11	-5.70	-5.72	-6.64	-6.84
12	-18.56	-16.95	-17.74	-18.38
13	-33.66	-30.86	-32.01	-33.42
14	-45.03	-40.88	-42.34	-44.48
15	-29.43	-28.60	-29.9	-31.29
16	-26.27	-25.79	-27.06	-27.41
17	-37.17	-37.16	-37.61	-38.45
18	-5.01	-5.06	-4.76	-5.19
19	-2.42	-2.04	-2.06	-2.10

Table A6.7.: Interaction energies in kcal mol^{-1} for the IONPI19 benchmark calculated with the listed composite (3C) DFT methods.

System	Ref.	PBEh-3c	B97-3c	r^2SCAN-3c
1	-39.09	-39.95	-37.34	-37.87
2	-25.63	-27.43	-24.94	-24.95
3	-19.9 0	-20.32	-15.82	-16.77
4	-14.81	-16.94	-15.82	-16.60
5	-25.65	-27.75	-26.90	-26.52
6	-19.74	-23.47	-21.97	-21.69
7	-21.51	-23.66	-25.51	-23.86
8	-14.57	-16.84	-15.56	-15.43
9	-10.41	-13.35	-11.45	-12.15
10	-1.93	-2.68	-2.52	-3.07
11	-5.70	-7.56	-6.30	-7.16
12	-18.56	-25.35	-18.58	-18.75
13	-33.66	-46.57	-35.02	-35.27
14	-45.03	-62.26	-46.49	-46.71
15	-29.43	-32.81	-27.75	-30.83
16	-26.27	-29.24	-26.39	-26.28
17	-37.17	-39.74	-39.76	-39.07
18	-5.01	-5.13	-5.20	-5.37
19	-2.42	-2.85	-2.22	-2.29

Table A6.8.: Interaction energies in kcal mol^{-1} for the IONPI19 benchmark calculated with the listed SQM and FF methods.

System	Ref.	PM6-D3H4X	PM7	GFN2-xTB	GFN1-xTB	GFN-FF
1	-39.09	-12.23	-36.61	-36.55	-40.77	–
2	-25.63	-29.04	-15.42	-28.48	-14.76	-30.84
3	-19.90	-20.35	-6.62	-24.72	-20.47	-60.7
4	-14.81	-26.78	-19.74	-23.60	-15.16	-17.98
5	-25.65	4.15	-23.89	-30.44	-41.87	–
6	-19.74	-33.71	-16.73	-29.43	-21.41	-12.13
7	-21.51	-14.63	-3.03	-20.11	-13.69	-7.31
8	-14.57	-9.63	-5.49	-19.75	-16.65	-14.14
9	-10.41	-8.45	-6.03	-9.60	-7.53	-1.9
10	-1.93	-0.17	-5.43	-1.12	2.37	3.54
11	-5.70	-3.55	-6.54	-4.76	-1.58	0.81
12	-18.56	-19.15	-6.54	-23.77	-10.73	-8.77
13	-33.66	-35.59	-13.16	-43.85	-20.47	-17.76
14	-45.03	-49.41	-19.08	-61.93	-29.31	-26.39
15	-29.43	-22.93	56.84	-26.20	-18.27	-11.09
16	-26.27	-20.40	-1.59	-24.90	-22.17	-10.73
17	-37.17	-37.15	11.39	-29.60	-27.24	-23.02
18	-5.01	-5.03	-3.88	-2.43	-1.73	0.86
19	-2.42	-4.01	0.10	-1.61	-0.58	0.30

Table A6.9.: CP uncorrected interaction energies in kcal mol^{-1} for systems **1**-**6** of the IONPI19 benchmark calculated with the listed methods. The MAD for the entire test set is also given.

Method	1	2	3	4	5	6	MAD
HF-NL	-41.66	-27.01	-19.82	-13.42	-26.02	-19.75	1.12
double hybrids							
PWPB95-D4	-40.79	-27.08	-18.93	-16.59	-27.43	-20.89	0.65
revDSD-BLYP-D4	-38.79	-26.97	-18.61	-16.68	-25.45	-21.06	0.72
revDSD-PBEP86-D4	-38.95	-26.72	-18.91	-16.44	-25.52	-21.04	0.45
B2PLYP-D4	-39.32	-26.63	-18.66	-16.25	-26.04	-20.93	0.89
hybrids							
ωB97M-D4	39.54	-25.62	-17.41	-14.69	-25.99	-19.56	0.73
ωB97X-V	-41.23	-26.60	-19.65	-15.18	-26.83	-20.58	0.75
B3LYP-NL	-40.29	-26.26	-19.34	-15.51	-26.89	-20.62	0.76
PW6B95-D4	-41.41	-27.43	-18.94	-16.76	-28.22	-20.83	0.85
PBE0-D4	-41.01	-26.92	-19.03	-16.27	-27.48	-20.92	0.90
MN15	-39.90	-25.73	-19.40	-14.97	-26.29	-19.78	0.71
M06-2x	-42.22	-27.26	-20.35	-15.96	-28.14	-21.13	1.07
(meta-)GGAs							
B97M-V	-37.44	-24.33	-18.24	-14.45	-25.35	-19.75	0.80
r^2SCAN-D4	-38.67	-26.08	-18.77	-15.98	-25.71	-20.82	0.72
TPSS-D4	-38.90	-25.88	-17.76	-16.33	-26.57	-20.30	1.02
M06-L	-35.51	-22.81	-16.93	-13.76	-24.26	-18.64	1.43
PBE-D4	-39.31	-26.17	-18.33	-16.65	-26.95	-20.83	1.06

Table A6.10.: DLPNO-CCSD(T)/*TightPNO*/def2-SVP diagnostic for nondynamical correlation and multi-reference character for the systems of the IONPI19 set.

System	T1 diagnostic			max. T2 amplitudes		
	AB	A	B	AB	A	B
1	0.0101	0.0106	0.0005	0.0386	0.0430	0.0446
2	0.0093	0.0105	0.0004	0.0388	0.0191	0.0667
3	0.0092	0.0104	0.0016	0.0385	0.0389	0.0409
4	0.0157	0.0172	0.0004	0.0513	0.0439	0.0667
5	0.0152	0.0151	0.0005	0.0430	0.0491	0.0446
6	0.0146	0.0155	0.0004	0.0340	0.0468	0.0667
7	0.0103	0.0106	0.0092	0.0106	0.0672	0.0031
8	0.0111	0.0120	0.0043	0.0409	0.0365	0.0354
9	0.0150	0.0159	0.0043	0.0510	0.0465	0.0354
10	0.0144	0.0151	0.0043	0.0528	0.0513	0.0354
11	0.0148	0.0155	0.0043	0.0437	0.0430	0.0354
12	0.0148	0.0155	0.0043	0.0465	0.0468	0.0354
13	0.0152	0.0155	0.0043	0.0466	0.0468	0.0354
14	0.0153	0.0156	0.0043	0.0495	0.0464	0.0354
15	0.0150	0.0144	0.0167	0.0622	0.0622	0.0500
16	0.0147	0.0144	0.0155	0.0622	0.0622	0.0608
17	0.0125	0.0126	0.0132	0.0125	0.0126	0.0362
18	–	0.0114	0.0114	–	0.0494	0.0507
19	–	0.0106	0.0106	–	0.0405	0.0398

IONPI17 subset statistics

Table A6.11.: Statistical evaluation of all tested low-cost DFT, SQM, and FF methods on the IONPI17 subset. The MD, MAD and SD are given in kcal mol^{-1}, the *rel*MAD is given in %.

Method	MD	MAD	SD	*rel*MAD
composite (3c) DFT				
PBEh-3c	-3.93	3.93	4.49	18.88
B97-3c	-0.53	1.50	1.86	8.69
r^2SCAN-3c	-0.82	1.41	1.41	10.86
SQM				
PM6-D3H4X	3.44	7.77	11.21	40.05
PM7	16.46	17.46	22.21	75.93
GFN2-xTB	-2.93	5.12	6.14	24.10
GFN1-xTB	4.58	7.26	7.76	42.87
FF				
GFN-FF*	6.07	11.86	14.07	80.28

Table A6.12.: Statistical evaluation of all tested WFT and DFT methods on the IONPI17 subset. The MD, MAD and SD are given in kcal mol^{-1}, the *rel*MAD is given in %.

Method	LD	MD	MAD	SD	*rel*MAD
MP2/CBS	–	-2.05	2.11	3.65	10.26
	D3	-0.44	2.14	3.13	16.38
HF	D4	-0.37	1.55	1.80	8.99
	NL	-0.31	1.19	1.95	7.88
double hybrids					
	D3	-0.81	1.18	1.41	5.42
PWPB95	D4	-0.38	0.57	0.67	3.27
	NL	-0.81	0.86	0.99	3.54
revDSD-BLYP	D4	-0.15	0.66	0.83	5.23
revDSD-PBEP86	D4	0.10	0.41	0.58	3.43
	D3	-0.74	1.03	1.43	4.49
B2PLYP	D4	-0.51	0.80	1.38	5.45
	NL	-0.73	0.86	1.24	4.81
hybrids					
	D3	0.05	1.20	1.58	6.34
ωB97M	D4	0.05	0.81	1.17	4.92
	NL	-1.08	1.11	1.23	6.11
	D3	1.18	1.65	1.62	8.50
ωB97X	D4	0.59	0.93	0.97	5.86
	NL	-0.49	0.78	1.03	3.55
	D3	-0.50	1.53	2.01	6.84
B3LYP	D4	-0.41	1.13	1.54	8.07
	NL	-0.67	0.81	0.77	5.23
	D3	-0.66	1.39	1.70	6.69
PW6B95	D4	-0.50	0.91	1.10	4.93
	NL	-0.85	0.92	1.10	4.26
	D3	-0.22	1.58	1.87	8.79
PBE0	D4	-0.29	0.99	1.17	4.08
	NL	-0.93	0.98	0.95	4.37
MN15	–	-0.59	0.75	0.87	5.84
M06-2x	–	-1.00	1.15	1.22	4.93
(meta-)GGAs					
	D3	-0.40	1.24	1.65	5.82
B97M	D4	0.24	1.34	1.64	10.07
	NL	0.01	0.89	1.12	6.56
	D3	-0.15	0.86	1.07	4.15
r^2SCAN	D4	-0.07	0.77	0.91	4.47
	NL	0.13	1.06	1.58	7.66
	D3	0.05	1.71	2.11	8.07
TPSS	D4	-0.28	1.09	1.39	7.73
	NL	-1.03	1.08	0.86	7.29
M06-L	–	1.58	1.61	1.26	7.40
	D3	0.06	1.43	1.92	6.32
PBE	D4	-0.25	1.12	1.34	8.76
	NL	-0.98	1.14	0.95	9.08

IONPI19 statistics

Table A6.13.: Statistical evaluation of all tested low-cost DFT, SQM, and FF methods on the IONPI19 set. The MD, MAD and SD are given in kcal mol^{-1}, the *rel*MAD is given in %.

Method	MD	MAD	SD	*rel*MAD
composite (3c) DFT				
PBEh-3c	-3.55	3.55	4.39	17.96
B97-3c	-0.48	1.36	1.76	8.41
r^2SCAN-3c	-0.75	1.29	1.35	10.37
SQM				
PM6-D3H4X	3.00	7.04	10.66	39.32
PM7	14.92	15.82	21.45	74.61
GFN2-xTB	-2.44	4.76	5.98	26.03
GFN1-xTB	4.37	6.76	7.34	45.82
FF				
GFN-FF*	6.06	11.86	14.07	80.28

Table A6.14.: Functional mean deviation from the reference values calculated as the average of all tested DFAs (3c methods excluded). For system **1**-**6** the values are given with and without CP correction.

System	Functional MD	
	CP	w/o CP
1	-0.38	-0.49
2	-0.15	-0.52
3	1.21	1.07
4	-0.53	-0.97
5	-0.63	-0.80
6	-0.27	-0.74
7	-0.62	–
8	0.23	–
9	-0.78	–
10	-0.41	–
11	-0.47	–
12	0.49	–
13	0.55	–
14	0.03	–
15	-0.59	–
16	-0.40	–
17	-1.16	–
18	-0.08	–
19	0.12	–

Table A6.15.: Statistical evaluation of all tested WFT and DFT methods on the IONPI19 set. The MD, MAD and SD are given in kcal mol^{-1}, the *rel*MAD is given in %.

Method	LD	MD	MAD	SD	*rel*MAD
MP2/CBS	–	-1.91	1.97	3.47	10.98
	D3	-0.45	1.97	2.95	16.32
HF	D4	-0.34	1.41	1.70	9.01
	NL	-0.33	1.12	1.84	8.62
double hybrids					
	D3	-0.75	1.09	1.35	5.48
PWPB95	D4	-0.35	0.51	0.64	3.11
	NL	-0.76	0.80	0.95	3.82
revDSD-BLYP	D4	-0.11	0.62	0.79	5.37
revDSD-PBEP86	D4	0.10	0.39	0.55	3.47
	D3	-0.67	0.94	1.36	4.32
B2PLYP	D4	-0.45	0.72	1.32	5.13
	NL	-0.66	0.78	1.19	4.72
hybrids					
	D3	0.00	1.12	1.50	6.76
ωB97M	D4	0.03	0.74	1.10	4.92
	NL	-1.01	1.04	1.18	5.59
	D3	-1.02	1.52	1.61	9.48
ωB97X	D4	0.52	0.84	0.94	5.43
	NL	-0.47	0.73	0.98	4.12
	D3	-0.46	1.39	1.90	6.87
B3LYP	D4	-0.34	1.04	1.46	7.85
	NL	-0.60	0.73	0.76	5.15
	D3	-0.57	1.26	1.61	6.44
PW6B95	D4	-0.44	0.83	1.05	4.68
	NL	-0.78	0.84	1.07	4.19
	D3	-0.21	1.43	1.77	8.25
PBE0	D4	-0.26	0.88	1.10	3.68
	NL	-0.85	0.90	0.93	4.32
MN15	–	-0.59	0.73	0.83	6.53
M06-2x	–	-0.92	1.06	1.17	5.07
(meta-)GGAs					
	D3	-0.34	1.13	1.57	6.24
B97M	D4	0.26	1.24	1.55	10.28
	NL	0.03	0.82	1.06	6.87
	D3	-0.13	0.80	1.01	4.68
r^2SCAN	D4	-0.04	0.71	0.87	4.79
	NL	0.13	0.97	1.49	7.63
	D3	0.06	1.56	1.99	8.20
TPSS	D4	-0.23	1.00	1.32	7.85
	NL	-0.92	1.00	0.88	7.42
M06-L	–	1.44	1.48	1.27	8.01
	D3	0.07	1.30	1.81	6.54
PBE	D4	-0.19	1.03	1.27	8.89
	NL	-0.87	1.05	0.96	9.01

A7. Supporting Information to Chapter 7

Appendix A7 contains:

- General remarks
- Computational details
- Statistical error measures
- Computational results
- Comparison of other theoretical methods

General remarks

The *xtb*[243] program package is available free of charge for academic use. The Supporting Information is available in addition:
Spicher, S.; Grimme, S. "Efficient Computation of Free Energy Contributions for Association Reactions of Large Molecules", *J. Phys. Chem. Lett* **2020**, *11*, 6606–6611.

Computational details

All DFT calculations are performed using the TURBOMOLE 7.2 program package.[335,336,612] The resolution-of-identity (RI) approximation for the Coulomb integrals is applied in all cases using matching default auxiliary basis sets.[341,342] For the integration of the exchange-correlation contribution, the numerical quadrature grid *m*5 is employed. The default convergence criteria (10^{-7} E_h for energies and 10^{-5} E_h/bohr for gradients) are used throughout. Computations of harmonic vibrational frequencies are performed analytically using the *aoforce* code of TURBOMOLE. All low-cost computations are performed with the *xtb* stand-alone program,[243] where GFN2-xTB as well as GFN-FF are implemented. PM6-D3H4 calculations have been performed via the *xtb* interface driver using MOPAC2016.[338] The same convergence criteria as before have been chosen in order to guarantee comparable results. For the computation of thermodynamic properties, PBEh-3c frequencies are scaled by a factor of 0.95 and GFN-FF frequencies by a factor of 1.03. A scaling factor slightly smaller and larger than unity has been tested for PM6-D3H4 but did not lead to any improvements. On the entire test set of 59 complexes (168 molecules), none of the tested methods showed more than twenty (small) imaginary frequencies in total, and not more than

one imaginary (less than $i20\,\mathrm{cm}^{-1}$) frequency per system. As mentioned in the text, all small imaginary frequencies are set real (multiplied by $-i$) and consistently included in the thermostatistical treatment.

Statistical error measures

In this work, the following statistical measures were used. p is an arbitrary property.

$$\delta p = p_{calc.} - p_{ref.} \tag{A7.1}$$

The error measures are defined by:

- Mean deviation (MD):

$$MD = \frac{1}{N}\sum_{i}^{N} \delta p_i \tag{A7.2}$$

- Mean absolute deviation (MAD):

$$MAD = \frac{1}{N}\sum_{i}^{N} |\delta p_i| \tag{A7.3}$$

- Root-mean-square deviation (RMSD):

$$RMSD = \sqrt{\frac{\sum_i^N |\delta p_i|^2}{N}} \tag{A7.4}$$

- relative MAD (relMAD):

$$relMAD = \frac{1}{N}\sum_{i}^{N} \frac{|\delta p_i|}{p_{ref}}. \tag{A7.5}$$

Table A7.1.: Statistical evaluation[a] of the enthalpic contribution ΔH_{298} to the free binding energy of 59 NCI complexes.

		S30L	S22	L7	Average
DFT Ref.	MD	±0.19	±0.05	±0.22	±0.01
	MAD	0.26	0.15	0.44	0.29
	RMSD	0.34	0.26	0.56	0.39
GFN2-xTB	MD	0.12	-0.14	-0.18	-0.07
	MAD	0.50	0.38	0.60	0.49
	RMSD	0.70	0.52	0.86	0.69
PM6-D3H4	MD	0.29	-0.19	0.43	0.18
	MAD	0.91	0.36	0.94	0.74
	RMSD	1.11	0.49	1.05	0.88
GFN-FF	MD	-0.23	0.40	-0.48	-0.10
	MAD	0.89	0.60	1.01	0.83
	RMSD	1.17	0.80	1.19	1.05

[a] The statistical measures are given with respect to the DFT reference for GFN2-xTB, PM6-D3H4 and GFN-FF. For the DFT reference, average values for PBEh-3c and B97-3c are taken and their RMSD of 0.39 kcal mol^{-1} is taken as a uncertainty estimate.

Computational results

Table A7.5.: $\Delta H_{\mathrm{mRRHO}}$ contribution to the association free energies[a] for the S22 benchmark set.

	PBEh-3c	B97-3c	Ref.	GFN2-xTB	PM6-D3H4	GFN-FF
1	1.54	0.91	1.23	0.79	1.72	1.32
2	1.71	1.64	1.68	1.55	1.57	2.05
3	1.46	1.32	1.39	0.75	1.05	2.81
4	1.88	1.83	1.86	1.38	1.24	2.62
5	1.39	1.34	1.37	0.92	0.80	2.51
6	1.34	1.19	1.27	0.70	1.07	2.64
7	1.45	1.79	1.62	0.71	0.99	3.23
8	1.18	1.23	1.21	1.21	1.24	1.20
9	1.23	1.28	1.26	1.25	0.08	0.03
10	1.19	-0.57	0.31	-0.56	-0.03	-0.57
11	1.13	-0.02	0.56	0.64	0.58	0.62
12	1.20	0.70	0.95	1.25	1.17	1.22
13	1.23	1.43	1.33	1.32	1.08	2.36
14	1.27	1.34	1.31	1.24	1.24	1.22
15	1.36	1.53	1.45	1.50	1.29	1.87
16	0.99	0.96	0.98	0.92	0.89	0.97
17	0.70	0.77	0.74	1.23	0.74	0.95
18	0.66	0.67	0.67	1.25	0.71	0.89
19	-0.27	-0.08	-0.17	0.93	0.84	1.00
20	0.02	0.14	0.08	0.05	-0.03	0.10
21	0.03	0.81	0.42	-0.55	-0.01	0.89
22	1.46	1.65	1.56	1.37	0.64	1.96

[a] Values are given in kcal mol^{-1}.

Table A7.2.: $\Delta G_{\mathrm{mRRHO}}$ contribution to the association free energies[a] for the S30L benchmark set.[120]

	PBEh-3c	B97-3c	Ref.	GFN2-xTB	PM6-D3H4	GFN-FF
1	16.38	15.92	16.15	16.02	16.25	17.16
2	14.33	14.11	14.22	14.63	14.86	15.85
3	16.59	16.81	16.70	16.65	16.22	15.80
4	14.66	15.17	14.91	14.70	13.18	14.52
5	17.53	18.44	17.98	18.25	16.98	16.84
6	15.53	16.72	16.12	16.63	16.57	16.14
7	17.58	18.65	18.11	19.07	21.13	18.77
8	18.17	19.80	18.98	20.69	20.09	16.23
9	16.22	16.32	16.27	18.23	18.29	17.99
10	17.00	17.61	17.30	18.37	18.37	18.47
11	15.57	16.91	16.24	18.14	19.18	18.49
12	16.36	17.69	17.02	17.95	19.36	18.53
13	14.10	15.39	14.74	14.48	14.72	16.07
14	14.91	15.42	15.16	14.64	13.10	14.58
15	15.29	15.23	15.26	17.34	19.41	16.80
16	16.55	16.72	16.63	18.80	19.70	16.62
17	16.88	16.66	16.77	16.59	17.60	17.64
18	14.64	16.94	15.79	16.12	16.87	17.36
19	13.73	14.78	14.25	14.76	16.60	15.54
20	15.20	16.18	15.69	16.19	17.27	16.02
21	18.45	18.55	18.50	15.37	16.94	18.01
22	20.22	18.06	19.14	18.10	18.42	21.18
23	–[b]	18.34	18.34	16.83	18.22	18.13
24	23.11	23.74	23.42	25.10	21.20	22.49
25	16.75	17.23	16.99	17.34	15.61	16.19
26	–[b]	16.72	16.72	17.43	15.09	15.76
27	16.06	15.56	15.81	17.68	18.52	16.91
28	14.33	15.41	14.87	17.01	18.71	16.76
29	–[b]	16.06	16.06	16.22	16.61	16.06
30	17.20	17.53	17.36	17.24	17.59	16.73

[a] Values are given in kcal mol^{-1}.
[b] The *aoforce* calculation did not converge for the host-guest complex system.

Table A7.3.: $\Delta G_{\mathrm{mRRHO}}$ contribution to the association free energies[a] for the S22 benchmark set.[537]

	PBEh-3c	B97-3c	Ref.	GFN2-xTB	PM6-D3H4	GFN-FF
1	7.17	7.56	7.365	7.38	8.63	7.66
2	7.76	7.70	7.73	8.55	7.38	8.89
3	12.94	12.90	12.92	11.8	11.62	13.87
4	13.09	13.58	13.33	12.55	11.61	13.63
5	13.53	13.64	13.58	12.92	12.29	14.32
6	12.52	12.65	12.58	12.31	12.05	14.25
7	14.10	13.33	13.71	13.87	12.74	14.64
8	3.61	3.82	3.71	5.36	6.06	5.31
9	7.04	6.87	6.95	7.64	9.35	8.60
10	4.95	7.19	6.07	9.56	8.56	9.53
11	7.65	9.77	8.71	11.25	11.19	11.15
12	9.34	10.12	9.73	10.94	10.95	10.78
13	12.77	13.28	13.02	12.73	12.05	14.32
14	9.72	9.59	9.65	11.29	11.52	10.47
15	14.00	13.75	13.87	14.81	13.56	14.00
16	6.10	6.04	6.07	6.10	5.85	6.44
17	7.01	6.97	6.99	7.41	9.12	9.71
18	6.34	6.38	6.36	7.23	8.54	9.82
19	7.80	7.95	7.875	7.68	6.93	8.25
20	9.09	9.37	9.23	11.24	11.75	12.04
21	11.14	10.21	10.67	13.35	12.75	12.43
22	11.66	11.98	11.82	11.64	12.27	12.63

[a] Values are given in kcal mol^{-1}.

Table A7.4.: $\Delta H_{\mathrm{mRRHO}}$ contribution to the association free energies[a] for the S30L benchmark set.

	PBEh-3c	B97-3c	Ref.	GFN2-xTB	PM6-D3H4	GFN-FF
1	1.71	2.75	2.23	2.16	1.04	1.64
2	1.45	2.60	2.03	2.17	1.17	1.63
3	1.06	1.22	1.14	0.66	1.49	1.40
4	0.88	0.97	0.93	0.51	1.37	1.15
5	1.20	1.82	1.51	0.61	1.15	1.33
6	1.15	1.84	1.50	1.31	0.49	1.37
7	0.44	0.38	0.41	1.37	0.15	2.83
8	1.18	0.40	0.79	1.40	1.94	1.71
9	1.18	1.54	1.36	1.30	0.71	0.95
10	1.34	1.92	1.63	1.35	0.74	0.97
11	1.77	3.03	2.40	1.71	0.71	0.79
12	1.83	3.01	2.42	1.72	0.70	0.79
13	1.45	1.36	1.41	1.60	1.61	1.33
14	1.39	1.10	1.25	1.44	1.83	1.62
15	2.20	2.49	2.35	2.56	1.65	1.49
16	2.46	2.77	2.62	3.09	1.54	1.66
17	1.25	1.99	1.62	2.06	2.01	3.87
18	1.36	1.38	1.37	0.70	1.94	3.88
19	1.19	1.81	1.50	1.46	3.41	2.29
20	1.45	2.00	1.73	1.46	2.95	2.36
21	1.12	2.36	1.74	1.41	2.58	2.15
22	2.17	3.05	2.61	0.46	0.15	3.62
23	–[b]	3.19	3.19	1.32	1.44	2.90
24	3.68	4.97	4.33	5.14	1.95	2.51
25	1.34	1.77	1.56	1.79	1.97	1.54
26	–[b]	1.79	1.79	1.83	1.89	1.43
27	2.05	1.89	1.97	1.75	0.96	2.00
28	1.45	0.62	1.04	1.73	1.10	2.18
29	–[b]	1.52	1.52	2.20	2.03	3.90
30	1.71	1.91	1.81	1.77	2.42	3.28

[a] Values are given in kcal mol^{-1}.
[b] The *aoforce* calculation did not converge for the host-guest complex system.

Table A7.6.: $\Delta G_{\mathrm{mRRHO}}$ contribution to the association free energies[a] for the L7 benchmark set.[538]

	PBEh-3c	B97-3c	Ref.	GFN2-xTB	PM6-D3H4	GFN-FF
1	18.75	20.59	19.67	16.73	20.03	16.91
2	13.41	15.40	14.40	15.81	15.40	15.75
3	16.02	15.20	15.61	15.09	15.63	14.22
4	17.02	16.49	16.75	16.84	16.94	15.18
5	16.49	16.75	16.62	16.61	16.06	15.09
6	14.67	14.96	14.81	16.77	13.72	16.73
7	16.42	16.54	16.48	16.58	15.86	19.00

[a] Values are given in kcal mol^{-1}.

Comparison of other theoretical methods

We also tested the performance of a second generic force-field, namely UFF.[114] Its performance on all three subsets and the averaged values are given below in Table A7.8.
For the subset with the smallest molecules (S22) UFF performs surprisingly well, whereas, for the more difficult S30L and L7 benchmark sets, the error becomes significantly larger. The averaged MAD is 2.05 kcal mol^{-1}, and thus the worst performing low-cost method tested in this work. For S22 the good result is surprising because many of the UFF optimized geometries are artificially distorted. As an example, two structure overlays between the DFT reference and the UFF optimized geometry are shown in A7.2.

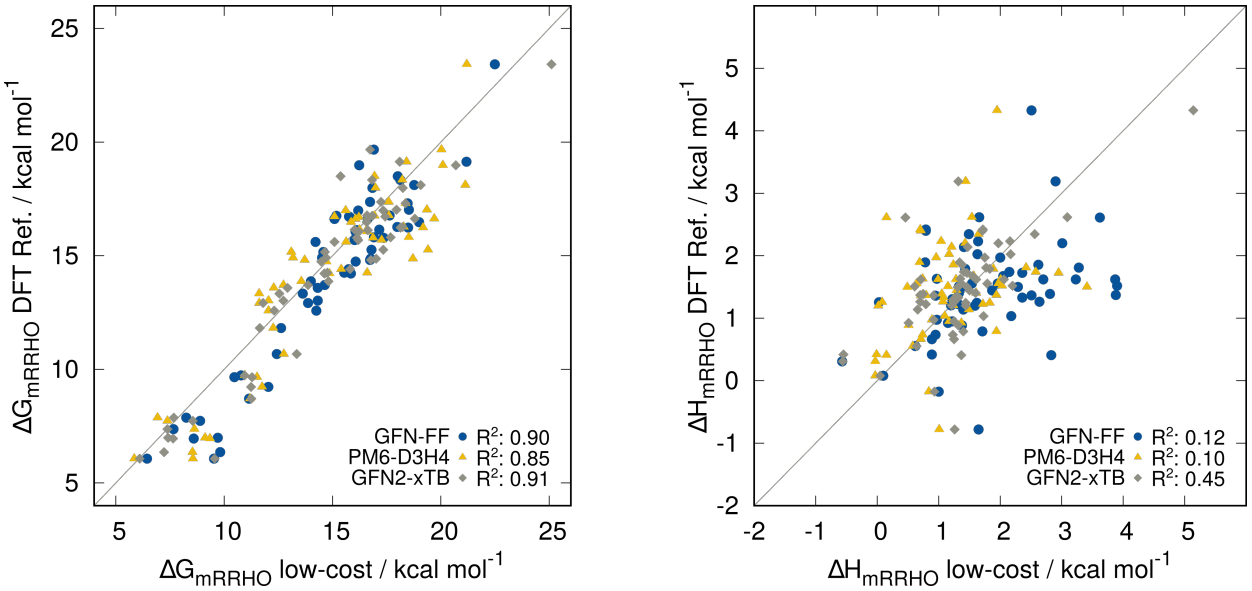


Figure A7.1.: Thermostatistical $\Delta G_{\mathrm{mRRHO}}$ and $\Delta H_{\mathrm{mRRHO}}$ values (in kcal mol^{-1}) to the association free energy of the complexes taken from the S30L, S22, and L7 benchmark set in a scatter plot. Results obtained by the low-cost methods GFN2-xtb, PM6-D3H4, and GFN-FF are plotted against DFT reference values. The correlation coefficients are given.

Table A7.7.: $\Delta H_{\mathrm{mRRHO}}$ contribution to the association free energies[a] for the L7 benchmark set.

	PBEh-3c	B97-3c	Ref.	GFN2-xTB	PM6-D3H4	GFN-FF
1	-0.18	-1.38	-0.78	1.26	1.01	1.65
2	1.19	2.60	1.90	1.34	0.69	0.78
3	0.68	1.72	1.20	1.44	0.02	1.59
4	1.00	0.77	0.89	1.32	0.51	1.38
5	1.14	3.14	2.14	1.47	1.21	1.41
6	1.66	1.58	1.62	1.61	1.28	2.70
7	2.11	2.29	2.20	1.97	1.43	3.01

[a] Values are given in kcal mol^{-1}.

Table A7.8.: Statistical evaluation (in kcal mol^{-1}) of the thermostatistical contribution $\Delta G_{\mathrm{mRRHO}}$ for the 59 complexes[a] with UFF. The *rel*MAD values are given in %.

		S30L	S22	L7	all sets
UFF	MD	0.29	-0.52	1.04	0.27
	MAD	2.04	0.82	3.30	2.05
	RMSD	2.57	1.05	4.35	2.66
	*rel*MAD	13.4	9.8	27.9	17.0

[a] The statistical measures are given with respect to the average DFT reference consisting of PBEh-3c and B97-3c values.

We further validated the accuracy of our averaged DFT reference values. Therefore we computed $\Delta G_{\mathrm{mRRHO}}$ values at the higher B3LYP-D3[252,658–660] functional in a large def2-TZVPP basis set[326] for the S22 benchmark set. A standard scaling factor of 0.96 for the vibrational frequencies[661] is applied. The results are shown in Figure A7.3 and Table A7.9, respectively. With a MAD of 0.31 kcal mol^{-1} and an

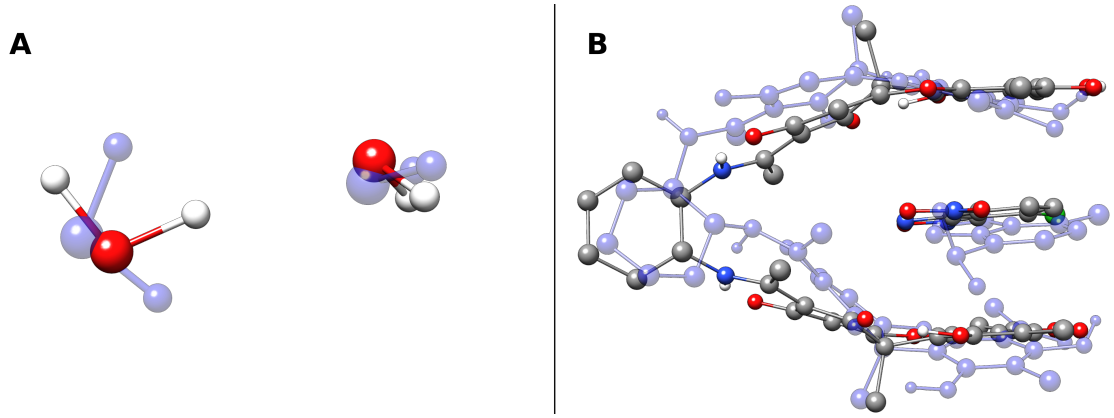


Figure A7.2.: (A) DFT structure of the water dimer (S22) and optimized UFF structure in transparent blue. (B) DFT structure of a molecular picer complex (S30L) and optimized UFF structure in transparent blue.

RMSD of 0.46 kcal mol^{-1}, the deviation between the B3LYP-D3 results and the averaged DFT reference is very similar to the uncertainty of the DFT reference itself. Consequently, the DFT reference used in this work can be considered as reliable within the given RMSD of about 0.4 kcal mol^{-1}.

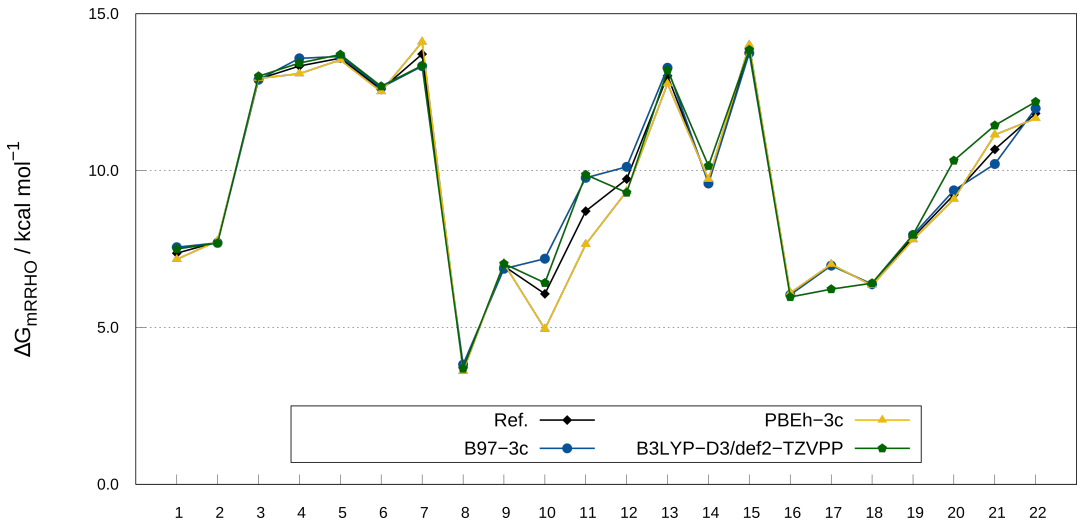


Figure A7.3.: Thermostatistical $\Delta G_{\mathrm{mRRHO}}$ values (in kcal mol^{-1}) to the association free energy of the complexes taken from the S22 benchmark set.

Table A7.9.: Statistical evaluation (in kcal mol^{-1}) of the thermostatistical contribution $\Delta G_{\mathrm{mRRHO}}$ for the 22 complexes[a] for B3LYP-D3/def2-TZVPP. The *rel*MAD values are given in %.

		S22
B3LYP-D3/def2-TZVPP	MD	0.15
	MAD	0.31
	RMSD	0.46
	*rel*MAD	3.4

[a] The statistical measures are given with respect to the average DFT reference consisting of averaged PBEh-3c and B97-3c values.

A8. Supporting Information to Chapter 8

Appendix A4 contains:

- General remarks
- Statistical error measures
- Computational results

General remarks

The *xtb*[515] program package is available free of charge for academic use. Collected atomic Cartesian coordinates (XYZ format) of all input and optimized structures are available in addition to the Supporting Information:
Spicher, S.; Grimme, S. "Single-Point Hessian Calculations for Improved Vibrational Frequencies and Rigid-Rotor-Harmonic-Oscillator Thermodynamics", *J. Chem. Theory Comput.* **2021**, *17*, 1701–1714.

Statistical error measures

In this work, the following statistical measures were used. p is an arbitrary property. Statistical measure for a set $x_1, \cdots, x_n$ of data points with references $r_1, \cdots, r_n$ are:

- Average:

$$\overline{x} = \frac{1}{n}\sum_{i}^{n} x_i \tag{A8.1}$$

- Mean deviation (MD):

$$MD = \frac{1}{n}\sum_{i}^{n}(x_i - r_i) \tag{A8.2}$$

- Mean absolute deviation (MAD):

$$MAD = \frac{1}{n}\sum_{i}^{n}|x_i - r_i| \tag{A8.3}$$

- relative MAD (relMAD):

$$relMAD = \frac{1}{n}\sum_{i}^{n}\frac{|x_i - r_i|}{r_i} \tag{A8.4}$$

- Standard deviation (SD):

$$SD = \sqrt{\frac{\sum_i^n |(x_i - r_i) - MD|^2}{n-1}} \tag{A8.5}$$

- Match-score

$$r_{match} = \frac{\left(\sum_i^k u_i v_i\right)^2}{\left(\sum_i^k u_i^2\right)\left(\sum_i^k v_i^2\right)} \tag{A8.6}$$

- Association energy:

$$\Delta E_{asso} = E(Complex) - E(Host) - E(Guest). \tag{A8.7}$$

Computational results

Table A8.1.: $\Delta G_{\mathrm{mRRHO}}$ contribution to the association free energies[a] for the S30L benchmark set[120] calculated by three different approaches with GFN2-xTB in comparison to B97-3c reference vales.

	HESS	OHESS	SPH	Reference
1	16.33	16.94	17.06	18.20
2	15.05	15.56	15.70	16.39
3	17.66	16.54	16.46	16.81
4	15.10	14.65	14.82	15.17
5	15.98	17.01	17.55	18.44
6	15.34	16.57	16.79	17.54
7	17.98	18.75	20.61	19.05
8	18.44	19.92	21.69	20.70
9	16.32	16.68	18.39	18.33
10	18.45	18.26	18.57	18.98
11	18.05	19.33	19.09	20.03
12	18.23	19.17	18.91	19.75
13	15.01	13.51	13.81	14.55
14	15.29	14.46	14.45	14.31
15	17.12	18.49	18.55	16.90
16	18.73	19.97	19.97	18.40
17	15.88	17.52	18.05	17.07
18	15.29	16.55	15.72	15.96
19	14.39	14.79	14.63	14.78
20	16.58	16.17	15.21	16.18
21	16.12	15.93	15.92	18.56
22	16.64	17.89	18.31	20.39
23	17.32	17.35	17.37	19.16
24	22.88	25.11	24.88	24.39
25	16.23	16.62	17.33	17.64
26	16.31	17.43	17.38	17.54
27	15.13	17.73	17.76	16.39
28	15.83	17.09	16.39	14.82
29	15.23	17.03	16.52	16.06
30	17.64	17.45	17.91	17.53

[a] Values are given in kcal mol^{-1}.

Table A8.2.: $\Delta G_{\mathrm{mRRHO}}$ contribution to the association free energies[a] for the S22 benchmark set[537] calculated by three different approaches with GFN2-xTB in comparison to B97-3c reference vales.

	HESS	OHESS	SPH	Reference
1	8.21	7.35	7.35	8.45
2	9.23	8.57	8.57	8.52
3	12.38	11.79	11.79	12.49
4	13.26	12.70	13.34	13.17
5	13.08	13.00	13.00	13.23
6	12.44	12.26	12.47	12.65
7	14.73	13.90	13.77	13.33
8	8.27	5.05	5.05	5.70
9	9.76	7.55	7.55	7.69
10	9.26	8.56	8.56	7.27
11	11.03	11.81	11.00	10.14
12	12.22	10.97	11.10	10.93
13	12.26	12.73	12.53	12.87
14	11.59	11.00	12.15	11.06
15	12.95	14.79	14.56	13.75
16	8.24	6.13	6.13	6.86
17	9.11	7.46	7.55	8.25
18	9.26	7.37	8.12	7.83
19	9.48	7.62	7.63	8.08
20	10.18	11.11	11.11	10.27
21	11.45	10.45	10.78	10.80
22	11.99	11.61	11.83	11.98

[a] Values are given in kcal mol^{-1}.

Table A8.3.: $\Delta G_{\mathrm{mRRHO}}$ contribution to the interaction energies[a] for the L7 benchmark set[538] calculated by three different approaches with GFN2-xTB in comparison to PBEh-3c reference vales.

	HESS	OHESS	SPH	Reference
1	19.52	16.75	17.34	18.07
2	15.00	15.91	16.14	16.05
3	14.91	15.09	15.09	15.00
4	15.01	16.87	16.87	17.15
5	14.19	16.88	16.63	16.19
6	15.61	15.83	15.24	14.80
7	17.48	16.68	16.22	16.57

[a] Values are given in kcal mol^{-1}.

Table A8.4.: G_{mRRHO} contributions[a] for five displacements of (S)-camphor calculated with and without previous optimization and by the SPH approaches at the GFN2-xTB and GFN1-xTB level of theory.

	GFN2-xTB			GFN1-xTB		
Scale	HESS	OHESS	SPH	HESS	OHESS	SPH
0.90	148.36	127.45	134.23	146.72	127.00	133.97
0.95	137.90	127.41	127.60	137.99	127.05	127.42
1.00	127.40	127.40	127.40	126.72	127.04	127.04
1.05	119.08	127.44	125.86	120.25	127.06	125.90
1.10	107.02	127.44	118.14	109.97	127.06	118.63

[a] Values are given in kcal mol^{-1}.

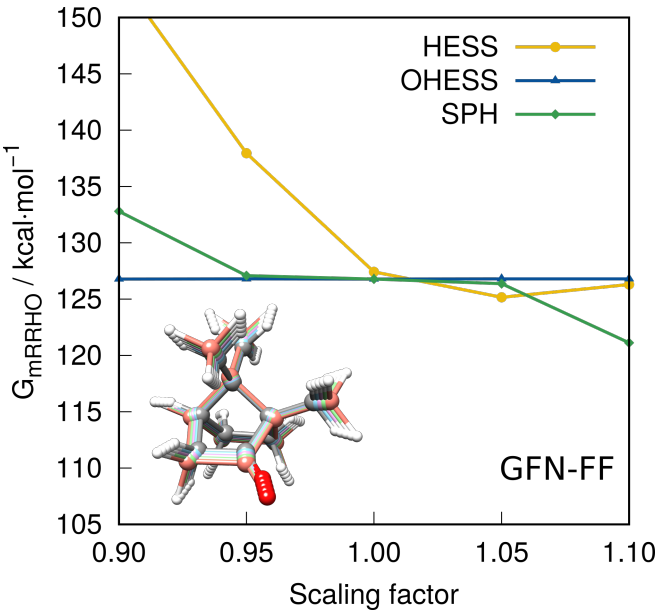


Figure A8.1.: Thermostatistical contributionG_{mRRHO} calculated with and without previous optimization and by the SPH approach for five displaced structures of (S)-camphor with GFN-FF.

Table A8.5.: $\Delta G_{\mathrm{mRRHO}}$ contributions[a] for the Diels-Alder reaction of two cyclopentadiene molecules calculated with and without previous optimization and by the SPH approaches at the GFN2-xTB and GFN-FF level of theory.

	GFN2-xTB			GFN-FF		
RC	HESS	OHESS	SPH	HESS	OHESS	SPH
1	90.64	90.64	90.64	91.78	89.27	89.27
2	90.86	90.71	90.71	91.91	89.28	89.28
3	91.13	90.65	90.65	92.10	89.38	89.90
4	91.37	90.60	90.79	92.33	89.28	90.05
5	91.73	90.55	91.10	92.66	89.41	90.27
6	92.12	90.56	91.33	92.34	89.27	90.55
7	92.50	90.54	91.58	92.75	89.33	90.89
8	92.76	90.50	91.89	92.28	89.25	91.29
9	93.82	90.52	92.25	92.91	89.27	91.76
10	94.32	90.51	92.64	93.16	89.27	92.22
11	94.83	90.53	93.01	93.49	89.36	92.59
12	95.32	99.61	95.95	94.03	89.27	93.07
13	96.90	99.62	97.83	94.89	94.34	93.52
14	98.02	99.62	98.50	95.47	94.35	94.72
15	98.63	99.62	98.80	96.47	97.64	96.70
16	99.43	99.62	99.03	97.28	97.64	96.84
17	99.59	99.62	99.30	97.74	97.65	97.07
18	99.66	99.62	99.62	98.13	97.65	97.34
19	99.64	99.62	99.62	98.30	97.64	97.59
20	99.63	99.63	99.62	98.30	97.64	97.64

[a] Values are given in kcal mol^{-1}.

Table A8.6.: G_{mRRHO} contributions[a] calculated along the MD trajectory of crambin[b] with and without previous optimization and by the SPH approaches at the GFN2-xTB and GFN-FF level of theory.

	GFN2-xTB			GFN-FF		
time / ps	HESS	OHESS	SPH	HESS	OHESS	SPH
51	3013.4	3007.6	2997.3	3056.8	3079.2	3048.1
52	3015.8	3007.1	2999.0	3055.0	3084.7	3052.9
53	3013.3	3005.8	2994.5	3059.6	3085.2	3049.1
54	3013.8	3003.7	3002.3	3058.3	3084.7	3052.5
55	3013.6	3004.5	2997.7	3061.0	3082.6	3047.3
56	3015.4	3007.0	3001.5	3058.2	3077.4	3050.0
57	3019.6	3006.5	3001.8	3058.6	3081.0	3047.4
58	3017.3	3006.4	3000.8	3062.9	3079.0	3050.5
59	3013.0	3005.2	2997.7	3058.6	3078.3	3051.4
60	3017.8	3004.9	3001.3	3059.5	3078.3	3056.5
61	3013.6	3005.1	2999.8	3050.3	3080.1	3052.0
62	3018.8	3005.7	3000.2	3063.3	3079.5	3050.0
63	3018.0	3005.8	2999.0	3066.8	3081.0	3052.2
64	3018.6	3008.3	3001.4	3061.2	3078.4	3051.9
65	3012.1	3008.5	2999.3	3053.3	3081.2	3048.4
66	3014.0	3008.8	3002.4	3057.8	3088.6	3047.1
67	3012.0	3009.5	3000.7	3055.2	3081.1	3048.2
68	3010.9	3008.7	2998.2	3054.0	3082.3	3046.9
69	3015.8	3008.3	3001.6	3049.6	3081.7	3049.4
70	3017.4	3008.6	3000.7	3059.2	3083.9	3049.7
71	3015.0	3006.2	3001.6	3049.3	3081.7	3049.6
72	3015.1	3005.7	3002.0	3057.0	3080.4	3052.1
73	3014.3	3005.6	2998.8	3055.1	3087.8	3050.5
74	3014.3	3007.8	3001.5	3044.2	3083.5	3045.5
75	3015.1	3008.2	3000.7	3053.6	3083.4	3050.3

[a] Values are given in kcal mol^{-1}.
[b] Values are given from 51 to 75 ps.

Table A8.7.: G_{mRRHO} contributions[a] calculated along the MD trajectory of crambin[b] with and without previous optimization and by the SPH approaches at the GFN2-xTB and GFN-FF level of theory.

	GFN2-xTB			GFN-FF		
time / ps	HESS	OHESS	SPH	HESS	OHESS	SPH
76	3016.2	3008.3	3002.3	3061.3	3084.2	3053.2
77	3014.6	3005.6	3003.5	3062.5	3088.4	3051.4
78	3013.7	3006.2	3000.9	3057.1	3090.9	3048.8
79	3015.7	3006.5	2999.0	3068.9	3085.3	3051.7
80	3018.4	3005.7	3001.4	3060.6	3089.0	3046.6
81	3015.8	3006.8	2997.8	3057.0	3082.8	3045.6
82	3011.7	3007.0	2996.1	3052.4	3081.8	3047.6
83	3015.4	3008.2	3000.1	3059.3	3085.4	3045.9
84	3016.1	3007.4	2998.5	3059.2	3086.1	3049.7
85	3013.7	3007.9	2998.9	3057.2	3082.4	3044.1
86	3019.0	3006.9	2999.7	3062.0	3084.5	3049.3
87	3016.5	3005.4	3000.4	3063.6	3087.4	3051.4
88	3017.5	3003.9	3001.7	3057.4	3086.7	3052.1
89	3019.1	3003.5	3003.3	3061.3	3079.9	3056.2
90	3019.8	3004.1	3001.1	3067.0	3078.5	3050.0
91	3019.0	3003.2	3002.5	3062.4	3079.9	3052.7
92	3018.8	3004.9	3000.7	3060.6	3086.0	3047.7
93	3015.5	3007.3	2995.5	3060.7	3081.5	3046.8
94	3014.7	3005.2	2998.7	3054.7	3080.2	3050.2
95	3016.3	3004.7	3002.0	3058.5	3082.8	3054.5
96	3011.4	3004.3	2998.7	3060.7	3083.4	3044.9
97	3016.8	3007.1	2999.7	3065.3	3077.5	3053.9
98	3015.5	3004.6	2998.5	3063.4	3080.7	3049.9
99	3018.5	3006.7	3000.0	3062.2	3083.5	3048.7
100	3012.9	3005.0	2999.6	3059.9	3079.8	3047.5

[a] Values are given in kcal mol^{-1}.
[b] Values are given from 76 to 100 ps.

A9. Supporting Information to Chapter 9

Appendix A9 contains:

- General remarks
- Computational results

General remarks

Intensities and frequencies of IR bands in the spectra of the test molecules and details of computations, an example for averaging of individual, weighted spectra, and computed $A_{\nu OH}/A_{\nu OD}$ and ν_{OH}/ν_{OD} values affected by rather strong coupling of vibrations of neighboring methanol molecules are available in addition to the Supporting Information:
Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. "Fast and Accurate Quantum Chemical Modeling of Infrared Spectra of Condensed-Phase Systems", *J. Phys. Chem. B* **2020**, *124*, 6664–6670.

Computational results

Table A9.1.: Experimental IR intensities (A, km·mol^{-1}) of all assessed molecules in the gas phase and the corresponding B97-3c computed values.

Molecule	Integration range	A^{gas}(exp.)	rel. A^{gas}(exp.)	A^{gas}(comp.)	A^{gas}(comp.)
cyclohexane-d_{12}	1240–1130	6.4	0.98	9.4	1.21
	1130–1045	6.5	1.00	7.8	1.00
	1045–950	5.9	0.91	6.4	0.82
	950–880	3.5	0.54	4.0	0.51
	770–600	3.5	0.54	4.4	0.56
	450–350	0.4	0.06	0.4	0.05
acetone	1738	145	1.00	170.2	1.00
	1456–1438	35.0	0.24	56.4	0.33
	1363–1360	69.0	0.48	81.9	0.48
	1218	66.5	0.46	77.9	0.46
	1093–1067	4.5	0.03	3.4	0.02
	896	7.6	0.05	11.5	0.07
	779	1.9	0.01	2.2	0.01
	528–483	17.0	0.12	12.8	0.08

Molecule	Integration range	A^{gas}(exp.)	rel. A^{gas}(exp.)	A^{gas}(comp.)	A^{gas}(comp.)
cyclohexane	1500–1330	24.0	1.00	30.9	1.00
	1310–1190	3.6	0.15	3.0	0.10
	1100–970	2.0	0.08	2.3	0.07
	970–800	6.6	0.28	8.5	0.28
	560–480	0.5	0.02	0.7	0.02
1,4-dioxane	1520–1410	16.2	0.09	13.2	0.07
	1410–1335	13.4	0.08	21.7	0.11
	1335–1200	48.6	0.28	38.0	0.19
	1200–1000	176.5	1.00	197.9	1.00
	960–800	84.9	0.48	88.1	0.45
	660–540	14.6	0.08	12.0	0.06
1,4-dioxane-d_8	1300–1070	231.7	1.00	224.2	1.00
	1070–1000	41.0	0.18	57.7	0.26
	1000–840	17.4	0.08	16.2	0.07
	840–700	59.7	0.26	47.6	0.21
	560–440	9.6	0.04	8.5	0.04
tetrahydropyrane	1550–1415	19.9	0.19	21.4	0.18
	1415–1325	11.7	0.11	14.4	0.12
	1325–1235	14.9	0.14	15.2	0.13
	1235–1165	41.2	0.40	26.4	0.22
	1165–1140	2.3	0.02	2.8	0.02
	1140–977	103.8	1.00	117.4	1.00
	977–930	2.4	0.02	3.7	0.03
	925–832	27.9	0.27	28.9	0.25
	832–780	4.0	0.04	5.1	0.04
	620–520	3.7	0.04	3.1	0.03
	450–340	7.2	0.07	5.3	0.05
	290–200	4.3	0.04	3.8	0.03
benzene	1484	13	0.13	13.8	0.12
	1038	7.5	0.07	12.2	0.10
	674	104.0	1.00	119.7	1.00
benzene-d_6	1330	2.9	0.04	13.8	0.12
	812	8.2	0.12	12.2	0.10
	496	66.0	1.00	119.7	1.00
pyridine	1540–1660	31.9	0.42	29.3	0.39
	1477–1512	2.5	0.03	2.2	0.03
	1472–1480	27.3	0.36	26.3	0.35
	1334–1400	1.0	0.01	0.0	0.00
	1256–1308	0.1	0.00	0.7	0.01
	1182–1246	3.7	0.05	3.4	0.05
	1110–1173	3.6	0.05	1.9	0.03
	1054–1100	2.8	0.04	4.4	0.06
	1009–1054	9.1	0.12	5.5	0.07
	958–1009	0.0	0.00	5.4	0.07
	671–735	75.7	1.00	74.9	1.00

Molecule	Integration range	A^{gas}(exp.)	rel. A^{gas}(exp.)	A^{gas}(comp.)	A^{gas}(comp.)
	650–659	0.0	0.00	0.3	0.00
	570–632	4.5	0.06	3.4	0.05
MeOD	775	0.47	0.47	0.57	0.57
	895	0.0	0.06	0.0	0.03
	980	1.0	1.0	1.0	1.0
	1028	0.2	0.23	0.4	0.44
	1069	0.1	0.08	0.0	0.02
	1078	0.1	0.08	0.2	0.15
	1135	0.6	0.55	0.7	0.65
MeOH	1034	1.0	1.0	1.0	1.0
	1075	0.0	0.02	0.0	0.04
	1145	0.0	0.02	0.0	0.0
	1336	0.1	0.13	0.2	0.21
	1455	0.1	0.09	0.0	0.04
acetonitrile	2267	1.00	0.11	1.00	0.33
	1448	9.1	1.00	3.0	1.00
	1382	0.8	0.09	0.6	0.21
	1042	1.2	0.14	0.7	0.24
	920	0.5	0.05	0.1	0.02

Table A9.2.: Experimental IR intensities (A, km·mol^{-1}) of all assessed molecules in solution and the corresponding B97-3c computed cluster values.

Molecule	Integration range	$A^{sol.}$(exp.)	rel. $A^{sol.}$(exp.)	$A^{sol.}$(comp.)	$A^{sol.}$(comp.)
cyclohexane-d_{12}	1240–1130	4.6	0.53	6.1	0.62
	1130–1045	8.7	1.00	9.9	1.00
	1045–950	5.4	0.62	4.7	0.47
	950–880	4.6	0.53	4.5	0.45
	770–600	5.6	0.64	8.8	0.89
	450–350	0.8	0.09	0.7	0.07
acetone	1740–1680	161.0	1.00	170.0	1.00
	1500–1380	55.0	0.34	67.0	0.39
	1380–1335	84.0	0.52	82.0	0.48
	1250–1210	44.0	0.27	49.0	0.29
	1100–1080	9.5	0.06	12.0	0.07
	940–840	5.0	0.03	7.0	0.04
	820–760	1.0	0.01	1.4	0.01
	545–524	12.8	0.08	12.2	0.07
cyclohexane(CCl_4)	1500–1330	28.7	1.00	32	1.00
	1310–1190	3.0	0.10	1.8	0.06
	1100–970	4.3	0.15	3.2	0.10
	970–800	7.3	0.25	9.3	0.29

Molecule	Integration range	$A^{sol.}$(exp.)	rel. $A^{sol.}$(exp.)	$A^{sol.}$(comp.)	$A^{sol.}$(comp.)
	560–480	1.1	0.04	1.1	0.03
cyclohexane(liq.)	1500–1330	33.1	1.00	32.5	1.00
	1310–1190	4.4	0.13	2.2	0.07
	1100–970	2.8	0.08	3.7	0.11
	970–800	7.3	0.22	11.1	0.34
	560–480	0.7	0.02	1.8	0.06
1,4-dioxane	1520–1410	22.0	0.10	19.1	0.09
	1410–1335	11.9	0.05	18.5	0.09
	1335–1200	55.5	0.26	42.6	0.21
	1200–1000	216.5	1.00	201.8	1.00
	960–800	108.8	0.50	106.3	0.53
	660–540	19.9	0.09	16.5	0.08
1,4-dioxane-d_8	1300–1070	234.9	1.00	217.6	1.00
	1070–1000	48.9	0.21	81.5	0.37
	1000–840	19.0	0.08	19.2	0.09
	840–700	83.4	0.36	48.1	0.22
	560–440	13.1	0.06	11.7	0.05
tetrahydropyrane	1550–1415	27.5	0.24	27.6	0.23
	1415–1325	12.2	0.10	15.1	0.12
	1325–1235	18.5	0.16	17.0	0.14
	1235–1165	33.2	0.28	22.7	0.19
	1165–1140	3.8	0.03	2.0	0.02
	1140–977	117.0	1.00	121.7	1.00
	977–930	4.40	0.04	6.1	0.05
	925–832	36.6	0.31	38.1	0.31
	832–780	6.1	0.05	9.1	0.07
	620–520	5.9	0.05	3.8	0.03
	450–340	11.2	0.10	7.9	0.06
	290–200	4.9	0.04	4.1	0.03
pyridine	1540–1660	26.0	0.38	30	0.48
	1477–1512	4.0	0.06	3.1	0.05
	1472–1480	31.0	0.46	34.0	0.54
	1334–1400	0.5	0.01	0.2	0.00
	1256–1308	0.0	0.00	0.4	0.01
	1182–1246	4.3	0.06	5.8	0.09
	1110–1173	3.6	0.05	2.8	0.04
	1054–1100	4.5	0.07	8.6	0.14
	1009–1054	7.7	0.11	4.5	0.07
	958–1009	5.4	0.08	7.0	0.11
	671–735	68.0	1.00	63.0	1.00
	650–659	1.1	0.02	1.0	0.02
	570–632	4.4	0.06	4.2	0.07
MeOD	824	0.1	0.33	0.03	0.18
	902	0.0	0.07	0.01	0.03
	979	0.2	1.00	0.17	1.00

Molecule	Integration range	$A^{sol.}$(exp.)	rel. $A^{sol.}$(exp.)	$A^{sol.}$(comp.)	$A^{sol.}$(comp.)
	1057	0.0	0.09	0.01	0.06
	1064	0.0	0.07	0.01	0.06
	1097	0.1	0.87	0.10	0.57
	1125	0.1	0.40	0.03	0.20
MeOH	1035	0.2	1.00	0.15	1.00
	1116	0.0	0.13	0.01	0.08
	1423	0.1	0.56	0.03	0.22
	1450	0.0	0.06	0.00	0.01
	1477	0.0	0.00	0.01	0.08
	1477	0.0	0.13	0.01	0.06
benzene	1508–1425	23.3	0.23	32.6	0.32
	1095–910	15.6	0.16	20.9	0.21
	730–615	100.0	1.00	101.0	1.00
benzene-d_6	1330	5.4	0.11	7.5	0.13
	812	11.4	0.23	18.7	0.33
	496	49.7	1.00	55.9	1.00

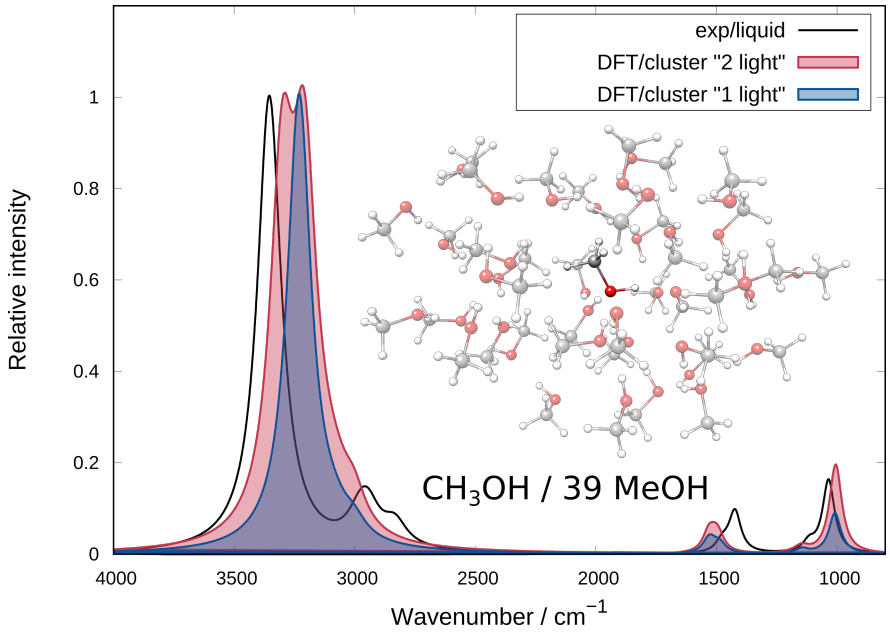


Figure A9.1.: IR spectra of CH_3OH in the spectral region of 4000–800 cm^{-1}. Experiment for liquid (black), and the corresponding cluster models (blue and red) comprising 40 methanol molecules. Vibrations of the central "solute" methanol molecule(s) are separated from other methanol molecules by setting the mass of all atoms of the "solvent" molecules to 400 amu. The blue spectrum corresponds to the case that only one "light" solute molecule is surrounded by 39 "heavy" molecules. The red spectrum belongs to a hydrogen bonded dimer of two "light" solute molecules (whose vibrations are strongly coupled) surrounded by 38 "heavy" molecules.

Table A9.3.: Experimental frequencies (cm^{-1}) of all assessed molecules in gas- and condensed-phase and the corresponding B97-3c computed values.

Molecule	Freq.^{gas}(exp)	Freq.^{gas}(comp.)	Freq.$^{sol.}$(exp)	Freq.$^{sol.}$(comp.)
acetone	1738	1753	1711	1711
	1456	1492	1442	1496
	1456	1476	1442	1486
	1438	1468	1419	1476
	1438	1467	1419	1464
	1363	1387	1364	1401
	1360	1385	1351	1395
	1218	1223	1229	1245
	1093	1113	1096	1119
	1067	1080	1071	1090
	896	879	897	912
	896	878	871	890
	779	775	791	797
	528	530	534	538
	483	488	534	503
MeOH	1034	1030.2	1035	1022
	1075	1079	1116	1130
	1145	1164	1423	1457
	1336	1386	1450	1483
	1455	1481	1477	1501
	2844	2943.	1477	1523
	3672	3760.1		
	1465	1498		
	1478.5	1517		
	2970	2987		
	2999	3073		
MeOD	775	780	824	834
	895	898	902	914
	980	977	979	972
	1028	1070	1064	1089
	1069	1084	1097	1113
	1078	1096	1125	1138
	1135	1149	2072	2144
	2074	2112	2216	2268
	2717	2737	2245	2313
	2213	2216	2493	2527
	2250	2276		

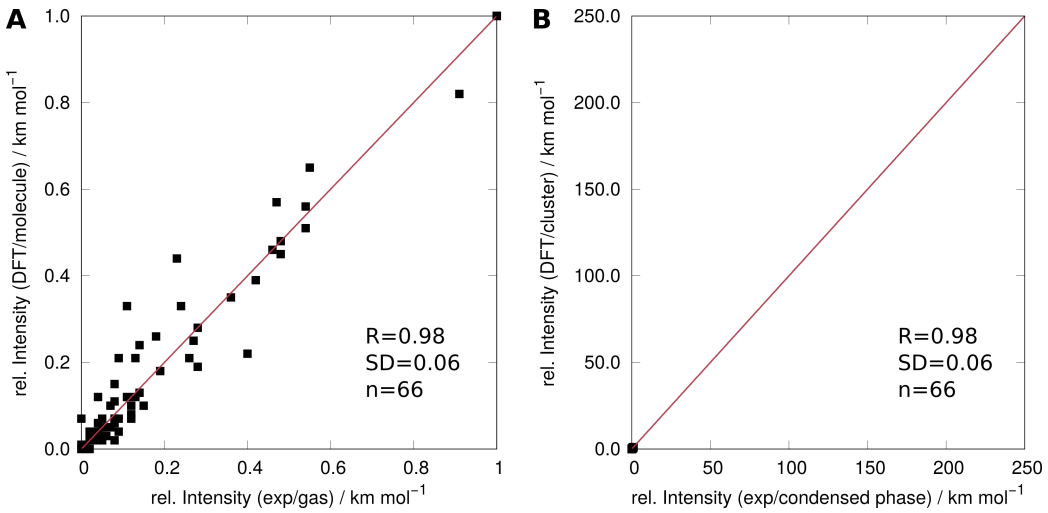


Figure A9.2.: Correlation plot of relative IR intensities calculated for isolated molecules of cyclohexane, cyclohexane-d_{12}, 1,4-dioxane, 1,4-dioxane-d_8, acetone, tetrahydropyrane, pyridine, benzene, benzene-d_6, acetonitrile (A) and the corresponding cluster models (B) *vs.* experimental gas- phase (A) and condensed-phase (B) relative intensities, respectively. *R*, correlation coefficient; SD, standard deviation; *n*, the total number of data included in the analysis.

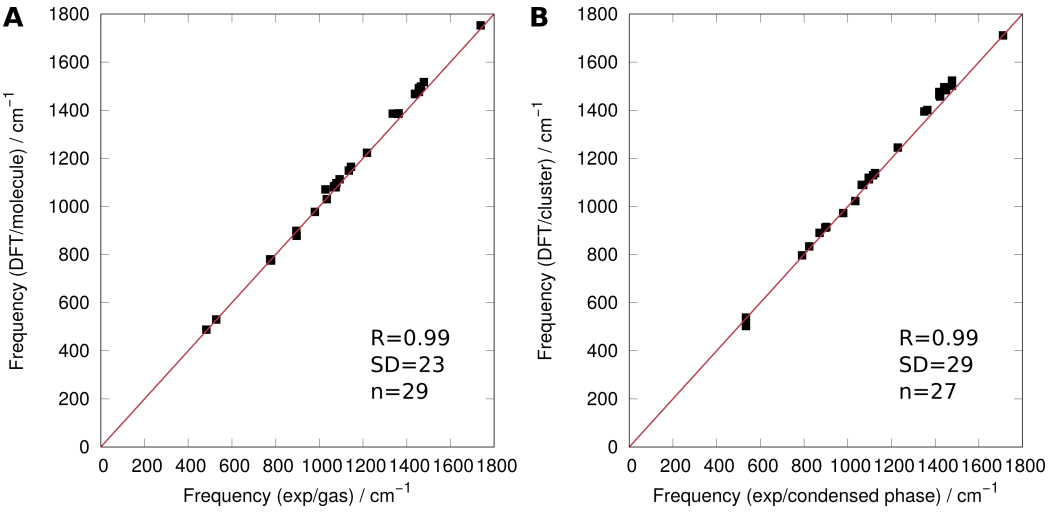
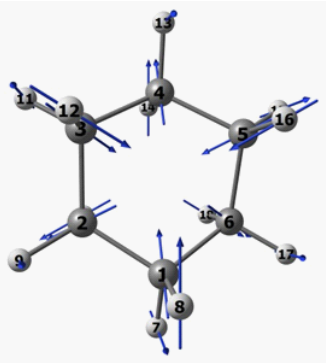


Figure A9.3.: Correlation plot of frequencies calculated for isolated molecules of acetone, methanol and methanol-d_4 (A) and corresponding cluster models (B) vs. experimental gas phase (A) and condensed-phase (B) frequencies, respectively. *R*, correlation coefficient; SD, standard deviation; *n*, total number of data included in the analysis.

Table A9.4.: Comparison of Mulliken charges of atoms (q) and normal modes for the isolated C_6D_{12} molecule and the same molecule in the C_6D_{12}:19CCl_4 cluster. Upper indices 1 and 2 refer to the cases of the isolated molecule and the cluster, respectively. Atomic amplitudes (R) are compared for two vibrations whose IR intensities (A in km·mol^{-}1) increase by ca. 10 % and ca. 70 %, respectively on passing from the isolated molecule to the cluster.



Atom	q^1	q^2	R^1	R^2	R^1-R^2	R^1	R^2	R^1-R^2
1C	-0.3029	-0.3221	0.16871	0.17005	-0.0013	0.0666	0.0762	-0.0096
2C	-0.3029	-0.3063	0.16871	0.1692	-0.0005	0.0666	0.0669	-0.0003
3C	-0.3029	-0.2218	0.16872	0.1711	-0.0024	0.0666	0.0563	0.0103
4C	-0.3029	-0.3167	0.16871	0.16991	-0.0012	0.0666	0.0540	0.0126
5C	-0.3029	-0.2711	0.16871	0.16793	0.0008	0.0666	0.0665	0.0001
6C	-0.3029	-0.2139	0.16872	0.16917	-0.0005	0.0666	0.0781	-0.0115
7H	0.1646	0.1543	0.36734	0.36868	-0.0013	0.2337	0.2468	-0.0131
8H	0.1383	0.1317	0.05717	0.05508	0.0021	0.3281	0.3534	-0.0253
9H	0.1646	0.1299	0.36734	0.37083	-0.0035	0.2337	0.2325	0.0012
10H	0.1383	0.1260	0.05717	0.06413	-0.0070	0.3281	0.3223	0.0058
11H	0.1646	0.1472	0.36733	0.36855	-0.0012	0.2337	0.2177	0.0160
12H	0.1383	0.1196	0.05717	0.05705	0.0001	0.3281	0.2971	0.0310
13H	0.1646	0.1788	0.36734	0.36707	0.0003	0.2337	0.2357	-0.0020
14H	0.1383	0.1214	0.05717	0.05093	0.0062	0.3281	0.3146	0.0135
15H	0.1646	0.1390	0.36734	0.36273	0.0046	0.2337	0.2335	0.0002
16H	0.1383	0.0956	0.05717	0.05672	0.0004	0.3281	0.3291	-0.0010
17H	0.1646	0.1376	0.36733	0.36402	0.0033	0.2337	0.2342	-0.0005
18H	0.1383	0.1099	0.05717	0.05523	0.0019	0.3281	0.3486	-0.0205

Acknowledgment

Without the support of several people, this thesis would not have been possible.
In particular, I wish to thank Prof. Dr. Stefan Grimme for the supervision of my doctoral thesis. He gave me the opportunity to work on many interesting projects and the freedom to develop and implement my own ideas. I am very thankful for his scientific support, his advice, and the many conversations we had about science and beyond.
Further, I would like to thank Prof. Dr. Thomas Bredow for being the second review of my thesis and the fruitful discussions as my mentor in the Bonn International Graduate School of Chemistry (BIGS Chemistry) program.
I especially thank Markus Bursch, as a long-time office colleague, who introduced me to the working group of Stefan Grimme. In this context, I also thank my colleagues Dr. Christoph Bannwarth, Dr. Gerit Brandenburg, Fabian Bohle, Sebastian Dohm, Hagen Neugebauer, Jeroen Koopman, Dr. Jakob Seibert, Dr. Eike Caldeweyher, Philipp Pracht, Julius Stückrath, Sebastian Ehlert, Dr. Jan Mewes and, Sarah Schmitz for the pleasant years, fruitful collaborations, and interesting discussions.
I also would like to thank the numerous collaborators for all the exciting projects and the pleasant cooperation during the last years. Here, I would like to thank Prof. Dr. Sigurd Höger, Prof. Dr. Fred Manby, Dr. Dinar Abdullin, Nico Fleck, and Prof. Dr. Olav Schiemann in particular.
Throughout my studies, I had the pleasure to work in a welcoming and scientifically excellent atmosphere. Hence, I want to thank all past and current members of the Grimme group for contributing to it in their own way. Something I will always remember is the ACS meeting in San Diego and the summer school in Sicily, which I attended together with my valued colleagues and friends Dr. Christoph Bannwarth, Philipp Pracht, Jeroen Koopman, and Hagen Neugebauer.
I am also grateful for the opportunity to work with Prof. Dr. Sergey A. Katsyuba during his stay in Bonn as well as for the steady technical and administrative support by Jens Mekelburger, and Claudia Kronz.
I would also like to give special thanks to Dr. Andreas Hansen, who has been a great colleague, advisor, and friend to me.
For a wonderful time together during my Bachelor and Master studies, I would like to thank Dominik Offermanns, Patrick Rössel, and Simon Steiner.
Finally, I want to thank my girlfriend Alena Tretiakova for her loving support and patience. I further want to take this opportunity to thank my brother Florian, my parents Annemarie and Frank, and my grandparents for their continuous support throughout my entire studies. Without my family, this thesis would not have been possible in this way.

Diese Dissertation ist im Zeitraum 10/2018 - 07/2021 entstanden

Mulliken Center for Theoretical Chemistry
Mathematisch-Naturwissenschaftliche Fakultät
Rheinische Friedrich-Wilhelms-Universität Bonn